# LEARNING AND INFERRING HUMAN ACTIONS WITH TEMPORAL PYRAMID FEATURES BASED ON CONDITIONAL RANDOM FIELDS

*Shih-Yao Lin*[1,2] *Yen-Yu Lin*[2] *Chu-Song Chen*[2] *Yi-Ping Hung*[1]

[1]National Taiwan University, Taiwan [2]Academia Sinica, Taiwan

## ABSTRACT

Finding an effective way to represent human actions is yet an open problem because it usually requires taking evidences extracted from various temporal resolutions into account. A conventional way of representing an action employs temporally ordered fine-grained movements, e.g., key poses or subtle motions. Many existing approaches model actions by directly learning the transitional relationships between those fine-grained features. Yet, an action data may have many similar observations with occasional and irregular changes, which make commonly used fine-grained features less reliable. This paper presents a set of temporal pyramid features that enriches action representation with various levels of semantic granularities. For learning and inferring the proposed pyramid features, we adopt a discriminative model with latent variables to capture the hidden dynamics in each layer of the pyramid. Our method is evaluated on a *Tai-Chi Chun* dataset and a *daily activities* dataset. Both of them are collected by us. Experimental results demonstrate that our approach achieves more favorable performance than existing methods.

***Index Terms***— human action recognition, conditional random fields, temporal pyramid representation

## 1. INTRODUCTION

Human action recognition has drawn increasing attention of researchers in last decades due to its wide range of applications, such as surveillance, health-care, and human-computer interactions, etc. Despite remarkable research efforts and encouraging advances [1, 2, 3, 4, 5, 6], accurate action recognition is still very challenging.

A conventional way to represent a human action is to employ a sequence of fine-grained movements, e.g., key poses or salient subtle-motions, and model the transition between them. However, actions may have many similar observations with occasional and irregular variations, which make commonly used fine-grained movements are no longer stable enough. Mid-level movements, e.g., sub-actions formed with a sequence of fine-grained movements, instead better characterize actions in some cases. Fig. 1 shows an action example of category `high jump` containing lots of fine-grained movements in its low-level representation. Modeling the transitional coherence between these fine-grained movements may not suffice for describing that action, since only short-term motions are extracted. It follows that using features from a single level of temporal representation is insufficient for describing complex human actions in general. Instead, that action can also be represented by mid-level sub-actions obtained by integrating several fine-grained movements. Longer-term motions are then included.

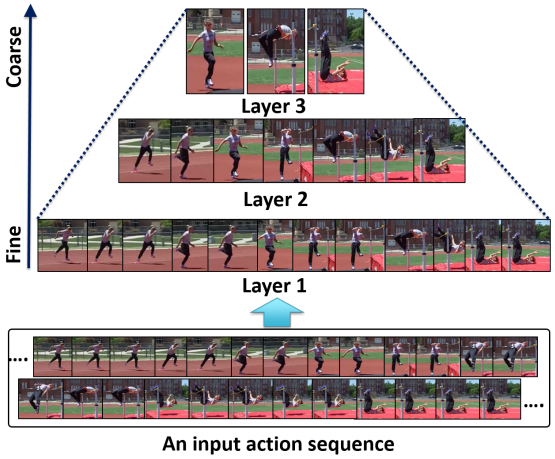



**Fig. 1**. An action in a temporal pyramid. Diverse information, such as long-term and short-term motions, are extracted by investigating the multi-level pyramid.

Recent studies have shown that learning and inferring from hierarchical feature representation often results in significant improvement in many visual learning tasks, such as spatial pyramids of image patches for scene recognition [7, 8] and temporal pyramids of video segments for action recognition [3, 9, 10, 11, 12, 13]. Wang *et al.* [14] proposed a temporal pattern representation, *Fourier temporal pyramid* (FTP), that represents an action as a hierarchical pyramid and handles both noisy data and temporal sequence misalignment. The above studies have shown their effectiveness. However, their adopted learning algorithms, such as *support vector machines* (SVM) or *multiple kernel learning* (MKL), ignore the temporal order of sequential data, and cannot make the most of temporal information for performance enhancement. Besides, most SVM-based methods cannot handle the issue of rate variations among actions. Hence they require a video

alignment process to pre-align actions before training and testing. The alignment process can be carried out by dynamic programming schemes, e.g., *dynamic time warping* (DTW). However, most alignment processes are sensitive to noise, and lead to extra computational cost.

Graphical models-based methods, *e.g.*, *conditional random fields* (CRFs) [15] and *hidden Markov model* (HMM) [16, 17], are widely used for modeling the temporal dynamic of action sequences. More importantly, the graphical model-based solutions do not need extra computational burden for temporal alignment process. Among various graphical models, the hidden-state CRFs (HCRFs) [18] have shown the expressive power for structured data prediction, and achieve superior performance to that HMM and CRFs [18, 19, 20].

Inspired by the FTP representation [14] for describing a temporal structure from fine to coarse, we propose a new temporal pyramid representation that expresses an action with various semantic granularities. Moreover, we introduce a method to learn and infer the temporal structure with various semantic granularities under conditional random fields. Our approach, termed as *multi-layer HCRFs* (MLHCRFs), is developed upon HCRFs. It leverages hidden variables to jointly learn the discriminative information at various temporal resolutions, and models the latent temporal structure between local descriptors in each layer of the pyramid. Our method is compared with the state-of-the-art methods on two datasets we collected, including *Tai-Chi Chun*3*D* and *Daily Activities*3*D* datasets. Superior results show its effectiveness.

## 2. THE PROPOSED APPROACH

In this section, a brief review of HCRFs is firstly given. The proposed temporal pyramid representation and its learning upon HCRFs are then depicted, respectively.

### 2.1. Action Recognition with HCRFs

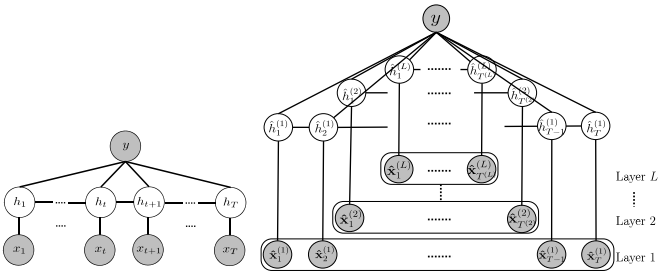
The main idea behind the HCRFs is to enrich CRFs [21] by augmenting hidden states to capture the implicit structure of the input features.

For an action instance $\mathbf{x} = \{x_1, x_2, ..., x_T\}$ of $T$ time stamps, a set of hidden variables, $\mathbf{h} = \{h_1, h_2, ..., h_T\} \in \mathcal{H}$, is created, where one variable for each time stamp. The hidden variables, whose states correspond to key poses in this work, are used to explore complex dependencies among action classes, key poses, and observations, and to model temporal coherence. The conditional probability $P(y|\mathbf{x}, \boldsymbol{\theta})$ in HCRFs is given by

$$P(y|\mathbf{x}, \boldsymbol{\theta}) = \sum_{\mathbf{h}\in\mathcal{H}} P(y, \mathbf{h}|\mathbf{x}, \boldsymbol{\theta}) \tag{1}$$

$$= \frac{\sum_{\mathbf{h}\in\mathcal{H}} \exp(\Psi(y, \mathbf{h}, \mathbf{x}, \boldsymbol{\theta}))}{\sum_{y'\in\mathcal{Y}, \mathbf{h}'\in\mathcal{H}} \exp(\Psi(y', \mathbf{h}', \mathbf{x}, \boldsymbol{\theta}))}, \tag{2}$$

where $\Psi$ is the *potential function*, and will be detailed later and $\boldsymbol{\theta}$ is the set of model parameters to be learned.




(a) (b)

**Fig. 2**. Graphical models. (a) HCRFs and (b) the proposed multi-layer HCRFs (MLHCRFs).

Like the original work of HCRFs [18], we adopt a chain structure shown in Fig. 2(a) to model the temporal relationships, and define the potential function as

$$\Psi(y, \mathbf{h}, \mathbf{x}, \boldsymbol{\theta}) = \sum_{t=1}^{T} \langle \phi(x_t), \theta_1(h_t) \rangle + \sum_{t=1}^{T} \theta_2(y, h_t) + \sum_{t=1}^{T-1} \theta_3(y, h_t, h_{t+1}), \tag{3}$$

where $\phi(x_t) \in \mathbb{R}^d$ is the feature representation of action $\mathbf{x}$ at time stamp $t$. $\phi(x_t)$ can be yielded by any features selected to characterize $x_t$. $\theta_1(h_t) \in \mathbb{R}^d$ is the parameter vector of the $t$th hidden variable. Inner product of $\langle \phi(x_t), \theta_1(h_t) \rangle$ represents the consensus between observation $x_t$ and hidden state $h_t$. Intuitively, $\theta_1(h_t)$ can be considered as the learned key pose to facilitate action classification. The number of states of each hidden variable $h_t$ corresponds to the number of key poses. $\theta_2(y, h_t) \in \mathbb{R}$ and $\theta_3(y, h_t, h_{t+1}) \in \mathbb{R}$ measure the compatibility among the corresponding variables.

Supposed that we are given a training set of $N$ actions, $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where each action instance $\mathbf{x}_i$ is temporally normalized, and consists of $T$ time stamps or frames, *i.e.*, $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, ..., x_{i,T}\}$, and $y_i \in \mathcal{Y}$ is its class label. $\mathcal{Y}$ is the class label set. The parameters $\boldsymbol{\theta}$ are derived with training set $D$ by *maximizing log likelihood*,

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log P(y_i|\mathbf{x}_i, \boldsymbol{\theta}) - \frac{\|\boldsymbol{\theta}\|^2}{2\sigma^2}, \tag{4}$$

where the first term is the log-likelihood of the training data, and the second term is used for regularization.

In our implementation, the gradient descent based L-BFG is used to optimize the parameter set $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3\}$. After optimization, the HCRFs model $\boldsymbol{\theta}^*$ is constructed. Given a testing action $\mathbf{x}$, its label $y$ is then inferred by

$$y = \arg\max_{y'\in\mathcal{Y}} \sum_{\mathbf{h}\in\mathcal{H}} P(y', \mathbf{h}|\mathbf{x}, \boldsymbol{\theta}^*). \tag{5}$$

### 2.2. Temporal Pyramid Feature Representation

The original work of FTP [14] was construct by a *top-down* scheme that recursively partitions an action into several video segments, and extracts features from all the segments. The

temporal structure of the action is represented as the pyramid structure. In contrast to the *top-down* scheme, we adopt a *button-up* scheme for building our pyramid. Our method is motivated by the observation that features computed from different temporal resolutions of actions tend to provide diverse and complementary information for recognition. An action sequence in our temporal pyramid representation can be summarized to $L$ layers, where each layer represents that action in a specific level of frame-wise feature quantization. The higher the layer, the coarser the features. For an input action $\mathbf{x} = \{\mathbf{x}_{1:T}\}$ of length $T$, we construct the pyramid features by merging $\alpha\,(l-1)$ frame-wise feature vectors into a concatenated feature vector in each layer $l$, where $\alpha$ denotes the merging parameter.

Let $T$ denote the length of an action video. The length of each layer $l$ is given by

$$T^{(l)} = \left\lceil \frac{T}{\alpha \cdot\ (l-1)} \right\rceil, \text{ for } 2 \leq l \leq L, \tag{6}$$

with $T^{(1)} \overset{def}{=} T$. Given an action $\mathbf{x} = \{x_{1:T}\}$, its temporal pyramid representation can be defined as

$$\hat{\mathbf{x}} = \left\{\hat{\mathbf{x}}^{(l)}\right\}_{l=1}^{L}, \text{ where } \hat{\mathbf{x}}^{(l)} = \hat{\mathbf{x}}^{(l)}_{1:T^{(l)}} \tag{7}$$

### 2.3. Learning HCRFs with Temporal Pyramid Features

For learning the proposed pyramid features, we adopt CRFs with latent variables to capture the hidden dynamics in each layer. Our method, multi-layer HCRFs (MLHCRFs), is developed upon HCRFs [18]. In MLHCRFs, the temporal pyramid representation of an action $\hat{\mathbf{x}}$ is associated with a set of hidden variables $\hat{\mathbf{h}} = \left\{\hat{\mathbf{h}}^{(l)}\right\}_{l=1}^{L}$, where $\hat{\mathbf{h}}^{(l)} = \hat{h}^{(l)}_{1:T^{(l)}}$ with one hidden variable $\hat{h}^{(l)}_t$ for each feature vector $\hat{\mathbf{x}}^{(l)}_t$ in layer $l$.

The hidden variables of our model are used not only to model the temporal structure of the observation in each layer but also to learn the favorite weights over all the layers. The proposed MLHCRFs model is shown in Fig. 2(b). Compared to the original work of HCRFs, our model augments a set of hidden variables in each layer of the pyramid features. The potential function is defined as follows:

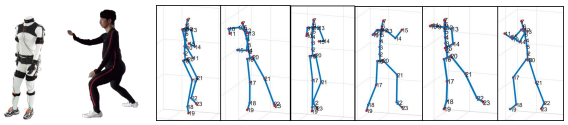
$$\Psi\left(y, \mathbf{h}, \mathbf{x}, \boldsymbol{\theta}\right) = \sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \langle \phi\left(\hat{\mathbf{x}}^{(l)}_t\right), \hat{\theta}^{(l)}_1\left(h^{(l)}_t\right)\rangle \tag{8}$$

$$+ \sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \hat{\theta}^{(l)}_2\left(y, h^{T^{(l)}}_t\right) + \sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}-1} \hat{\theta}^{(l)}_3\left(y, h^{(l)}_t, h^{(l)}_{t+1}\right),$$

where $\hat{\theta} = \left\{\theta^{(l)}_1, \theta^{(l)}_2, \theta^{(l)}_3\right\}_{l=1}^{L}$ denotes the parameter set which can be optimized by solving Eq. (4).

## 3. EXPERIMENTS

In this section, we firstly introduce the setting of the conducted experiments, including the two used datasets, the feature representations, and the evaluation metrics. We then depict the experimental results and the analysis.



(a) (b) (c)

**Fig. 3**. *Tai-Chi Chun3D* dataset we collected. (a) wearable mocap, (b) Tai-Chi Chun expert, and (c) some sample frames.

**Fig. 4**. Some samples of our *Daily Activities3D* dataset.

### 3.1. Datasets for Evaluation

Our method is evaluated on the *Tai-Chi Chun3D* and the *Daily Activities 3D* datasets. Both of them are collected by us.

#### *3.1.1. Tai-Chi Chun3D database:*

This database is captured by `XSens MVN` *motion capture* (Mocap), which is shown in Fig. 3(a) and can provide 3D locations of 23 body joints estimation in real time.

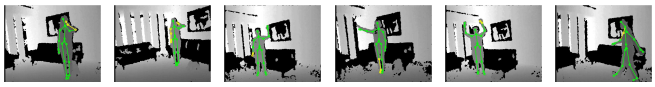
The database contains 21 Tai-Chi Chun actions, which were performed one time by a Tai-Chi Chun expert shown in Fig. 3(b). The frame rate is 200 fps. Thus, the collected videos are having very high temporal resolution. The durations of the collected actions range from 3 to 10 seconds. Hence, each of them contains from 600 to 2,000 frames.

To increase the diversity of the data, we generate four additional synthetic actions for each action. Specifically, we randomly select several frames from the original action, and add Gaussian noise to the body joint locations of the selected frames. The total number of action instances is 105. Some action examples of this dataset are shown in Fig. 3(c). More examples of this dataset can be found in our supplementary video: `https://youtu.be/dyNFTpIP3Tw`

#### *3.1.2. Daily Activities 3D dataset:*

This database contains 15 daily activities, including *Walk, Sit down, Sit still, Use a TV remote, Stand up, Stand still, Pick up books, Carry books, Put down books, Carry a backpack, Drop a backpack, Make a phone call, Drink water, Wave hand*, and *Clap*. Fig. 4 shows some frame examples of this dataset. A `Microsoft Kinect` is used in the collection so that the RGB video, the depth maps and the inferred skeletons [22] of each sequence are available simultaneously. Each skeleton data represents by using 3D locations of 20 body joints. The RGB and depth videos are captured by using a at frame rate 20 fps. Ten actors were employed to perform 15 daily activities in the construction of this dataset. Each actor perform each activity two times. This dataset contains 300 action instances.

Recognition difficulties, such as large intra-class variations, high inter-class similarity, and different perspective settings, make this dataset quite challenging. Fig. 5



(a) (b) (c) (d) (e) (f)

**Fig. 5**. Challenges in our *Daily Activities3D* dataset. (a)∼(b): high inter-class similarity. The skeleton structures of (a) an activity *make a phone call* and (b) activity *drink water* look very similar. (c)∼(f): large intra-class variations. For activity *wave hand*, actors may wave their left, right, or both hands.

shows some challenges examples. More challenge examples can be found in our supplementary video: `https://youtu.be/2ObOaxIa7l0`

### 3.2. Feature Representation and Evaluation Metrics

For both of the two databases we collected, each action is represented by the absolute 3D body joint positions (JP) in the skeletal streams. Each action instance consists of thirty skeleton ($T = 30$) frames which are uniformly sampled from each action. The normalization process in [3] is adopted for making the skeletons invariant to absolute location of actors.

For our *Tai-Chi Chun3D* dataset, we use two-fold cross validation for performance measure. The action instances is randomly partitioned into two equal-size groups. The action instance from one group serve as the training data, while the rest act as the testing data. For our *Daily Activities3D* dataset, we adopt the cross-subject test setting [23], where half of the subjects were used for training and the other half were used for testing. We then switch their roles, and report the average performance. A three-layer temporal pyramid is adopted in both of these two datasets.

### 3.3. Experimental Results

For the two datasets collected by us, we choose nine existing approaches for comparison, including *k-nearest neighbor* (kNN), *naive Bayes classifier* (NBC), *recurrent neural networks* (RNN) [24], *action graph* (AG) [23], *hidden Markov model* (HMM) [17], *hidden-CRFs* (HCRFs) [18], *hiddn conditional neural fields* (HCNFs) [20], *hierarchical sequence summarization* model (HSS) [20], and the method by Gowayye *et al.* [13]. Except [13], all the methods adopt the 3D JP features that we compiled. The method by Gawayye *et al.* uses the features based on body joint trajectories and applies Fourier temporal pyramid, as described in [13].

The recognition rates of all methods on our *Tai-Chi Chun3D* dataset are reported in Table 1. The baseline methods, kNN, and NBC give the accuracy of $46.0\%$ and $71.4\%$, respectively. RNN [24] obtain performance of $84.1\%$. Graphical model-based methods, HMM [17], AG [23], HCRFs [18], HCNFs [20], HSS [20] give the performance between $80.1\%$ and $93.0\%$. The state-of-the-art method [13] reaches $93.2\%$. Our method achieves the recognition rate of $96.2\%$, and is superior to the all competing approaches.

**Table 1**. Results on *Tai-Chi Chun3D* dataset

| Method | Accuracy (%) |
|---|---|
| $k$-NN Classifier | 46.0 |
| Naïve Bayes Classifier (NBC) | 71.4 |
| Action Graph [23] | 74.3 |
| Hidden Markov Model (AG) [17] | 80.1 |
| Recurrent Neural Networks (RNN) [24] | 84.1 |
| Hidden-State CRFs (HCRFs) [18] | 91.3 |
| Hidden Conditional Neural Fields (HCNFs) [20] | 92.3 |
| Hierarchical Sequence Summarization Model (HSS) [20] | 93.0 |
| Method by Gowayyed *et al.* [13] | 93.2 |
| Ours | **96.2** |

**Table 2**. Results on *Daily Activities3D* dataset

| Method | Accuracy (%) |
|---|---|
| $k$-NN Classifier | 69.6 |
| Naïve Bayes Classifier (NBC) | 73.3 |
| Action Graph (AG) [23] | 73.5 |
| Hidden Markov Model (HMM) [17] | 75.3 |
| Recurrent Neural Networks (RNN) [24] | 77.3 |
| Hidden-State CRFs (HCRFs) [18] | 80.3 |
| Hidden Conditional Neural Fields (HCNFs) [20] | 81.3 |
| Hierarchical Sequence Summarization Model (HSS) [20] | 82.3 |
| Method by Gowayyed *et al.* [13] | 83.0 |
| Ours | **86.6** |

The recognition accuracy of all methods on our *Daily Activities3D* dataset are shown in Table 2. The baseline approaches, kNN and NBC give the accuracy of $66.6\%$ and $73.3\%$, respectively. RNN [24] give the accuracy of $77.3\%$. The graphical model-based methods, HMM [17], AG [23], HCRFs [18], HCNFs [20], HSS [20] get recognition accuracy between $73.5\%$ and $82.3\%$. The state-of-the-art method [13] achieves an accuracy of $83.0\%$. Our method achieves a recognition rate of $86.6\%$, which outperforms all the approaches.

Our approach leverages multi-level temporal evidences, and integrates them based on hidden variables, the experimental results on both datasets show its robust and effectiveness.

## 4. CONCLUSIONS

In this paper, we have presented a set of temporal pyramid features that enrich action representation with various levels of semantic granularities. We have also proposed a multi-layer conditional random fields (MLHCRF) with latent states to learn and infer the temporal pyramid features. The hidden variables in our model are designed to select the favorable concatenations, and hence enhance the recognition performance. We have evaluated our approach on two datasets we collected, and compared it with both baseline and the state-of-the-art methods. The experimental results have shown that our approach achieve more favorable performance than the competing methods.

## 5. ACKNOWLEDGEMENTS

This work was supported in part by Ministry of Science and Technology under the grants MOST 104-2628-E-001-001-MY2, MOST 105-2221-E-001-030-MY2, MOST 104-2627-E-002-006-, and MOST 105-2627-E-002-003-.

## 6. REFERENCES


[1] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," in *Proc. Int' Conf. Acoustics, Speech and Signal Processing*, 2016, pp. 2712–2716.

[2] Fairouz Hussein, Sari Awwad, and Massimo Piccardi, "Joint action recognition and summarization by sub-modular inference," in *Proc. Int' Conf. Acoustics, Speech and Signal Processing*, 2016, pp. 2697–2701.

[3] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2014, pp. 588–595.

[4] I Rodomagoulakis, N Kardaris, V Pitsikalis, E Mavroudi, A Katsamanis, A Tsiami, and P Maragos, "Multimodal human action recognition in assistive human-robot interaction," in *Proc. Int' Conf. Acoustics, Speech and Signal Processing*, 2016, pp. 2702–2706.

[5] Nick C Tang, Yen-Yu Lin, Ju-Hsuan Hua, Shih-En Wei, Ming-Fang Weng, and Hong-Yuan Mark Liao, "Robust action recognition via borrowing information across video modalities," *IEEE Trans. on Image Processing*, vol. 24, no. 2, pp. 709–723, 2015.

[6] Yen-Yu Lin, Ju-Hsuan Hua, Nick C Tang, Min-Hung Chen, and Hong-Yuan Mark Liao, "Depth and skeleton associated action recognition without online accessible rgb-d cameras," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2014, pp. 2617–2624.

[7] Kristen Grauman and Trevor Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proc. Int' Conf. Computer Vision*, 2005, vol. 2, pp. 1458–1465.

[8] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 2169–2178.

[9] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese, "A hierarchical representation for future action prediction," in *Proc. Euro. Conf. Computer Vision*, 2014, pp. 689–704.

[10] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proc. Euro. Conf. Computer Vision*, pp. 392–405. 2010.

[11] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1290–1297.

[12] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu, "Robust 3d action recognition with random occupancy patterns," in *Proc. Euro. Conf. Computer Vision*, 2012, pp. 872–885.

[13] Mohammad A Gowayyed, Marwan Torki, Mohamed E Hussein, and Motaz El-Saban, "Histogram of oriented displacements (hod): describing trajectories of human joints for action recognition," in *Proc. Int' Joint Conf. on Artificial Intelligence*, 2013, pp. 1351–1357.

[14] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Learning actionlet ensemble for 3d human action recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 914–927, 2014.

[15] Liang Wang and David Suter, "Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[16] Chia-Chih Chen and JK Aggarwal, "Modeling human activities as speech," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 3425–3432.

[17] Fengjun Lv and Ramakant Nevatia, "Recognition and segmentation of 3-d human action using hmm and multi-class adaboost," in *Proc. Euro. Conf. Computer Vision*, 2006, pp. 359–372.

[18] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell, "Hidden conditional random fields," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 1848–1852, 2007.

[19] Yale Song, Louis-Philippe Morency, and Randall Davis, "Multi-view latent variable discriminative models for action recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2120–2127.

[20] Yale Song, Louis-Philippe Morency, and Ronald W Davis, "Action recognition by hierarchical sequence summarization," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2013, pp. 3562–3569.

[21] C. Sutton and A. McCallum, *An Introduction to Conditional Random Fields for Relational Learning*, MIT Press, 2007.

[22] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, et al., "Efficient human pose estimation from single depth images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2821–2840, 2013.

[23] Wanqing Li, Zhengyou Zhang, and Zicheng Liu, "Action recognition based on a bag of 3d points," in *Proc. Int' Conf. Computer Vision Workshops*, 2010, pp. 9–14.

[24] James Martens and Ilya Sutskever, "Learning recurrent neural networks with hessian-free optimization," in *Proc. Int' Conf. Machine Learning*, 2011, pp. 1033–1040.