

Probabilistic Models for Concurrent Chatting Activity Recognition

Chia-chun Lian and Jane Yung-jen Hsu

Department of Computer Science and Information Engineering
National Taiwan University
yjhsu@csie.ntu.edu.tw

Abstract

Recognition of chatting activities in social interactions is useful for constructing human social networks. However, the existence of multiple people involved in multiple dialogues presents special challenges. To model the conversational dynamics of concurrent chatting behaviors, this paper advocates Factorial Conditional Random Fields (FCRFs) as a model to accommodate co-temporal relationships among multiple activity states. In addition, to avoid the use of inefficient Loopy Belief Propagation (LBP) algorithm, we propose using Iterative Classification Algorithm (ICA) as the inference method for FCRFs. We designed experiments to compare our FCRFs model with two dynamic probabilistic models, Parallel Condition Random Fields (PCRFs) and Hidden Markov Models (HMMs), in learning and decoding based on auditory data. The experimental results show that FCRFs outperform PCRFs and HMM-like models. We also discover that FCRFs using the ICA inference approach not only improves the recognition accuracy but also takes significantly less time than the LBP inference method.

1 Introduction

Human social organizations often have complex structures. The organization of a group defines the roles and functions of each individual, thereby ensuring the functions of the group as a whole. *Ad hoc social organization* may develop during a public gathering due to common interests or goals. For example, at an academic conference, attendees tend to interact with people who share similar background or research interests. As a result, conversation patterns among the attendees can be used to map out the human networks [Gips and Pentland, 2006]. Such insights can help provide important support for timely services, e.g. sending announcements to attendees with relevant interests, or tracking conversations on hot topics at the meeting.

This research aims to develop the methodology for recognition of concurrent chatting activities from multiple audio streams. The main challenge is the existence of multiple participants involved in multiple conversations. To model the

dynamic interactions, we adopt a probabilistic framework to learn and recognize concurrent chatting activities.

This paper proposes using Factorial Conditional Random Fields (FCRFs) [Sutton *et al.*, 2007] to detect and learn from patterns of multiple chatting activities. First, we survey several related projects and technologies for chatting activity recognition. Section 3 defines the FCRFs model, and section 4 introduces its learning and decoding, combined with LBP or ICA inference methods. Finally, we present our experiments to evaluate the performance of different models and inference methods, followed by the conclusions and future work.

2 Background

It is intuitively reasonable to infer activities of daily living using dynamic probabilistic models, such as Hidden Markov Models (HMMs) and complex Dynamic Bayesian Networks (DBNs). To recognize chatting activities, researchers tried to examine the *Mutual Information* between each person's voicing segments as a matching measure [Basu, 2002; Choudhury and Basu, 2004]. However, since multiple conversational groups usually exist concurrently in public occasions, it is very likely that the matching result does not correspond to the actual pair of conversational partners. To avoid ambiguity, model-based methods are used to monitor the chatting activities among different social groups. To capture the turn-taking behaviors of conversations, HMMs are used to describe the transition possibility of dynamic changes among group configurations [Brdiczka *et al.*, 2005]. Factored DBNs were introduced in [Wyatt *et al.*, 2007] to separate speakers in a multi-person environment relying on privacy-sensitive acoustic features, where the state factorization makes it relatively simple to express complex dependencies among variables.

Conditional Random Fields (CRFs) [Lafferty *et al.*, 2001] provide a powerful probabilistic framework for labeling and segmenting structured data by relaxing the *Markov independence assumption*. Various extensions of CRFs have been successfully applied to learning temporal patterns of complex human behaviors [Liao *et al.*, 2007]. Multi-Task CRFs (MCRFs), a generalization of LCRFs, are proposed to do multitasking sequence labeling for human motion recognition [Shimosaka *et al.*, 2007]. FCRFs are adopted by [Wu *et al.*, 2007] to recognize multiple concurrent activities from the MIT House.n dataset [Intille *et al.*, 2006]. Researchers at

MIT presented Dynamic CRFs (DCRFs) combined with factored approach to capture the complex interactions between NP and POS labels in natural-language chunking task [Sutton *et al.*, 2007]. A hierarchical CRFs model was proposed in [Liao *et al.*, 2007] for automatic activity and location inference from the traces of GPS data.

To solve inference problems on generalized CRFs with loops, LBP algorithm is proposed based on local messages passing [Yedidia *et al.*, 2002]. Unfortunately, the LBP algorithm is neither an exact inference algorithm nor guaranteed to converge except when the model structure is a tree [Sutton and McCallum, 2007]. In addition, its updating procedure is very time-consuming to train and decode generalized CRFs models. Other researchers from MIT presented ICA to iteratively infer the states of variables given the neighboring variables as observed information [Neville and Jensen, 2000]. ICA inference was shown empirically to help improve classification accuracy and robustness of the LBP algorithm for graph structures with high link density [Sen and Getoor, 2007].

3 Model Design

Let X be the set of observed variables representing acoustic feature values and Y be the set of hidden variables representing the chatting activity labels. In addition, let (x, y) denote the values assigned to the variables (X, Y) , N denote a given number of concurrent chatting activities, and T denote a given total time steps. Specifically, we use a 3-state variable $Y_i^t \in Y$ whose value is y_i^t to represent the state of chatting activity i at time step t . We also use $Y_i = \{Y_i^t\}_{t=1}^T$ whose values are y_i to represent the state sequence of chatting activity i over time. Similarly, we use one observed variable X_i^t assigned with x_i^t to denote single observed feature value related to Y_i^t . As for multiple feature values, each of them will be represented as an independent observed variable.

Let $G = (V, E)$ be an undirected graph structure consisting of two vertex sets $\{(X, Y)\} \equiv V$. For any given time slice t and chatting activity i , we build edges $(Y_i^t, Y_i^{t+1}) \in E$ to represent the possibility of activity state transition across time slices. We also build edges $(X_i^t, Y_i^t) \in E$ to represent the possible relationships between activity labels and acoustic observations. Most importantly, edges $(Y_i^t, Y_j^t) \in E$ are built to represent the possibility of co-temporal relationships between any two concurrent chatting activities i and j . That is, all the hidden nodes within the same time slice are fully connected. Figure 1 shows a sample FCRFs model for the recognition of 3 concurrent chatting activities in a dynamic form by unrolling the structure of two time slices.

Secondly, we let C be the set of maximum cliques in G , where each clique $c(i, j, t) \in C$ is composed of the vertices based on the indexes (i, j, t) and chatting activities $i \neq j$. Figure 1 presents 3 sample cliques, where the *local clique* consists of $(X_i^t, Y_i^t) \in C$, the *temporal clique* consists of $(X_i^t, Y_i^t, X_i^{t+1}, Y_i^{t+1}) \in C$, and the *co-temporal clique* consists of $(X_i^t, Y_i^t, X_j^t, Y_j^t) \in C$. Meanwhile, several non-negative potential functions are also defined on these cliques, which are shown as follows:

- We use $\phi_i^A(x_i^t, y_i^t, t) =$

$$\exp\left(\sum_{p=1}^P w_i^{(p)} f_i^{(p)}(x_i^t, y_i^t, t)\right)$$

to denote the *local potential function* defined on every local clique, where $f_i^{(p)}(\dots)$ is a function to indicate whether the state values are equal to the p^{th} state combination within this clique, and $w_i^{(p)}$ is its corresponding weight assigned to $W_i^{(p)}$. Particularly, $f_i^{(p)}(\dots) = x_i^t$, if X_i^t represents a numerical feature value.

- We use $\phi_i^B(x_i^t, y_i^t, x_i^{t+1}, y_i^{t+1}, t) =$

$$\exp\left(\sum_{q=1}^Q w_i^{(q)} f_i^{(q)}(x_i^t, y_i^t, x_i^{t+1}, y_i^{t+1}, t)\right)$$

to denote the *temporal potential function* defined on every temporal clique, where $f_i^{(q)}(\dots)$ is a function to indicate whether the state values are equal to the q^{th} state combination within this clique, and $w_i^{(q)}$ is its corresponding weight assigned to $W_i^{(q)}$. Particularly, $f_i^{(q)}(\dots) = |x_i^{t+1} - x_i^t|$, if X_i^t represents a numerical feature value.

- We use $\phi_{ij}^\Delta(x_i^t, y_i^t, x_j^t, y_j^t, t) =$

$$\exp\left(\sum_{r=1}^R w_{ij}^{(r)} f_{ij}^{(r)}(x_i^t, y_i^t, x_j^t, y_j^t, t)\right)$$

to denote the *co-temporal potential function* defined on every co-temporal clique, where $f_{ij}^{(r)}(\dots)$ is a function to indicate whether the state values are equal to the r^{th} state combination within this clique, and $w_{ij}^{(r)}$ is its corresponding weight assigned to $W_{ij}^{(r)}$. Particularly, $f_{ij}^{(r)}(\dots) = |x_i^t - x_j^t|$, if X_i^t represents a numerical feature value.

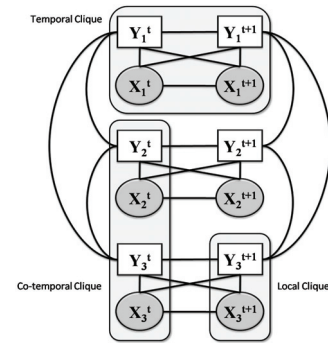


Figure 1: A sample FCRF of 3 concurrent chatting activities.

Finally, we use $W = \{W^{(k)}\}_{k=1}^K$ whose assigning values are $w = \{w^{(k)}\}_{k=1}^K$ to denote the combined set for all of the model parameters $\{W_i^A \cup W_i^B \cup W_{ij}^\Delta\}_{i \neq j}^N$, where $W_i^A =$

$\{W_i^{(p)}\}_{p=1}^P$, $W_i^B = \{W_i^{(q)}\}_{q=1}^Q$, $W_{ij}^\Delta = \{W_{ij}^{(r)}\}_{r=1}^R$ and K is the number of model parameters. In addition, we use $D = \{x^{(m)}, y^{(m)}\}_{m=1}^M$ to denote the data set used for learning and decoding processes, where M is the number of data. Now we can formally define the mathematical formulation of FCRFs model as follows:

$$P(y|x, w) = \frac{1}{Z(x)} \left(\prod_{t=1}^T \prod_{i=1}^N \phi_i^A(x_i^t, y_i^t, t) \right) \left(\prod_{t=1}^{T-1} \prod_{i=1}^N \phi_i^B(x_i^t, y_i^t, x_i^{t+1}, y_i^{t+1}, t) \right) \left(\prod_{t=1}^T \prod_{i,j} \phi_{ij}^\Delta(x_i^t, y_i^t, x_j^t, y_j^t, t) \right) \quad (1)$$

where $Z(x)$ is the normalization constant.

4 Inference Methods

Before describing the necessary inference tasks in our FCRFs model, let us use $f^{(k)}(x_{c(i,j,t)}, y_{c(i,j,t)}, t)$ to denote either $f_i^{(p)}(\dots)$, $f_i^{(q)}(\dots)$ or $f_{ij}^{(r)}(\dots)$ for simplification, where $(x_{c(i,j,t)}, y_{c(i,j,t)})$ represents the assigning values of nodes in the specific clique $c(i, j, t)$. The following sections describe the learning and decoding processes in detail and how we utilize the ICA inference method to improve the efficiency.

4.1 Learning and Decoding Processes

The purpose of the learning process is to determine the weight $w^{(k)}$ corresponding to each feature function $f^{(k)}$. To do this, we can maximize the log-likelihood relying on the training data set D , where the log-likelihood function is shown as follows:

$$L(D|w) = \sum_{m=1}^M \log P(y^{(m)}|x^{(m)}, w) \quad (2)$$

which is the log value of $P(D|w)$ defined as follows:

$$\begin{aligned} P(D|w) &= \prod_{m=1}^M P(y^{(m)}, x^{(m)}|w) \\ &= \prod_{m=1}^M P(y^{(m)}|x^{(m)}, w) \prod_{m=1}^M P(x^{(m)}|w) \\ &\cong \prod_{m=1}^M P(y^{(m)}|x^{(m)}, w) \end{aligned}$$

where $\prod_{m=1}^M P(x^{(m)}|w)$ is constant and can be ignored, because we reasonably assume that $P(X|W)$ is a uniform distribution. As a result, we can derive the partial derivative of

log-likelihood with respect to $w^{(k)}$ from Eq. (1) and Eq. (2) as follows:

$$\begin{aligned} \frac{\partial L(D|w)}{\partial w^{(k)}} &= \sum_{m=1}^M \sum_{t=1}^T \sum_{i,j}^N f^{(k)}(x_{c(i,j,t)}^{(m)}, y_{c(i,j,t)}^{(m)}, t) - \\ &\sum_{m=1}^M \sum_{t=1}^T \sum_{i,j}^N P(y_{c(i,j,t)}^{(m)}|x^{(m)}, w) f^{(k)}(x_{c(i,j,t)}^{(m)}, y_{c(i,j,t)}^{(m)}, t) \end{aligned} \quad (3)$$

In this way, we can learn the weights w by satisfying the equation $\partial L(D|w)/\partial w^{(k)} = 0$. To solve such an optimization problem, we use L-BFGS method to conduct the learning process [Sutton and McCallum, 2007]. Noticeably, the marginal probability $P(y_{c(i,j,t)}^{(m)}|x^{(m)}, w)$ in Eq. (3) can be difficult to calculate. However, we cannot use the Forward-Backward Algorithm (FBA) [Rabiner, 1989] to efficiently compute it, because such a DP method can only be used in linear-chain graph structures like HMMs and LCRFs models. Therefore, we decide to use LBP sum-product algorithm with random schedule strategy to approximate the marginal probability [Yedidia *et al.*, 2002]. Although LBP simply conducts approximate inference, it has often been used for loopy CRFs inference [Sutton *et al.*, 2007; Vishwanathan *et al.*, 2006; Liao *et al.*, 2007].

Unfortunately, as for the objective value of log-likelihood required by the L-BFGS optimization process, it is infeasible for LBP algorithm to calculate the normalization constant $Z(x)$ in Eq. (1). Therefore, we decide to use Bethe free energy [Yedidia *et al.*, 2005] to approximate the normalization constant. Furthermore, to avoid the over-fitting problem, what we actually do is to maximize the penalized log-likelihood $L(w|D) = L(D|w) + \sum_{k=1}^K \log P(w^{(k)})$ by taking into consideration a zero-mean Gaussian prior distribution $P(w^{(k)}) = \exp(-(w^{(k)})^2/2\sigma^2)$ with variance $\sigma = 1.5$ for each parameter $W^{(k)}$. As a result, the original partial derivative in Eq. (3) becomes a new penalized form which is shown as follows:

$$\frac{\partial L(w|D)}{\partial w^{(k)}} = \frac{\partial L(D|w)}{\partial w^{(k)}} - \frac{w^{(k)}}{\sigma^2}$$

As regards the decoding process, LBP algorithm can be also used for performing the MAP (Maximum A Priori) inference which can decode the most possible sequences of activity states [Yedidia *et al.*, 2002]. To conduct the MAP inference, we simply propagate the max value during the message updating procedure in the original LBP sum-product algorithm. After such an LBP max-product algorithm converges and the MAP probability of each hidden variable is estimated, we can label every hidden variable by choosing the most likely value according to the MAP probability.

4.2 Decomposition

Another interpretation for the FCRFs model is to separate the factored graph structure into several linear-chain structures. That is, the original FCRFs model is considered to be composed of several LCRFs models, where each of them

represents single chatting activity. To further maintain the co-temporal relationships, each hidden node Y_i^t depends not only on its own observed node X_i^t but also on other hidden nodes Y_j^t as observations. As a result, the new form for each LCRFs model is expressed as follows:

$$P(y_i|o_i, w_i) = \frac{1}{Z(o_i)} \left(\prod_{t=1}^T \phi_i^A(o_i^t, y_i^t, t) \right) \left(\prod_{t=1}^{T-1} \phi_i^B(o_i^t, y_i^t, o_i^{t+1}, y_i^{t+1}, t) \right)$$

where o_i^t is a value set assigned to new observed variables $O_i^t = X_i^t \cup \{Y_j^t\}_{i \neq j}^N$, o_i is a value set assigned to $O_i = \{O_i^t\}_{t=1}^T$, and w_i is a value set assigned to the LCRFs model parameters $W_i = W_i^A \cup W_i^B$ only for chatting activity i . Noticeably, $\{Y_j^t\}_{i \neq j}^N$ represents the neighbor nodes of Y_i^t in the same time slice t . In this way, the learning process for the original FCRFs structure is to train individual LCRFs models, where each LCRFs model considers other activity states as additional observations. As a result, the main inference tasks, including the calculation of marginal probability $P(y_{c(i,j,t)}^{(m)}|o_i^{(m)}, w_i)$ and the normalization constant $Z(o_i^{(m)})$, can be efficiently obtained through the use of FBA, such that the exact inference is still likely to be applied in FCRFs learning.

4.3 ICA Inference

Most importantly, ICA provides another inference approach in an iterative fashion for decoding process. The basic idea of ICA is to decode every target hidden variable based on the assigning labels of its neighbor variables, where the labels might be dynamically updated throughout the run of iterative process. Compared with LBP algorithm, ICA does the inference not only based on observed variables but also based on hidden variables as observations. After a given number of iterative cycles, the classification process will eventually terminate and all the hidden variables will be assigned with fixed labels.

Therefore, we can use the ICA inference method to label multiple concurrent sequences, as long as we priorly follow the decomposition procedure, letting the LCRFs model parameters W_i for every chatting activity i be learned as w_i . Algorithm 1 formally provides the detailed ICA procedure for our FCRFs decoding, where an upper limit $\Lambda = 10^3$ for the number of iterations is set to avoid infinite repeat, followed by an ICA example with 2 concurrent chatting activities as shown in Figure 2. Noticeably, since the LCRFs model provides an efficient FBA, the ICA inference method also has an opportunity to help accelerate the decoding process.

5 Experiments

Before carrying out the experiments, we asked every participant to wear an audio recorder around his or her neck to collect the audio data. During a given period of time, the participants can randomly determine their conversational partners

Algorithm 1 Iterative Classification Algorithm (Λ)

- 1: **repeat**
 - 2: {Decode new labels}
 - 3: **for all** $Y_i \subseteq Y$ **do**
 - 4: Store $\bar{y}_i \leftarrow \arg \max_{y_i} P(y_i|o_i, w_i)$
 - 5: **end for**
 - 6: {Update new labels}
 - 7: **for all** $Y_i \subseteq Y$ **do**
 - 8: Assign \bar{y}_i to Y_i as new labels
 - 9: **end for**
 - 10: **until** $\bar{y}_i = y_i$, for all $Y_i \subseteq Y$
-

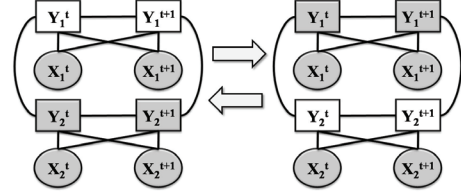


Figure 2: An ICA example for FCRFs model decoding.

and the audio recorders can record their conversations. As a result, the collected audio recordings can be used for the annotation of participants' conversational states as well as the extraction of acoustic feature values. To do annotation, researchers should listen very carefully to the audio recordings to know who has spoken with whom, relying on the content of what was said.

To evaluate the performance, we used the cross-validation method to test the experimental data. However, to deal with sequential data, this method suffers from the problem that the important activities to be recognized may only occur in specific time periods, such that the possible patterns in the testing data cannot be learned in the training data, or the learned results in training data cannot help recognize the testing data. To address the problem, the sequential data should be fragmented into several small segments and each of the adjacent segments along the temporal axis will be systematically reassigned to each fold. In practice, we define the time duration between two consecutive time steps as 1 second and each segment contains 10 time slices ($T = 10$).

Finally, we measure the recognition accuracy as the percentage of correctly predicted conversations by applying the 10-fold cross-validation to compare the performance of the various types of models and inference methods. More specifically, the comparisons for recognition accuracy include *accuracy*, *recall*, *precision* and *F-score*. In addition, to analyze the efficiency of FCRFs learning and decoding by using different inference methods, including LBP and ICA inferences, we measure the learning time as the accumulated training time during the process of cross-validation on an Xeon 5130 2.0G PC, and measure the decoding time as the accumulated decoding time in a similar way. The following sections present several experiments to verify the advantages of our FCRFs models with ICA inference method.

5.1 FCRFs Models vs. PCRFs Models

In this experiment, the primary purpose is to analyze the recognition accuracy of two different FCRFs models by using LBP and ICA inference methods respectively. In addition, another Parallel CRFs (PCRFs) model is also trained for comparison, which is composed of several LCRFs models parallelly and no co-temporal connections between hidden nodes are built. That is, the co-temporal relationships among chatting activities in FCRFs model are eliminated. As a result, we calculate the accuracy, recall, precision and F-score for each chatting activity and then average them respectively. We've designed 2 common scenarios to collect auditory data for experiments, which are shown as follows:

1. **Meeting activity** The data is a set of 80-minute audio streams recorded by 3 participants who are sitting in fixed seats and holding discussion on a project. During the period, they can switch their conversational partners and no others would interrupt their meeting. Noticeably, the audio recorders simply record the voices from the participants rather than other non-human noises, so we choose the volume features as our primary observations. Although most of the time only one person is speaking, the lack of voice recognition makes it relatively difficult to disambiguate the speakers. The recognition accuracy is summarized in Table 1.

Table 1: Recognition accuracy (%) in meeting activity.

Results (%)	Accuracy	Precision	Recall	F-score
PCRFs	80.71	78.94	73.45	75.22
FCRFs + LBP	80.31	78.60	81.27	79.77
FCRFs + ICA	82.28	81.53	77.74	79.41

2. **Public occasion** The data is a set of 45-minute audio streams recorded by 3 participants who can walk around and chat with others in a real-world laboratory. Noticeably, not only these participants who wear audio recorders are involved in our experimental environment, but also other occupants who do not wear audio recorders are allowed to exist. In this way, all of them can randomly chat with each other or choose to do their personal activities, thereby creating a naturalistic scenario varies in background noise. Therefore, in addition to the feature values of volume and MI, we consider extra features used for human voice detection. The recognition accuracy is summarized in Table 2.

Table 2: Recognition accuracy (%) in public occasion.

Results (%)	Accuracy	Precision	Recall	F-score
PCRFs	80.62	69.46	39.67	50.47
FCRFs + LBP	84.57	66.11	63.20	64.09
FCRFs + ICA	86.62	82.16	61.37	68.94

In both the above scenarios, we can observe a consistent phenomenon that all the FCRFs models significantly outperform the PCRFs model in the comparison of F-score. This

result provides us the conclusion that it is helpful to utilize the co-temporal relationship for chatting activity recognition. However, the FCRFs model with LBP inference performs even more badly than the PCRFs model in the comparison of precision, while the FCRFs model with ICA inference still performs the best.

5.2 CRF-like Models vs. HMM-like Models

In this experiment, two DBN models, Parallel HMMs (PHMMs) [Vogler and Metaxas, 2001] and Coupled HMMs (CHMMs) [Brand *et al.*, 1997], are trained for comparison. PHMMs model is very similar to the PCRFs model, where all of the multiple chatting activities are independent temporal processes and each of them is modeled as a linear-chain HMMs. Furthermore, the CHMM model assumes that each hidden variable is conditionally dependent on all hidden variables in the previous time slice. Therefore, the CHMM model also has the ability to capture the interactive relationships among chatting activities, which can be compared with the FCRFs model. By analyzing the audio data collected in the public occasion, the recognition accuracy of HMM-like models is summarized in Table 3.

Table 3: Recognition accuracy (%) of various HMMs.

Results (%)	Accuracy	Precision	Recall	F-score
PHMMs	42.90	35.01	42.63	36.69
CHMMs	49.95	42.29	32.83	36.86

Taken as a whole, we can discover that all the CRF-like models significantly outperform the HMM-like models in all the comparisons of recognition accuracy, which concludes that the CRF-like models indeed have the ability to accommodate overlapping features and are much more powerful to capture complex co-temporal relationships among multiple chatting activities.

5.3 LBP Efficiency vs. ICA Efficiency

The efficiency analysis based on the audio data collected in the public occasion is summarized in Table 4, which compares the learning time and decoding time by using the various types of CRFs models as well as inference methods.

Table 4: Performance comparisons of learning time (sec.) and decoding time (sec.) using the various types of CRFs models and inference methods in the public occasion.

Results (sec.)	Learning Time	Decoding Time
PCRFs	2879.57	1.20
FCRFs + LBP	113320.00	40.05
FCRFs + ICA	6437.16	19.00

In this experiment, we can come to an important conclusion that the FCRFs model with ICA inference takes much less time than LBP method to complete the learning and decoding processes. Especially in the comparison of learning time, we can observe that there is even no significant difference between the PCRFs model and the FCRFs model with

ICA inference. Given the difference in convergence property of LBP and ICA in various conversation contexts, we chose to compare them when effective convergence is reached – just to be fair. Additional experiments may indeed offer potentially interesting observations by stopping LBP early at increasing rounds of message passing.

6 Conclusion and Future Work

This paper proposes the FCRFs model for joint recognition of multiple concurrent chatting activities by using the ICA inference method. We designed the experiments based on the collected auditory data to compare our FCRFs model with other dynamic probabilistic models, including PCRFs model and HMMs-like models. The initial experiment showed that the FCRFs model with ICA inference method, which is capable of accommodating the co-temporal relationships, can help improve the recognition accuracy in the presence of co-existing conversations. Most importantly, the FCRFs model using ICA inference approach significantly takes much less time to conduct the learning and decoding processes than the LBP inference method. FCRF with ICA can be generalized to recognize more general conversations or activities among a larger group of participants with properly designed FCRF graph structure capturing co-temporal relationships.

Acknowledgements

This research was supported by grants from the National Science Council NSC 96-2628-E-002-173-MY3, Ministry of Education in Taiwan 97R0062-06, and Intel Corporation.

References

- [Basu, 2002] S. Basu. *Conversational Scene Analysis*. PhD thesis, Massachusetts Institute of Technology, 2002.
- [Brand *et al.*, 1997] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Proceedings of the IEEE Conference on CVPR'97*, 1997.
- [Brdiczka *et al.*, 2005] O. Brdiczka, J. Maisonnasse, and P. Reignier. Automatic detection of interaction groups. In *Proceedings of the 7th International Conference on Multimodal Interfaces*, 2005.
- [Choudhury and Basu, 2004] T. Choudhury and S. Basu. Modeling conversational dynamics as a mixed-memory markov process. In *Proceedings of the Advances in Neural Information Processing Systems 17*, 2004.
- [Gips and Pentland, 2006] J. Gips and A. Pentland. Mapping human networks. In *Proceedings of the 4th Annual IEEE International Conference on Pervasive Computing and Communications*, 2006.
- [Intille *et al.*, 2006] S. S. Intille, K. Larson, E. M. Tapia, J. S. Beaudin, P. Kaushik, J. Nawyn, and R. Rockinson. House_n placelab data set, 2006.
- [Lafferty *et al.*, 2001] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.
- [Liao *et al.*, 2007] L. Liao, D. Fox, and H. Kautz. Hierarchical conditional random fields for GPS-based activity recognition. In *Springer Tracts in Advanced Robotics: Robotics Research*, volume 28, pages 487–506. Springer, 2007.
- [Neville and Jensen, 2000] J. Neville and D. Jensen. Iterative classification in relational data. In *Proceedings of the AAAI 2000 Workshop Learning Statistical Models from Relational Data*, 2000.
- [Rabiner, 1989] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the 1989 IEEE International Conference*, 1989.
- [Sen and Getoor, 2007] P. Sen and L. Getoor. Link-based classification. Technical Report CS-TR-4858, University of Maryland, February 2007.
- [Shimosaka *et al.*, 2007] M. Shimosaka, T. Mori, and T. Sato. Robust action recognition and segmentation with multi-task conditional random fields. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, 2007.
- [Sutton and McCallum, 2007] C. Sutton and A. McCallum. *Introduction to Statistical Relational Learning*, chapter 4, pages 93–126. The MIT Press, 2007.
- [Sutton *et al.*, 2007] C. Sutton, A. McCallum, and K. Rohanimanesh. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8, 2007.
- [Vishwanathan *et al.*, 2006] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. Murphy. Accelerated training of conditional random fields with stochastic meta-descent. In *Proceedings of the 23rd International Conference on Machine Learning*, June 2006.
- [Vogler and Metaxas, 2001] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. In *Proceedings of the Computer Vision and Image Understanding 2001*, 2001.
- [Wu *et al.*, 2007] T. Wu, C. Lian, and J. Y. Hsu. Joint recognition of multiple concurrent activities using factorial conditional random fields. In *Proceedings of the AAAI 2007 Workshop Plan, Activity, and Intent Recognition*, 2007.
- [Wyatt *et al.*, 2007] D. Wyatt, T. Choudhury, J. Bilmes, and H. Kautz. A privacy-sensitive approach to modeling multi-person conversations. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007.
- [Yedidia *et al.*, 2002] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millennium*, pages 239–269. Morgan Kaufmann, 2002.
- [Yedidia *et al.*, 2005] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7), July 2005.