International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 24 (2015) pp 45483-454837

INNOVATIVE HUMAN COMPUTER INTERACTION USING MOVING OBJECT EXPERT TEMPORAL CONDITIONAL RANDOM FIELDS (ETCRF) ALGORITHM

V.ANAND, B.E., M.E.,
Research Scholar(Anna University)
Associate Professor, PSV College of Engineering and Technology,
Krishnagiri, Tamil Nadu, India

R.S.D. WAHIDHA BANU
Government College of Engineering, Salem, India

ABSTRACT
In our work, we use a Expert temporal conditional random field based(ETMRF) on to model the spatio-temporal structure of present an anticipatory temporal conditional random field (ETCRF), where we model the past with the CRF discussed but augmented with the trajectories and with nodes/ edges representing the object affordances, sub-activities, and trajectories in the future. Since there are many possible futures, each ETCRF represents only one of them. In order to find the most likely ones, we consider each ETCRF as a particle and propagate them over time, using the set of particles to represent the distribution over the future possible activities. One challenge is to use the discriminative power of the CRFs (where the observations are continuous and labels are discrete) for also producing the generative anticipation—labels over sub-activities, affordances, and spatial trajectories. We then show that for new ETCRP training set improves an accuracy level of 89% with reduced time interval in seconds respectively.

1. INTRODUCTION

There has been a significant amount of work in detecting human activities from 2D RGB videos from inertial/ location sensors, and more recently from RGB-D videos. The primary approach in these works is to first convert the input sensor stream into a spatio-temporal representation, and then to infer labels over the inputs. These works use different types of information, such as human pose, interaction with objects, object shape and appearance features. However, these methods can be used only to predict the labeling of an observed activity and cannot be used to anticipate what can happen next and how.

Utilizing computers had always begged the question of interfacing. The methods by which human has been interacting with computers has travelled a long way. The journey still continues and new designs of technologies and systems appear more and more every day and the research in this area has been growing very fast in the last few decades.

However, in the area of HCI, where a typical requirement is to have prompt status of the speakers in the scene, on-line processing is necessary (implying in access to only current and past information). Considering the real-time challenge and also the aforementioned common adversities of a realistic video scenario, it is commonly beneficial to use information that is complementary to the video modality.

2. RELATED WORKS

A number of algorithms and designs have been proposed in literature we shall discuss few of them here according to Human computer Interaction systems.

In recent years, much effort has been made to detect human activities from still images as well as videos . Many methods have been proposed to model the temporal structure of low-level features extracted from video, e.g., histograms of spatiotemporal filter responses[1]. This includes both discriminative and generative models. Another approach is to represent activities as collections of semantic attributes. These methods use an intermediate level of representation such as the presence or absence of semantic concepts (e.g., scene types, actions, objects, etc.) in order to generalize to unseen instances. There are also a few recent works which address the task of early recognition . We refer the reader to for a comprehensive survey of the field and discuss works that are closely related to ours[2-3].

Temporal segmentation[4] In activity detection from 2D videos, much previous work has focussed on short video clips, assuming that temporal segmentation has been done apriori. It has been observed that temporal boundaries of actions are not precisely defined in practice, whether they are obtained automatically using weak-supervision [5] or by hand [6]. These works represent the action clips by an orderless bag-of-features and try to improve classification of the action clips by refining their temporal boundaries.

However, they only model the temporal extent of actions, not their temporal structure. Some recent effort in recognizing actions from longer video sequences take an event detection approach [6-9], where they evaluate a classifier function at many different segments of the video and then predict event presence. Similarly, change point detection methods [10-11], perform a sequence of change-point analysis in a sliding window along the time dimension. However, these methods only detect local boundaries and tend to over-segment complex actions which often contain many changes in local motion statistics.

Koppula, Gupta and Saxena (KGS, [12]) proposed a model to jointly predict sub-activities and object affordances by taking into account spatio-temporal interactions between human poses and objects over longer time periods. However, KGS found that not knowing the graph structure (i.e., the correct temporal segmentation) decreased the performance significantly. This is because the boundary between two sub-

activities is often not very clear. Previous work [13]–[15] only considers interaction classes without close physical contact (e.g., handshaking, talking, and queueing) and uses a detector or tracker to extract each interacting person. However, body part trackers and human detectors perform poorly when there are diverse categories of human motion that contain significant pose variations, limiting the performance of their interaction classifiers. Moreover, there are large variations in videos, including changes in subject appearance, scale, viewpoint, moving people and objects in the background, etc. These variations make the motion patterns of human interactions much noisier and thus a robust interaction recognition algorithm is required.

Such class of approaches have the advantage of not requiring a pre-trained model for the Video task, making the algorithm less dependent on the characteristics of the external data, and more robust to different acoustic scenarios. However, depending on the nature of the system, bottom-up approaches may be impracticable. This is mainly the case of real-time interfaces, which mandatorily require on-line processing (that is, to solve "who is speaking now?" without the knowledge of future data). When dealing with on-line SD, two important extra constraints must be considered: short-time analysis of the input streams, and no access to the upcoming data. The first is necessary so that a low-latency system is presented to the users, and the later is a natural limitation of any real-time system.

In [18], Noulas and Krose developed a multimodal system for HCI purposes. Potential speakers are found using a face detector, and then their audio-visual behavior is modeled as states of a dynamic Bayesian network. For the observations, Scale-invariant feature transform (SIFT) is applied to the facial features extracted from the images and the MFCC features are extracted from the audio. The models of each speaker are updated through hierarchical model selection [19] as more data arrive. The authors claim satisfactory results, but the on-line processing is still slower than real-time and it does not deal with overlapping speech. Furthermore, experiments were limited to two short test scenarios.

3. PROPOSED METHODOLOGY OVERVIEW

Our goal is to anticipate what a human will do next given the current observation of his pose and the surrounding environment. These observations are from RGB-D videos

recorded with a Kinect sensor. From these videos, we obtain the human pose using the Openni's skeleton tracker and extract the tracked object point clouds using SIFT feature

matching Algorithm. In our work, we infer the object affordances based on its usage in the activity and do not require the object category labels. We discuss the effect of knowing the object categories on the anticipation performance. Since activities happen over a long time horizon, with each activity being composed of sub-activities involving different number of objects. We model the activity using a spatio-temporal graph (a CRF), as shown in Fig. 1. The extracted human pose and objects form the nodes in this graph, and the edges between them represent their interactions are described based on our proposed algorithm.

Anticipated temporal segments are generated based on the available object affordances and the current configuration of the 3D scene. For example, if a person has picked up a coffee mug, one possible outcome could be drinking from it. Therefore, for each object, we sample possible locations at the end of the anticipated sub-activity and several trajectories based on the selected affordance. The temporal segmentation determines the structure of HCI. It is quite challenging to estimate this structure because of two reasons. First, an activity comprises several sub-activities of varying temporal length, with an ambiguity in the temporal boundaries. Thus a single graph structure may not explain the activity well. Second, there can be several possible graph structures when we are reasoning about activities in the future (i.e., when the goal is to anticipate future activities, different from just detecting the past activities). Multiple spatio-temporal graphs are possible in these cases and we need to reason over them in our learning algorithm .

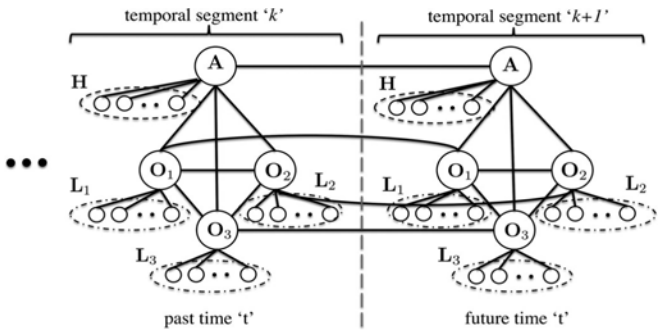


Fig. 1. An ETCRF that models the human poses H, object affordance labels O, object locations L, and sub-activity labels A, over past time 't', and future time 'd'. Two temporal segments are shown in this figure: kth for the recent past, and ðk þ 1Þth for the future. Each temporal segment has three objects for illustration in the figure.

4 . MOVING OBJECT CLASSIFICATION USING ETCRF

The concept of affordances as all "action possibilities" provided by the environment. Many recent works in computer vision and robotics reason about object functionality (e.g., sittable, drinkable, etc.) instead of object identities (e.g., chairs, mugs, etc.). These works take a recognition based approach to identify the semantic affordance labels . Few recent works explore the physical aspects of affordances based on human interactions . For example, detect the functionality of the object (specifically, chairs) with respect to possible human poses. In our work, we consider semantic affordances with spatio-temporal grounding which help in anticipating the future activities. Here, we describe how we model the spatio-temporal aspects of affordances.

Given the observations of a scene containing a human and objects for time t in the past, and its goal is to anticipate future possibilities for time d. However, for the future d frames, we do not even know the structure of the graph—there may be different number

of objects being interacted with depending on which subactivity is performed in the future. Our goal is to compute a distribution over the possible future states (i.e., sub-activity, human poses and object locations). We will do so by sampling several possible graph structures by

augmenting the graph in time, each of which we will call an Expet temporal conditional random field (ETCRF). We first describe an ETCRF below.

4.1 Modeling Past with a CRF

MRFs/CRFs are a workhorse of machine learning and have been applied to a variety of applications. Recently, with RGB-D data they have been applied to scene labeling and activity detection . Conditioned on a variety of features as input, the CRFs model rich contextual relations. Learning and inference is tractable in these methods when the label space is discrete and small.

Following , we discretize time to the frames of the video4 and group the frames into temporal segments, where each temporal segment spans a set of contiguous frames corresponding to a single sub-activity. Therefore, at time 't' we have observed 't' frames of the activity that are grouped into 'k' temporal segments. For the past t frames, we know the nodes of the CRF but we do not know the temporal segmentation, i.e., which frame level nodes are connected to each of the segment level node. The node labels are also unknown. For a given temporal segmentation, we represent the graph until time t as: $\mathcal{G}^t = (\mathcal{V}^t, \mathcal{E}^t)$, where $\mathcal{E}^t$ represents the edges, and V^t represents the nodes $\{\mathcal{H}^t, \mathcal{O}^t, \mathcal{L}^t, \mathcal{A}^t\}$: human pose nodes H^t, object affordance nodes O^t, objectlocation nodes Lt, and sub-activity nodes.

Our goal is to model the $P(\mathcal{H}^t, \mathcal{O}^t, \mathcal{L}^t, \mathcal{A}^t | \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}})$, where Ft H and Ft L are the observations for the human poses and object locations until time t. Using the independencies expressed as:

$$P_{Gt}(H^t, O^t, L^t, A^t | \Phi^t_H, \Phi^t_L) = P(O^t, A^t | H^t, L^t) P(H^t, L^t | \Phi^t_H, \Phi^t_L)$$

The second term $P(\mathcal{H}^t, \mathcal{L}^t | \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}})$ models the distribution of true human pose and object locations (both are continuous trajectories) given the observations from the RGB-3D Kinect sensor. We model it using a Gaussian distribution. The first term $P(\mathcal{O}^t, \mathcal{A}^t | \mathcal{H}^t, \mathcal{L}^t)$ predicts the object affordances and the sub-activities that are discrete labels—this term further factorizes following the graph structure as:

$$P(\mathcal{O}^t, \mathcal{A}^t | \mathcal{H}^t, \mathcal{L}^t) \propto \overbrace{\prod_{o_i \in \mathcal{O}} \Psi_{\mathcal{O}}(o_i | \ell_{o_i})}^{\text{objectaffordance}} \overbrace{\prod_{a_i \in \mathcal{A}} \Psi_{\mathcal{A}}(a_i | h_{a_i})}^{\text{sub-activity}} \overbrace{\prod_{v_i, v_j \in \mathcal{E}} \Psi_{\mathcal{E}}(v_i, v_j | \cdot)}^{\text{edgeterms}}.$$

The energy function expressed as

$$E(\mathbf{y} | \Phi(\mathbf{x}); \mathbf{w}) = \sum_{i \in \mathcal{V}} \sum_{k \in K} y_i^k [w_n^k \cdot \phi_n(i)] + \sum_{(i,j) \in \mathcal{E}} \sum_{(l,k) \in K \times K} y_i^l y_j^k [w_e^{lk} \cdot \phi_e(i,j)]$$

4.2.Modeling one Possible Future with an Augmented Temporal CRF (ETCRF)

We defined the anticipatory temporal conditional random field as an augmented graph $\mathcal{G}^{t,d} = (\mathcal{V}^{t,d}, \mathcal{E}^{t,d})$, where t is observed time and d is the future anticipation time. $\mathcal{V}^{t,d} = \{\mathcal{H}^{t,d}, \mathcal{O}^{t,d}, \mathcal{L}^{t,d}, \mathcal{A}^{t,d}\}$ represents the set of nodes in the past time t as well as in the future time d. $\mathcal{E}^{t,d}$ represents the set of all edges in the graph see Fig. 1. The observations (not shown in the figure) are represented as set of features, $\Phi^t_{\mathcal{H}}$ and $\Phi^t_{\mathcal{O}'}$ extracted from the t observed video frames. Note that we do not have observations for the future frames. In the augmented graph $\mathcal{G}^{t,d}$, we have:

$$P_{\mathcal{G}^{t,d}}(\mathcal{H}^{t,d}, \mathcal{O}^{t,d}, \mathcal{L}^{t,d}, \mathcal{A}^{t,d} | \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}}) = P(\mathcal{O}^{t,d}, \mathcal{A}^{t,d} | \mathcal{H}^{t,d}, \mathcal{L}^{t,d}) P(\mathcal{H}^{t,d}, \mathcal{L}^{t,d} | \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}})$$

the weights using importance sampling as shown as

$$p(\mathbf{g}^{t,d} | \Phi_t) \approx \sum_{s=1}^{S} \hat{w}_t^s \delta_{\mathbf{g}^{t,d(s)}}(\mathbf{g}^{t,d}),$$

$$\hat{w}_t^s \propto \frac{p(\mathbf{g}^{t,d(s)} | \Phi_t)}{q(\mathbf{g}^{t,d(s)} | \Phi_t)}$$

Here, $\delta_x(y)$ is the Kronecker delta function which takes the value 1 if x equals y and 0 otherwise, $\hat{w}_t^s$ is the weight of the sample s after observing t frames, and $q(\mathbf{g}^{t,d} | \Phi_t)$ is the proposal distribution.

5.Pseudocode for ETCRF

Data: RGB-D video frames
Result: Future sub-activity and affordance anticipations
$t = 0, P = \{\}$.

while new frame ft observed **do**
Generate frame features for frame ft
if temporal segmentation not given then
Find best segmentation using additive energy $E(\mathbf{y} | \Phi^A(\mathbf{x}); \mathbf{w})$
Sample segmentations by split and merge moves
end
Compute segment features ϕ_n and ϕ_e
Compute $\hat{\mathbf{y}}$, best labeling of the past-CRF ;
for each object do
Sample possible future affordance and sub-activity from the discrete distribution $P(\mathcal{O}^d, \mathcal{A}^d | \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}})$;
Sample future object location based on the affordance heatmaps ψ_o;
Generate corresponding object trajectory and human poses for d future frames;
Augment the past-CRF to generate an ETCRF particle

$\mathbf{g}^{t,d(s)}$;
$P = P \cup \{\mathbf{g}^{t,d(s)}\}$
end
for each particle $\mathbf{g}^{t,d(s)} \in P$ do
for each augmented frame do
Generate frame features
end
if temporal segmentation not given then
Find best segmentation using additive energy
$E(\mathbf{y}|\Phi^A(\mathbf{x}); \tilde{\mathbf{w}})$
Sample segmentations by split and merge moves
end
Compute segment features ϕ_n and ϕ_e
Compute ^y, best labeling for the ETCRF particle
Compute weight $\hat{w}_t^s$
end
P= top-k scored particles in P;
At= future sub-activity and affordance labels of top-3
particles based on $E(\mathbf{y}|\Phi(\mathbf{x}); \mathbf{w})$;
t =t + 1;
end

6. RESULTS AND DISCUSSION

In this section we describe the detailed evaluation of our Proposed approach on both offline data as well as HCI experiments.

6.1 Data SET

We use CAD-120 dataset , which has 120 RGB-D videos of four different subjects performing 10 high-level activities. The data is annotated with object affordance and sub-activity labels and includes ground-truth object categories, tracked object bounding boxes and human skeletons. The set of high-level activities are: {making cereal, taking medicine, stacking objects, unstacking objects, microwaving food, picking objects, cleaning objects, taking food, arranging objects, having a meal}, the set of sub-activity labels are: {reaching, moving, pouring, eating, drinking, opening, placing, closing, scrubbing, null} and the set of affordance labels are: {reachable, movable, pourable, pourto, containable, drinkable, openable, placeable, closable, scrubbable, scrubber, stationary}. We use all sub-activity classes for prediction of observed frames but do not anticipate null sub-activity.

6.2 Detection Results

For comparison, we follow the same train-test split described in KGS and train our model on activities performed by three subjects and test on activities of a new subject. We report the results obtained by four-fold cross validation by averaging across the folds. We consider the overall micro accuracy (P/R), macro precision and macro recall of the detected sub-activities, affordances and overall activity. Micro accuracy is the percentage of correctly classified labels. Macro precision and recall are the averages of precision and recall respectively for all classes.

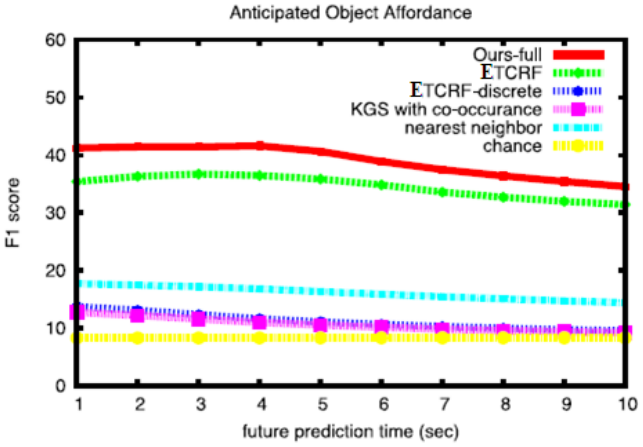
Table 1 shows the performance of our proposed approach on object affordance, sub-activity and high-level activity labeling for past activities. Rows 3-5 show the performance for the case where ground-truth temporal segmentation is provided and rows 6-9 show the performance for the different methods when no temporal segmentation is provided. With known graph structure, the model using the the full set of features (row 4) outperforms the model which uses only the additive features (row 5): macro precision and recall improve by 5 and 10.1 percent for labeling object affordance respectively and by 3.7 and 6.2 percent for labeling sub-activities respectively.

TABLE 1

Results on CAD-120 Dataset for detection, Showing Average Micro Precision/Recall, and Average Macro Precision and Recall for Affordances, Sub-Activities and High-Level Activities

	Object Affordance			Sub-activity			High-level Activity		
	micro	macro		micro	macro		micro	macro	
method	P/R	Prec.	Recall	P/R	Prec.	Recall	P/R	Prec.	Recall
With ground-truth segmentation.									
chance	8.3 (0.0)	8.3 (0.0)	8.3 (0.0)	10.0 (0.0)	10.0 (0.0)	10.0 (0.0)	10.0 (0.0)	10.0 (0.0)	10.0 (0.0)
max class	65.7 (1.0)	65.7 (1.0)	8.3 (0.0)	29.2 (0.2)	29.2 (0.2)	10.0 (0.0)	10.0 (0.0)	10.0 (0.0)	10.0 (0.0)
KGS [5]	91.8 (0.4)	**90.4** (2.5)	74.2 (3.1)	86.0 (0.9)	84.2 (1.3)	76.9 (2.6)	84.7 (2.4)	85.3 (2.0)	84.2 (2.5)
Our model: all features	**93.9** (0.4)	89.2 (1.3)	**82.5** (2.0)	**89.3** (0.9)	**87.9** (1.8)	**84.9** (1.5)	**93.5** (3.0)	**95.0** (2.3)	**93.3** (3.1)
Our model: only additive features	92.0 (0.5)	84.2 (2.2)	72.4 (1.2)	86.5 (0.6)	84.2 (1.3)	78.7 (1.9)	90.3 (3.8)	92.8 (2.7)	90.0 (3.9)
Without ground-truth segmentation.									
Our DP seg.	83.6 (1.1)	70.5 (2.3)	53.6 (4.0)	**71.5** (1.4)	71.0 (3.2)	60.1 (3.7)	80.6 (4.1)	86.1 (2.5)	80.0 (4.2)
Our DP seg. + moves	84.2 (0.9)	72.6 (2.3)	58.4 (5.3)	71.2 (1.1)	70.6 (3.7)	61.5 (4.5)	83.1 (5.2)	**88.0** (3.4)	**82.5** (5.4)
heuristic seg. (KGS)	83.9 (1.5)	75.9 (4.6)	64.2 (4.0)	68.2 (0.3)	71.1 (1.9)	62.2 (4.1)	80.6 (1.1)	81.8 (2.2)	80.0 (1.2)
Our DP seg. + moves + heuristic seg.	**85.4** (0.7)	**77.0** (2.9)	**67.4** (3.3)	70.3 (0.6)	**74.8** (1.6)	**66.2** (3.4)	83.1 (3.0)	87.0 (3.6)	82.7 (3.1)

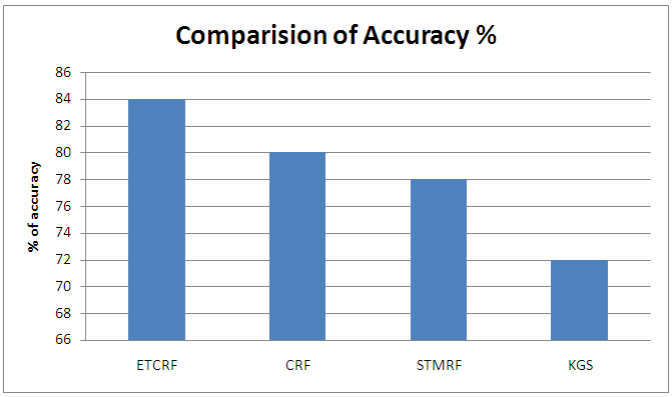
This shows that additive features bring us close, but not quite, to the optimal graph structure. When the graph structure is not known, the performance drops significantly. Our graph sampling approach based on the additive energy function (row 6) achieves 83.6 and 71.5 percent micro precision for labeling object affordance and sub-activities, respectively. This is improved by sampling additional graph structures based on the Split and Merge moves (row 7).



Graph 1. Plots showing how object affordance levels.

Graph 1 shows how the macro F1 score and the anticipation metric changes with the anticipation time. The average

duration of a sub-activity in the -120 dataset is around 3.6 seconds, therefore, an anticipation duration of 10 seconds is over two to three subactivities. With the increase in anticipation duration, performance of the others approach that of a random chance baseline, the performance of our ETCRF declines. It still outperforms other baselines for all anticipation times.



Graph 2: Plots showing Accuracy levels.

7. CONCLUSION

In this work, we considered the problem of detecting the past human activities as well as anticipating the future using object affordances. We showed how the anticipation
of future activities can be used by a robot to perform lookahead planning of its reactive responses. We modeled the human activities and object affordances in the past using a
rich graphical model (CRF), and extended it to include future possible scenarios. Each possibility was represented as a potential graph structure and labeling over the graph (which includes discrete labels as well as human and object trajectories), which we called ETCRF. We used importance sampling techniques for estimating and evaluating the most likely future scenarios. The structure of the ETCRF was obtained by first considering the potential graph structures that are close to the ground-truth ones by approximating the graph with only additive features. We then designed moves to explore the space of likely graph structures. We showed that anticipation can improve performance of detection of even past activities and affordances. We also extensively evaluated our algorithm, against baselines, on the tasks of anticipating activity and affordance labels as well as the object trajectories.

REFERENCES

[1] K. Tang, L. Fei-Fei, and D. Koller, “Learning latent temporal structure for complex event detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2012, pp. 1250–1257.

[2] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, “A database for fine grained activity detection of cooking activities,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2012, pp. 1194– 1201.

[3] H. Pirsiavash and D. Ramanan, “Detecting activities of daily living in first-person camera views,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2012, pp. 2847–2854.

[4] J.-K. Min and S.-B. Cho, “Activity recognition based on wearable sensors using selection/fusion hybrid ensemble,” in Proc. IEEE Int. Conf. Syst. Man, Cybern., 2011, pp. 1319–1324.

[5] H. S. Koppula, R. Gupta, and A. Saxena, “Learning human activities and object affordances from RGB-D videos,” Int. J. Robot. Res., vol. 32, pp. 951–970, 2013.

[6] J. Sung, C. Ponce, B. Selman, and A. Saxena, “Unstructured human activity detection from RGBD images,” in Proc. IEEE Int. Conf. Robot. Autom., 2012, pp. 842–849.

[7] B. Ni, G. Wang, and P. Moulin, “RGBD-HuDaAct: A color-depth video database for human daily activity recognition,” in Proc. IEEE Int. Conf. Comput. Vis. Workshop, 2011, pp. 1147–1153.

[8] E. Guizzo and E. Ackerman, “The rise of the robot worker,” IEEE Spectr., vol. 49, no. 10, pp. 34–41, Oct. 2012.

[9] [9] S. Nikolaidis and J. Shah, “Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy,” in Proc. IEEE 8th Int. Conf. Human-Robot Interact., 2013, pp. 33–40.

[10] J. Gibson, The Ecological Approach to Visual Perception. Boston, MA, USA: Houghton Mifflin, 1979.

[11] S. Maji, L. Bourdev, and J. Malik, “Action recognition from a distributed representation of pose and appearance,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2011, pp. 3177–3184.

[12] Z. Xing, J. Pei, G. Dong, and P. S. Yu, “Mining sequence classifiers for early prediction,” in Proc. SIAM Int. Conf. Data Mining, 2008, pp. 644–655.

[13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2008, pp. 1–8.

[14] J. Niebles, C. Chen, and L. Fei-fei, “Modeling temporal structure of decomposable motion segments for activity classification,” in Proc. Eur. Conf. Comput. Vis., 2010, pp. 392–405.

[15] A. Gaidon, Z. Harchaoui, and C. Schmid, “Actom sequence models for efficient action detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2011, pp. 3201–3208.

[16] B. Laxton, L. Jongwoo, and D. Kriegman, “Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2007, pp. 1–8.