Trajectory-Based Human Activity Recognition Using Hidden Conditional Random Fields

QINGBIN GAO, SHILIANG SUN

Department of Computer Science and Technology, East China Normal University
500 Dongchuan Road, Shanghai 200241, P.R. China
E-MAIL: qbgao10@gmail.com, slsun@cs.ecnu.edu.cn

Abstract:

This paper presents a new method for recognizing trajectory-based human activities. We use a discriminative latent variable model in our proposed method, which considers that human trajectories are made up of some specific motion regimes, and different activities have different switching patterns among the motion regimes. We model the trajectories using Hidden Conditional Random Fields (HCRFs) and the motion regimes act as sub-structures in the model. Experiments using both synthetic and real data sets demonstrate the superiority of our model in comparison with other methods, including Hidden Markov Models (HMM) and Conditional Random Fields (CRFs).

Keywords:

human activity recognition, trajectory classification, hidden conditional random field

1. Introduction

The goal of human activity recognition (HAR) is to understand what people are doing from their position [1], figure [2], motion [3], or other spatiotemporal information derived from video sequences. With the potential for wide applications, HAR has been actively investigated for tens of years. A focus of recent interest is the use of trajectory data, to learn to recognize human behaviors in which a person is engaged over a long period of time [1] [4] [5]. An important application area in this domain is automatic surveillance which is used in busy public places, such as parks, airports, campus, etc. In a surveillance case, HAR aims at characterizing human behaviors and alarming at any illegal or abnormal activities being performed [6]. Other examples in this area include human robot interaction [7], intelligent environment [8], etc.

The challenge of this research is how to recognize trajectories accurately. Methods based on Hidden Markov Models (HMMs) have been widely used for this problem [1] [5] [9]. In these methods, a restrictive, usually unrealistic assumption is made to ensure that observations are conditionally independent given the values of latent variables. However, since human behaviors are complex, it is often more accurately modeled by incorporating long range dependencies and allowing latent variables to depend on several local features.

Conditional Random Fields (CRFs) have proven to be a successful tool for labeling sequence data and have been successfully used for tasks such as part-of-speech tagging and gesture recognition [3] [10]. CRFs condition on the observations without modeling them, and therefore they avoid the independence assumption and can accommodate long range dependencies among observations at different steps. However, CRFs assign each observation in a sequence a label, and they neither capture hidden states nor directly provide a way to estimate the conditional probability of a class label for an entire sequence [11]. This situation leads to their unfitness for trajectory classification tasks.

From daily experience we know that complex human behaviors usually consist of simple motion regimes. For example, the behavior of a person "crossing a park" may be decomposed into "moving east first" and "then moving north". This observation underlies the use of models including hidden states, which have a capacity for capturing intrinsic sub-structures. Hidden Conditional Random Fields (HCRFs) are discriminative latent variable models. HCRFs are based on CRFs, and moreover, they use intermediate hidden variables to model the latent structures of the input domain [12]. Therefore they avoid the independence assumption and have a capacity for capturing sub-structures.

In this paper, we propose a method for trajectory-based human activity recognition based on HCRFs. In our

method, a set of latent variables is introduced to model the unobservable motion regimes and different activities are recognized based on different switching patterns. Our work is related to the switched dynamical HMM (SD-HMM) [1]. However, there are important differences. One most significant difference is that SD-HMM is a generative model while ours is a discriminative one. Another difference is that, in [1], different activities share identical motion regimes, while in our method, the potential of motion regimes of different activities are differently parameterized. We examine our model on both synthetic and real data sets and compare its performance against HMM-based and CRF-based methods. Experimental results show the superiority of our model.

The remainder of this paper is organized as follows. Section 2 gives a brief introduction of CRFs. Section 3 presents the detailed model for human trajectories, including the parameter estimation and inference techniques. Section 4 reports experimental results on both synthetic and real data sets including comparisons with two other methods, HMMs and CRFs. Finally, Section 5 gives conclusions and future research directions.

2. CRFs: A Nutshell

Before describing our model, we give a review of CRFs proposed by [10], which will make HCRFs easier to understand.

CRFs are undirected graphical models (UGMs) which aim at mapping a sequence of observations $\mathbf{X} = \{x_1, x_2, ..., x_m\}$ to a sequence of labels $\mathbf{Y} = \{y_1, y_2, ..., y_m\}$. Let G = (V, E) be a UGM and $\mathbf{Y}$ be indexed by the vertices of G, $\mathbf{Y} = (\mathbf{Y}_v)_{v \in V}$. $(i, j) \in E$ is an edge when there exists a link between nodes y_i and y_j. By defining different edge structures, CRFs can be applied to different tasks. If when conditioned on $\mathbf{X}$, each y_v obeys the Markov property with respect to G, then ($\mathbf{Y}$, $\mathbf{X}$) is a CRF. To define the conditional distribution $P(\mathbf{Y}|\mathbf{X})$, we formulate in terms of maximal cliques, which are the fully connected sub-graphs in a CRF. Let *C* be the set of all maximal cliques of G, and the non-negative potential function of clique *c* be represented as $\phi_c(\mathbf{Y}_c, \mathbf{X}_c)$, then the conditional distribution can be written as

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \prod_{c \in C} \phi_c(\mathbf{Y}_c, \mathbf{X}_c), \tag{1}$$

where $Z(\mathbf{X})$ is a normalization factor which guarantees that the distribution sums to one. Specifically, $Z(\mathbf{X})$ can be computed by summing over all possible configurations of $\mathbf{Y}$

$$Z(\mathbf{X}) = \sum_{\mathbf{Y}} \prod_{c \in C} \phi_c(\mathbf{Y}_c, \mathbf{X}_c). \tag{2}$$

The potential function can be defined arbitrarily according to special tasks. A widely used form is

$$\begin{aligned} \phi_c(\mathbf{Y}_c, \mathbf{X}_c) = \exp(&\sum_{i \in V_c} \lambda_i f_{1,c}(y_i, \mathbf{X}_c) \\ &+ \sum_{(i,j) \in E_c} \beta_{i,j} f_{2,c}(y_i, y_j, \mathbf{X}_c)), \end{aligned} \tag{3}$$

where $f_{1,c}$ is a state feature function which models the observation-label correlations, $f_{2,c}$ is a transition feature function which models the label-label dependencies, and λ_i and $\beta_{i,j}$ are weights to be estimated.

To simplify the formula, we use a feature function $\mathbf{F}_c(\mathbf{Y}_c, \mathbf{X}_c)$ to represent either a state function or a transition function, and λ_i, $\beta_{i,j}$ are represented by a set of weights $\mathbf{w}_c$. Then the forms of potential functions turn to

$$\phi_c(\mathbf{Y}_c, \mathbf{X}_c) = \exp(\mathbf{w}_c \mathbf{F}_c(\mathbf{Y}_c, \mathbf{X}_c)). \tag{4}$$

Put 4 into 1 and 2, and the conditional distribution turns to

$$\begin{aligned} P(Y|X) &= \frac{1}{Z(\mathbf{X})} \prod_{c \in C} \exp(\mathbf{w}_c \mathbf{F}_c(\mathbf{Y}_c, \mathbf{X}_c)) \\ &= \frac{1}{Z(\mathbf{X})} \exp(\sum_{c \in C} (\mathbf{w}_c \mathbf{F}_c(\mathbf{Y}_c, \mathbf{X}_c))), \end{aligned} \tag{5}$$

where $Z(\mathbf{X})$ is

$$Z(\mathbf{X}) = \sum_{\mathbf{Y}} \exp(\sum_{c \in C} \mathbf{w}_c \mathbf{F}_c(\mathbf{Y}_c, \mathbf{X}_c)). \tag{6}$$

3. Human Activity Recognition

3.1. Trajectory Model

Our task is to learn a mapping from a sequential trajectory $\mathbf{X}$ to a single activity label y. Formally, each trajectory $\mathbf{X}$ is a vector of observations, $\mathbf{X} = \{x_1, x_2, ..., x_T\}$, and each observation x_t implies the displacement of a person from time t-1 to time t ($t = 1, ..., T$). x_t is represented by a D-dimension local feature, $\phi(x_t) \in R^D$. Each y is one of the activity labels represented by a set of constants. Assume we have $\mathcal{Y}$ activities, then $y \in \{1, 2, ..., \mathcal{Y}\}$. Based on the fully observable CRFs described in previous section, we introduce a vector of latent variables $\mathbf{H} = \{h_1, h_2, ..., h_T\}$ to model the intermediate motion regimes contained in complex activities [12]. Each h_t is a member of a finite set $\mathcal{H}$, which is the collection of all possible motion regimes. For example, if we assume that all trajectories are made up of five motion regimes, which are "stopped", "moving east", "moving west", "moving south", "moving north",

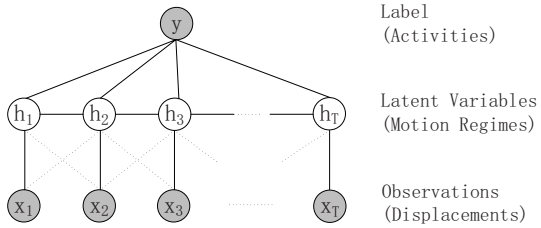


Figure 1. The chain structure HCFR for trajectory recognition.

then $\mathcal{H}$ contains all five of them and each h_t corresponds to one of them. Let's still consider the UMG G = (V, E), In a HCRF, the latent variables $\mathbf{H} = \{h_1, h_2, ..., h_T\}$ correspond to vertices in the graph and $(i, j) \in E$ is an edge when there exists a link between variables h_i and h_j. It's worth noticing that the presence of an edge between two vertices in an UMG implies the dependencies between the random variables represented by these vertices. By defining different edge structures, HCRFs can be applied to different domains. Returning to our activity recognition task, which intrinsically is a temporal classification problem. Based on the general HCRFs and considering the specific characters of our task, we define a linear-chain structure in order to capture the temporal dynamics (see Figure 1). In this structure, the maximal cliques include pairs of neighboring states (h_{t-1}, h_t). The connectivity between each latent state and observations, which implies the long range dependencies among observations, is unrestricted. We introduce a window size w to define the connectivity. $w = 0$ indicates that the current state is only depend on the current observation, while $w > 0$ indicates that neighbor observations from $t - w$ to $t + w$ are also used.

Given the above definitions, first we model human trajectories in a CRF way as

$$P(y, \mathbf{H}|\mathbf{X}; \theta) = \frac{1}{Z(\mathbf{X}; \theta)} \exp(\sum_{t=1}^{T} F(y, h_{t-1}, h_t, \mathbf{X}; \theta)), \tag{7}$$

marginalizing out the latent variables $\mathbf{H} = \{h_1, h_2, ..., h_T\}$ yields the following HCRF form

$$\begin{aligned} P(y|\mathbf{X}; \theta) &= \sum_{\mathbf{H}} P(y, \mathbf{H}|\mathbf{X}; \theta) \\ &= \frac{1}{Z(\mathbf{X}; \theta)} \sum_{\mathbf{H}} (\exp(\sum_{t=1}^{T} F(y, h_{t-1}, h_t, \mathbf{X}; \theta))), \end{aligned} \tag{8}$$

where the normalization factor $Z(\mathbf{X})$ take the form as

$$Z(\mathbf{X}; \theta) = \sum_{y', \mathbf{H}} \exp(\sum_{t=1}^{T} F(y', h_{t-1}, h_t, \mathbf{X}; \theta)). \tag{9}$$

We define the feature function F as follows

$$\begin{aligned} F(y, h_{t-1}, h_t, \mathbf{X}; \theta) = &\sum_{a \in A} \theta_a f_a(y, h_{t-1}, h_t, \mathbf{X}) \\ &+ \sum_{b \in B} \theta_b f_b(y, h_t, \mathbf{X}), \end{aligned} \tag{10}$$

where A is the set of edge features and B is the set of node features, f_a is a predefined transition function which depends on a pair of latent variables and f_b is a predefined state function which depends on a single latent variable in the model. $\theta = \{\theta_a, \theta_b\}$ are parameters to be estimated from training data.

3.2. Parameter Estimation

Our training data set consists of N labeled trajectories, $\mathcal{T} = \{(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), ..., (\mathbf{X}_N, y_N)\}$. The parameters can be obtained by optimizing the conditional log-likelihood of the training data

$$L(\theta) = \sum_{i=1}^{N} L_i(\theta) = \sum_{i=1}^{N} logP(y_i|\mathbf{X}_i; \theta). \tag{11}$$

While in practice, we often regularize the problem by optimizing a penalized likelihood: $L(\theta) + R(\theta)$, where $R(\theta)$ is the log of a Gaussian prior with variance σ^2, i.e., $R(\theta) \sim \exp(-\frac{1}{2\sigma^2} \parallel \theta \parallel^2)$ [13].

Likelihood maximization leads to an optimization task, which can be solved using gradient ascent methods. In our paper, we solve this problem using a limited-memory variable-metric gradient ascent method (BFGS) [14].

3.3. Classification

For testing, given a new observed trajectory X, we want to classify it into one of the activities $y^* \in \mathcal{Y}$ which maximizes the conditional probability

$$y^* = \arg\max_{y \in \mathcal{Y}} P(y|X, \theta^*), \tag{12}$$

where the values of θ^* are learned from the training data.

Since HCRFs can be considered as UMGs, the inference tasks can be solved using belief propagation [15].

4. Experiments

We run a variety of experiments using both synthetic and real data. To evaluate the performance of our model, comparisons with other approaches are also given.

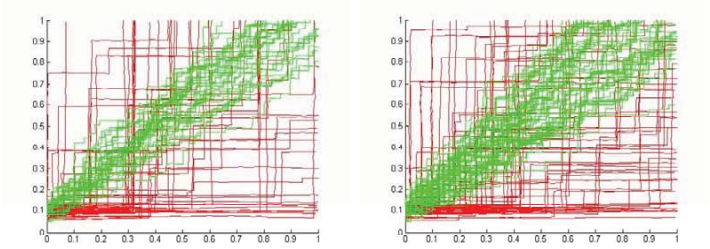
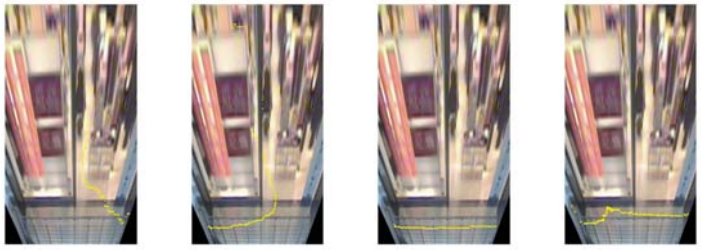
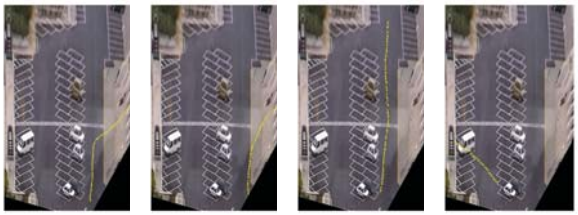


Figure 2. Two synthetic activities sharing the same motion regimes, with different switching patterns. Training data(left), testing data(right).

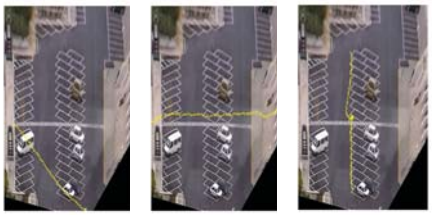


(a) (b) (c) (d)

Figure 3. Examples of the four activities defined for the shopping scenario: (a) entering; (b) leaving; (c) passing; (d) browsing.



(a) (b) (c) (d)



(e) (f) (g)

Figure 4. Examples of the seven activities defined for the campus scenario: (a) entering building; (b) leaving building; (c) walking along; (d) crossing park up; (e) crossing park down; (f) passing through; (g) wandering.

4.1. Synthetic Data

We first run a simple synthetic example in an ideal scenario, which aims at demonstrating the effectiveness of our model. In this experiment, we consider two activities shown in Figure 2. The two activities depicted in red and green share two motion regimes: moving horizontally and moving vertically. The mean of horizontal displacements is $\mathbf{T}_1 = \begin{bmatrix}0.02 & 0\end{bmatrix}^T$, and the mean of vertical displacements is $\mathbf{T}_2 = \begin{bmatrix}0 & 0.02\end{bmatrix}^T$. Corresponding covariances are $\mathbf{Q}_1 = \mathbf{Q}_2 = 10^{-3}\mathtt{I}$. The only difference between the two activities resides on the switching patterns. The red activity has a low probability of switching between different motion regimes, while the green activity has identical probabilities of switching at all instants. Respectively, for the red and green activities, the transition matrices are

$$\mathbf{B}_1 = \begin{bmatrix}0.95 & 0.05\\ 0.05 & 0.95\end{bmatrix} \quad \mathbf{B}_2 = \begin{bmatrix}0.5 & 0.5\\ 0.5 & 0.5\end{bmatrix}$$

Given the above parameters, we generate 100 training trajectories and 100 testing trajectories using HMMs. The reason why we use HMMs to generate the synthetic data is that, discriminative models condition on observations without modeling them, thus, without knowledge of observations, they are unable to generate data.

4.2. Experimental Results on Synthetic Data

From the way we generate this synthetic data set, it is clear to see that each frame in a trajectory corresponds to only one motion regime. Thus, we only run the experiment using a HCRF with window size $w = 0$. Finally, the classification accuracy obtained on the testing data is 100%, showing that our model possibly have a capacity to recognize trajectories.

4.3. Real Data

We consider two scenarios in our experiments with real data, which include a shopping center and a university campus. In the shopping center scenario, four human activities have been predefined. While in the campus scenario, seven human activities have been predefined. Figure 3 shows examples of trajectories in the shopping center scenario and Figure 4 shows examples of trajectories in the campus

scenario.

A notable point in the experiments with real data is that, since formula (8) has to marginalize out the latent variables, our model works with a finite number of motion regimes. Estimating the number of motion regimes is a model selection task, and lots of exact methods have already existed for this task [16]. Since model selection is not the focus of our paper, we employ the model selection result of [1]. Thus, for the shopping data, we define five motion regimes: "stopped", "moving north", "moving south", "moving east", and "moving west". While for the campus data, we define nine motion regimes: "stopped", "moving north", "moving north-east", "moving east", "moving south-east", "moving south", "moving south-west", "moving west", and "moving north-west".

Another notable point is that, in original data sets, each element in a trajectory sequence implies the position of a human. In order to use the displacement features, we perform some preprocessing. Representing an original trajectory by $\mathbf{P} = \{p_0, p_1, ..., p_T\}$, thus our input trajectory will be $\mathbf{X} = \{x_1, x_2, ..., x_T\}$, where $x_t = p_t - p_{t-1}$ ($t = 1, ..., T$).

After preprocessing, we get 53 available trajectories in the shopping scenario and 143 available trajectories in the campus scenario.

4.4. Experimental Results on Real Data

We consider two different procedures for splitting the available data into training and testing sets: 1) a single training/testing splitting; 2) a complete p-fold cross validation. For The shopping scenario, the first procedure picks three samples of each activity to generate the training set, and the rest samples generate the testing set. While for the campus scenario, the first procedure splits all available data into two disjoint sets with each set containing 50% of all data. The second procedure performs a complete ten-fold cross validation for both scenarios.

Experiment on same data sets, we evaluate our model with varying levels of long range dependencies (with different window size) and compare the performance with HMM and CRF models.

In our HMM experiments, we consider the switched dynamical HMM (SD-HMM) proposed in [1], which is actually a two layer hierarchical HMM. The lower layer consists of a bank of Gaussians which imply the motion regimes and the higher layer models the switching among the motion regimes.

Though CRFs do not directly provide a way to map an entire sequence to a class label, with some tricks, they still work. In our CRF experiments, each input trajectory sequence $\mathbf{X} = \{x_1, x_2, ..., x_T\}$ is associated with a sequence of labels $\mathbf{Y} = \{y_1, y_2, ..., y_T\}$. In training data, the label sequences are generated by repeating the target activity label y T times. For a testing trajectory sequence, the final activity label assigned is the label which appeared most frequently in the decoded sequence. We come up with this idea from the literature of gesture recognition [12].

Table 1. Comparison of Recognition Performance for Shopping Scenario

Methods	1st split procedure	2nd split procedure
HMM	70.73%	70.73%
CRF w=0	12.20%	12.77%
CRF w=1	17.07%	10.67%
HCRF w=0	**85.37**%	**85.11**%
HCRF w=1	80.49%	76.60%
HCRF w=2	80.49%	80.85%
HCRF w=3	75.61%	78.72%

Table 2. Comparison of Recognition Performance for Campus Scenario

Methods	1st split procedure	2nd split procedure
HMM	82.61%	87.60%
CRF w=0	10.14%	9.302%
CRF w=1	13.04%	10.85%
HCRF w=0	88.41%	92.25%
HCRF w=1	**91.30**%	**93.02**%
HCRF w=2	78.26%	89.92%
HCRF w=3	68.12%	87.60%

Table 1 shows the results for the shopping experiments and Table 2 shows the results for the campus experiments. As we can see, our approach performs better than the HMM-based and CRF-based methods.

From the results in Table 1, we can see that our approach performs best at window size 0. Though this implies that the independence assumption is correct, our model still performs better than HMMs. From the results in Table 2, we can see that increasing the window size from 0 to 1 improves the performance of our model. This implies that incorporating appropriate degree of long range dependencies is helpful. However, we also see that further increasing the window size does not improve the performance.

It is a foregone conclusion that CRFs achieve bad results. We try to recognize human activities by modeling the intermediate motion regimes, but CRFs have no capacity to capture sub-structures.

5. Conclusions

In this work, we have presented a method for recognizing trajectory-based human activities. Our method models trajectories using HCRFs while shared motion regimes act as latent variables. Thus different trajectories are recognized based on different switching patterns. To validate our model, we run a variety of experiments using both synthetic and real data and compare the performance with other methods. Experimental results have shown that our method outperforms both HMM-based methods and CRF-based methods.

For future research, the proposed method can be embedded with model selection methods. In this way, the number of latent variables can be obtained automatically and the model will be more flexible. Another possible direction is extending the proposed method to infinite Gaussian mixture models [17]. In this way, techniques of variational inference will play an important role.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Project 61075005, and the Fundamental Research Funds for the Central Universities.

References

[1] J. C. Nascimento, A. T. Figueiredo and J. S. Marques, "Trajectory classification using switched dynamical hidden Markov models", IEEE Transactions on Image Processing, pp. 1338-1348, 2010

[2] N. Vaswani, A. R. Chowdhury and R. Chellappa, "Shape activity: A continuous state HMM for moving/deforming shapes with application to abnormal activity detection", IEEE Transactions on Image Processing, pp. 1603-1616, 2005

[3] C. Sminchisescu, A. Kanaujia, Z. Li and D. Metaxas, "Conditional models for contextual human motion recognition", Proceedings of the 10th IEEE International Conference on Computer Vision, pp. 1808-1815, 2005

[4] L. Liao, D. Fox and H. Kautz, "Hierarchical conditional random fields for GPS-based activity recognition", Springer Tracts in Advanced Robotics, pp. 487-506, 2007

[5] F. Bashir, A. Khokhar and D.Schonfeld, "Object trajectory-based activity classification and recognition using hidden Markov models", IEEE Transactions on Image Processing, pp. 1912-1919, 2007

[6] W. Hu, T. Tan, L. Wang and S. Maybank, "A survey on visual surveillance of object motion and behaviors", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, pp. 334-352, 2004

[7] M. Bennewitz, W. Burgard, G. Cielniak and S. Thrun, "Learning motion patterns of people for compliant robot motion", International Journal of Robotics Research, pp. 31-48, 2005

[8] B. Brumitt, B. Meyers, J. Krumm, A. Kern and S. Shafer, "Easyliving: Technologies for intelligent environments", Handheld and Ubiquitous Computing, pp. 97-119, 2000

[9] H. H. Bui, D. Q. Phung and S. Venkatesh, "Hierarchical hidden Markov models with general state hierarchy", Proceedings of the 9th National Conference on Artificial Intelligence, pp. 324-329, 2004

[10] J. Lafferty, A. McCallum and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", Proceedings of the 18th International Conference on Machine Learning, pp. 282-289, 2001

[11] S. B. Wang, A. Quattoni, L. P. Morency and D. Demirdjian, "Hidden conditional random fields for gesture recognition", Proceedings of the 19th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1521-1527, 2006

[12] A. Quattoni, S. Wang, L. P. Morency, M. Collins and T.Darrell, "Hidden conditional random fields", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1848-1853, 2007

[13] L. P. Morency, A. Quattoni and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition", Proceedings of the 20th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2007

[14] A. McCallum, "Efficiently inducing features of conditional random fields", Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence, pp. 403-410, 2003

[15] C. M. Bishop, Pattern Recognition and Machine Learning, Springer Science, Spring Street, New York, USA, 2006.

[16] A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 381-396, 2002

[17] S. Sun and X. Xu, "Variational inference for infinite mixtures of gaussian processes with applications to traffic flow prediction", IEEE Transactions on Intelligent Transportation Systems, pp. 466-475, 2011