TRAJECTORY-BASED HUMAN ACTIVITY RECOGNITION USING HIDDEN CONDITIONAL RANDOM FIELDS

Qingbin Gao, Shiliang Sun

Department of Computer Science and Technology
East China Normal University, Shanghai, China
qbgao10@gmail.com, slsun@ecnu.edu.cn

ABSTRACT

This paper presents a 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 demonstrate the superiority of our model in comparison with other methods, including Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs).

Index Terms— 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]. From daily experience we know that 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 Conditional Random Fields (CRFs) [6], and moreover, they use intermediate hidden variables to model the latent structures of the input domain [7]. Therefore they avoid the unrealistic independence assumption of Hidden Markov Models (HMMs) 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, we use a set of latent variables to model the unobservable motion regimes and different activities are recognized which use different switching patterns among the motion regimes. 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.

2. HUMAN ACTIVITY RECOGNITION

2.1. 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 [6], 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 [7]. Each h_t is a member of a finite set $\mathcal{H}$, which is the collection of all possible motion regimes. From above definitions, it is clear to see that a trajectory recognition task is intrinsically a temporal classification problem. Based on the general HCRFs [7] and considering the specific characters of our task, we define a linear-chain structure in order to capture the temporal dynamics (see Fig.1). In this structure, the fully connected 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 trajec-

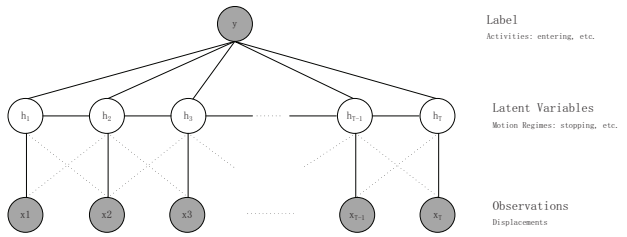


Fig. 1. The chain structure HCFR for trajectory recognition.

tories in a CRF way as

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

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

$$\begin{aligned} P(y|\mathbf{X}; \theta) &= \sum_{\mathbf{H}} P(y, \mathbf{H}|\mathbf{X}; \theta) \\ &= \frac{1}{Z(\mathbf{X}; \theta)} \sum_{\mathbf{H}} (\exp(\sum_{t=1}^{T} F(y, h_{t-1}, h_t, \mathbf{X}; \theta))), \end{aligned} \tag{2}$$

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

$$Z(\mathbf{X}; \theta) = \sum_{y', \mathbf{H}} \exp(\sum_{t=1}^{T} F(y', h_{t-1}, h_t, \mathbf{X}; \theta)). \tag{3}$$

We define the feature function F as follows

$$\begin{aligned} F(y, h_{t-1}, h_t, \mathbf{X}; \theta) = &\sum_{a \in A} \theta_a f_a(y, h_{t-1}, h_t, \mathbf{X}) \\ &+ \sum_{b \in B} \theta_b f_b(y, h_t, \mathbf{X}), \end{aligned} \tag{4}$$

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.

2.2. 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

$$L(\theta) = \sum_{i=1}^{N} L_i(\theta) = \sum_{i=1}^{N} logP(y_i|\mathbf{X}_i; \theta). \tag{5}$$

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., $R(\theta) \sim \exp(-\frac{1}{2\sigma^2} \parallel \theta \parallel^2)$ [8].

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) [3].

2.3. 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

$$y^* = \arg\max_{y \in \mathcal{Y}} P(y|X, \theta^*), \tag{6}$$

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

Since HCRFs can be considered as undirected graphical models (UMGs), the inference tasks can be solved using belief propagation.

3. 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.

3.1. Synthetic Data

We first run a simple synthetic example in an ideal scenario [1], which aims at demonstrating the effectiveness of our model. In this experiment, we consider two activities shown in Fig.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. 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.

From the way we generate this synthetic data set, it is clear to see that each frame in a trajectory corresponds to a motion regime. Thus, we only run the experiment using a HCRF with window size $w = 0$. Finally, the classification accuracy

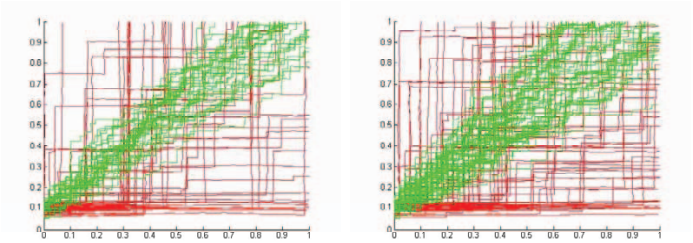
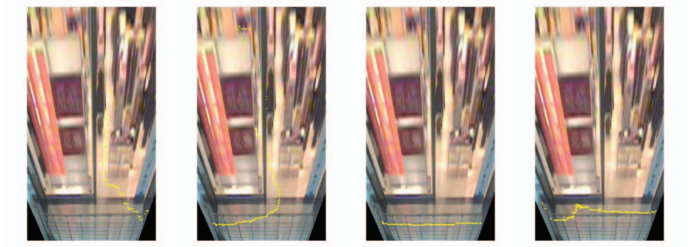


Fig. 2. Two synthetic activities sharing the same motion regimes, with different switching patterns. Training data(left), testing data(right).



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

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

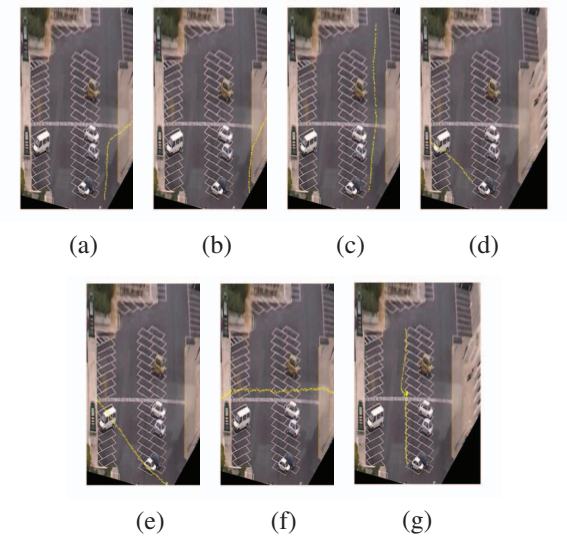
obtained on the testing data is 100%, showing that our model possibly have a capacity to recognize trajectories.

3.2. Real Data

3.2.1. Description

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. Fig.3 shows examples of trajectories in the shopping center scenario and Fig.4 shows examples of trajectories in the campus scenario.

A notable point in the experiments with real data is that, since equation (2) 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 exact methods have already existed for this task [9]. 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".



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

(e) (f) (g)

Fig. 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.

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

3.2.2. Classification Results

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.

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

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 1. Comparison of Recognition Performance for the Shopping 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 2. Comparison of Recognition Performance for the Campus Scenario

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 [7].

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.

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.

4. 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. 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. In this way, techniques of variational inference will play an important role.

5. 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*, vol. 19, no. 5, 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*, vol. 14, no. 10, 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*, vol. 28, 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*, vol. 16, no. 7, pp. 1912–1919, 2007.

[6] 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.

[7] A. Quattoni, S. Wang, L. P. Morency, M. Collins, and T.Darrell, "Hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1848–1853, 2007.

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

[9] A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 381–396, 2002.