Continuously Evolvable Bayesian Nets for Human Action Analysis in Videos

Nirmalaya Ghosh, Bir Bhanu, and Giovanni Denina
Center for Research in Intelligent Systems
University of California at Riverside, Riverside, CA 92521, USA
Email: nirmalya@ee.ucr.edu, bhanu@cris.ucr.edu, gdenina@vislab.ee.ucr.edu

***Abstract*—This paper proposes a novel data driven continuously evolvable Bayesian Net (BN) framework to analyze human actions in video. In unpredictable video streams, only a few generic causal relations and their interrelations together with the dynamic changes of these interrelations are used to probabilistically estimate relatively complex human activities. Based on the available evidences in streaming videos, the proposed BN can dynamically change the number of nodes in every frame and different relations for the same nodes in different frames. The performance of the proposed BN framework is shown for complex movie clips where actions like hand on head or waist, standing close, and holding hands take place among multiple individuals under changing pose conditions. The proposed BN can represent and recognize the human activities in a scalable manner**

Keywords-human action recognition; Bayesian Nets; interactions of multiple people; behavior analysis;

I. INTRODUCTION

For most situations the human action in an uncontrolled scenario is unpredictable and valid only probabilistically. To account for the uncertainty we need probabilistic models to represent and recognize human action. Bayesian Net (BN) is widely known as one of the frameworks to systematically handle uncertainty and they have received some success in human action recognition, specifically dynamic BN (DBN) in recognizing a small set of predefined activities in heavily constrained videos [4].

Due to the unpredictability in streaming data, it is not possible to account for every possible continuous action pattern. Furthermore, it is also not a scalable solution. For example, in a college patio, typical student activities may be walking, standing, talking to a person, talking over phone, biking, skateboarding, etc. The complexity in such a scenario comes from several factors. (a) Each action has a different model. (b) These models may be slightly different for different persons, i.e., different persons do the same actions differently. (c) Every person does not perform all the activities at the same time (e.g., a person skateboarding cannot do biking at that time). In such a monitored environment, the set of persons/activities may change dynamically, unpredictably and uncontrollably.

In such an unconstrained situation, analyzing human actions are very challenging computer vision and pattern recognition task. For BN based probabilistic action analysis, the brute force method with exhaustive model of all persons/activities considered, will have a large number of redundant nodes for the persons or activities not actually present or occurred. The-state-of-the-art BNs suffer from the combinatorial explosion [5] for the case with unpredictably changing set of performers.

For such an application of BN, we need a new framework where, (1) the number of (micro) event and (macro) action nodes for a particular time and particular person can be instantiated online based on current evidences, (2) the relations among these nodes can change over the time and from person to person, and (3) there are causal dependencies across the time for dynamic events and activities. None of the state-of-the-art Bayesian Net frameworks satisfies all these properties simultaneously.

This paper proposes a novel Bayesian Net, namely Structure Modifiable Adaptive Reason-building Temporal Bayesian Network (SmartBN) that has the following unique properties.

(1) It is flexible in the number of nodes in individual frames despite temporal connections across the frames. DBN cannot accommodate it.
(2) It is flexible in the relations between nodes, as the same two nodes may have different causal relations in two consecutive frames (DBN does not allow it).
(3) Its structure is continuously evolved online with the evidences that are present in the streaming data.

Key contributions: (1) A novel BN framework (SmartBN) suitable for unpredictable dynamic processes, (2) Use of SmartBN in a challenging computer vision/pattern recognition task, a video based event and human action recognition, and (3) Experimental results on various complex video datasets.

II. RELATED WORK AND MOTIVATIONS

A. *Related Works*

Table 1 shows probabilistic graphical models that have been used for human actions modeling.

Some of the non-graph-based approaches use periodicity [11], motion history [8], angular relations [12], or discontinuity in optical flow [14] and keep a log of activities [10] in multi-activity cases [9].

Non-graphical approaches do not handle uncertainty in systematic manner and use single periodic activity per video [11, 8, 12] and/or very constrained environment with known performers/activities [9, 10, 14]. Graphical approaches of Table 1 handle uncertainty of occurrence of known activities and known number of performers. But in typical activity analysis, like the student activity mentioned earlier, neither we

978-1-4244-4620-9/09/$25.00

know the number of performers, nor the activities actually taking place. Hence fixed structures of graphical models in Table 1 are too constrained to fit unpredictability in streaming activity data. Note that, EBN [16], although expandable for a single frame, does not support temporal causality across them, and hence not applicable for activity analysis. It does not also allow evolution of relationships (see Sec 3.1) between the same two nodes in consecutive frames.

TABLE I. REVIEW OF GRAPH-BASED VIDEO ACTIVITY ANALYSIS

Working Principle	Limitations
DBN, switching linear dynamic systems, learning of parameters, walking/jogging [1]	Structure fixed for known cases
Triangulated graphs, walk modeling, entropy based structure, greedy search [2]	Variable set is known a priori
Body-part BNs added to get hierarchical BN of the body, temporal links [3]	2 humans, fixed structure
DBN, sports, classify by skeleton-tips [4]	Known structure
Probabilistic hierarchy, blob clustering, Kalman tracking, HMM, single periodic activity per video, parameter learning [5]	Periodic single activity, known structure.
Parallel HMM, body part activity, parameter learning, graph pattern matching [6]	Known activities & fixed nodes
Trajectories in high dimension, learning densities in augmented HMM [7]	Expected actions with constraints
PCA based projected trajectories, HMM pattern learning, tracking, analysis [13]	Known node structures
Conditional Random Fields (CRF) for gesture recognition [21]	Fixed nodes & relations.

B. *Motivations for this work*

There are two types of uncertainties for human activity analysis in streaming videos in uncontrolled environment:

Case (a): The uncertainty of the presence or absence of human performers and their activities, where *we have prior knowledge with certainty about the entities*, i.e. who are expected to be present/absent and exactly what activity types we are expecting from each of the performers. In this case, even with the absence of some performers or activities, we can have dummy nodes or place holders for them in the DBN and keep on considering them for all the time with the expectation that they can appear any time. This method is not efficient and not scalable in complexity [5] or personal subtleties in activities [8].

Case (b): Besides the uncertainty of the presence or absence of humans and their activities (as in case (a)), there is uncertainty in the identity and number of the performers and the activities. That means, we have *no prior knowledge even about the entities*, i.e. we do not know who are expected to be present in the videos, and what they are expected to do. So here, we cannot even use the place holder nodes in the DBN as we don't have any idea how many places to hold. In these cases, we often have very generic causal patterns (lowest level events like, entity moving, two entities are close, etc.) that can gradually integrate over multiple abstraction levels to form the complete activity model.

In this paper we consider case (b). As DBN or other state-of-the-art BNs are inadequate [19], we use a novel continuously evolvable SmartBN framework, and an Expansion-Instantiation (EI) principle to analyze human activities in video. The proposed framework provides unique flexibility of online instantiation and continuous evolution with the streaming evidences. It is scalable and can also handle the case (a).

III. SMARTBN FRAMEWORK FOR HUMAN ACTIVITIES

We consider a typical uncontrolled environment, where we *do not have prior knowledge on the following*:

(1) Numbers of persons *expected* to be present,
(2) Number of persons *actually* present,
(3) Types of activities *expected* in videos,
(4) Types of activities *actually* performed by a person,
(5) Order of the events and activities,
(6) Interrelations between the activities.

The lowest level events are entity moving, entities close/distant, etc. and their causal structures. Particle filters are used for tracking body parts. Generic causalities at multiple abstraction levels integrate lowest level events to model complex human activities, like pose change, shifting person, talking, grabbing, etc.

The evidences from streaming videos at any stage provide a dynamic set of random variables comprising of: (a) *visible* 2D features (body-parts) of the *currently present* persons, (b) inter-feature distances between *visible* features (of the same and/or different persons), (c) low-level events *actually observed*, and (d) high level activities *actually detected* in the current frame. This dynamic set forms the "individual", as referred by the proposed SmartBN framework in the rest of this section.

A. *SmartBN & Expansion-Instantiation*

The key ideas of SmartBN framework and the principle of Expansion-Instantiation (EI) are:

(1) Identifying generic causal relations (we call them causal templates) that best represent the relational building blocks of a process.
(2) Expanding/replicating these causal templates to represent each individual with current evidences.
(3) Instantiating the SmartBN structure for a sliding window of few frames (to accommodate temporal causalities) over the streaming sequence of frames.

In a broader sense, there may be four possible types of causal templates, as described below.

Expandable templates: These are the uncertainty relations based on the evidences for a single individual in a single frame only; for example, the interrelations between body-parts of the same person.

Dynamic templates: These are the uncertainty relations that model the causalities from multiple "individuals" or frames to model dynamic (e.g., body-part motions) or inter-person (one grabbing other) processes.

Evolvable templates: These are uncertain relations that have the capability of online evolution over time, based on changing evidences in the streaming data; for example, when a person is explaining, he/she may explain to different persons over the frames. Thus, *listening* activity may evolve with time.

Interrelation templates: These are *conditional* causal

relations. They may influence themselves or others to be present or absent or change; for example, the absence of the head of a person in one of the two consecutive frames initiates "exit" or "entry" event for the person, while absence of other body parts initiate "body-part pose change" event.

In the rest of this section, we describe models of different activities and give the causal template type they belong to in the SmartBN for human activity analysis in video.

B. Human Activities Considered

In this work, we have considered the following human activities, in ascending order of complexity:

Events: Entry (Y) or exit (X) of a person; moving (V) or changing pose (P) of body-parts; hand on own head/waist (O); and standing close to other person (C).

Activities: Person moving (S), changing pose (Q), holding something (K), grabbing other person by hand(s) (G), talking (T), listening (B), explaining (E).

In SmartBN instantiation for these events/activities we take a bottom-up approach to show how the low-level inter-feature distances are integrated to define first the micro-level events and then the macro-level human activities in a hierarchical fashion.

As most of the events and activities discussed in this paper in some way depend on distances between body-parts and (dis)similarities among them, we start with two threshold based probability models and the subscript conventions (for persons/body-parts) used in this paper.

C. Notations and Probability Models

Subscript notations: In this paper, the subscripts denote the persons (ID) and body-parts (BP) involved in the event/activity; the numbers (ID: 1, 2, 3) denote persons, and the letters (BP: Head: H, Right-hand: R, Left-hand: L, Waist: W) denote the body-parts.

Superscript notations: The superscripts in the event or activity notations denote the frame number concerned.

Probability Models: We use $\boldsymbol{\delta}$, the distance (between two body-parts) or the dissimilarity (between two magnitudes or directions), to define two thresholds ($\boldsymbol{\theta}$) based probability models, in eqn (1) and (2). From data, $\boldsymbol{\delta}$ is computed automatically and $\boldsymbol{\theta}$ is decided from anthropometry [15]. To define a probability model for "how close or similar two entities are" we use the probability model TPM1 in eqn (1). Note that Pr (.) = 0 for $\delta > \theta$, and Pr (.) monotonically increases inside the range (0, 1) for $\delta \leq \theta$.

$$TPM1(\delta,\theta) = \Pr(event \mid \delta,\theta) = \min\left[1,\ \max\left[0,\ \left(1-2*\left(1+e^{-0.1*(\delta-\theta)}\right)^{-1}\right)\right]\right] \quad (1)$$

To define a probability model for "how distant or dissimilar two entities are" we use probability model TPM2 in eqn (2) where, Pr (.) = 0 for $\delta < \theta$ and Pr(.) increases monotonically in the range (0, 1) for $\delta \geq \theta$.

$$TPM2(\delta,\theta) = \Pr(event \mid \delta,\theta) = \min\left[1,\ \max\left[0,\ \left(2*\left(1+e^{-0.1*(\delta-\theta)}\right)^{-1}-1\right)\right]\right] \quad (2)$$

D. Features, Relations and Interrelations

- *Visible body parts (BP): 2D features*

Four body-parts represent each person (see Fig 1(a)): head (*H*), waist (*W*) right hand (*R*) and left hand (*L*). We take the 2D positions of *visible* body parts to be the root nodes (see Fig 1(b)) with probability 1. The probability is defined by fitting 2D Gaussian to the 2D histogram of the positions in real time.

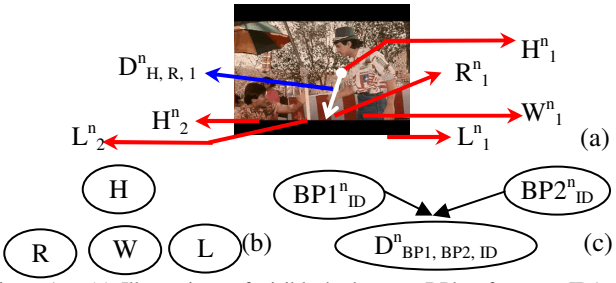

Figure 1. (a) Illustrations of visible body-parts BP's of person ID1 (Hn1, Rn1, Ln1, and Wn1) and person ID2 (Hn2, and Ln2) and inter-body-part vector DnH,R,1 between head and right hand of person ID1 in nth frame. (b) Body-parts (BP: H: head, R: right hand, L: left hand, W: waist). (c) Expandable causal template for relations between body-parts. DnBP1, BP2, ID is the difference vector between body-parts BP1 and BP2 of ID1 in nth frame.

- ***Interrelations (D):*** *among the body-parts*

The difference vectors (as illustrated in Fig 1(a)), defined by the distances (*D*) and angular directions ($\angle D$), between two body-parts (*BP*) of the same person define the pose of the person. These distances (D) are normalized by the distance between the head and the waist (*N*) to make the D's invariant to distance of the person from the camera. The causal template of D's is shown in Fig 1(c), with conditional probabilities equal to 1. Note that, this is an expandable causal template, and based on the visible body-parts of a person, this template is expanded or replicated and instantiated.

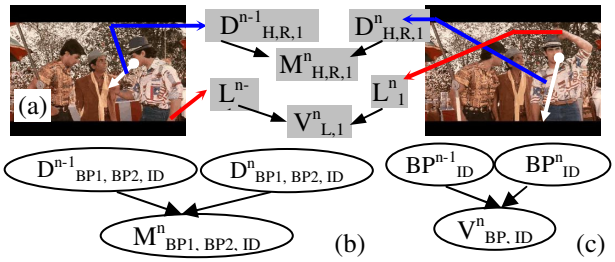

Figure 2. (a) Illustrations of event MnH,R,1 (for changing relations from Dn-1H,R,1 and DnH,R,1 between the head and the right hand of person ID1) and event VnL,1 (for moving left hand of person ID1 from Ln-11 to Ln1) in nth frame. Expandable and Dynamic causal templates for body-parts: (b) MnBP1, BP2, ID for changing interrelations from Dn-1BP1, BP2, ID to DnBP1, BP2, ID, (c) VnBP, ID for moving body-part from BPn-1ID to BPnID across the frames.

- ***Changing interrelations (M):*** *among the body-parts*

If the body-part distances (D's) change across the frames by *more* than a threshold, then corresponding relations are changing (M's). We use TPM2 of eqn (2) for the conditional probability and the expandable and dynamic (as it takes information from the previous frame also) causal template as shown in Fig 2(b). The template will be used to instantiate in SmartBN online, only when the event occurs. DBN does not have such flexibility.

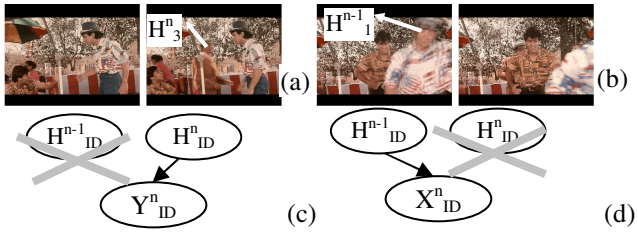

Figure 3. Illustrations of events (a) entry of person ID3 (appearing head Hn3) and (b) exit of person ID1 (disappearing head Hn-11); interrelation causal templates of (c) Event YnID, and (d) Event XnID for person ID in the nth frame.

E. Low-level Events

- ***Event Y**: Entry of a person*

When we don't see the head of a person, we say that person is absent in the frame. And we define entry (Y^n_{ID}) of a person ID, when we have head (H^n_{ID}) of the person in current n^{th} frame, but no (H^{n-1}_{ID}) in the previous $(n-1)^{th}$ frame. It is an interrelation causal template, because the absence of the evidence in another frame instantiates the causal relation in the current frame (see Fig 3(a) & (c)). So event Y is Boolean with conditional probability 1 or 0.

- ***Event X**: Exit of a person*

Similarly, if head was seen in the last frame (H^{n-1}_{ID}), but not in the current frame (Fig 3(b)), then the person ID has exited from the scene (event X^n_{ID}). It has an interrelation (like Y) and dynamic (multi-frame evidences) causal template (Fig 3(d)) with Boolean conditional probability.

- ***Event V**: Moving body parts*

When 2D positions of the body parts (BP's) of a person change across the frames by *more* than a threshold, moving-body-part events (V's) take place (see Fig 2(a)). The causal template of V's (in Fig 2(c)) is expandable and dynamic (just like M's). TPM2 (eq. (2)) defines the distance-based conditional probability of V's.

- ***Event P**: Pose change of a particular body part*

If the changing interrelations (M's) between the same body-parts (BP1 and BP2) persist for more than one consecutive frames, then the poses of these particular body-parts have changed (event P's). As evidences from multiple frames are concerned, P's have an expandable and dynamic causal template (see Fig 4(a) & (b)) with the probability model in (3).

$$\Pr\left(P^n_{BP1,BP2,ID} \mid M^{n-1}_{BP1,BP2,ID},\ M^n_{BP1,BP2,ID}\right) = \sqrt[2]{\prod_{k:(n-1),n} \Pr\left(M^k_{BP1,BP2,ID} \mid D^{k-1}_{BP1,BP2,ID},\ D^k_{BP1,BP2,\ ID}\right)} \quad (3)$$

where, $BP1, BP2 \in \{H, W, R, L\}_{ID}$

- ***Event O**: Hand on own head/waist*

When the right/left hand and the head/waist for the same person are *close*, hand-on-head/waist event (O's) take place. The evidences being only from the current frame, O's have expandable causal template (see Fig 4(a) & (c)) instantiated online for the involved body-parts. The closeness and eqn. (1) define the conditional probability.

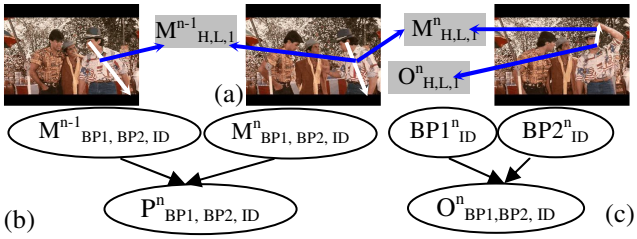

Figure 4. (a) Illustrations of pose change (from $M^{n-1}_{H,L,1}$ to $M^{n-1}_{H,L,1}$) between the head and the left hand; and the event $O^n_{H,L,1}$ for the left hand on the head for person ID1 in n^{th} frame; (b) Expandable and dynamic causal template of the event $P^n_{BP1,BP2,ID}$ for the pose change between body-parts BP1 and BP2 of person ID. (c) Expandable causal template of the event $O^n_{BP1,BP2,\ ID}$ for the hand BP1 on the body-part BP2 (head or waist) of person ID.

- ***Event C**: Standing close to other person*

When the heads and/or waists of two persons are *close* (see Fig 5(a)), the standing-closely events (C's) take place. As this depends on different persons, although from the same frame, the causal template (in Fig 5(b)) is dynamic. This is also expandable because both head and waist pairs may be close. Unlike DBNs, SmartBN can instantiate dynamic links, even for different *visible* body-part structures of different persons. The closeness and eqn (1) define its conditional probability. Its value is raised to 1 when both heads and waists are close.

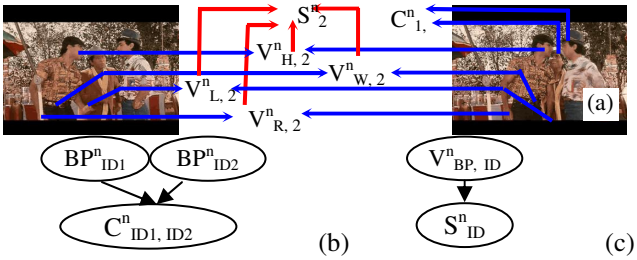

Figure 5. (a) Illustrations of the event Cn1,3 for persons ID1 and ID3 standing closely and the actiivity Sn2 for person ID2 moving from the moving body-part events (VnH,2, VnR,2, VnL,2, & VnW,2). (b) Dynamic and expandable causal template of the standing-closely event CnID1, ID2 for persons ID1 & ID2, (c) Expandable causal template of the shifting activity SnID for person ID from the body-part motion (VnBP, ID) for the nth frame.

F. Higher level human activities

- ***Activity S**: Person moving or shifting*

For a person ID, if *every* visible body parts (BP's) are moving ($V^n_{BP,ID}$) (see Fig 5(a)), then the person is shifting (activity S^n_{ID}). The activities S's have an expandable causal template (see Fig 5(c)) that expands with the number of visible body parts. Its probability in eqn. (4) is the geometric mean of the conditional probabilities of V's (see Sec 3.5.3) of all visible BP's.

$$\Pr\left(S^n_{ID} \mid \{V^n_{BP\,ID}\}\ \forall\ all\ visible\ BP's\ of\ ID\right) = \sqrt[N]{\prod_{N:all\ visible\ BPs} \Pr\left(V^n_{BP,ID} \mid BP^{n-1}_{ID}, BP^n_{ID}\right)} \quad (4)$$

- ***Activity Q**: Pose change of a person*

If some of the body parts (BP's) of a person ID appears or disappears or the changing interrelation (M's) between BP's in the last frame is discontinued (see illustration in Fig 6(a)), then the pose of ID is changing (activity Q's). Due to conditional

relations across the frames, it has interrelation causal templates (shown in Fig 6(b)-(d)). The conditional probability is the number of *current* evidences (BP's or M's) supporting the activity Q, normalized by *all possible* such evidences for visible body-parts.

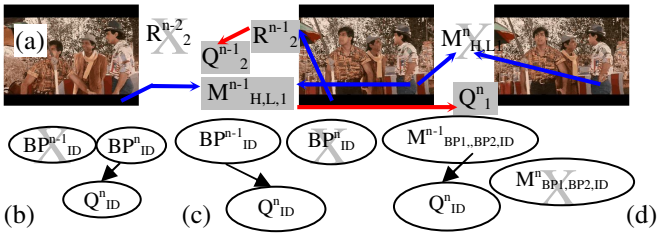

Figure 6. Illustrations of pose changes, Qn-12 of person ID2 (due to appearing right hand Rn-12) in (n-1)th frame, and Qn1 of person ID1 (due to transient interrelation change Mn-1H,L,1 between the head and the left hand) in nth frame. Interrelation causal templates of the activity QnID for person ID in nth frame when (b) body part BPnID appears, or (c) body part BPn-1ID disappears, or (d) transient change Mn-1BP1,BP2,ID in the last frame discontinued.

- ***Activity K:*** *Holding something by both hands*

The activity K's are encountered when both hands of a person are *close* to each other (see Fig 7(a)) and EI instantiates the expandable causal template in Fig 7(c) in the SmartBN. The closeness of hands and TPM1 in eqn (1) defines conditional probabilities of the activities, K's.

- ***Activity G****: Grabbing other person by hand(s)*

When the right and/or left hand of a person ID1 is *close* to any body-part (BP's) of another person ID2 (see Fig 7(b)), we say that ID1 grabs ID2 (G's). Due to the evidences coming from two persons, G's have dynamic causal template (Fig 7(d)). It can expand for two hands of ID1 or two body-parts of ID2. DBN cannot handle such online evolution, while SamrtBN can. The conditional probability uses TPM1 in (1).

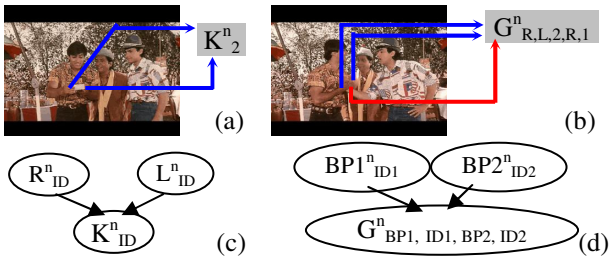

Figure 7. Illustrations of (a) Kn2: ID2 holding something and (b) GnR,L,2,R,1: ID2 grabbing ID1's right hand by his both hands. (c) Expandable causal template of the activity KnID for the person ID. (d) Dynamic and expandable causal template of the activity GnBP1,ID1,BP2,ID2 for ID1 grabbing ID2's BP2 by ID1's BP1.

- ***Activity T:*** *Person talking*

It is hard to define talking by only video data, while much easier with audio. We use the fact that, often, when a person talks his hand(s) move in an explaining fashion, continuously. Hence we relate continuous body-part pose change (P's) of the hands to the talking activities (T's) (see Fig 8(a)). The causal template of T's (shown in Fig 8(b)) expands with number of supporting evidences (*EV1*) of P's. Conditional probability (or belief) of T is the number of *EV1*, normalized by the number of all possible evidences (*EV2*) for T, (i.e., D's with the talker's hand(s)).

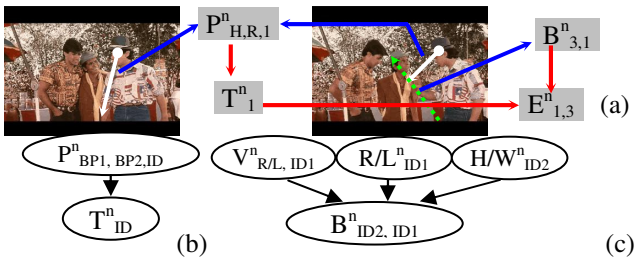

Figure 8. (a) Illustrations of pose change of right-hand, PnH,R,1 (in white arrows) defining talking activity Tn1 of person ID1; motion direction of ID1's right hand (in green arrow) detecting listening activity Bn3,1 of person ID3; and Tn1 and Bn3,1 defining activity En1,3 for ID1 explaining ID3. (b) Expandable causal template of the activity TnID from hand-pose changes PnBP1,BP2,ID for person ID in the nth frame. (c) Evolvable and interrelation causal template for the activity BnID2,ID1: ID2 is listening to ID1, from VnR/L,ID1 motion direction of ID1's hand and the vector direction between ID1's hand R/LnID1 and ID2's head or waist H/WnID2.

- ***Activity B:*** *Person listening*

If a person ID1 is talking, then the listener ID2 is decided by *closeness* of the direction of the hand-motion, to the directions of the vectors between the moving hand and the other persons' heads/waists. The person(s) ID2 satisfying a directional *similarity* is the listener for the listening activities (B's). This is an indirect method, but often works fine (see Fig 8(a)). Directional similarity and TPM1 (in eqn (1)) define the conditional probability in eqn (5). The set of listeners for the same talker may change/evolve and expand (based on which hand(s) is/are moving and towards whom), and B's have an evolvable and interrelation (because it is instantiated only when there is a node T) template of Fig 8(c).

$$Be\left(B^n_{ID2,ID1}\right) = \Pr\left(B^n_{ID2,ID1} \mid V^n_{BP1,ID1}\,,\, BP1^n_{ID1},\, BP2^n_{ID2}\right)$$
$$= TPM1\left(\left(\angle V^n_{BP1,ID1} \;-\; \angle\left(BP1^n_{ID1} - BP2^n_{ID2}\right)\right),\, \theta_B\right) \qquad (5)$$
$$\text{where, } BP1 \in \{R,L\}_{ID1}\,,\; BP2 \in \{H,W\}_{ID2 \neq ID1}$$

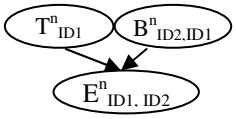

Figure 9. Evolvable causal tempale of the activity EnID1,ID2: talker ID1 is explaning to listener ID2.

- ***Activity E:*** *Explaining to the listener*

If a person ID1 is talking and person ID2 is listening to ID1 (see Fig 8(a)) then ID1 is explaining to ID2 (activity $E^n_{ID1,\,ID2}$). With the variable set of detectable listeners, the evolvable causal template of activity E (in Fig 9) can change across the frames. Its conditional probability is the geometric mean of the beliefs of its parent nodes (from eqn (5) and Sec 3.6.5).

IV. Experimental Results

A. *Activity Data*

Most of the publicly available human activity data suffers from (more than) one of the following: (a) single activity per video sequence, (b) constrained environment, (c) single person, (d) activities with known structures, and (e) unnatural breaks in multi-activity flow. To the authors, movie clips are

the most natural and unbiased video with uncontrolled and unconstrained continuous flow of multiple human activities without any predictable patterns. Hence we have used movie clips, taken in uncontrolled outdoor environment with minimum zooming effects and with all the activities described in Sec 3.2 (see Fig 10 and Fig. 11).

B. Tracking of Body Parts and Parameters

The method for tracking body parts is sequential Monte Carlo (SMC) also known as Particle filter [20]. The main inputs are AVI video and the initial location of the point to be tracked. Other parameters (fixed for all the examples 1-6) are: (1) number of particles: 1200; (2) color cues parameters (i.e. amount of noise in the image and number of bins) : 8 Bins for HSV, Measurement Color Noise: 0.2; (3) redistribution threshold : 7*1200/10; (4) initial state covariance : $\sigma_{x_1} = 30$, $\sigma_{vx_1} = 2$, $\sigma_{y_1} = 30$, $\sigma_{vy_1} = 2$, $\sigma_{Hx_1} = 2$, $\sigma_{Hy_1} = 2$; (5) position covariance : $\sigma_y = 0.5$. Tracking of body parts [17] is implemented using color HSV values and computing its likelihood based on Bhattacharya distance.

From generic anthropometric relations [15], we decided the thresholds for the activity definitions. For the events O and C and the activities K and G, we have used (θ_O = 30), (θ_C = 50), (θ_K = 20) and (θ_G = 30) respectively, all in normalized pixel distance. For the event M, threshold ($\theta_M = \theta_F * \theta_D$ = 15*30) is used, where θ_F is in normalized pixels for magnitude changes in D's and θ_D is in degrees for directional changes. For the activity B, we use directional similarity threshold of θ_B = 60º. All these parameters are fixed for all the examples 1-6.

C. Examples 1-3:

The data shown in Fig. 10 is used for activity recognition. We use video clips of 100 frames with interactions among two, three and four people. The results of tracking body parts by particle filters are shown in Figure 10. This method is robust in tracking head and waist. However, it fails to track both hands accurately since these parts experiences quick changes. The failure stems from particles not spreading out in a larger area and particles not able to keep up with the pixel movement. Other causes of failures are occlusion and contact between other body parts (i.e. hand touching head and particles stays on the head). The results of some activity (O, C, K) recognition are shown in Table 2. While these results are acceptable it is to be noted that the activity recognition results depend on the quality of tracking body parts.

Figure 10. Video Data and Tracking Results of actions using Particle Filters. (a) 2 person (b) 3 person (c) 4 person. Four body parts(head, two hands, and waist)are tracked by Particle Filters. The approximate location of each body part for each person is shown in different color (person 1: white, person 2: yellow, person 3: blue, person 4: red)

TABLE II. PERFORMANCE COMPARISON: NUMBER OF TIMES DETECTED

Data	Example 1: 2 PERSON Video Data (100 Frames)				
Activity/Event	Ground truth	Correctly Detected		False-alarm	Missed
O: Hand on H/W	240	211	88%	34	29
C: Standing close	160	114	71%	28	46
K: Holding	32	17	53%	12	15
Data	**Example 2: 3PERSON Video Data (100 Frames)**				
O: Hand on H/W	470	317	67%	126	153
C: Standing close	196	100	51%	64	96
K: Holding	180	162	90%	74	18
Data	**Example 3: 4 PERSON Video Data (100 Frames)**				
O: Hand on H/W	386	276	72%	114	110
C: Standing close	178	124	70%	68	54
K: Holding	290	279	96%	98	11

Figure 11. Sample Frames from a video data with 1855 frames.

D. *Evolutions of SmartBN Structures*

Due to the flexibility of the proposed SmartBN, and the data-driven EI instantiations, SmartBN structure varies widely across the frames. Unlike DBN or HMM, we do *not* have to keep place holders for activities or events that are absent in the current frame, but may be encountered in future frames. Note that, DBN or HMM does not support this flexibility. We show three examples of SmartBNs in Fig 12-14. We use a video clip with 1855 frames (see samples in Fig. 11) of size 640 x 480 pixels. Here we used only the 42 key frames [18] for tracking body parts. The square nodes are the new coming nodes for the current frame or the missing nodes from the previous frame, showing the evolution of the SmartBN structure. For better display, we drop the superscript and add the subscript 'p' for the previous frame or 'c' for the current frame.

Example 4: The key activities in the frames shown in Fig 12 are: person ID2 moves his hands (V's), and grabs ID3 by his neck ($G_{L,2,H,3}$), that changes pose of ID3 (Q_3) also. The SmartBN for the n^{th} frame is shown, with types of the nodes identified with different colors for different persons. Note that: (a) ID3 has different sets of body-parts (BP's) and inter-relations (D's) for $(n-1)^{th}$ and n^{th} frames. (b) There is pose change node for ID3 but not for ID1. (c) The events/activities like shifting, talking, listening, explaining have no nodes due to the absence of these activities in the current frame.

Example 5: The key activities in Fig 13 are: ID3 talks (T_3) with hand pose changing continuously (P's) with ID2 changing pose (Q_2). In the SmartBN for the n^{th} frame, note that: (a) some of the nodes of example 1, like closely-standing node ($C_{2,3}$) are now absent, (b) talking activity node T_3 is added, (c) but no particular listener is decided and hence no nodes for listening (B's) or explaining (E's).

Example 6: The key activities in Fig 14 are: ID1 talks (T_1) and explains ($E_{1,3}$) to ID3 ($B_{3,1}$); ID1 and ID2 change pose (Q_1, Q_2). In the SmartBN for n^{th} frame, note that, (1) node Q_3, and T_3 disappears (compared to Fig 13), and (2) nodes like Q_1, T_1, $B_{3,1}$, and $E_{1,3}$ newly appear.

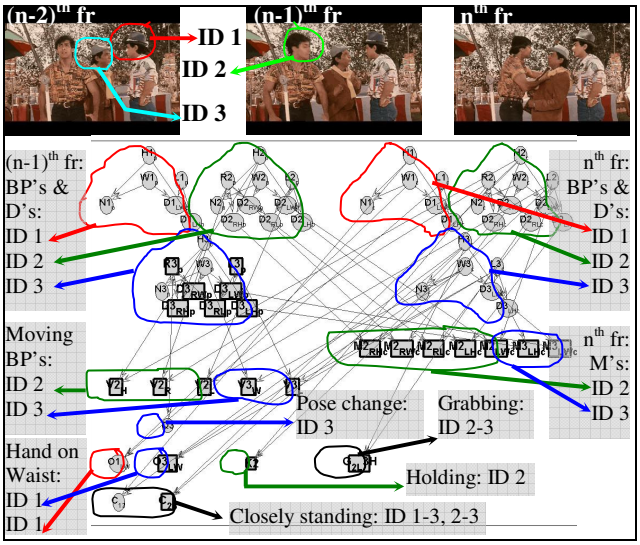

Figure 12. Example 1: SmartBN for nth frame shown. Activities: moving body-parts, pose change, holding, grabbing, closely standing, and hand on waist. Color-coding: Red: ID1, Green: ID2, Blue: ID3, Black: inter-person activities.

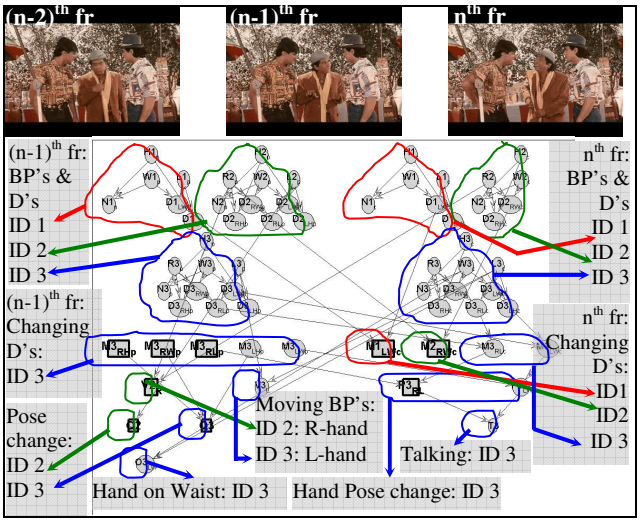

Figure 13. Example 2: SmartBN for nth frame shown. Activities: hand pose change, hand on waist, and talking. Color-coding: Red: ID1, Green: ID2, Blue: ID3, Black: inter-person activities.

Discussion: From the examples 4-6 it is clear that, unlike DBN, proposed SmartBN framework supports dynamic causal links like one from $M1_{LWp}$ (in the previous frame) and $M1_{LWc}$ (in the current frame) to $P1_{LW}$ (in the current frame) in Fig 14, despite different node structures for these consecutive frames. The scalability advantage of SmartBN becomes more evident when we consider the entire activity structures of different frames, as shown in Fig 12-14. The event or activity nodes of SmartBN are instantiated *only when those actually occur* (see

Examples above). Unlike DBN or HMM, SmartBN could continuously evolve online and model the video activity in a very scalable manner.

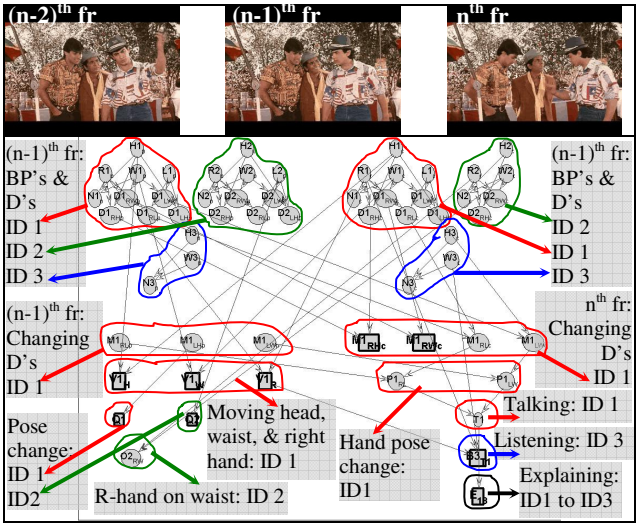

Figure 14. Example 3: SmartBN for nth frame shown. Activities: hand pose changing, hand on waist, body pose changing, talking, listening, and explaining. Color-coding: Red: ID1, Green: ID2, Blue: ID3, Black: inter-person activities.

TABLE III. PERFORMANCE COMPARISON: NUMBER OF TIMES DETECTED

Activity/Event	Ground truth	Correctly Detected		False-alarm	Missed
Y: Entry	2	2	100%	0	0
X: Exit	3	3	100%	0	0
O: Hand on H/W	20	10	50%	7	10
C: Standing close	21	21	100%	6	0
S: Shifting person	10	6	60%	0	4
Q: Pose change	53	39	74%	5	14
K: Holding	6	2	33%	1	4
G: Grabbing	7	5	71%	3	2
T: Talking	34	14	41%	4	20
B: Listening	33	3	9%	1	30
E: Explaining	14	3	21%	1	11

The performance of SmartBN for human events and activities for examples 4-6 is shown in Table 3. For simple events (Y, X, O, C) the results are very good (100% for all except O); for the activities with direct observations (S, Q, K, G) the results are acceptable (more than 60% for all except K); and for the activities (T, B, E) that are indirectly observed the results are fair. Performance of O and K can be improved if we take the *tip* (instead of the *wrist*) of the hands as 2D features, while considering more directional tolerance (θ_D = 60°) will reduce the number of missed Q's. Detection of indirectly observed activities T, B, and E can be improved with higher spatial/temporal resolutions since for examples 4-6 we only used the key frames.

V. CONCLUSIONS

We proposed a novel continuously evolvable BN framework, SmartBN that is scalable and can self-modify to represent an unpredictable dynamic process. We use it to analyze human activities in real video clips of movies which are a significant challenging task. Various experiments are presented to demonstrate the efficacy of the proposed approach for detecting continuous unpredictable human activities, including inferring some high level semantic information from video only. We have shown the SmartBN variation over the image-frames in Fig 12-14. In the future we plan to work on automated learning of (a) causal templates and (b) event thresholds.

VI. ACKNOWLEDGEMENT

This work was supported in part by NSF grant 0551741, ARO grant W911NF-07-1-0485 and ONR grant on Aware Building.

REFERENCES

[1] V. Pavlovic, J.M. Rehg, T Cham, & K. Murphy, "A Dynamic Bayesian Network Approach to Figure Tracking Using Learning Dynamic Models", Proc. ICCV 1999, pp 94-101.

[2] Y. Song, L. Goncalves, & P. Perona, "Unsupervised Learning of Human Motion", PAMI 25(7) 2003, pp 814-827.

[3] S. Park, & J.K. Aggarwal, "A hierarchical Bayesian network for event recognition of human actions and interactions", Multimedia Systems, Vol. 10, 2004, pp 164-179.

[4] Y. Luo, T.-D. Wu, & J.-N. Hwang, "Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks", CVIU 2003, Vol. 92, pp 196-216.

[5] C. Bregler, "Learning and recognizing human dynamics in video sequences", Proc. CVPR, 1997, pp 568-574.

[6] B. Ozer, T. Lv, & W. Wolf, "A bottom-up approach for activity recognition in smart rooms", Proc. IEEE Intl. Conf. Multimedia Expo, 2002, Vol. 1, pp 917-920.

[7] A. Psarrou, S. Gong, & M. Walter, "Recognition of human gestures and behaviors based on motion trajectories", Intl. J. Image & Vision Computing, 20 (2002) pp 349-358.

[8] O. Masoud, & N. Papanikolopoulos, "A method for human action recognition", Intl. J. Img. Vis. Comp. 21(8) 2003 pp 729-743.

[9] B. Song, N. Vaswani, & A.K. Chowdhury, "Closed-loop tracking and change detection in Multi-Activity Sequences", CVPR 2007.

[10] P. Peixoto, J. Batista, & H.J. Araujo, "Real-time human activity monitoring exploring multiple vision sensors", Intl. J. Robotics & Autonomous Systems, 35 (2001) pp 221-228.

[11] F. Cheng, W.J. Christmas, & J. Kittler, "Detection and description of human running behavior in sports video multimedia database", Proc. IEEE 11th Intl. Conf. Img. Anal & Proc., 2001, pp 366-371.

[12] A. Ali, & J.K. Aggarwal, "Segmentation and recognition of continuous human activity", IEEE Wkshp. Detec. Recog. Events in Video, 2001.

[13] N. Krahnstover, M. Yeasin, & R. Sharma, "Towards a unified framework for tracking and analysis of human motion", IEEE Wkshp. Detec. Recog. Events in Video, 2001, pp 47-54.

[14] Y. Rui, & P. Anandan, "Segmenting visual actions based on spatio-temporal motion patterns", Proc. CVPR 2000, Vol. 1, pp 111-118.

[15] S. Pheasant, "Anthropometry, Ergonomics and Design", Taylor & Francis Publication, 1986.

[16] Z.W. Kim, & R. Nevatia, "Expandable Bayesian networks for 3D object description from multiple views and multiple mode inputs", PAMI, 25(6) 2003, pp 769-774.

[17] S.C.W. Ong, & S. Ranganath, "Automatic sign language analysis: a survey and the future beyond lexical meaning", PAMI 27(6) 2005.

[18] H.-C. Lee & S.-D. Kim, "Iterative key-frame selection in the rate-constraint environment," *Sig. Proc.: Img. Comm.*: 18 (2003), 1-15.

[19] K.B. Korb, & A.E. Nicholson, "Bayesian Artificial Intelligence", Chapman and Hall Publication, 2004.

[20] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," IEEE Trans. on Signal Processing, Vol. 50, No. 2, Feb. 2002.

[21] S. Wang et al., "Hidden conditional random fields for gesture recognition," IEEE CVPR, 2006.