Unstructured Human Activity Detection from RGDB Images

Authors : Jaeyong Sung, Colin Ponce, Bart Selman and Ashutosh Saxena

Presented By:
Ankit Goyal and Avi Singh

Introduction

Detecting and recognizing human activities is of great interest in several fields, with one being personal robotics. There are numerous challenges in activity recognition. Most of the activities occur in cluttered environments, in an uncontrolled manner. Different people have different speeds and mannerisms while performing these activities. This paper presents an approach to detect activities under such conditions. Some of the activities detected and recognized are cooking, drinking water, brushing teeth.

Previous work has focused mainly on RGB videos, or on the usage of RFID sensors. RGB videos lead to poor accuracy even in case of uncluttered environments, while the RFID methods are too intrusive as they require the placement of RFID tags on people and objects.

The model used in this a paper is a two-layered maximum entropy Markov model. This model exploits the inherent hierarchical nature of human activities. For example, brushing a teeth involves several sub-activities such as picking up the toothbrush, squeezing the tooth-paste, actual brushing etc. The graphical model is not fixed, and an on-the-fly graph structure selection techniques is described.

Related Work

A lot of work has been done in the field of human activity recognition. Some of the common approaches and their limitations have been listed below:

1. One approach is to use space-time features to model points of interest in the video. Some authors have suggested methods to add more information to these features.However, this approach is only capable of classifying, rather than detecting activities.
2. Other approaches include filtering techniques and sampling of video patches.
3. Hierarchical techniques for activity recognition have been used as well, but these typically focus on neurologically-inspired visual cortex-type models. There is a blind adhere to models of the visual cortex which may not always be correct.
4. Other approaches includes the use of Hidden Markov Models(HMMs). However it has been argued in literature that CRFs and MEMMs overcome limitations posed by HMMs. CRFs and MEMMs enables longer term interaction among observations which HMMs don't.

Proposed Model

The model should incorporate different nuances in the human activity. An activity comprises of a series of sub-activities done in some particular order. In order to incorporate the hierarchical nature of human activity, a maximum entropy Markov model is proposed [Fig 1.] and its salient features are explained as below:

- x^tdenote the features extracted from the articulated skeleton model at time frame t.
- Every frame is connected to high-level activities through the mid-level sub-activities. High-level activities do not change every frame, we do not index them by time. Rather, we simply write z^i to denote the i^{th} high-level activity. Activity i occurs from time t_{i1+1} to time t_i.
- Every frame is connected to a sub activity.y^t represents the sub activity connected to frame at time t. The sub activities are interconnected in sets to an activity. Thus, $\{y^{t_{i-1}+1}, \ldots, y^{t_i}\}$is the set of sub-activities connected to activity z^i

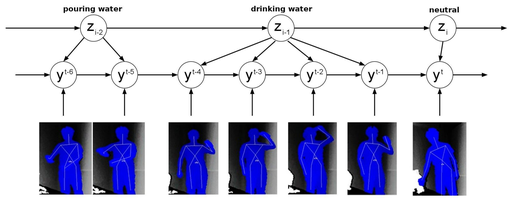

Figure 1: Proposed Model

Feature Extraction

Skeletal Features

PrimeSense provides a system for skeleton tracking from RGBD data.It gives us the three-dimensional Euclidean coordinates for fifteen joints, and rotation matrices for their orientations. All measurements are with respect to the frame of the sensor. Features computed from this data are as follows:

1. **Body pose:** Since features with respect to the sensor are not useful, the rotation matrices for 10 joints are calculated with respect to the torso, and represented using quaternions. Whether a person is standing or sitting, or leaning over is computed using the position each foot with with respect to the torso, and by computing head and hip joint angles with respect to the vertical.
2. **Hand Position:** The position of the two hands are computed with respect to the head and the torso. The hand positions are also observed for the last six frames, and the maximum and minimum hand positions are extracted from them.
3. **Motion Information** Nine frames are selected from the last three seconds, and the joint rotations that have occurred are computed with respect to each these nine frames.

HOG Features

Histogram of Oriented Gradient (HOG) features are computed, which give us a count of how frequently a particular gradient is seen in the ROI of an image. Using the skeletal data from PrimeSense, a bounding box is drawn around the head, torso, left arm, right arm. HOG features are then computed inside each of these bounding boxes.

Learning and Inference

Learning

A Gaussian Mixture Model is used to cluster the original data, and each individual cluster is treated as a separate sub-activity. Clusters are also generated from some negative examples (no activity happening), so that the system is not prone to errors on observing random activities. We need to evaluate the values for the following terms from the training data in order to perform inference.

1. $P(y^t|x^t)$: This term models the dependence of the sub-activity label y^t on the input features x^t. The GMM used in the previous step is used here too.
2. $P(y^{t_i-m}|y^{t_i-m-1,z_i})$:For all activities except neutral, this table is built from observing the transition of posterior probability from the soft cluster of Gaussian Mixture model at each frame. For neutral activities, $P(y^{t_i-m}|y^{t_i-m-1,z_i=N}) \propto 1 - \sigma_{z_i \neq N} P(y^{t_i-m}|y^{t_i-m-1,z_i=N})$
3. $P(z_i|z_{i-1})$ Set manually.

Inference

The joint probability is computed as follows:

$$\begin{aligned} &P(z_i, y^{t_{i-1}+1} \cdots y^{t_i} | O_i, z_{i-1}) \\ &\quad = P(z_i|O_i, z_{i-1}) P(y^{t_{i-1}+1} \cdots y^{t_i} | z_i, O_i, z_{i-1}) \\ &\quad = P(z_i|z_{i1}) \cdot \prod_{t=t_{i-1}+2}^{t_i} P(y^t|y^{t-1}, z_i, x^t) \\ &\qquad \cdot \sum_{y^{t_{i-1}}} P(y^{t_{i-1}+1}|y^{t_{i-1}}, z_i, x^{t_{i-1}+1}) P(y^{t_{i-1}}) \end{aligned}$$

The unknown terms in the equation above are derived as :

$$P(y^t|y^{t-1}, z_i, x^t) = \frac{P(y^{t-1}, z_i, x^t|y^t) P(y^t)}{P(y^{t-1}, z_i, x^t)}$$

From the model, the following conditional independence assumption are made: y^{t1} and z^i are independent from x^t given y^t is made and under this we get

$$P(y^t|y^{t-1}, z_i, x^t) = \frac{P(y^t|y^{t-1}, z_i) P(y^t|x^t)}{P(y^t)}$$

Finally the entire formula can be written as :

$$\begin{aligned} &P(z_i, y^{t_{i-1}+1} \cdots y^{t_i} | O_i, z_{i-1}) \\ &\quad = P(z_i|z_{i-1}) \\ &\qquad \cdot \sum_{y^{t_i-1}} \frac{P(y^{t_{i-1}+1}|y^{t_i-1}, z_i) P(y^{t_{i-1}+1}|x^{t_{i-1}+1})}{P(y^{t_{i-1}+1})} P(y^{t_{i-1}}) \\ &\qquad \cdot \prod_{t=t_{i-1}+2}^{t_i} \frac{P(y^t|y^{t-1}, z_i) P(y^t|x^t)}{P(y^t)} \end{aligned}$$

This formula can be factorized as follows :

$$P(z_i, y^{t_{i-1}+1} \cdots y^{t_i} | O_i, z_{i-1}) = \mathcal{A} \cdot \prod_{t=t_{i-1}+2}^{t_i} \mathcal{B}(y^{t-1}, y^t)$$

To maximize the probability the individual terms are maximized as:

$$\begin{aligned} \max P(z_i, y^{t_{i-1}+1} \cdots y^{t_i} | O_i, z_{i-1}) &= \max_{y^{t_{i-1}+1}} \mathcal{A} \\ &\cdot \max_{y^{t_{i-1}+2}} \mathcal{B}(y^{t_{i-1}+1}, y^{t_{i-1}+2}) \cdots \max_{y^{t_i}} \mathcal{B}(y^{t_i-1}, y^{t_i}) \end{aligned}$$

Graph Selection

Using the above results we can find the set of y^t 's that maximize the joint probability $P(z_i, y^{t_{i-1}+1} \cdots y^{t_i} | O_i, z_{i-1})$, the probability of an activity z^i being associated with the i^{th} substructure and the previous activity. Our task is to use that to compute the probability of z^i given all observations up to this point. Simply trying all the possibilities would be intractable and so we use a dynamic programming approach which is explained as follows :

- We are at some time t; we wish to select the optimal graph structure given everything we have seen so far. We will define the graph structure inductively based on graph structures that were chosen at previous points in time. Let $G_{t'}$ represent the graph structure that was chosen at some time $t' < t$. As a base case, G_0 is always the empty graph.
- For every $t' < t$, define a candidate graph structure $\tilde{G}_t^{t'}$ consisting of $G_{t'}$, followed by a single substructure from time $t' + 1$ to time t connected to a single high-level node z^i.
- Given the set of candidate structures $\{\tilde{G}_t^{t'} | 1 \leq t' < t\}$, the plan is to find the graph structure and high-level activity to maximize the likelihood given the set of observations so far $z_i \in Z$.
- Let O be the set of all observations so far. Then $P(z_i|O; \tilde{G}_t^{t'})$ is given by the following equation :

$$\begin{aligned} P(z_i|O; \tilde{G}_t^{t'}) &= \sum_{z_{i-1}} P(z_i, z_{i-1}|O; \tilde{G}_t^{t'}) \\ &= \sum_{z_{i-1}} P(z_{i-1}|O; \tilde{G}_t^{t'}) P(z_i|O, z_{i-1}; \tilde{G}_t^{t'}) \\ &= \sum_{z_{i-1}} P(z_{i-1}|O; G_{t'}) P(z_i|O_i, z_{i-1}) \end{aligned}$$

- The first factor inside the summation is calculated through dynamic programming and the second factor is calculated from $P(z_i, y^{t_{i-1}+1} \cdots y^{t_i} | O_i, z_{i-1})$ as described earlier.
- The optimal probability of having node i be a specific activity z^i is computed as follows and is stored for the purpose of dynamic programming.

$$P(z_i|O; G_t) = \max_{t' < t} P(z_i|O; \tilde{G}_t^{t'})$$

- Thus the prediction at time t is made by the following :

$$\text{activity}_t = \arg\max_{z_i} P(z_i|O) = \arg\max_{z_i} \max_{t' < t} P(z_i|O; \tilde{G}_t^{t'})$$

The selected graph is shown to be optimal and the time complexity for the entire calculation is $O(n \cdot m \cdot T^2 \cdot t)$.

Results

The results show an average precision/recall of 84.7%/83.2% in detecting the correct activity when the person was seen before in the training set and 67.9%/155.5% when the person was not seen before. The results are summarized in the table below.

		"New Person"								"Have Seen"							
		Naive Classifier		One-layer MEMM		Full Model RGB HOG		Full Model RGBD HOG		Naive Classifier		One-layer MEMM		Full Model RGB HOG		Full Model Skel.+Skel HOG	
Location	Activity	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
bathroom	rinsing mouth	77.7	49.3	71.8	63.2	42.2	73.3	49.1	97.3	51.1	51.4	73.3	49.7	70.7	53.1	61.4	70.9
	brushing teeth	81.5	65.1	83.3	57.1	50.7	30.8	73.4	16.6	80.5	55.3	81.5	65.1	81.5	75.6	96.7	77.1
	wearing contact lens	82.0	89.7	81.5	89.7	44.2	40.6	52.5	59.5	78.6	88.3	87.8	71.9	87.8	71.9	79.2	94.7
	Average	74.7	53.1	78.9	70.2	45.7	48.2	58.3	57.8	72.7	65.0	80.9	62.2	80.0	66.9	79.1	80.9
bedroom	talking on the phone	82.0	32.6	82.0	32.6	0.0	0.0	15.6	3.8	93.2	38.3	70.2	67.2	70.2	69.0	88.7	90.8
	drinking water	19.2	12.1	19.1	12.1	0.0	0.0	3.0	0.1	70.0	71.7	64.1	31.6	64.1	39.6	83.3	81.7
	opening pill container	95.6	65.9	95.6	65.9	60.6	34.8	33.8	36.5	95.0	57.4	46.7	52.3	48.7	54.8	93.3	77.4
	Average	65.6	36.9	65.6	36.9	20.2	11.6	17.4	15.2	76.1	59.2	61.0	50.4	61.0	54.5	88.4	83.3
kitchen	cooking (chopping)	33.3	56.9	33.3	57.4	56.1	90.0	59.9	74.2	45.6	43.3	78.9	28.7	78.9	29.0	70.3	85.7
	cooking (stirring)	44.2	29.3	45.6	31.4	58.0	4.0	94.5	11.1	24.8	17.7	44.6	45.8	44.6	45.8	74.3	47.3
	drinking water	72.5	21.3	71.6	23.9	0.0	0.0	91.8	23.9	99.4	75.3	52.2	51.5	52.2	52.4	88.8	86.3
	opening pill container	76.9	6.2	75.8	6.2	83.6	33.5	54.1	35.0	91.9	55.2	17.9	62.4	17.9	62.4	91.0	77.4
	Average	56.8	28.4	56.6	29.7	49.4	31.9	75.1	36.1	68.4	47.9	48.4	47.2	48.4	47.4	81.1	74.3
living room	talking on the phone	69.7	0.9	83.3	25.0	0.0	0.0	31.0	11.8	51.5	48.5	34.1	67.7	34.1	67.7	88.8	90.6
	drinking water	57.1	53.1	52.8	55.8	0.0	0.0	1.2	0.0	54.3	69.3	80.2	48.7	71.0	53.8	80.2	82.6
	talking on couch	71.5	35.4	57.4	91.3	42.7	59.4	53.2	63.2	73.2	43.7	91.4	50.7	91.4	50.7	98.8	94.7
	relaxing on couch	97.2	76.4	95.8	78.6	0.0	0.0	100.0	21.5	31.3	21.1	95.7	96.5	95.7	96.5	86.8	82.7
	Average	73.9	41.5	72.3	62.7	10.7	14.9	46.4	24.1	52.6	45.7	75.4	65.9	73.1	67.2	88.7	87.7
office	talking on the phone	90.5	31.0	60.6	31.5	17.5	6.7	2.7	0.6	69.4	48.2	80.4	52.2	80.4	52.2	87.6	92.0
	writing on whiteboard	47.1	73.3	45.2	74.1	41.2	25.1	94.0	97.0	75.5	81.3	42.5	59.3	42.5	59.3	85.5	91.9
	drinking water	41.1	12.4	51.2	23.2	0.0	0.0	0.0	0.0	67.1	68.8	53.4	36.7	53.4	36.7	82.3	81.5
	working on computer	93.5	76.8	93.5	76.8	100.0	11.9	100.0	29.0	83.4	40.7	89.2	69.3	89.2	69.3	89.5	93.8
	Average	68.5	48.4	62.6	51.4	39.7	10.9	49.2	31.7	73.8	59.8	66.4	54.4	66.4	54.4	86.2	89.8
Overall Average		**66.3**	**41.7**	**67.2**	**50.2**	**33.1**	**23.5**	**49.3**	**33.0**	**67.9**	**55.5**	**66.4**	**56.0**	**65.8**	**58.1**	**84.7**	**83.2**

Figure 2: Comparison of different models