Spatio-Temporal Triangular-Chain CRF for Activity Recognition

Congqi Cao
National Laboratory of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
Beijing, China, 100190
congqi.cao@nlpr.ia.ac.cn

Yifan Zhang
National Laboratory of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
Beijing, China, 100190
yfzhang@nlpr.ia.ac.cn

Hanqing Lu
National Laboratory of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
Beijing, China, 100190
luhq@nlpr.ia.ac.cn

ABSTRACT

Understanding human activities in video is a fundamental problem in computer vision. In real life, human activities are composed of temporal and spatial arrangement of actions. Understanding such complex activities requires recognizing not only each individual action, but more importantly, capturing their spatio-temporal relationships. This paper addresses the problem of complex activity recognition with a unified hierarchical model. We expand triangular-chain CRFs (TriCRFs) to the spatial dimension. The proposed architecture can be perceived as a spatio-temporal version of the TriCRFs, in which the labels of actions and activity are modeled jointly and their complex dependencies are exploited. Experiments show that our model generates promising results, outperforming competing methods significantly. The framework also can be applied to model other structured sequential data.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

Keywords

Activity recognition, Hierarchical model, Joint learning, Spatio-temporal dependencies, Triangular-chain CRFs

1. INTRODUCTION

Activity recognition is one of the most popular research fields in computer vision which has spawned a rich literature [11, 4, 16, 3] due to its promising application in video monitoring, behavioral analysis and artificial intelligence. In real life, human activities are complex since humans are capable of performing multiple simple actions simultaneously. Complex human activities are composed of temporal and spatial arrangement of atomic actions and each atomic action is composed by a temporal arrangement of body poses [9]. Understanding such complex activities requires recognizing not only each individual action, but more importantly, capturing their spatio-temporal relationships.

Much of the initial work in activity recognition has been focused on analyzing either high-level coarse grained activities or mid-level fine grained actions. However, the recognition result of each level is useful in practice which allows multiple layers of abstraction. Besides, research in congnitive psychology has shown that human perceive activities as hierarchical structures [7]. Therefore it is necessary to recognize activities and actions jointly in a hierarchical model. Recently, there has been significant interest in it. [13] presented a framework for modeling complex composite activity using stochastic grammar. However, the grammar induction is very important for this model. It will be more robust and more flexible if the model could be learned from data fully automatically, instead of using expert's knowledge. [9] proposed a compositional hierarchical model to recognize human activities and actions by formulating an energy minimization problem which is similar to a latent structural SVM case. The energy terms are associated to the activity, actions and poses, as well as temporal transitions between actions and poses. However, the transitions between actions in [9] are independent of activity, and there is no energy terms to model the influence of pose observations to activity recognition. [7] trained a hierarchical model based on HMMs and a context-free grammar using HTK toolkit which is a speech recognition engine to model human activities as temporally structured processes. [7] demonstrated that applying techniques borrowed from speech recognition is feasible since human activities and speech have similar inherently hierarchical nature. The limitation of this model is that it can not capture spatial composition of actions. Besides, it is based on HMMs which make strong independence assumption of observations. Such assumption ignores the multiple interacting features and long-range dependencies of the observation [8].

Different from the existing deep learning methods [15, 12] which mainly focused on feature extraction and representation, our work aims at modeling spatial and temporal relationships of poses and actions in video. For modeling sequential data, [5] proposed TriCRFs shown in Figure 1, a unified probabilistic model jointly representing the sequence and meta-sequence labels. TriCRFs have been successfully used in named entity recognition and dialog act classification

MM'15, October 26–30, 2015, Brisbane, Australia.
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806304.

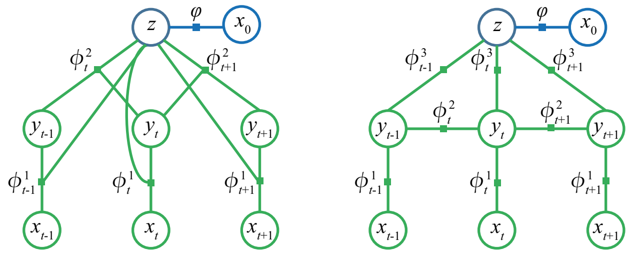

Figure 1: Two structures of triangular-chain CRFs. The transition potentials and observation potentials are dependent of z in the left structure and independent of z in the right structure.

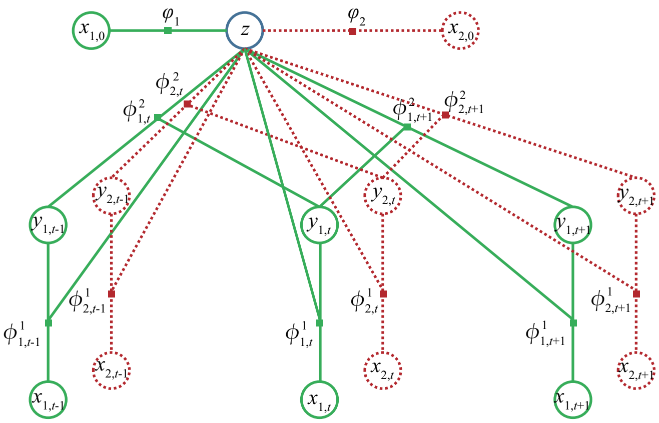

Figure 2: The structure of spatio-temporal triangular-chain CRFs which could capture spatial and temporal information in sequential data.

for spoken language understanding [5, 6, 14]. However, no applications on computer vision tasks have been observed.

As mentioned above, modeling complex human activity is still a challenging work because of its complicated spatio-temporal relationships and multiple variabilities among actions. In this paper, we propose a novel framework for activity and action recognition which could model the spatio-temporal relationships of multi-level labels jointly in a unified hierarchical model shown in Figure 2. Our model is based on triangular-chain CRFs. Furthermore, we expand traditional temporal TriCRFs to the spatial dimension. The model could both explicitly encode dependencies and preserves uncertainty between actions and activity. As far as we know, we are the first to utilize spatio-temporal TriCRFs in complex activity recognition. Experiments on composable human activity dataset show that our approach outperforms other methods which demonstrates the effectiveness of our framework for modeling spatio-temporal relationships in activity recognition. This model could also be applied to other structured sequential modeling problems.

2. FRAMEWORK

2.1 Background

Human activity, speech and natural language are all sequential data. Many problems of sequential data can be treated as sequential labeling or sequence classification [5]. More specifically, sequential labeling is a problem of predicting a sequence of label $\mathbf{y}$ given an input sequence of observations $\mathbf{x}$ which could be formulated as learning $p(\mathbf{y}|\mathbf{x})$. Sequence classification is a problem of predicting a single label z given $\mathbf{x}$, formulated as learning $p(z|\mathbf{x})$. TriCRFs solve the two correlated problems jointly by predicting the best labels $(\hat{\mathbf{y}}, \hat{z})$ using $p(\mathbf{y}, z|\mathbf{x})$ trained from data $(\mathbf{x}^{(n)}, \mathbf{y}^{(n)}, z^{(n)})_{n=1}^N$.

Linear-chain conditional random fields (CRFs) was proposed to segment and label data by Lafferty in 2001 [8]. As being discriminative that directly model the global conditional distribution, CRFs offer several advantages over generative models such as hidden Markov models and stochastic grammars which use non discriminant criterion, including the ability to relax strong independence assumptions. CRFs could also avoid the fundamental label bias limitation of maximum entropy Markov models (MEMMs) and other locally normalized discriminative models [8].

Spatio-temporal triangular-chain CRFs have all the advantages of CRFs. Besides, Spatio-temporal triangular-chain CRFs can capture the spatial and temporal relationships of atomic actions to model multi-level labels of activity at the same time. It could model the complex dependencies between activity, actions and pose observations. Especially, the transitions between actions are dependent of activity, and there is potential terms to model the influence of pose observations to activity recognition. Our model is shown in Figure 2. Let R be the number of spatial regions. For each region, the complex relationships between actions and activities are modeled by temporal tiangular-chain CRFs. Different regions are fused together in the activity level. The figure illustrates the $R = 2$ case.

2.2 Formulation

In order to take spatial information into account, we extend TriCRFs to spatial dimension. Spatio-temporal triangular-chain CRFs are defined as follows.

$$p_\lambda(\mathbf{y}, z|\mathbf{x}) \quad = \quad \prod_{r=1}^{R} p_{\lambda_r}(\mathbf{y}_r, z|\mathbf{x}_r) \qquad (1)$$

where we assume that the probability $p_\lambda(\mathbf{y}, z|\mathbf{x})$ is proportional to the product of every region's probability $p_{\lambda_r}(\mathbf{y}_r, z|\mathbf{x}_r)$.

$$p_{\lambda_r}(y_r, z|x_r) = \frac{1}{Z(x_r)} \cdot \prod_{t=1}^{T} (\phi_{r,t}(z, y_{r,t}, y_{r,t-1}, x_r))\varphi_r(z, x_r) \qquad (2)$$

where $\phi_{r,t}$ and φ_r are the potentials of spatio-temporal TriCRFs. $Z(\mathbf{x}_r)$ is for normalization which defined as follows.

$$Z(\mathbf{x}_r) \triangleq \sum_{\mathbf{y}_r, z} \prod_{t=1}^{T} \phi_{r,t}(z, y_{r,t}, y_{r,t-1}, \mathbf{x}_r)\varphi_r(z, \mathbf{x}_r) \qquad (3)$$

As illustrated in Figure 2 and formulated with the equations, $\phi_{r,t}$ is time-dependent and φ_r is time-independent. Specifically, $\phi_{r,t}$ could be partitioned into z-dependent and z-independent factors. φ_r plays as a prior role.

$$\begin{aligned} \varphi_r(z, \mathbf{x}_r) \quad &= \quad \exp(\sum_k \lambda_{r,k}^0 f_{r,k}^0(z, \mathbf{x}_{r,0})) && (4) \\ \phi_{r,t}(z, y_{r,t}, y_{r,t-1}, \mathbf{x}_r) \quad &= \quad \phi_{r,t}^d(z, y_{r,t}, y_{r,t-1}, \mathbf{x}_r) \\ &\quad\quad \cdot \phi_{r,t}^i(y_{r,t}, y_{r,t-1}, \mathbf{x}_r) && (5) \end{aligned}$$

where $\mathbf{x}_{r,0}$ is an observation feature vector for classifying z. $f_{r,k}$ stands for a real-valued feature function and $\lambda_{r,k}$ is a weight parameter.

z-dependent factor $\phi^d_{r,t}$ can be further partitioned into observation item and transition item dependent of z.

$$\phi^d_{r,t}(z, y_{r,t}, y_{r,t-1}, \mathbf{x}_r) = \phi^1_{r,t}(z, y_{r,t}, \mathbf{x}_r)\phi^2_{r,t}(z, y_{r,t}, y_{r,t-1}) \tag{6}$$

$$\phi^1_{r,t}(z, y_{r,t}\mathbf{x}_r) = \exp(\sum_k \lambda^1_{r,k} f^1_{r,k}(z, y_{r,t}, \mathbf{x}_r)) \tag{7}$$

$$\phi^2_{r,t}(z, y_{r,t}, y_{r,t-1}) = \exp(\sum_k \lambda^2_{r,k} f^2_{r,k}(z, y_{r,t}, y_{r,t-1})) \tag{8}$$

z-independent factor $\phi^i_{r,t}$ also can be partitioned into observation item and transition item independent of z.

$$\phi^i_{r,t}(y_{r,t}, y_{r,t-1}, \mathbf{x}_r) = \phi^1_{r,t}(y_{r,t}, \mathbf{x}_r)\phi^2_{r,t}(y_{r,t}, y_{r,t-1}) \tag{9}$$

$$\phi^1_{r,t}(y_{r,t}, \mathbf{x}_r) = \exp(\sum_k \lambda^1_{r,k} f^1_{r,k}(y_{r,t}, \mathbf{x}_r)) \tag{10}$$

$$\phi^2_{r,t}(y_{r,t}, y_{r,t-1}) = \exp(\sum_k \lambda^2_{r,k} f^2_{r,k}(y_{r,t}, y_{r,t-1})) \tag{11}$$

2.3 Inference and Parameter Estimation

Spatio-temporal TriCRFs have efficient training and decoding algorithms based on dynamic programming. And being a convex optimization problem, parameter estimation is guaranteed to find the global optimum.

For inference, the marginal probability distributions $p_{\lambda_r}(z, y_{r,t}|\mathbf{x}_r)$, $p_{\lambda_r}(z, y_{r,t}, y_{r,t-1}|\mathbf{x}_r)$, $p_{\lambda_r}(z|x_r)$ and the partition function $Z(\mathbf{x}_r)$ are calculated via the forward-backward algorithm introduced in [5]. The viterbi decoding algorithm for each chain is also the same as [5].

For parameter estimation, using conditional maximum log-likelihood criterion, the objective function is formulated as follows.

$$\begin{aligned} L(\lambda) &= \sum_{r=1}^{R}\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_{r,k} f_{r,k}(z^{(n)}, y^{(n)}_{r,t}, y^{(n)}_{r,t-1}, \mathbf{x}^{(n)}_r) \\ &- \sum_{r=1}^{R}\sum_{n=1}^{N} \log Z(\mathbf{x}^{(n)}_r) - \sum_{r=1}^{R}\sum_{k=1}^{K} \frac{\lambda^2_{r,k}}{2\sigma^2} \end{aligned} \tag{12}$$

where the last term functions as a regularization term to avoid over-fitting.

The parameters are learned through differentiating the objective function. A limited memory version of the quasi-Newton method is used to optimize the parameters.

$$\begin{aligned} \frac{\partial L}{\partial \lambda_{r,k}} &= \sum_{n=1}^{N}\sum_{t=1}^{T} f_{r,k}(z^{(n)}, y^{(n)}_{r,t}, y^{(n)}_{r,t-1}, \mathbf{x}^{(n)}_r) \\ &- \sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{z', y_r, y'_r} f_{r,k}(z', y_r, y'_{r,}, \mathbf{x}^{(n)}_r) p_{\lambda_r}(z, y_r, y'_{r,}|\mathbf{x}^{(n)}_r) \\ &- \frac{\lambda_{r,k}}{\sigma^2} \end{aligned} \tag{13}$$

The pseudo-likelihood parameters are used for initialization [1]. We modified the code published by Minwoo Jeong [5, 6] to construct our spatio-temporal TriCRFs by taking spatial information into account.

3. EXPERIMENTS

3.1 Composable Activity Dataset

We evaluate our method on Composable Activity dataset which is introduced by Ivan et al. in [9]. The dataset consists of 693 videos containing activities in 16 classes performed by 14 actors. Each activity is composed by spatio-temporal combinations of atomic actions. For instance, *composed activity 2* is composed by *walking*, *picking object*, *put an object*, *erasing board* in space and time. There are total 26 types of atomic actions. The dataset offers a global annotation at the activity level for each video, as well as per-frame annotations of the atomic actions with an array indicates the associated body region (right arm, left arm, right leg and left leg).

Table 1: Recognition accuracy of different R

Algorithm	Region number	Accuracy
Our method	1	0.759
Our method	2	**0.790**
Our method	4	0.772

Table 2: Recognition accuracy comparison

Algorithm	Codebook size	Accuracy
Our method	200(fixed)	**0.790**
BoW	200(fixed)	0.672
BoW	600(fixed)	0.623
H-BoW [9]	200(fixed)	0.742
H-BoW [9]	600(fixed)	0.716
HMM	200(fixed)	0.765
HMM	600(fixed)	0.723

3.2 Features

In order to facilitate a fair comparison, we follow the same experimental settings as [9]. Performance is evaluated in leave-one-subject-out experiment setup. Observed features are extracted using the method in [2, 10] from RGB-D videos which include relative location between body joints, angles between limbs and angles between limbs and plans spanned by body parts. Using the feature extraction code in [9], right arm, left arm, right leg and left leg are represented by a 21-dimension feature vector respectively. [9] quantized the observed features of each body part into M clusters of poses. The three techniques for comparison are a BoW representation plus a lineal SVM classifier (BoW-approach), a version of authors' hierarchical model without learning the pose dictionary (H-BoW approach), and a Hidden Markov Model approach (HMM approach). It is more equitable to compare our spatio-temporal model with the three techniques which also quantize the observations using k-means. The classification result of our model is higher than the other methods including the version of model in [9] with fixed pose dictionary. Experiments in [9] showed that jointly learning pose dictionary with actions and activities could improve the accuracy. Our model also can include a third semantic level. It is straightforward to add another layer in our model to represent pose and it is feasible to use other more appropriate methods to obtain a pose dictionary. The performance could probably be improved further.

3.3 Results and Analysis

We divide the whole body into R regions, where $R = \{1, 2, 4\}$. For the case $R = 4$, the body is divided into 4 regions which are right arm, left arm, right leg and left leg; There are 4 chains in spatio-temporal TriCRFs. For the case $R = 2$, the body is divided into upper body and lower body; The upper body is composed of right arm and left arm, while the lower body is composed of right leg and left leg; We concatenate the right arm observed features and left

Table 3: The comparison of robustness. The table lists the decreases of accuracies by randomly select a part to be occluded in every testing sequence.

Algorithm	Our method	Method in [9]	BoW	HMM
Decrease	**0.0527**	0.072	0.125	0.103

arm observed features together to represent the upper body part; The same setting is for the lower body part; There are 2 chains in spatio-temporal TriCRFs. For the case $R = 1$, the body is treated as a whole part and there is only one temporal TriCRFs chain.

For spatio-temporal TriCRFs, the accuracy of action labeling is 33.8%, 55.7%, 56.6% for R=1,2,4 respectively. The accuracies of activity recognition are shown in Table 1. Dividing the whole body into spatial regions does increase the recognition accuracies of actions and activities at the same time, which on the other side demonstrates that our spatio-temporal TriCRFs model captures the relationships in space successfully. The reason for why 2-region setting is better than 4-region setting is probably because the action labels are not fine-grained for each body part. In different actions, some body parts may show the same pose, resulting in that there is relatively small between-action-class distance, which is less discriminative to predict the action and activity label. The unsupervised clustering method used to quantize observations also should be improved.

As shown in Table 2. The accuracy of spatio-temporal TriCRFs shows clear improvement over the competing methods. The strength of our model can be measured against BoW, H-BoW and HMM with the same setting to obtain a fixed pose dictionary by k-means without dictionary learning. Our spatio-temporal TriCRFs model does better in modeling dependencies between multi-level labels. Adding pose learning method in our model with extra layer could probably further increase performance.

To evaluate the robustness, we randomly select a part to be occluded in every testing sequence. The accuracy of our model decreases by 5.27 %, while the method in [9] decreases by 7.2 %, BoW decreases by 12.5 % and HMM decreases by 10.3 % as shown in Table 3.

4. CONCLUSION

We have presented spatio-temporal triangular-chain CRF-s, a unified hierarchical model that is powerful in exploiting dependencies between multiple layers for complex activity recognition. Spatio-temporal TriCRFs could model the complex relationships between activity, actions and pose observations. Our model takes more information into account than other existing methods. The transitions between actions are dependent of activity, and there is potential terms to model the influence of pose observations to activity recognition. Furthermore, interactions between different spatial regions are included. Experiments demonstrates the effectiveness of our model. The framework is also applicable in other sequential modeling problems.

For future work, we consider to add another layer to learn pose representations jointly with actions and activity. We can add several constraints in inference to make sure action labels of different body parts are consistent with each other.

5. ACKNOWLEDGEMENTS

We thank the anonymous reviews for their valuable comments. This work was partly supported by 863 Program (2014AA015104) and National Natural Science Foundation of China (61332016, 61202325, 61379100).

6. REFERENCES

[1] J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):pp. 179–195, 1975.

[2] C. Chen, Y. Zhuang, F. Nie, Y. Yang, F. Wu, and J. Xiao. Learning a 3d human pose distance metric from geometric pose descriptor. *IEEE Trans. Vis. Comput. Graph.*, 17(11):1676–1689, 2011.

[3] H. Chu, W. Lin, J. Wu, X. Zhou, Y. Chen, and H. Li. A new heat-map-based algorithm for human group activity recognition. In *Proceedings of the 20th ACM Multimedia Conference, October 29 - November 02, 2012*, pages 1069–1072. ACM, 2012.

[4] A. Fathi and J. M. Rehg. Modeling actions through state changes. In *CVPR 2013, June 23-28, 2013*, pages 2579–2586. IEEE, 2013.

[5] M. Jeong and G. G. Lee. Triangular-chain conditional random fields. *IEEE Transactions on Audio, Speech & Language Processing*, 16(7):1287–1302, 2008.

[6] M. Jeong and G. G. Lee. Multi-domain spoken language understanding with transfer learning. *Speech Communication*, 51(5):412–424, 2009.

[7] H. Kuehne, A. B. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR 2014, June 23-28, 2014*, pages 780–787. IEEE, 2014.

[8] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML 2001, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann, 2001.

[9] I. Lillo, A. Soto, and J. C. Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *CVPR 2014, June 23-28, 2014*, pages 812–819, 2014.

[10] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011, 20-25 June 2011*, pages 1297–1304. IEEE Computer Society, 2011.

[11] C. Sminchisescu, A. Kanaujia, and D. N. Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, 104(2-3):210–220, 2006.

[12] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3D: generic features for video analysis. *CoRR*, abs/1412.0767, 2014.

[13] N. N. Vo and A. F. Bobick. From stochastic grammar to bayes network: Probabilistic parsing of complex activity. In *CVPR 2014, June 23-28, 2014*, pages 2641–2648. IEEE, 2014.

[14] P. Xu and R. Sarikaya. Convolutional neural network based triangular CRF for joint intent detection and slot filling. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, December 8-12, 2013*, pages 78–83. IEEE, 2013.

[15] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative CNN video representation for event detection. *CoRR*, abs/1411.4006, 2014.

[16] Y. Zhang, Y. Zhang, E. Swears, N. Larios, Z. Wang, and Q. Ji. Modeling temporal interactions with interval temporal bayesian networks for complex activity recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(10):2468–2483, 2013.