A Network of Dynamic Probabilistic Models for Human Interaction Analysis

Heung-Il Suk, *Student Member, IEEE,* Anil K. Jain, *Fellow, IEEE,* and Seong-Whan Lee, *Fellow, IEEE*

***Abstract*—We propose a novel method of analyzing human interactions based on the walking trajectories of human subjects, which provide elementary and necessary components for understanding and interpretation of complex human interactions in visual surveillance tasks. Our principal assumption is that an interaction episode is composed of meaningful small unit interactions, which we call "sub-interactions." We model each sub-interaction by a dynamic probabilistic model and propose a modified factorial hidden Markov model (HMM) with factored observations. The complete interaction is represented with a network of dynamic probabilistic models (DPMs) by an ordered concatenation of sub-interaction models. The rationale for this approach is that it is more effective in utilizing common components, i.e., sub-interaction models, to describe complex interaction patterns. By assembling these sub-interaction models in a network, possibly with a mixture of different types of DPMs, such as standard HMMs, variants of HMMs, dynamic Bayesian networks, and so on, we can design a robust model for the analysis of human interactions. We show the feasibility and effectiveness of the proposed method by analyzing the structure of network of DPMs and its success on four different databases: a self-collected dataset, Tsinghua University's dataset, the public domain CAVIAR dataset, and the Edinburgh Informatics Forum Pedestrian dataset.**

***Index Terms*—Dynamic Bayesian network, human interaction analysis, network of dynamic probabilistic models, sub-interactions, video surveillance.**

I. Introduction

UNDERSTANDING and analyzing human activities in video is one of the challenging issues in computer vision. During the past couple of decades, many research groups have addressed this problem and achieved good results mostly for recognizing single person actions [2]–[6]. With growing interest in vision-based surveillance systems, many researchers are now devoting their efforts to the analysis of human activities with interactions.

Most of the existing work on the recognition of human activity or interaction is focused on representing and learning spatiotemporal patterns embedded in human activities. To model spatiotemporal patterns, the hidden Markov model (HMM) or its variants [7]–[15], dynamic Bayesian networks [16]–[19], and other approaches [20]–[22] have been proposed in the literature. A comprehensive review of modeling, recognition, and analysis of human actions and interactions is available in [23].

There are two main approaches for modeling human interactions: 1) represent an interaction pattern by a single model [8]–[10], [16]–[18], [21], [24] with a large hidden state space, and 2) consider an interaction as a series of small unit interactions, which we call "sub-interactions" [7], [19], [25], [26]. Assume, e.g., that there are two interactions, *Approach + Meet + GoTogether* and *Approach + Meet + GoSeparately* as shown in Fig. 1. The first approach represents each interaction with a separate model, whereas the second one considers each interaction as a combination of four sub-interactions, i.e., *Approach*, *Meet*, *GoTogether*, and *GoSeparately* and then recognize an interaction by taking results from the sub-interaction models.

Table I gives a brief comparison of various methods in the literature for detection or recognition of person-to-person interactions in video sequence. Oliver *et al.* [9] proposed a system for detection of two person interactions. They utilized coupled hidden Markov models (CHMMs), a variant of the HMM integrating two or more information streams, for modeling and recognizing human behavior by employing a Bayesian approach in a visual surveillance task. For the classification of three-agent activities, Liu and Chua [10] proposed an observation decomposed hidden Markov model (ODHMM) introducing a parameter to assign a role to each agent for the problem of variable number of agents in an interaction. Du *et al.* [17] decomposed an interaction into multiple interacting stochastic processes and proposed a coupled hierarchical durational-state dynamic Bayesian network (DBN).

Unlike the work mentioned above, some groups have approached this problem in a hierarchical manner. Hongeng *et al.* [21] decomposed multiagent events into simple and complex single and multithread events. Park and Trivedi [16] considered the problem of vision-based surveillance tasks in more realistic

Manuscript received October 12, 2010; revised January 5, 2011; accepted January 24, 2011. Date of publication March 28, 2011; date of current version July 7, 2011. A preliminary partial version of this work was presented at the IEEE Workshop on Applications on Computer Vision (WACV), Snowbird, UT, December 2009 [1]. This work was supported in part by the World Class University Program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology, under Grant R31-10008-0, and in part by the Korea Science and Engineering Foundation (KOSEF) Grant funded by the Korean Government (MEST), under Grant 2009-0060113. All correspondence should be directed to S.-W. Lee. This paper was recommended by Associate Editor L. Zhang.

H.-I. Suk is with the Department of Computer Science and Engineering, Korea University, Seoul 136-713, Korea.

A. K. Jain is with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824-1226 USA, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, Korea.

S.-W. Lee is with the Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, Korea (e-mail: swlee@image.korea.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2011.2133570

(a) (b)

Fig. 1. Examples of a person-to-person interaction of interest. (a) *Approach + Meet + GoTogether*. (b) *Approach + Meet + GoSeparately*.

environments. They proposed a track-level and body-level analysis framework in which the two levels are adaptively switched depending on the inter-person configuration and imaging fidelity.

Although these previous methods have worked well in their respective problems, their disadvantage is that they do not make an effective use of the common small-unit interactions to build the person-to-person interaction models. In other words, when the target interactions share some common sub-interactions, they are represented in multiple interaction models repeatedly and independently. Accordingly, they require large training sets for reliable estimation of the parameters of interaction models.

Motivated by this, Ivanov and Bobick [7] proposed a two-level method for detection and recognition of temporally extended activities and interactions between two agents. In their system, the lower-level detects primitive events using HMMs and then the upper-level exploits a stochastic context-free grammar parsing mechanism to handle the uncertainty of the stream symbols passed from the lower-level. Nguyen *et al.* [28] proposed a sampling-based approximation algorithm for inference in a hierarchical hidden Markov model (HHMM), which is widely used for human activity recognition [27]–[29], for real-time recognition of single-person activity.

Pinhanez and Bobick [31] proposed a method of modeling temporal constraints of past, now, and future, called PNF-network, based on Allen's interval algebra [32], for the detection of actions and sub-actions. Shi *et al.* [33] proposed a propagation networks (P-Nets) to represent sequentially ordered activities by constraining temporal and logical ordering and duration of the activity intervals. These two methods basically utilize a deterministic approach in traversing a network. That is, a node in the network can only be activated if and only if the preceding node is deactivated. This approach may cause a poor performance in terms of an event detection or recognition in video because of the ambiguities and inaccuracies in the observations from low-level image processing. In order to deal with these problems, Albanese *et al.* proposed a probabilistic Petri net (PPN) suited to express uncertainty in the state of a node or associate a probability to a particular transition in a network for human activity detection in video [34]. However, the PPN takes a threshold-based method to fire ensuing nodes in traversing a network. As stated in their paper, finding an optimal threshold for decision of a transition is another big challenging problem.

Hakeem and Shah [26] proposed a new approach to learn, detect, and represent events in video by modeling multiagent events in terms of a temporally varying sequence of sub-events. Using the sub-event dependency graph, encoded from the training dataset, they clustered the maximally correlated sub-events using normalized cuts. Xiang and Gong [19] proposed a framework for automatic behavior profiling and online anomaly detection with no manual labeling of the training dataset. They modeled each event pattern using a multiple observation HMM (MOHMM) and measured the similarity between behavior patterns based on the likelihood computed in MOHMMs. By applying a spectral clustering algorithm to the similarity matrix, they discovered natural grouping of behavior patterns in a manner similar to Hakeem and Shah's method [26].

Ryoo and Aggarwal [22] proposed a description-based approach for recognition of human actions and interactions by extending Park and Aggarwal's work [35]. A context-free grammar was employed for the formal representation of the structure of high-level composite human activities.

In this paper, we are concerned with recognizing and analyzing various scenarios of person-to-person interactions based on the trajectories of human subjects. We propose a generic approach for representing the complete interaction patterns as a network of sub-interaction models, for which we also design a variant of the factorial hidden Markov model (FHMM) with factored observations for modeling sub-interactions.

The primary contribution of this paper is as follows. We propose a novel method of modeling interactions with a network of dynamic probabilistic models (NDPM) in order to represent complex patterns with a combination of simpler sub-interaction models. By representing complex interactions with a network of simple sub-interactions and independently training these sub-interaction models, we can greatly reduce the computational complexity of a model and simplify the training task. Another benefit of the proposed NDPM is that it can be extended to incorporate new interactions by simply adding extra sub-interaction models and paths as needed. We also propose a dynamic programming (DP)-based inference algorithm, which has a linear computational complexity and allows us to construct a network linking different types of dynamic probabilistic models. That is, we can build a network of mixture models by linking standard HMMs, variants of HMMs, and other types of dynamic probabilistic models for different sub-interactions. However, *a priori* knowledge about the structure of interactions is required to build the corresponding NDPM, which results in robust models. Another feature of our method is the modified factorial hidden Markov model (MFHMM), which characterizes the dynamic patterns by dividing observations into independent and shared components. Compared to the preliminary version of this paper that appeared in [1], we have extended our work by: 1) generalizing the inference algorithm in a NDPM; 2) developing a new sub-interaction model; and 3) carrying out more extensive experiments and analysis on three public domain databases.

Due to the representational characteristics of the HHMM and the proposed NDPM, they appear to be similar to each other. However, the proposed NDPM is different from the HHMM in many aspects. First, the proposed NDPM can share common small unit models by building a more compact

TABLE I
COMPARISON OF METHODS FOR PERSON-TO-PERSON INTERACTION DETECTION OR RECOGNITION

Approach	Authors	Methods	Limitations	Database (NTA)
SG	[9]	CHMM	Prior knowledge by the use of synthetic prior models	SCD (5)
	[10]	ODHMM, agents' role assignment	Difficult to distinguish classes having similar decomposed sub-observations	SCD (7)
	[17]	Coupled hierarchical durational-state DBN	High computational complexity	SCD (7)
	[21]	Decompose activities into single and multiple-thread events	Require prior knowledge over target events	SCD (5)
	[24]	Nearest neighbor classifier, Hausdorff distance	Applicable to detect only small unit interactions not complex interactions	CAVIAR (5) [30]
	[35]	Hierarchical Bayesian network, standard HMMs	Detailed representation of human body, tracking multiple body parts	SCD (4)
DCP	[7]	HMMs for primitive events, stochastic context free grammar	Require prior knowledge over target activities	SCD (5)
	[19]	Multi-observation HMM, spectral clustering	Retrain for inclusion of new behaviors	SCD (6)
	[22]	HMM for body part's gesture, context free grammar	Manually described production rules	SCD (8)
	[26]	Sub-event dependency graph, spectral clustering	Restructure a sub-event graph for inclusion of new events	SCD (6)
	Proposed method	A single network of dynamic probabilistic models by sharing common atomic models	Require prior knowledge over target interactions	SCD (5), CAVIAR (2) [30] Tsinghua DB (4) [17] EIFP DB [44]

SG, single model for each human interaction or activity; DCP, decomposed modeling of the entire interaction into small units; SCD, self-collected dataset; NTA, number of target activities.

model while the HHMM is restricted to a tree structure, partially allowing to share common small unit models only through explicit parameter tying which is not efficient [27]. Bui and Venkatesh [27] proposed a lattice-like state hierarchy to overcome this limitation of HHMM. Thanks to the structural characteristics, the NDPM can be considered as a *cyclic model*, which allows to perform inference by traversing a network in a cyclic path while the HHMM is a *non-cyclic model*. Furthermore, it is possible for the proposed NDPM to design a mixture of different dynamic probabilistic models, e.g., linking HMMs, CHMMs, DBNs, and others, without modifying the inference algorithm and to extend the structure whenever it needs more patterns without destroying the current network structure or retraining the current small unit models. These features also differentiate our method from Hakeem and Shah's work [26]. Unlike Albanese *et al.*'s work, transitions and segmentation points between sub-interaction models in the proposed NDPM are determined based on the likelihoods.

The rest of this paper is organized as follows. A novel network-based interaction model composed of dynamic probabilistic models and a DP-based linear inference algorithm in an NDPM is described in Section II. In Section II, we also design a new type of dynamic probabilistic model for sub-interactions. Experimental results and performance comparisons with competing methods are presented in Section III. We conclude this paper by summarizing the proposed method and providing directions for future work in Section IV.

II. PROPOSED NETWORK-BASED INTERACTION MODEL

The goal of this paper is to design a method that recognizes person-to-person interactions and segments an input sequence into meaningful small units for further analysis. For example, suppose there are two interactions as shown in Fig. 1. Given a sequence of video frames for the interaction of Fig. 1, our method not only classifies it to the appropriate interaction class but also detects the start and end points of sub-interactions, i.e., *Approach*, *Meet*, and *GoTogether*.

In this section, we first describe how to represent target interactions in a single network which we call a NDPM. We then propose an inference algorithm to efficiently compute the likelihood of each interaction and to segment a given observation sequence into atomic sub-interactions. Note that the description of the proposed method is based on a network in which standard HMMs are injected. However, our method of designing an interaction network and an inference algorithm in NDPM is not limited to standard HMMs. In other words, we can replace standard HMMs with other dynamic probabilistic models. That is why the proposed model is referred to as a *network of dynamic probabilistic models*.

A. Representation of Interactions in a Network

Interactions of interest in this paper are similar to those in [7], [9], [10], [17], [21], [24], [26], namely, *Follow + Meet + GoTogether*, *Follow + Meet + GoSeparately*, *Approach + Meet + GoTogether*, *Approach + Meet + GoSeparately*, and *Approach + PassBy*. These interactions can be described as follows:

$$Interaction \triangleq (Follow|Approach) \cdot (PassBy|(Meet \cdot (GoSeparately|GoTogether)))$$

where the symbol "|" means disjunction (or) and "·" means concatenation.

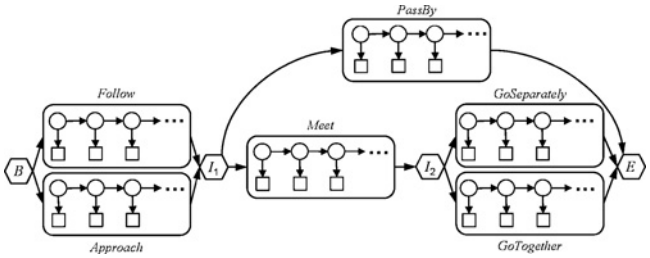


Fig. 2. Example of a network-based interaction model by concatenating sub-interaction standard HMMs to represent five interactions: *Follow* + *Meet* + *GoTogether*, *Follow* + *Meet* + *GoSeparately*, *Approach* + *Meet* + *GoTogether*, *Approach* + *Meet* + *GoSeparately*, and *Approach* + *PassBy*.

A directed graph or network is an elegant structure to embody the relationship or the order of sub-interactions for representing the target interactions in a single model. Fig. 2 shows an interaction network designed for five different interactions. The introduction of the dummy network nodes, i.e., B, I_1, I_2, and E in Fig. 2, makes modeling and inference conceptually simple and easy; they represent the milestones in the history of the interaction sequence.

There are two kinds of arcs in the network: internal and external. Internal arcs, within rounded rectangles, represent causal relations between random variables in each sub-interaction model. Although we have depicted only three time slices for each sub-interaction model in Fig. 2, there may be as many slices as the number of frame sequences related to the sub-interactions. External arcs link a dummy node to a sub-interaction model or vice versa. They represent a transition of sub-interactions where one ends and a new one begins. During an inference, each link from a sub-interaction model to a dummy node passes the likelihood computed by the sub-interaction model, and each dummy node chooses the maximum among the incoming likelihoods. On the contrary, the links from a dummy node to sub-interaction models simply propagate the likelihood to the following sub-interaction models without needing their own additional computation.

For the recognition of interactions we compute the likelihood for the given observations, considering all possible paths of sub-interaction models in the network, and state transitions within the sub-interaction models. When the last observation arrives, we finish the forward computation and retrieve the sequence of sub-interactions, which maximizes the likelihood for the given observations in the network. From this result, we can recognize and detect interactions in a video sequence. Through this compact and intuitive way, the proposed NDPM represents various interactions and can be used to decode them in order to segment unit-interactions of interest. The detailed explanation of the computation is given in the forthcoming section.

Given the structural characteristics of the proposed NDPM, we can easily insert additional interactions into the NDPM. Assume that a new interaction contains sub-interactions not included in the current NDPM. Then it is sufficient to model those sub-interactions independently and insert them into the NDPM with appropriate links and dummy nodes, if necessary. But if the new interaction is only composed of the sub-interactions already included in the NDPM, then we need to only include the links to represent the new interaction. Thus, the proposed network-based interaction model is extensible.

B. Inference in NDPM

Since we model interactions as a sequence of sub-interactions, the process requires simultaneous solution to problems of finding the start and end points of sub-interactions, determining their class labels, and finally classifying the interaction for the given input sequence based on the sequence of sub-interaction labels. This requires finding the best alignment between a given input sequence and a complete state sequence. The solution is based on the *best model* and *state sequence* that jointly maximize the likelihood along with the segmental sequence. In order to tackle this problem, we exploit the *one-pass DP search* [36].

1) *Problem Formulation:* Let $\mathbf{M} = \{M_1, \ldots, M_k, \ldots, M_K\}$, where M_k is a label of sub-interaction, i.e., $M_k \in \{$*Follow*, *Approach*, *Meet*, *GoTogether*, *GoSeparately*, *PassBy*$\}$ and $K \geq 1$ be the number of sub-interaction models which need to be determined from a given input frame sequence $\mathbf{Y} = \mathbf{Y}_1\mathbf{Y}_2, \ldots, \mathbf{Y}_T$. Note that the number of sub-interactions K is not known *a priori* and depends on the complexity of interaction in video. In the case of Fig. 2, it can take a value of either two or three. Our goal is to find the best alignment of $\mathbf{Y}$ to the best path $\hat{\mathbf{M}}$ that maximizes the likelihood from the network as follows:

$$P(\mathbf{Y}|G) \triangleq \max_{K,\mathbf{M}} P(\mathbf{Y}, \mathbf{M}|G) \tag{1}$$

where G denotes a network representing interactions of interest, e.g., as shown in Fig. 2. Let one possible segmentation of $\mathbf{Y}$ that aligns to $\mathbf{M}$ be

$$\begin{aligned} \mathbf{V} &= V_1 V_2, \ldots, V_K \\ &= (y(1, t_1),\ y(t_1 + 1,\ t_2),\ \ldots,\ y(t_{K-1} + 1,\ t_K)) \end{aligned}$$

where $y(t_{k-1} + 1, t_k) = \{\mathbf{Y}_{t_{k-1}+1}, \ldots, \mathbf{Y}_{t_k}\}$ denotes a segment aligned to a sub-interaction M_k and $1 = t_0 < t_1 < \ldots < t_K = T$. Then (1) can be formulated as the maximization problem as follows:

$$P(\mathbf{Y}|G) = \max_{K,\mathbf{V},\mathbf{M}} P(\mathbf{V}, \mathbf{M}|G). \tag{2}$$

This is a joint optimization problem of computing the maximum likelihood for the given observations by determining the optimum number of sub-interactions K, the best sequence of sub-interactions $\mathbf{M}$, and the best segmentation $\mathbf{V}$ of $\mathbf{Y}$ aligned to $\mathbf{M}$.

Now, let us assume that all the paths among the sub-interaction models in the network are equally probable, i.e., assume a uniform distribution for transitions from one sub-interaction to the others. Then we can write the probability on the right-hand side of (2) as a product of the likelihood of individual sub-interaction models in $\mathbf{M}$ as follows:

$$P(\mathbf{V}, \mathbf{M}|G) = P(\mathbf{V}|\mathbf{M}) \cdot P(\mathbf{M}|G) \tag{3}$$

$$\triangleq \prod_{k=1}^{K} P(V_k|M_k). \tag{4}$$

Here, we ignore the second term in (3) due to the assumption of uniformity in transitions among the sub-interactions. If we further apply the concept of Viterbi path alignment (V_k, S_k) [37] inside a sub-interaction model M_k then we can rewrite (2) as follows:

$$P(\mathbf{Y}|G) \triangleq \max_{K,\mathbf{V},\mathbf{M}} \left[\prod_{k=1}^{K} \max_{S_k} P(V_k, S_k|M_k) \right] \tag{5}$$

where $S_k = s(t_{k-1}+1, t_k) = \{s_{t_{k-1}+1}, \ldots, s_{t_k}\}$ denotes a legal state sequence within the sub-interaction model M_k.

2) *Global DP in NDPM:* Let us denote the current network node in a path as $gr \in \{I_1, I_2, E\}$ and the other node that immediately precedes it as gl. The pair gl and gr is connected via a set of parallel standard HMMs, as shown in Fig. 2. This can be regarded as a conceptual link with the label of standard HMMs. Let $L(gl, gr)$ be a set of sub-interaction models, i.e., standard HMMs in the path from node gl to gr. In addition, let us define the likelihood of the initial partial sequence of length t at the network node gr as follows:

$$\Delta_t(gr) = P(\mathbf{Y}_1 \cdots \mathbf{Y}_t, s_1 \cdots s_t, s_t \rightarrow gr|G). \tag{6}$$

This is the accumulated joint likelihood of the partial sequence $\mathbf{Y}_1, \ldots, \mathbf{Y}_t$ and the best state sequence $s_1, \ldots, s_t$ reaching the node gr at time t. The probability on the right-hand side of (6) is equivalent to $P(\mathbf{Y}_1 \cdots \mathbf{Y}_t, s_1 \cdots s_t|G)$ because no computation is needed during propagation of likelihood from a sub-interaction model to a network dummy node, i.e., $s_t \rightarrow gr$, as explained in Section II-A. Based on the DP principle, we can rewrite $\Delta_t(gr)$ as a recurrence relation as follows:

$$\Delta_t(gr) = \max_{(gl,M) \text{ s.t. } M \in L(gl,gr)} \Delta_{t'}(gl) \cdot P\left(\begin{array}{c} y(t'+1, t) \\ s(t'+1, t) \end{array} \middle| M \right) \tag{7}$$

where $t \in \{1, \ldots, T\}$, $gr \in \{I_1, I_2, E\}$. From here on, we use M instead of M_k to keep the notation simple. The notation (gl, M) means traversal from gl to other network node gr via the sub-interaction model M, $y(t'+1, t) = \mathbf{Y}_{t'+1} \cdots \mathbf{Y}_t$, $s(t'+1, t) = s_{t'+1} \cdots s_t$, and $t' < t$. This is what we refer to as *global DP* which performs maximization at the level of sub-interaction models. The second factor on the right-hand side in (7), which is the likelihood of the partial sequence of observations $y(t'+1, t)$ and states $s(t'+1, t)$, can be computed within a sub-interaction model M. The start and end points of the corresponding sub-interaction, i.e., $t'+1$ and t, respectively, are determined probabilistically in what we refer to as *local DP* described below.

3) *Local DP Within Sub-Interaction Models:* In a manner similar to (7), we define a measure for the internal state i of each sub-interaction model as

$$\begin{aligned} \delta_t^M(i) &= \Delta_{t'}(gl) \cdot P(y(t'+1, t), s(t'+1, t-1), s_t = i|M) \\ &= \max_j \left\{ \delta_{t-1}^M(j) \cdot A^M_{s_t=i|s_{t-1}=j} \right\} \cdot B^M_{s_t=i}(\mathbf{Y}_t) \end{aligned} \tag{8}$$

where $t \in \{1, \ldots, T\}$ and M denotes a label of a sub-interaction model. Equation (8) gives the joint likelihood of the partial sequence $\mathbf{Y}_1, \ldots, \mathbf{Y}_t$ and the best state sequence $s_1, \ldots, s_t$, where $s_t = i$ in the model M. In (8), $A^M_{s_t=i|s_{t-1}=j}$ and $B^M_{s_t=i}(\mathbf{Y}_t)$ denote, respectively, the state transition probability $P(s_t = i|s_{t-1} = j, M)$ and the probability of observing $\mathbf{Y}_t$ at the state i in the model M, $P(\mathbf{Y}_t|s_t = i, M)$. This is the second recurrence relation called as *local DP*. Its aim is to find the best state transitions to a state i at time t for the partial sequence $\mathbf{Y}_1, \ldots, \mathbf{Y}_t$ in the model M.

In order to retrieve the state sequences for segmentation and recognition of interactions, we need to keep track of the source state, which maximizes (8) at each time t in each sub-interaction model M, for efficient segmentation as follows:

$$\psi_t^M(i) = \operatorname*{argmax}_j \left\{ \delta_{t-1}^M(j) \cdot A^M_{s_t=i|s_{t-1}=j} \right\} \cdot B^M_{s_t=i}(\mathbf{Y}_t). \tag{9}$$

Furthermore, we also utilize one more variable $\varphi_t^M(i)$ in order to maintain the duration of the best path to the state i since it entered the sub-interaction model M as follows:

$$\varphi_t^M(i) = \varphi_{t-1}^M(\psi_t^M(i)) + 1. \tag{10}$$

For recognizing interactions and segmenting to meaningful sub-interactions in a video sequence, the segmentation boundaries of sub-interactions are not known *a priori*. Let us consider a likely segment $y(t, t+d)$ for any $d > 0$. At time $t-1$, the likelihood of a model corresponding to the sub-interaction being completed will be high. During the global DP in (7), the likelihood is propagated to a given network node gr. Then the decision on picking the starting boundary of a new segment $y(t, t+d)$ in the model M, where d denotes the possible number of frames devoted to the sub-interaction, is computed by

$$\delta_t^M(i) = \max \left\{ \begin{array}{c} \delta_{t-1}^M(j) \cdot A^M_{s_t=i|s_{t-1}=j} \\ \Delta_{t-1}(gl) \cdot \pi^M_{s_t=i} \end{array} \right\} \cdot B^M_{S_t=i}(\mathbf{Y}_t) \tag{11}$$

where $\pi^M_{s_t=i}$ denotes the initial state probability of staying the state i in the model M. Here, the two expressions in the braces correspond, respectively, to staying in the state i itself and to making a transition to a sub-interaction model M from outside implying that a next sub-interaction or a segment begins at this point. Unlike the other state transitions within a sub-interaction model M, we need to keep track of the source and duration in the model, which maximize (11) for the state i, because it is also possible for it to have the likelihood forwarded from outside the sub-interaction model M, i.e., a network dummy node gl as follows:

$$\psi_t^M(i) = \operatorname*{argmax}_{i,gl} \{\delta_{t-1}^M(i) \cdot A^M_{s_t=i|s_{t-1}=i}, \Delta_{t-1}(gl) \cdot \pi^M_{s_t=i}\} \tag{12}$$

$$\varphi_t^M(i) = \begin{cases} 1 & \text{if, } \Gamma_1 < \Gamma_2 \\ \varphi_{t-1}^M(\psi_t^M(i)) + 1 & \text{otherwise} \end{cases} \tag{13}$$

where $\Gamma_1 = \delta_{t-1}^M(i) \cdot A^M_{s_t=i|s_{t-1}=i}$ and $\Gamma_2 = \Delta_{t-1}(gl) \cdot \pi^M_{s_t=i}$. The left-hand term within the braces in (12) and the lower term on the right-hand side of (13) imply that a transition has occurred within a model M.

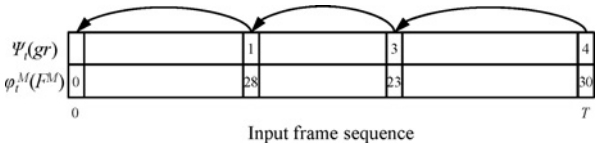


Fig. 3. Example of the backtracking procedure. The numbers in the upper row denote the label of sub-interaction models and those in the lower row represent the time duration in the corresponding sub-interaction models. Refer to the contexts for the description of the two variables Ψ_t and φ_t^M.

4) *Interaction Classification Via Backtracking:* Based on the computation of the local DP, the global DP in (7) is straightforward. Since the likelihood from the final state F^M in a sub-interaction model M is equal to the right-hand side of (7) for the pair M and gl, the likelihood propagated to the network node gr in global DP is given by

$$\Delta_t(gr) = \max_{(gl,M) \text{ s.t. } M \in L(gl,gr)} \delta_t^M(F^M) \tag{14}$$

where $t \in \{1, \ldots, T\}$, $gr \in \{I_1, I_2, E\}$, and F^M denotes the final state in the model M. Here, we introduce the variable $\Psi_t(gr)$, which records the sub-interaction model's label M and the source network node gl, which together produce the maximum value $\Delta_t(gr)$ as follows:

$$\Psi_t(gr) = \operatorname*{argmax}_{(gl,M) \text{ s.t. } M \in L(gl,gr)} \delta_t^M(F^M). \tag{15}$$

After the forward pass, i.e., computing the recurrences of (7) and (8) until the last observation $\mathbf{Y}_T$, the algorithm traces back the current result of the forward pass in order to recover the best sequence of sub-interaction models starting from the rightmost network node "E" in Fig. 2. The backtracking is based on the information about the transitions among network nodes in Ψ_t, and the elapsed time of the sub-interaction model M in φ_t^M, which evaluated the maximum likelihood at time t. An example of the backtracking procedure is given in Fig. 3. No computation is required during backtracking since all we need to do is to access the array of two variables, Ψ_t and φ_t^M.

Without any constraints on traversing the network, it is possible for the NDPM to output an undefined sequence of sub-interactions as an output interaction, e.g., *Follow + PassBy*. In (11), we can filter out the forwarding likelihood coming from the network node gl in the case of an unallowed transition from the previous model.

C. Computational Complexity of NDPM

The majority of the computation in the proposed NDPM is carried out in the local DP while the only computation performed in the global DP is to choose the maximum likelihood passed to network nodes. Since we have built simpler sub-interaction models by decomposing the original complex interactions, the number of hidden states to represent atomic sub-interactions is much less than that of the full interaction models, resulting in lower computational complexity of dynamic probabilistic models.

Let the average number of states for a hidden variable be N. For K possible sub-interaction models, assuming that they are all standard HMMs in an NDPM, the time complexity for a whole sequence of observations of length T is $\mathcal{O}(KN^2T)$, which is linear in both the length of observations T and the number of sub-interaction models K, and quadratic in the number of hidden states N that is much smaller than that of the full interaction models. This complexity is promising compared to the exponential complexity of the inference algorithm for a hierarchical HMM, $\mathcal{O}(T^3N^D)$, where D is the number of levels in the hierarchy [27], and that of the probabilistic Petri net, $\mathcal{O}(TRK^k)$, where R is the number of transition nodes related to both sub-interactions and skips, and k is a bounded constant for the number of tokens in the network at any marking [34].

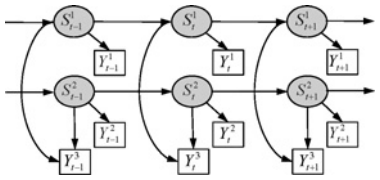


Fig. 4. Graphical representation of a MFHMM.

D. MFHMM

We now propose a new type of dynamic probabilistic model for representation of sub-interactions. For the representation of dynamic patterns in human motions, we introduce two hidden variables, one for each human subject involved in the interaction. Modeling with two hidden variables rather than a single hidden variable results in reducing the computational complexity of the model due to a decomposed hidden state space. In addition, it makes it easy to analyze and interpret the model. We decompose a feature vector $\mathbf{Y}_t$ into three parts according to the correlation: Y_t^1 and Y_t^2 represent individual motions of the two subjects and Y_t^3 represents the change in the relative distance between two subjects.

The resulting model has been extended to the temporal dimension by introducing the first-order Markov assumption describing the motion dynamics as shown in Fig. 4. The shaded circles denote hidden variables and the white rectangles denote the observation variables. Since the model is graphically and conceptually similar to the FHMM [38], we call it a "MFHMM." The sequence of feature vectors Y_t^1, Y_t^2, and Y_t^3 represent an observation of an interaction that we analyze by the proposed interaction model. We represent the distributions of the observations in a state with a Gaussian distribution.

It is also possible to represent the proposed model with an FHMM by combining all the observations into a single vector and by assuming statistical independence among the observations. The proposed model, however, is conceptually more intuitive to understand in graphical representation and has low computational cost in terms of inference due to the smaller size of the covariance matrices. In effect, it introduces a fine-grained factorization of the observations.

Given an observation sequence $\mathbf{Y} = \left[Y_{1:T}^1, Y_{1:T}^2, Y_{1:T}^3\right]^T$, where $Y_{1:T}^i = Y_1^i, \ldots, Y_T^i$, $i \in \{1, 2, 3\}$, and a model λ, the probability of observing the sequence is defined as

$$\begin{aligned} P(\mathbf{Y}|\lambda) &= \sum_{\mathbf{S}} P(\mathbf{Y}, \mathbf{S}|\lambda) \\ &= \sum_{S_{1:T}^{1:2}} P(\mathbf{Y}|S_{1:T}^1, S_{1:T}^2) P(S_{1:T}^1, S_{1:T}^2|\lambda) \end{aligned} \tag{16}$$

where $\mathbf{S} = S_{1:T}^{1:2} = [S_{1:T}^1, S_{1:T}^2]^T$ is a sequence of hidden states. In (16), the first term on the right-hand side computes the density of the observation sequence given the values of the two hidden variables S^1 and S^2 from time 1 to T, while the second term computes the joint probability of the sequences of two hidden variables.

Since the values of the two hidden variables are unobservable, we marginalize out all possible values in the states space. Hence, we need to compute the joint probability of $\mathbf{Y}$ and $\mathbf{S}$ within the model λ. The joint probability can be efficiently computed by factorizing it into a product of local conditional probabilities, one for each random variable, by utilizing conditional independencies or d-separation [39] as follows:

$$P(\mathbf{Y}, \mathbf{S}) = P(\mathbf{S}_1) \prod_{t=2}^{T} P(\mathbf{S}_t|\mathbf{S}_{t-1}) \prod_{t=1}^{T} P(\mathbf{Y}_t|\mathbf{S}_t) \quad (17)$$

$$= \pi_{S_1^{1:2}} \prod_{t=2}^{T} A_{S_t^{1:2}|S_{t-1}^{1:2}} \prod_{t=1}^{T} B_{S_t^{1:2}}(\mathbf{Y}_t) \quad (18)$$

where $P(\mathbf{S}_1) = \pi_{S_1^{1:2}} = P(S_1^1)P(S_1^2)$ represents the initial state probability, $P(\mathbf{S}_t|\mathbf{S}_{t-1}) = A_{S_t^{1:2}|S_{t-1}^{1:2}} = P(S_t^1|S_{t-1}^1)P(S_t^2|S_{t-1}^2)$ the state transition probability, and $P(\mathbf{Y}_t|\mathbf{S}_t) = B_{S_t^{1:2}}(\mathbf{Y}_t) = P(Y_t^1|S_t^1)P(Y_t^2|S_t^2)P(Y_t^3|S_t^1, S_t^2)$ the probability of observing $\mathbf{Y}_t = [Y_t^1, Y_t^2, Y_t^3]^T$ from the state $\mathbf{S}_t = [S_t^1, S_t^2]^T$. In (17), we omitted λ to keep it uncluttered. This factored joint probability has the same form as other dynamic state-space models, e.g., HMMs or Kalman filters.

III. Experimental Results and Analysis

In this section, we describe experiments on the recognition and analysis of person-to-person interactions based on the trajectories of human subjects. Since detection and tracking of human subjects in a video sequence are not of interest here, we performed these processes manually. We perform experiments on four different databases: a self-collected database, Tsinghua University's dataset [17], the public domain CAVIAR dataset [30], and the Edinburgh Informatics Forum Pedestrian dataset [44] to show the feasibility of our approach. The proposed models, MFHMMs and NDPM, as well as the other competing models, MOHMM [19], parallel HMM (PaHMM) [43], CHMM [9], ODHMM [10], and FHMM [38]), were implemented in MATLAB by using the Bayes net toolbox (BNT) [41], [42].

A. Data Collection and Feature Extraction

In our database, we define five interaction scenarios similar to those in [7], [9], [10], [17], [21], [24], and [26] as follows.

1) *Follow, meet, and go together* (Interaction 1).
2) *Follow, meet, and go separately* (Interaction 2).
3) *Approach, meet, and go together* (Interaction 3).
4) *Approach, meet, and go separately* (Interaction 4).
5) *Approach and pass by* (Interaction 5).

The five interactions of interest are composed of either two or three sub-interactions as shown in Table II. We collected 75 outdoor video sequences with 15 sequences for each interaction involving five subjects. The number of frames in a sequence is 296, on average. All the video sequences were captured using a JAI CV-S3300 camera, which was installed on the second floor of a building facing down. It operated at 30 f/s, with a frame resolution of 320×240, and 24-bit color.

TABLE II
Interaction Scenarios and the Associated Sub-Interactions

	Sub-Interactions					
	FL	*AP*	*MT*	*GT*	*GS*	*PB*
Interaction 1	○		○	○		
Interaction 2	○		○		○	
Interaction 3		○	○	○		
Interaction 4		○	○		○	
Interaction 5		○				○

"*Follow*" (*FL*), "*Approach*" (*AP*), "*Meet*" (*MT*), "*GoTogether*" (*GT*), "*GoSeparately*" (*GS*), "*PassBy*" (*PB*).

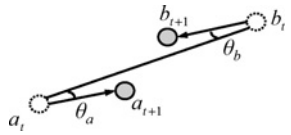


Fig. 5. Features tolerant to changes in distance between the camera and the subjects and subject's initial position and orientation [10].

Spatiotemporal patterns of many types of human interactions are independent of the initial position, moving direction of human subjects and the distance from the camera. Liu and Chua [10] mentioned these problems and used feature vectors composed of five elements as follows:

$$Y_t = \left[\frac{d(a_{t-1}, a_t)}{d(a_{t-1}, b_{t-1})}, \frac{d(b_{t-1}, b_t)}{d(a_{t-1}, b_{t-1})}, \frac{d(a_t, b_t)}{d(a_{t-1}, b_{t-1})}, cos(\theta_a), cos(\theta_b) \right]^T \quad (19)$$

where a_t and b_t denote the positions of subjects a and b at time t, $d(a_t, b_t)$ is the Euclidean distance between two points a_t and b_t, θ_α is the angle between the lines of (a_{t-1}, b_{t-1}) and (α_{t-1}, α_t), and $cos(\theta_\alpha)$ is the motion direction of a subject α based on the relative positions in the previous frame, $\alpha \in \{a, b\}$, respectively. Fig. 5 illustrates these variables.

B. Modeling Sub-Interactions

1) *Automatic Segmentation of Interaction Video Sequences:* The entire interaction sequence is automatically segmented into sub-interactions using the Viterbi algorithm [37].

For each interaction sequence, we train the parameters of HMM, $\hat{\lambda}$, to an observation sequence $\mathbf{Y} = y_1 y_2, \ldots, y_T$. The number of states is equal to the number of sub-interactions, which is assumed to be known *a priori*. We choose a left-to-right topology for an HMM, because there are no recursive or repeated sub-interactions in any of the interactions of interest. We find the best sequence of states $\mathbf{Q}^* = q_1 q_2, \ldots, q_T$ which maximizes $P(\mathbf{Q}, \mathbf{Y}|\hat{\lambda})$, i.e., $\mathbf{Q}^* = \text{argmax}_{\mathbf{Q}} P(\mathbf{Q}, \mathbf{Y}|\hat{\lambda})$. An example of segmentation result for the interaction *Follow + Meet + GoTogether* is presented in Fig. 6. The middle graph represents the change in the likelihood for the three states in time. The likelihood of a state dedicated to the current

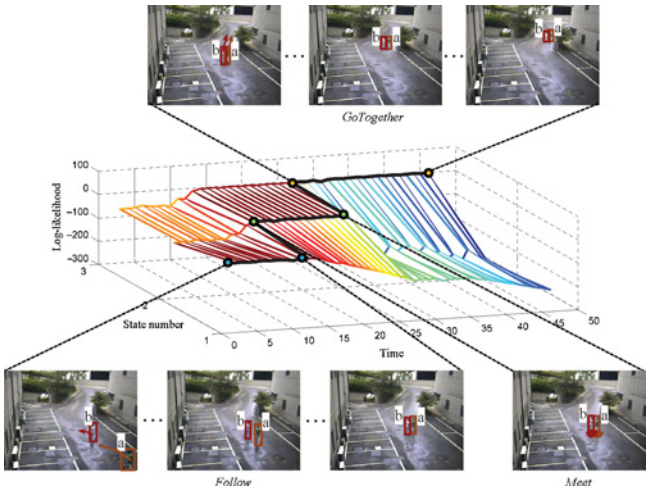


Fig. 6. Segmentation of an interaction sequence into sub-interactions via the Viterbi algorithm.

TABLE III
CONFUSION MATRIX FOR RECOGNITION OF SUB-INTERACTIONS

	FL	*AP*	*MT*	*GT*	*GS*	*PB*
Follow (FL)	5	0	0	1	0	0
Approach (AP)	0	5	0	0	0	1
Meet (MT)	0	0	6	0	0	0
GoTogether (GT)	1	0	0	5	0	0
GoSeparately (GS)	0	0	0	0	3	3
PassBy (PB)	0	0	0	0	2	4

observation is greater than that of the other states not related to it. In Fig. 6, the state transitions are marked with circles and solid lines in the graph. The video frames corresponding to each sub-interaction are also presented at the top and the bottom of the figure.

2) *Recognition of Sub-Interactions:* The hyperparameters, i.e., the number of states which each hidden variable in an MFHMM can take, were determined by cross-validation. We varied the number of states from two to five considering the complexity of the target sub-interaction patterns. We assume a left-to-right topology with a single Gaussian distribution for states.

We train sub-interaction models with 30 randomly selected interaction video sequences. Six out of the 15 sequences are selected for each interaction, and six test clips per sub-interaction are used to determine whether the sub-interactions are separable. The confusion matrix of the recognition result is given in Table III. These results show that the two sub-interactions, *GoSeparately* and *PassBy*, are confusing and appear to have very similar trajectories. So we combine these two sub-interactions and build a single model for them, except that the labels in the NDPM are different in order to ensure that they have a valid interpretation.

Since we have a limited number of training and test data of sub-interactions, we believe that it is not fair to compare the performance of the proposed MFHMM with that of a factorial HMM [38]. Besides our main objective is interaction recognition.

C. Recognition of Person-to-Person Interactions

We create a network of MFHMMs for recognition of the five interactions by replacing the standard HMMs in Fig. 2 with MFHMMs. Given a test video, we first find the sequence of sub-interaction models which maximizes the likelihood for the given observation $\mathbf{Y}$ of the test video, according to the function as follows:

$$\text{Interaction} = \underset{\hat{\mathbf{M}}}{\text{argmax}}\, P(\mathbf{Y}, \mathbf{M}|G, \Theta) \tag{20}$$

where $\Theta = \{\theta_{FL}, \theta_{AP}, \theta_{MT}, \theta_{GT}, \theta_{GS}\}$, $\theta_i = \{\pi_i, A_i, B_i\}$ denotes a set of parameters for a dynamic probabilistic model, $i \in \{$*Follow (FL)*, *Approach (AP)*, *Meet (MT)*, *GoTogether (GT)*, *GoSeparately (GS)*$\}$, and $\mathbf{M}$ represents the sequence of the labels of sub-interaction models in the network G.

1) *Performance of the Proposed Interaction Model:* We exploit the cross-validation technique with 20 interaction sequences for training and the remaining 55 interaction sequences for test. We further exploit 20 interaction sequences out of the 55 test sequences for validation. From ten repetitions of this process, where all the training, test, and validation datasets were selected randomly, the proposed NDPM showed an average recognition rate of 87.82% with a standard deviation of 2.71%.

For the analysis and understanding of the proposed NDPM, we illustrate the detailed results on a test video of *Follow* + *Meet* + *GoTogether* in Fig. 7. It presents the temporal evolution of each sub-interaction model's likelihood and the segmental results marked on the trajectories of human subjects in Fig. 7(a) and (b), respectively. Note that in Fig. 7(a), we plotted the likelihood values for every seventh frame starting from the 124th frame for a clear view. So the unit on the time-axis corresponds to the $(7 \times t + 124)$th frame in the input video. Fig. 7(a) explains the effectiveness of the proposed method by showing that the likelihood of the sub-interaction model dedicated to the observation sequences increases as more evidence becomes available. However, the likelihood of the other models unrelated to the observations stays low until the evidence represented by those model becomes available. That is, the *Follow* model outputs the highest likelihood for the input sequence up to the 215th frame, but after the 216th frame its likelihood reduces and that of *Meet* model dramatically increases. Last, the *GoTogether* model has the highest likelihood between the 271st and the 341st frames. In the likelihood graph, the slopes of the sub-interaction models connected via the network nodes are very similar with a time delay. This phenomenon is a result of passing the maximum likelihood at a network node to the ensuing sub-interaction models. We also illustrate the optimal sub-interactions and internal states transitions in Fig. 8. The vertical axis is the concatenation of the five sub-interaction models and the ticks for each sub-interaction denote the indices of the state pairs.

2) *Performance Comparisons with Other Models:* In this experiment, we compare the performance of our method to that of five other widely used methods in the literature: MOHMM [19], PaHMM [43], CHMM [9], ODHMM [10], and FHMM [38]. These dynamic probabilistic models, for which

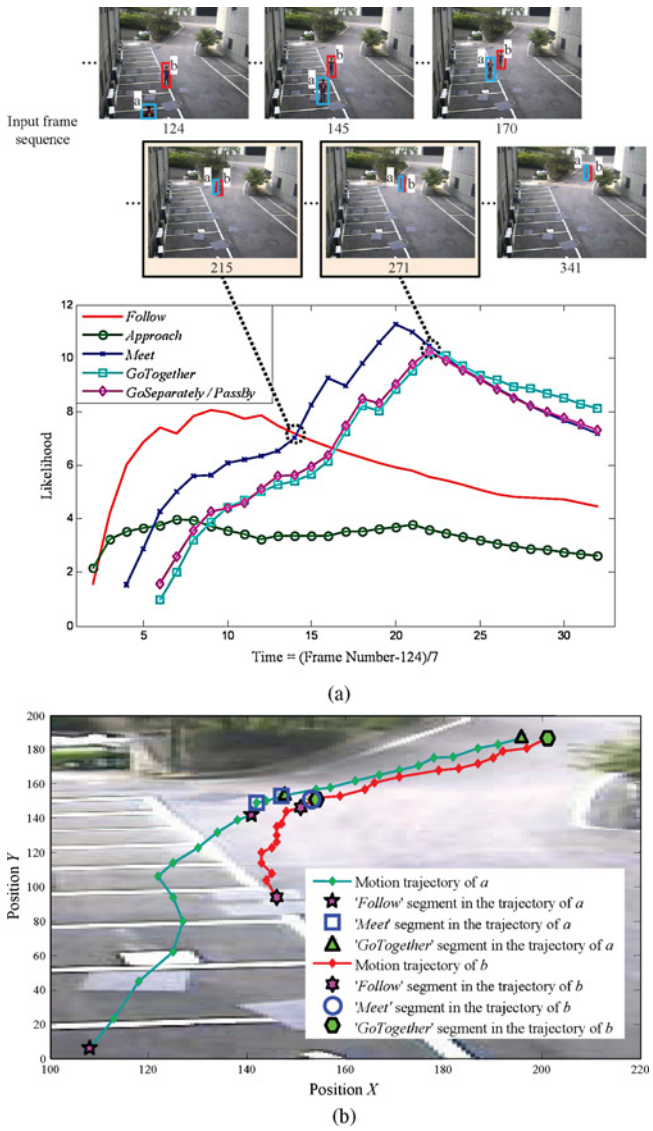


Fig. 7. Analysis of the results for the interaction *Follow* + *Meet* + *GoTogether*. (a) Temporal evolution of likelihood of the five sub-interaction models. (b) Segmentation results marked on the trajectories of the subjects.

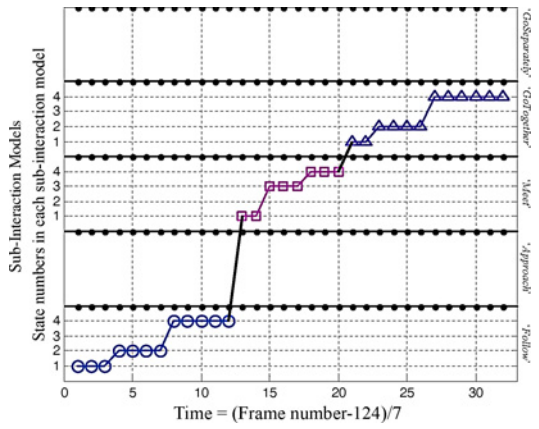


Fig. 8. Optimal sequence of sub-interactions and states within the sub-interaction models which produced the maximum likelihood for the video sequence shown in Fig. 7(a). The state numbers on the vertical axis denote the state pair indices in each sub-interaction model.

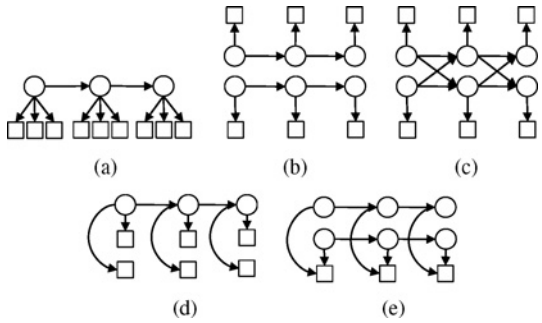


Fig. 9. Graphical representation of five different models in the literature which are used for performance comparison with the proposed model. (a) MOHMM. (b) PaHMM. (c) CHMM. (d) ODHMM. (e) FHMM.

the graphical representations are shown in Fig. 9, are trained for each interaction. We trained these models with the same training data, which were used for the training of the proposed NDPM. Note that at this time, each model represents the whole interaction, not sub-interactions. The hyperparameters of these models, i.e., the number of hidden states, are determined with a cross-validation technique, varying from five to nine. We implemented all these models in MATLAB using BNT [42]. More specifically, we converted them into DBNs and performed an exact inference with junction-tree [40] and interface [41] algorithms.

The performance of each model is given in Table IV. The proposed NDPM outperforms all the other models. The main reason for the low performance of the competing methods could be the limited size of the training data. The more parameters in a model, the larger the amount of training data required. In summary, we can say that the proposed method is effective in the sense that it can find optimal parameters with a small number of training samples due to the decomposed representation of the complete interaction.

3) *Mixture of Sub-Interaction Models in NDPM:* It is also possible to mix different types of sub-interaction models in a network. That is, we can use different models for each of the sub-interactions. For example, we can apply an MOHMM for a *Meet* sub-interaction and MFHMMs for the rest, reflecting the fact that the temporal pattern of *Meet* is relatively simpler than that of the others. In the following experiment, we build a number of NDPMs by considering all the possible combinations of dynamic probabilistic models used in the previous experiment, i.e., six different types of models for five sub-interactions, in total 5^6 different NDPMs. A combination of FHMM for *Follow*, MOHMM for *Meet*, and MFHMMs for the rest as presented in Fig. 10 resulted in the highest recognition of 98.18%, correctly classifying 54 out of the 55 video sequences.

4) *Test on Tsinghua University's Dataset:* We conduct experiments on a subset of the Tsinghua University's dataset [17], made available to us (44 video clips captured by a single static camera). The interactions are (TU-I) two persons walk in opposite direction and pass by, (TU-II) two persons run in opposite direction and pass by, (TU-III) two persons walk and approach from opposite directions and when meeting, they stand and chat and then resume walking in their initial

TABLE IV
PERFORMANCE COMPARISON WITH COMPETING METHODS IN THE LITERATURE ON IN-HOUSE DATABASE

	MOHMM [19]	PaHMM [43]	CHMM [9]	ODHMM [10]	FHMM [38]	MFHMM	NDPM
Recognition rate (%)	82.22	68.89	75.56	80	66.67	75.56	87.63

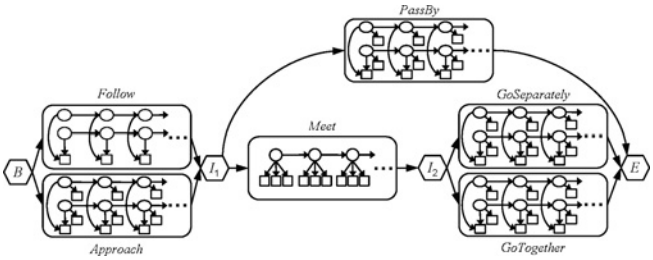


Fig. 10. Mixture of dynamic probabilistic models, FHMM for *Follow*, MOHMM for *Meet*, and MFHMMs for the rest, in a network which gave the best performance on the dataset.

Fig. 11. Representative frames for the four interactions in Tsinghua University's dataset [17].

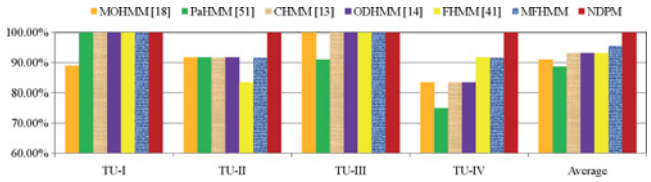


Fig. 12. Performance comparison of the proposed NDPM with competing methods in the literature on Tsinghua University's dataset [17].

directions, (TU-IV) two persons approach and meet, one puts an objects on the ground and goes away, and after a short while, the other takes the object and goes away. The representative frames for each of the four interactions are presented in Fig. 11. Based on the characteristics of subjects' trajectories in the video clips, we consider the TU-III as *Approach* + *Meet* + *GoSeparately*, and TU-I and TU-II as *Approach* + *PassBy*. In the case of TU-IV, two out of the 12 video clips belong to *Approach* + *Meet* + *GoSeparately* and the rest belong to *Approach* + *PassBy*. We ignore the intervention of an object, otherwise our sub-interaction models would need to be redesigned with appropriate features.

For recognition, we exploit the sub-interaction models trained on our self-collected dataset. Fig. 12 summarizes the performance of the proposed method and the competing methods. Except for the NDPM, each model represented a whole interaction using a single model. The proposed NDPM with MFHMMs for sub-interactions showed the highest performance among the seven different models; the proposed model correctly classifies all the video clips in the Tsinghua University's dataset.

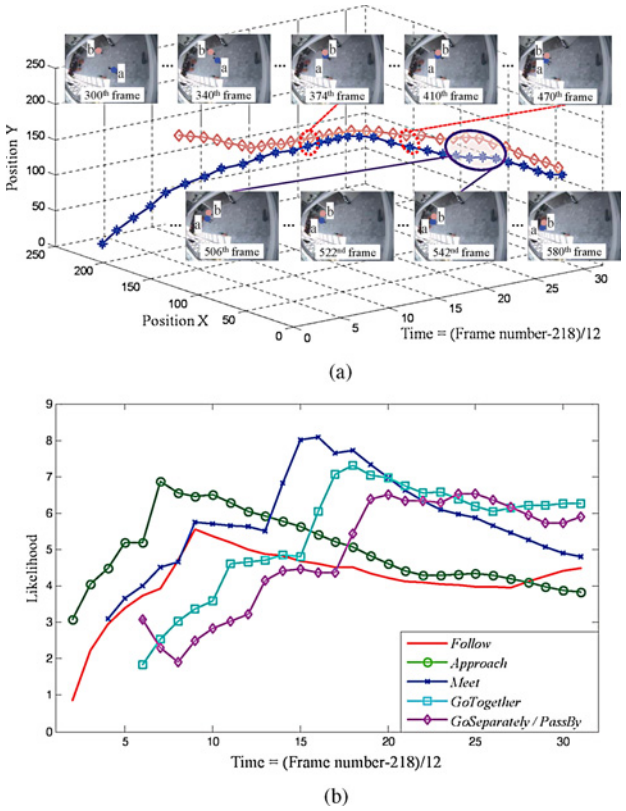


Fig. 13. Results on the interaction *Approach* + *Meet* + *GoTogether* in the CAVIAR dataset [30]. (a) *Approach* + *Meet* + *GoTogether* interaction in CAVIAR dataset [30]: motion trajectories of two subjects (middle), input frame sequences (top and bottom), and the time instants (red dotted circles) at which new sub-interactions begin to occur. The online color version provides a clearer view. (b) Temporal evolution of sub-interaction models' likelihoods.

5) *Test on CAVIAR Dataset:* We also apply our method to the *people/groups meeting, walking together*, and *splitting up* sequences in the CAVIAR project [30]. We again utilize the sub-interaction models trained on our self-collected dataset without any modification for the recognition and analysis of the CAVIAR dataset.

Because there are only three video clips in CAVIAR, which have similar interactions to our database, we do not compare the performances of different models, but only present the results obtained by the proposed NDPM with MFHMMs for sub-interactions. The first result is from the *Meet-WalkTogether2.mpeg* video clip (see Fig. 13), which shows that the proposed NDPM responds well to the change in subjects' motion trajectories. Since two of the three video clips include the same kind of interaction, we demonstrate the result for one of them. In Fig. 13, the two subjects *a* and *b* stayed at the same position for a while between the 374th and 410th frame. Then they walked together in the same direction. But around the 506th frame, subject *b* suddenly changed his/her direction

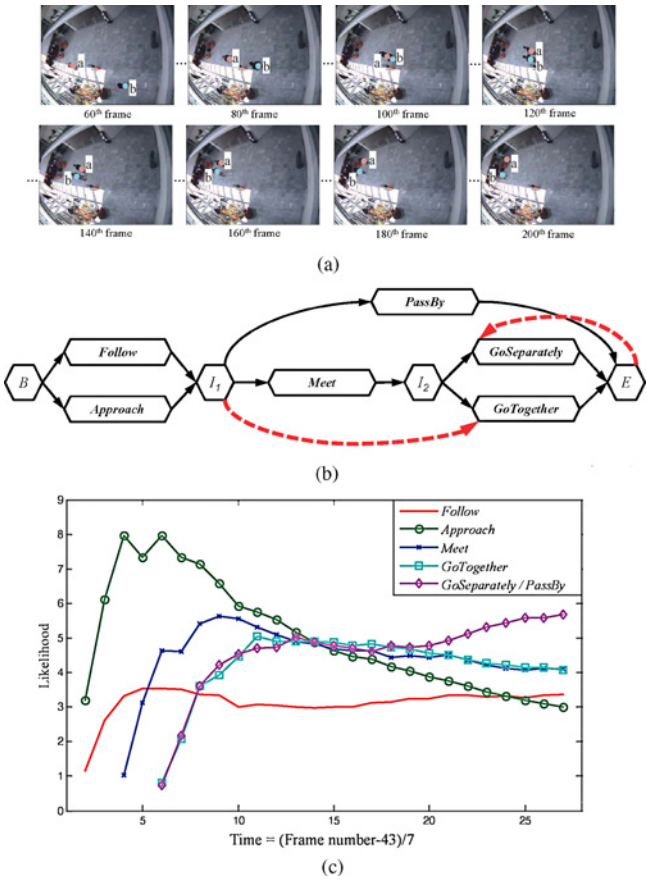


Fig. 14. Representing a new interaction by minimally adding paths to the trained NDPM. (a) Frame sequences of the interaction *Approach* + *GoTogether* + *GoSeparately* in the CAVIAR dataset [30]. (b) Modified NDPM for recognition and analysis of a new interaction (not available during training the model based on the self-collected dataset) in the CAVIAR dataset [30]. The dotted red arcs allow us to represent the new interaction making a cyclic path in the network. (c) Temporal evolution of likelihood of sub-interaction models in the modified NDPM.

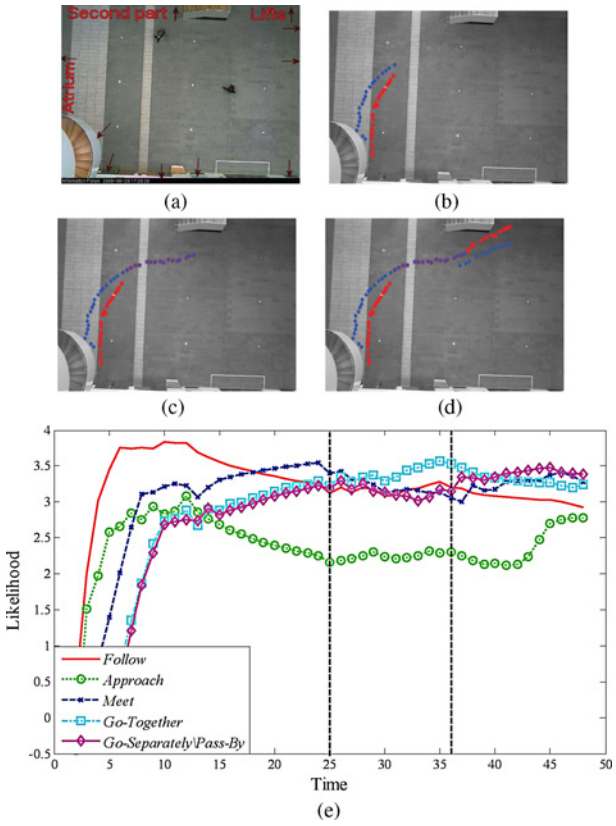


Fig. 15. Trajectories of two subjects in a file *track.Jun08.txt* of the EIFP database. The vertical dotted lines denote the time points that segment two consecutive sub-interactions. (a) View of the scene, (b) *Follow*, (c) *GoTogether*, (d) *GoSeparately*, and (e) temporal evolution of likelihoods of sub-interaction models in the NDPM of Fig. 14(b).

and went a different way, which caused the subject to become distant from subject *a*. This fact is reflected in the temporal evolution of the likelihood of sub-interaction models between time instants 24 and 27 in Fig. 13(b). We should mention that the ticks in the time-axis do not match the real frame numbers, because we plotted the figure with the likelihood computed every 12 frames for clarity of presentation.

The second result is from the *Meet-WalkSplit.mpeg* video clip. This video clip contains the interaction *Approach* + *GoTogether* + *GoSeparately*; the corresponding frame sequences are shown in Fig. 14(a) for every 20th frame. This interaction is not defined in our interaction scenarios and therefore, it cannot be recognized correctly with the current NDPM shown in Fig. 2, but with MFHMMs instead of the standard HMMs. However, we can accommodate this new interaction by simply adding new paths from the network node I_1 to the *GoTogether* model and *E* to *GoSeparately* model. The resulting NDPM is shown in Fig. 14(b), where the broken red arcs make a path of the successive occurrence of *GoTogether* and *GoSeparately* allowing a cyclic traverse. The change in the likelihood for each sub-interaction model is shown in Fig. 14(c), where the ticks in the time-axis do not match the frame numbers due to computation performed every seventh frame.

The fact that we can represent a new interaction without any modification or retraining of the sub-interaction models is one of the strong points of the proposed method. If sub-interactions not contained in the current NDPM are included in new interactions, we only need to learn those ones. The new interactions can then be recognized by incorporating the newly trained sub-interaction models into the NDPM with appropriate arcs. The reusability of the sub-interaction models for recognition of new interactions is another feature of the proposed method.

6) *Test on Edinburgh Informatics Forum Pedestrian (EIFP) Dataset:* Our final experiment was performed on the EIFP database [44] that consists of a set of detected targets of people walking through the Informatics Forum at the University of Edinburgh from the videos captured by a camera fixed overhead approximately 23 m above the ground [45]. A view of the scene and one of the sample video sequences detected by the proposed method is presented in Fig. 15. For the detection of the interaction *Follow* + *GoTogether* + *GoSeparately* the NDPM designed in Fig. 14(b) is used. Similar to our earlier experiments, the likelihoods of sub-interaction models of *Follow*, *GoTogether*, and *GoSeparately* responded well to the observations showing high likelihoods in the corresponding time points.

IV. Conclusion and Future Research

Understanding and analyzing human behavior in video is one of the challenging issues in computer vision. Vision-based surveillance systems have received a lot of interest and many studies have been conducted on understanding human-centric events in video sequences.

In this paper, we are concerned with analyzing various scenarios of person-to-person interactions based on the trajectories of human subjects. The primary contribution of this paper is the development of a novel framework, i.e., a NDPM, to represent complex interaction patterns. By representing complex patterns with a sequence of simple sub-interaction models and independently building these models, we can greatly reduce the computational complexity of the model and simplify the training problem. One of the advantages of the proposed method is that it can be extended to incorporate new interaction patterns, not seen during training, by minimally adding extra sub-interaction models and arcs. Furthermore, it is also possible to inject different types of dynamic probabilistic models for different sub-interactions in a network. That is, a standard HMM for one sub-interaction and a dynamic Bayesian network for another sub-interaction can be linked together in a network to represent interactions which include those sub-interactions. We also proposed a one-pass DP search [36] based inference algorithm for NDPM. In the proposed NDPM, it is natural to build a cyclic model to represent repetitive sub-interactions in a complex interaction without increasing the computational cost and redesigning the interaction models. While moving from an acyclic model to a cyclic model can be considered easy in terms of a graphical representation, when it comes to inference it is a challenging problem in machine learning. We have demonstrated the capability of the proposed NDPM in modeling cyclic patterns based on the experiments on the CAVIAR [30] and EIPF dataset [44]. Another feature of the method is the modified FHMM which characterizes the dynamic patterns by dividing observations into independent and shared components.

In our experiments with self-collected 75 video sequences, with 15 sequences per interaction, the proposed methods provided an average recognition rate of 87.82%, outperforming five other models proposed in the literature. Unlike previous methods, the proposed NDPM could combine different types of dynamic probabilistic models for different sub-interactions resulting in a mixture model. We also demonstrated the effectiveness and robustness of the proposed method by analyzing the internal organization of the NDPM and successfully applying our approach to Tsinghua University's dataset [17], the public CAVIAR dataset [30], and the EIFP dataset [44] with no retraining of the model that was trained on our self-collected interaction sequences.

Even though the scenarios considered in this paper are relatively simpler than encountered in real situations, we believe they provide elementary and necessary components for the understanding and interpretation of more complex human interactions in a visual surveillance task. Furthermore, it is possible to apply the proposed NDPM for detection of interested events or actions from a continuous video sequence. For instance, it can be used to detect person-to-person contact or secret rendezvous in a huge amount of video streams. The proposed method can be used in other domains. Sports video analysis [46] is one of the potential applications. The proposed method can be used in a system for summarization or classification of a play sequence into predefined tactical patterns and recognition of unknown patterns based on the analysis of players' movement in sports videos.

Limitations of the proposed approach are that it can only recognize and analyze predefined interactions and the proposed NDPM requires *a priori* knowledge about the structure of patterns which we want to model. It may be desirable to automatically build an NDPM which represents various interactions or, more generally, events occurring in various environments based on a large dataset. A data-driven method of creating an NDPM will be the focus of our future work.

Acknowledgment

The authors would like to thank Prof. F. Chen for providing Tsinghua University's dataset for performance comparison.

References

[1] H.-I. Suk, B.-K. Sin, and S.-W. Lee, "Analyzing human interactions with a network of dynamic probabilistic models," in *Proc. IEEE Workshop Applicat. Comput. Vision*, Dec. 2009, pp. 319–324.

[2] S. Ali, A. Basharat, and M. Shah, "Chaotic invariants for human action recognition," in *Proc. 11th IEEE Int. Conf. Comput. Vision*, Oct. 2007, pp. 1–8.

[3] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Patt. Recog.*, Jun. 2008, pp. 1–8.

[4] M. Ahmad and S.-W. Lee, "Human action recognition using shape and CLG-motion flow from multi-view image sequence," *Patt. Recog.*, vol. 41, no. 7, pp. 2237–2252, 2008.

[5] W. Lin, M.-T. Sun, R. Poovendran, and Z. Zhang, "Activity recognition using a combination of category components and local models for video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1128–1139, Aug. 2008.

[6] W. Li, Z. Zhang, and Z. Liu, "Expandable data-driven graphical modeling of human actions based on salient postures," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1499–1510, Nov. 2008.

[7] Y. Ivanov and A. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 22, no. 8, pp. 852–872, Aug. 2000.

[8] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[9] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, Aug. 2000.

[10] X. Liu and C. Chua, "Multi-agent activity recognition using observation decomposed hidden Markov models," *Image Vision Comput.*, vol. 24, pp. 166–175, Feb. 2006.

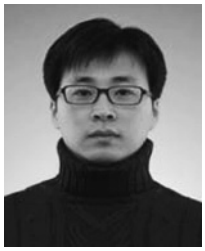
[11] M. Al-Hames, C. Lenz, S. Reiter, J. Schenk, F. Wallhoff, and G. Rigoll, "Robust multi-modal group action recognition in meetings from disturbed videos with the asynchronous hidden Markov model," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2. Sep. 2007, pp. 213–216.

[12] S. Fine, Y. Singer, and N. Tishby, "The hierarchical hidden Markov model: Analysis and applications," *Mach. Learn.*, vol. 32, no. 1, pp. 41–72, Jul. 1998.

[13] N. Oliver, A. Garg, and E. Horvits, "Layered representations for learning and inferring office activity from multiple sensory channels," *Comput. Vision Image Understand.*, vol. 96, no. 2, pp. 163–180, Nov. 2004.

[14] T. Duong, D. Phung, H. Bui, and S. Venkatesh, "Efficient duration and hierarchical modeling for human activity recognition," *Artif. Intell.*, vol. 173, nos. 7–8, pp. 830–856, May 2009.

[15] W. Lin, M.-T. Sun, R. Poovendran, and Z. Zhang, "Group event detection with a varying number of group members for video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 8, pp. 1057–1067, Aug. 2010.
[16] S. Park and M. Trivedi, "Understanding human interactions with track and body synergies (TBS) captured from multiple views," *Comput. Vision Image Understand.*, vol. 111, pp. 2–20, Jul. 2008.
[17] Y. Du, F. Chen, and W. Xu, "Human interaction representation and recognition through motion decomposition," *IEEE Signal Process. Lett.*, vol. 14, no. 12, pp. 952–955, Dec. 2007.
[18] W. Zhang, F. Chen, W. Xy, and Y. Du, "Hierarchical group process representation in multi-agent activity recognition," *Signal Process.: Image Commun.*, vol. 23, pp. 739–753, Nov. 2008.
[19] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 30, no. 5, pp. 893–908, May 2008.
[20] J. Sherrah, S. Gong, A. Howell, and H. Buxton, "Interpretation of group behaviour in visually mediated interaction," in *Proc. 16th Int. Conf. Patt. Recog.*, vol. 1. Sep. 2000, pp. 266–269.
[21] S. Hongeng, R. Nevatia, and F. Bremond, "Video-based event recognition: Activity representation and probabilistic recognition methods," *Comput. Vision Image Understand.*, vol. 96, no. 2, pp. 129–162, Nov. 2004.
[22] M. Ryoo and J. Aggarwal, "Semantic representation and recognition of continued and recursive human activities," *Int. J. Comput. Vision*, vol. 82, no. 1, pp. 1–24, Apr. 2009.
[23] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.
[24] S. Blunsden, E. Andrade, and R. Fisher, "Non parametric classification of human interaction," in *Proc. 3rd Iberian Conf. Patt. Recog. Image Anal.*, vol. 2. Jun. 2007, pp. 347–354.
[25] B. Takács, S. Butler, and Y. Demiris, "Multi-agent behaviour segmentation via spectral clustering," in *Proc. AAAI Workshop Plan, Activity Intention Recog.*, Jul. 2007, pp. 74–81.
[26] A. Hakeem and M. Shah, "Learning detection and representation of multi-agent events in videos," *Artif. Intell.*, vol. 171, pp. 586–605, Jun. 2007.
[27] H. Bui, D. Phung, and S. Venkatesh, "Hierarchical hidden Markov models with general state hierarchy," in *Proc. 9th Nat. Conf. Artif. Intell.*, Jul. 2004, pp. 324–329.
[28] N. Nguyen, D. Phung, H. Bui, and S. Venkatesh, "Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Patt. Recog.*, vol. 1. Jun. 2005, pp. 955–960.
[29] K. Daiki, O. Takayuki, and D. Koichiro, "HHMM based recognition of human activity," *IEICE Trans. Inform. Syst.*, vol. E89-D, no. 7, pp. 2180–2185, 2006.
[30] *CAVIAR Test Case Scenarios*. (2005, Mar.) [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1
[31] C. Pinhanez and A. Bobick, "Human action detection using PNF propagation of temporal constraints," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Patt. Recog.*, Jun. 1998, pp. 898–904.
[32] J. Allen, "Toward a general theory of action and time," *Artif. Intell.*, vol. 23, no. 2, pp. 123–154, Jul. 1984.
[33] Y. Shi, Y. Huang, D. Minnen, A. Bobick, and I. Essa, "Propagation networks for recognition of partially ordered sequential action," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Patt. Recog.*, vol. 2. Jun. 2004, pp. 862–869.
[34] M. Albanese, V. Moscato, R. Chellappa, A. Picariello, V. Subrahmanian, P. Turaga, and O. Udrea, "Constrained probabilistic Petri-net framework for human activity detection in video," *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1429–1443, Dec. 2008.
[35] S. Park and J. Aggarwal, "A hierarchical Bayesian network for event recognition of human actions and interactions," *Multimedia Syst.*, vol. 10, no. 2, pp. 164–179, Aug. 2004.
[36] H. Ney and S. Ortmanns, "Progress in dynamic programming search for LVCSR," *Proc. IEEE*, vol. 88, no. 8, pp. 1224–1240, Aug. 2000.
[37] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–285, Feb. 1989.
[38] Z. Ghahramani and M. Jordan, "Factorial hidden Markov model," *Mach. Learn.*, vol. 29, nos. 2–3, pp. 245–273, 1997.
[39] F. Jensen, *Bayesian Networks and Decision Graphs*. New York: Springer, 2001, ch. 1, pp. 3–34.
[40] C. Huang and A. Darwiche, "Inference in belief networks: A procedural guide," *Int. J. Approximate Reason.*, vol. 15, no. 3, pp. 225–263, 1996.
[41] K. Murphy, "Dynamic Bayesian network: Representation, inference and learning," Ph.D. dissertation, Comput. Sci. Division, Univ. California, Berkeley, Jul. 2002.
[42] *Bayes Net Toolbox for MATLAB* [Online]. Available: http://bnt.sourceforge.net
[43] C. Vogler and D. Metaxas, "A framework for recognizing the simultaneous aspects of American sign language," *Comput. Vision Image Understand.*, vol. 81, no. 3, pp. 358–384, Mar. 2001.
[44] *Edinburgh Informatics Forum Pedestrian Database* [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING
[45] B. Majecka, "Statistical models of pedestrian behaviour in the forum," M.S. dissertation, School Inform., Univ. Edinburgh, Edinburgh, U.K., 2009.
[46] D. Sadlier and N. O'Connor, "Event detection in field sports video using audio-visual features and a support vector machine," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 10, pp. 1225–1233, Oct. 2005.



Heung-Il Suk (S'08) received the B.S. and M.S. degrees in computer engineering from Pukyong National University, Busan, Korea, in 2004 and 2007, respectively. He is currently pursuing the Ph.D. degree from the Department of Computer Science and Engineering, Korea University, Seoul, Korea.

From 2004 to 2005, he stayed at the Computer and Vision Research Center, University of Texas at Austin, Austin, as a Visiting Researcher. His current research interests include machine learning, computer vision, and brain–computer interfaces.

Anil K. Jain (S'70–M'72–SM'86–F'91) is a University Distinguished Professor with the Department of Computer Science and Engineering, Michigan State University, East Lansing. His current research interests include pattern recognition and biometric authentication. He holds six patents in the area of fingerprints. He is the author of a number of books, including *Handbook of Fingerprint Recognition* in 2009, *Handbook of Biometrics* in 2007, *Handbook of Multibiometrics* in 2006, *Handbook of Face Recognition* in 2005, *BIOMETRICS: Personal Identification in Networked Society* in 1999, and *Algorithms for Clustering Data* in 1988. ISI has designated him a highly cited researcher. According to Citeseer, his book *Algorithms for Clustering Data* (Englewood Cliffs, NJ: Prentice-Hall, 1988) is ranked 93 in most cited articles in computer science.

He received the IEEE TRANSACTIONS ON NEURAL NETWORKS Outstanding Paper Award in 1996 and the Pattern Recognition Society Best Paper Awards in 1987, 1991, and 2005. He was the Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE from 1991 to 1994. He is a Fellow of the AAAS, ACM, IEEE, IAPR, and SPIE. He has received Fulbright, Guggenheim, Alexander von Humboldt, IEEE Computer Society Technical Achievement, IEEE Wallace McDowell, ICDM Research Contributions, and IAPR King-Sun Fu Awards. He was a member of the Defense Science Board and the National Academies Committees on Whither Biometrics and Improvised Explosive Devices.

Seong-Whan Lee (S'84–M'89–SM'96–F'10) received the B.S. degree in computer science and statistics from Seoul National University, Seoul, Korea, in 1984, and the M.S. and Ph.D. degrees in computer science from Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 1986 and 1989, respectively.

He is currently the Hyundai Motor Chair Professor at Korea University, Seoul, where he is the Head of the Department of Brain and Cognitive Engineering and the Director of the Institute for Brain and Cognitive Engineering. From 1989 to 1995, he was an Assistant Professor with the Department of Computer Science, Chungbuk National University, Cheongju, Korea. In 1995, he joined the Faculty of the Department of Computer Science and Engineering, Korea University, as a Full Professor. His current research interests include pattern recognition, computer vision, and brain informatics. He has more than 250 publications in international journals and conference proceedings, and authored ten books.

Dr. Lee was the winner of the Annual Best Student Paper Award of the Korea Information Science Society in 1986. He received the First Outstanding Young Researcher Award at the Second International Conference on Document Analysis and Recognition in 1993, and the First Distinguished Research Award from Chungbuk National University in 1994. He also received the Outstanding Research Award from the Korea Information Science Society in 1996. He is a Chairman or Governing Board Member of several professional societies. He was the founding Co-Editor-in-Chief of the *International Journal of Document Analysis and Recognition*. He has been an Associate Editor of several international journals including *Pattern Recognition*, *ACM Transactions on Applied Perception*, the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, IMAGE AND VISION COMPUTING, *International Journal of Pattern Recognition and Artificial Intelligence*, and *International Journal of Image and Graphics*. He was a General or Program Chair of many international conferences and workshops and was also on the program committees of numerous conferences and workshops.