Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Bi-dimension decomposed hidden Markov models for multi-person activity recognition*

Wei-dong ZHANG†, Feng CHEN, Wen-li XU

(*Department of Automation, Tsinghua University, Beijing 100084, China*)

†E-mail: zwd03@mails.tsinghua.edu.cn

Received May 22, 2008; Revision accepted Oct. 18, 2008; Crosschecked Apr. 10, 2009

Abstract: We present a novel model for recognizing long-term complex activities involving multiple persons. The proposed model, named 'decomposed hidden Markov model' (DHMM), combines spatial decomposition and hierarchical abstraction to capture multi-modal, long-term dependent and multi-scale characteristics of activities. Decomposition in space and time offers conceptual advantages of compaction and clarity, and greatly reduces the size of state space as well as the number of parameters. DHMMs are efficient even when the number of persons is variable. We also introduce an efficient approximation algorithm for inference and parameter estimation. Experiments on multi-person activities and multi-modal individual activities demonstrate that DHMMs are more efficient and reliable than familiar models, such as coupled HMMs, hierarchical HMMs, and multi-observation HMMs.

Key words: Multi-channel setting, Hierarchical modeling, Hidden Markov model, Activity recognition
doi:10.1631/jzus.A0820388 **Document code:** A **CLC number:** TP391.4

INTRODUCTION

Activity recognition has been one of the most active topics in computer vision for its huge number of potential applications, such as visual surveillance, human-computer interface, motion-based diagnosis and identification (Moeslund *et al*., 2006). As efficient models for learning sequential characteristics of data sequence, dynamic Bayesian networks (DBNs), especially hidden Markov models (HMMs), have been widely used in recognition of individual activities and simple interacting activities. However, few models have been proposed for modeling long-term complex activities involving interactions of multiple persons performing multiple actions simultaneously, which presents both of spatial and temporal difficulties for traditional models.

From spatial aspect, real scenes of interest often contain multiple persons and complex interactions between them (Intille and Bobick, 2001), which brings several challenges. First, large feature space for multiple persons suffers from dimension disaster. Second, the variable number of persons makes most exsiting methods unsuitable because of the uncertain dimension of feature vectors and uncertain correlations between persons (Wada and Matsuyama, 2000; Liu and Chua, 2006). Third, properly modeling asynchrony and correlations between multiple dynamic processes is a challenging problem. Standard HMMs suffer from feature space and state space exploring. Although some multi-modal models were presented (Brand *et al.*, 1997; Gong and Xiang, 2003; Du *et al.*, 2008), they are computationally intractable when the model consists of more than two chains.

From temporal aspect, how to capture the inherent hierarchical structure of activity, which presents long-term dependency and multi-scale characteristics, is a challenging problem. From perceptual psychology viewpoint (Zacks and Tversky, 2001), activity can be viewed as composition of actions having some orderly relations. The 'part of' relations

* Project (No. 60772050) supported by the National Natural Science Foundation of China

between parts and sub-parts constitute a partonomic hierarchy. Orderly relations present the sequential characteristics. Usually, correlations between sub-parts of activities do not decay as quickly as expected by standard HMMs, which is difficult to handle under the Markov assumption. Activities can also be viewed as existing at various levels of abstraction, which is named 'taxonomic hierarchy' reflecting multi-scale characteristics. If flat models such as HMMs and coupled HMMs (CHMMs) are used, the complexity of the model will become intractable with increasing errors and over-fitting in training.

Furthermore, it is necessary and beneficial to model both of spatial and temporal characteristics of multi-person activities. However, existing models can reflect only one aspect directly. Hierarchical HMMs (Fine *et al*., 1998; Bui *et al*., 2002; Oliver *et al*., 2004) solve the hierarchical structure to some extent, but involve a combinatoric number of states to recognize multi-person activities. Multi-channel HMMs (Brand *et al*., 1997; Gong and Xiang, 2003) and multi-observation HMMs (Liu and Chua, 2006) adopt a compositional representation of two or more variables for multiple processes, but cannot represent the hierarchical structure of activities. Therefore, it is necessary to present some new tractable models to handle the two aspects simultaneously.

In this paper, we present a new network that can recognize long-term complex multi-person activities even when the number of persons is variable. The proposed model, named 'decomposed hidden Markov model' (DHMM), decomposes the states of traditional HMMs in multi-scale time and multi-modal space. The compositional states bring conceptual advantages of parsimony and clarity, with consequent computational benefits in efficiency and accuracy. In DHMMs, decomposition in space generates multiple coupled HMM chains to model multi-person interactions, and the number of chains varies as the number of persons does (In general, spatial decomposition cannot be uniquely decided by the person number. Sometimes several persons share one chain or one person uses several chains.). A relation layer uncouples these chains for a more compact representation while keeping the causal-temporal influences. Higher multi-level abstracts capture long-term dependency and multi-scale characteristics of multi-person activities. The general structure is shown in Fig.1.

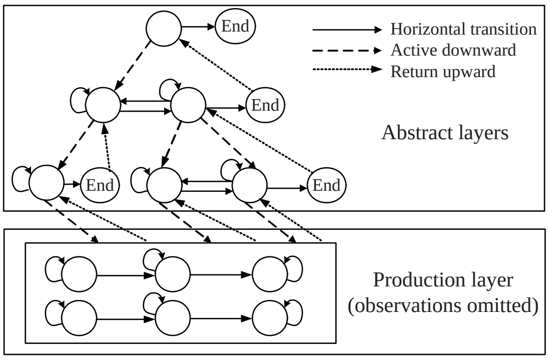

Fig.1 General structure of the decomposed hidden Markov model

RELATED WORK

In the past two decades, there were significant researches on human activity recognition based on DBNs, especially HMMs and their extensions (Moeslund *et al*., 2006). It is well known that standard HMMs suffer from feature space and state space exploring for multi-modal problems, and that the performance of HMMs tends to degrade when long-term dependence and multi-scale characteristics are presented as the sequence length increases.

Some models were presented to model long-term dependence and multi-scale characteristics. Abstract HMMs (Bui *et al*., 2002) and hierarchical HMMs (HHMMs) (Fine *et al*., 1998; Nguyen *et al*., 2005) exploit the hierarchical characteristics of activities by recursive layer structure, and use a special 'end' state to control the return to the higher level. Layered HMMs (Oliver *et al*., 2004) classify activities level by level using standard HMMs to capture different levels of abstraction and corresponding duration. Du *et al*. (2008) presented hierarchical DBNs with duration states at high levels to represent multi-scale characteristics of activities. However, it is necessary to merge multiple feature vectors into a high-dimensional vector when multi-person activities are concerned. Hierarchical networks still suffer from the large feature/state space problem.

To address multi-channel interactions, Brand *et al*.(1997) improved linked HMMs consisting of two coupled chains evolving in lockstep by CHMMs. They introduced coupling between time slices to capture causal-temporal influences between multiple interactive chains. Dynamically multi-linked HMMs

(DML-HMMs) (Gong and Xiang, 2003) consist of a more optimized factorization of state transition matrices and have fewer state connections than CHMMs. Observation decomposed HMMs (ODHMMs) (Liu and Chua, 2006) decompose the original observation into a set of sub-observations to recognize multi-agent activities. Zhang D. *et al*.(2006) used a two-layer HMM to recognize meeting activities. Du *et al*.(2007) used level-by-level interacting networks to recognize multi-scale dynamics of interactions. These models handle multi-modal problems to some extent. However, these models are very complex networks that can be intractable for more than two chains (persons), and they cannot represent the hierarchical structure of activities.

None of these models can represent spatial and temporal characteristics simultaneously. We present a new network by decomposing a state in two dimensions of space and time, to combine multiple sequences and represent the hierarchical structure of activities. The proposed model has a more compact structure, and is more efficient and reliable for multi-person activity recognition.

DECOMPOSED HIDDEN MARKOV MODEL

DHMMs decompose the state of standard HMMs in multi-modal space and multi-scale time, and decompose the observations in space. One example represented as a DBN is shown in Fig.2. In DHMM, there is a production layer that consists of several independent state chains. Each chain evolves on its own dynamics, produces its own observations, and can start/end at any time. So the number of chains varies flexibly. Above the production layer, the DHMM has multiple abstract layers like HHMMs (Murphy and Paskin, 2001) to represent activities at multiple scales. We especially name the lowest abstract layer as the 'relation layer', which has influences on all the chains at the production layer.

The original observation is decomposed into a set of sub-observations, each depending on only one production-state chain. The model of Fig.2 has two layers of hierarchical abstraction and three chains in the production layer, of which one starts at time *t*. The state and observation at chain *k* and time *t* are represented by Q_t^k and O_t^k, respectively. H_t^k denotes the state at abstract layer *k* and time *t*, and F_t^k is an indicator variable indicating when the Markov chain at layer *k* enters an end state and returns to the (*k*+1)th layer. The downward arcs between the *H* variables represent vertical transitions, while the downward arcs between the relation layer and the production layer represent influences.

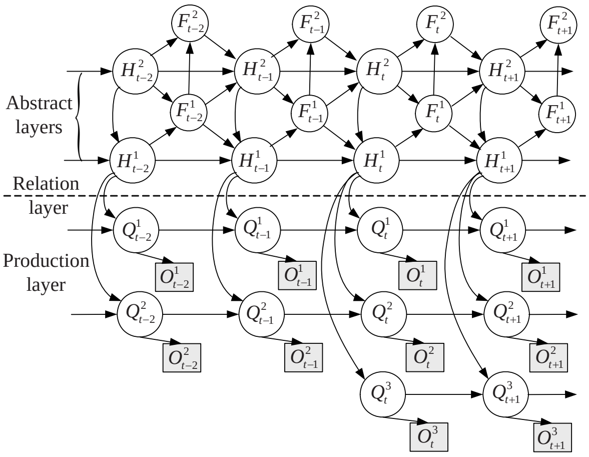

Fig.2 Dynamic Bayesian network representation of a decomposed hidden Markov model example

Spatial decomposition

Asynchrony of multi-channel sequences will lead state and feature space of single-observation models (such as HMMs and hierarchically structural HMMs) growing rapidly to become intractable. Some multi-modal models such as CHMMs and DML-HMMs are proposed. These models are tractable in space with *KN* states as opposed to N^K of combinatoric HMMs, where *K* is the number of chains and *N* is the number of chain states. However, these models are not suitable for multi-person activity recognition when the person number is variable, or for hierarchical representation. Furthermore, they are densely connected models and take an $O(KN^{K+1})$ parameter set. Training will suffer over-fitting with increasing errors when *K* becomes large.

To solve the multi-modal problem using compact models, we introduce a novel multi-channel setting, which reduces the model complexity greatly. A DHMM decomposes dynamically the state of HMMs in space into several coupled variables to capture the spatial structure of activities. The correlations between variables are then uncoupled by a relation layer; that is, there are only influences of the relation layer on production state chains. These chains evolve on their own dynamics, and can start/end at any time, while chains in CHMMs have influences on each other and start/end simultaneously. In multi-person activity recognition, we consider multiple persons interacting to form a group, and the group has influences on each person. The dynamics of the group are captured by the relation layer, and each person is modeled by an HMM chain. DHMMs reflect the influences of the group on individuals, while CHMMs reflect direct influences between individuals. Since the relation layer captures the dynamics of all the chains, it has a combinatoric state space of these chains theoretically. Fortunately, not all correlations exist all the time but a few of them need to be modeled in a moment, so combinations in reality usually distribute in a sparse manner, which keeps the state space in a low complexity. Observation is also decomposed and each of sub-observations depends only on the corresponding state chain.

The spatial decomposition of DHMMs brings several advantages. Firstly, after observation decomposition, DHMMs avoid the problem of large feature space and the problem of transitions between different feature spaces, and then feature selection of each chain becomes more flexible. DHMMs can model feature sets at different time scales, which makes models suitable for individual activities with multi-scale features, such as the global features and the local features used in (Du *et al*., 2008). Secondly, decomposition in space reduces the complexity of the model. Relation uncoupling reduces further the number of parameters. The models take KN states and $O(KN^3)$ parameters, while HMMs take N^K states and CHMMs take $O(KN^{K+1})$ parameters. Thirdly, using one or several HMM chains to model one person provides more details about the internal dynamics of individual actions. Because of independence between individual channels, loss in one channel does not affect others. DHMMs are not sensitive to the varying of the person number. Finally, complexity reduction and independence in model simplify the inference and learning, and reduce the chance of over-fitting and errors in training.

Temporally hierarchical abstraction

Activities have naturally a hierarchical structure that presents multi-scale characteristics and long-term dependencies. The dependency of a sequence does not decay as the Markov assumption expects. Flat models, such as HMMs, CHMMs and ODHMMs are not competent anymore. The number of parameters rapidly becomes intractable as the number of scales increases and the probability of a sequence decreases exponentially with the increase of sequence length. If an orderly characteristic is followed in activities, we think that a high activity can be decomposed recursively into a sequence of simpler activities until primitive actions are reached. To represent multi-scale and long-term dependent characteristics, we introduce the hierarchical network.

Above the production layer, DHMMs are structured multi-level stochastic processes to capture the hierarchical structure of activities. We usually assume abstract layers have a tree structure (Fig.2). The lowest layer, i.e., the relation layer, represents the causes (goal) of multi-person interactions which provide a higher level of description than individual actions. DHMMs model further group activities at multiple levels of abstraction to represent the taxonomic characteristic. Every state in the high level of abstract layers consists of a sequence of lower-level states. The transitions in the same layer indicate the partially ordered relations of activity parts.

Fig.3 shows why the hierarchical model is curial and how it handles the long-term dependency in activities. Person 1 interacts with persons 2 and 3 at conjoint intervals to finish a sub-activity. Sometimes activities are performed following different global orders but keeping local orders. Assuming that each sub-activity consists of N states, the flat model implements transition between sub-activities with N^2 possible one-step paths. The hierarchical model abstracts a long chain of simpler actions into a single sub-activity that carries all needed historical information. It takes $O(N)$ transition parameters. The higher layer represents larger-scale characteristics, and captures the correlations between states far away; that is, DHMMs capture long-term dependences using higher layers. The hierarchical structure reduces the model complexity greatly, compared to standard HMMs. Furthermore, it can also simplify parameter learning by sub-models learning separately.

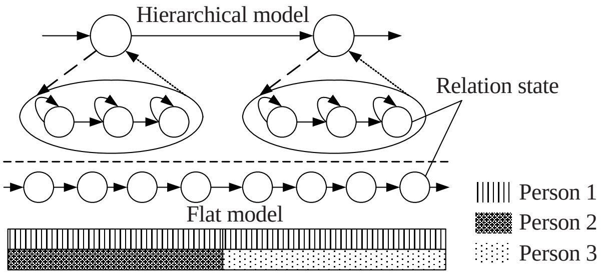

Fig.3 Demonstration of the flat model and the hierarchical model for multi-person activities

Special cases and related models

A Bayesian network is a graphical way to represent the conditional independence of a joint distribution. There are many ways to factorize a joint distribution, and consequently there are many Bayesian networks consistent with a particular joint (Ghahramani, 2001). When the internal dynamics of individual actions is ignored, we obtain a multi-observation hierarchical HMM (MOHHMM), as shown in Fig.4. Learning with a large dataset, they would gain better performance for a complex sequence than ODHMMs in theory because of their hierarchical network. If the model has only one abstract layer (i.e., the relation layer), it will be a two-layer influence model similar to that in (Zhang D. *et al.*, 2006). However, the model does not consider the evolution of group process, and thus cannot be extended to hierarchical modeling easily.

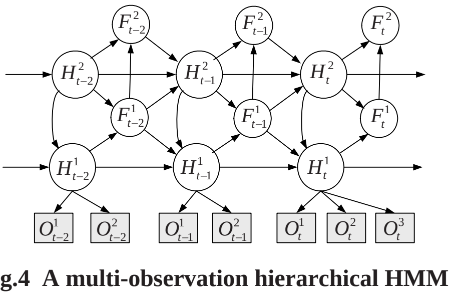

Fig.4 A multi-observation hierarchical HMM

The biggest difference between DHMMs and other models is that DHMMs represent both of spatial and temporal characteristics while other models represent only one of them. DHMMs are also less complex than others for multi-person activity recognition with consequent computational advantages of less cost and smaller errors in training.

Once sub-observations are merged into large ones when the features extracted from multiple persons have similar distributions, the models have the same structure as HHMMs. But DHMMs are more flexible to model multi-modal sequences with a more compact model. DHMMs are also related to CHMMs. CHMMs model interactions of multiple HMM chains by directly linking the cross time slices and cross chain states as $P(S_t^i \mid S_t^1, S_t^2, \cdots, S_t^K)$, while in DHMMs the current state of each chain is only influenced by the previous state of the same chain and the relation state as $P(Q_t^k \mid Q_{t-1}^1, H_t^1)$. Considering the special implementation of a DHMM as shown in Fig.5, we set H as the related parts of S^1 and S^2, and Q^1 and Q^2 as the unrelated parts, and then we have $H=S^1 \cap S^2$ and $Q^1 \perp Q^2|H$. The model is exactly equivalent to CHMMs but with fewer parameters and lower complexity. When the dependency of observations on the relation layer is small, the model can be simplified to standard DHMMs. However, the DHMM has a more compact structure than the CHMM, and can represent the hierarchical structure of acidities.

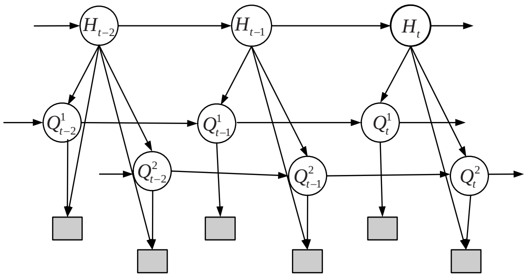

Fig.5 A special DHMM with observations depending on the relation layer

LEARNING AND INFERENCE

Parameter estimation in DHMMs

In a DHMM, the set of parameters, denoted as λ, includes three parts: initial parameters Π, transition parameters A, and observation distributions B. We assume K_h to be the index of the highest abstract layer, K_q the number of HMM chains at the production layer, and N the maximum number of states of each chain at all layers. The state variables in abstract layers can be encoded as $H_t = (H_t^1, H_t^2, \cdots, H_t^{K_h})$, and the state variables in the production layer are encoded as Q_t with varying dimensions. Define $R_t=(H_t, Q_t)$. The parameters Π, A, and B are defined as follows.

(1) Initial parameters *Π*.

π_i^{h,K_h} : the probability that the top activity begins at sub-activity *i*. $\pi_{i|l}^{h,k}$: the probability that the *k*th layer activities begin at sub-activity *i* conditioned on the higher activity *l*. $\pi_{i|l}^{q,k}$: the probability that the *k*th person firstly does action *i* conditioned on relation *l*.

(2) State transition parameters *A*.

First, we define $A_{i,j|l}^{h,k}$ as the transition probability from activities *i* to *j* conditioned on relation *l*, $\tau_{i|l}^{h,k} = A_{i,\mathrm{end}|l}^{h,k}$ the probability of terminating from state *i*. The transition parameters $a_{i,j|l}^{h,k}$ in the original HHMMs (Fine *et al*., 1998) can be obtained as $a_{i,j|l}^{h,k} = A_{i,j|l}^{h,k}(1-\tau_{i|l}^{h,k})$. Then we define $P(F_t^k \mid H_t^k, H_t^{k+1}, F_t^{k-1})$ and $P(H_t^k \mid H_{t-1}^k, F_{t-1}^{k-1}, F_{t-1}^k, H_t^{k+1})$ the same as in (Murphy and Paskin, 2001).

To formulate the transition probabilities of production layer states, we introduce another indicator $\{V_t^k\}$ decided by the feature extraction module. $V_t^k = 1$ indicates the *k*th person is observable, $V_t^k = 0$ indicates the *k*th person appears or reappears at time *t*, and $V_t^k = -1$ indicates the *k*th person is unobservable. Then we define $A_{i,j|l}^{q,k}$ being the transition probability from *i* to *j* conditioned on *l*. Then, the conditional probability distributions of the production layer are defined as

$$P(Q_t^k = j \mid Q_{t-1}^k = i, H_t^1 = l, V_t^k = v) = \begin{cases} 1, & \text{if } v = -1, \\ \pi_{i|l}^{q,k}, & \text{if } v = 1, \\ A_{i,j|l}^{q,k}, & \text{if } v = 0. \end{cases} \tag{1}$$

(3) Observation distribution $B = \{b_{O_t^k|i}\}$.

$$b_{O_t^k|i} = \sum_{n=1}^{C^k} w_{i,n}^k \cdot N(O_t^k, \mu_{i,n}^k, \Sigma_{i,n}^k). \tag{2}$$

Then, the observation distributions are defined as

$$P(O_t^k \mid Q_t^k = i,\ V_t^k = v) = \begin{cases} 1, & \text{if } v = -1, \\ b_{O_t^k|i}, & \text{else}, \end{cases} \tag{3}$$

where $N(O_t^k, \mu_{i,n}^k, \Sigma_{i,n}^k)$ denotes a Gaussian distribution, $\mu_{i,n}^k$ and $\Sigma_{i,n}^k$ are the mean and covariance respectively, C^k is the number of components of the mixture Gaussian distribution, $w_{i,n}^k$ is the weight, and O_t^k is the sub-observation of the *k*th production state chain at time *t*.

The number of parameters of the production state transition model is $O(K_qN^3)$, and that of the abstract state transition model is $O(K_hN^3)$. A general maximum a posterior expectation maximization (MAP-EM) algorithm is used for parameter estimation based on the generalized forward-backward iterations. This algorithm is derived in a similar fashion to the learning algorithms of HMMs. In multi-person activity recognition, it is hard to collect a set of sequences to capture all the possible transitions while the segmented data seem to be collected easily. We can learn the sub-HMMs separately and combine them together.

Calculating the likelihood

Murphy (2002) presented an $O(T)$ inference algorithm for hierarchical networks with an equivalent DBN representation instead of the original inference algorithm described in (Fine *et al*., 1998) for HHMMs, which takes $O(T^3N^K)$ time by looping over all possible lengths of subsequences generated by each Markov model at each level, where *T* is the sequence length, *N* is the maximum number of states at each layer, and *K* is the depth of the hierarchy. As long-term multi-person activity recognition is concerned, *T* tends to increase rapidly. While *N* does take a large value in applications because of the balance of performance and parsimony or simplicity (Forster, 2000), we adapt and improve Murphy’s algorithm in our inference.

Similar to the way we define the inference of HMMs, we define global transition probabilities $a_{r,r'}^f = P(R_{t+1} = r' \mid R_t = r, F_t = f)$ and forward operators $\alpha_t(r)$ as the probabilities of the observations up to time *t* and the DHMM in state *r*. $\alpha_t(r)=P(O_{1:t}, R_t=r|\lambda)$, and we can solve it inductively:

$$\alpha_{t+1}(r') = b_{r'}(O_{t+1})\sum_r \sum_f \alpha_t(r) a_{r,r'}^f. \tag{4}$$

Similarly, we can obtain the generalized backward algorithm. This takes $O(TN^{2(K_h+K_q)})$, where *T* is the sequence length, K_q is the number of persons, K_h is the number of abstract layers, and *N* is the (maximum) number of states at each channel and each level, and this becomes computationally difficult as the number of chains increases. This inference cannot be used directly for a variable person number. We assume that forward operators satisfy the following condition:

$$\begin{aligned}\alpha_t(r) &= \phi_t P(H_t = h)P(O_{1:t}, Q_t = q, H_t^1 = h^1) \\ &= \phi_t' P(H_t = h)\prod_k P(O_{1:t}^k, Q_t^k = q^k, H_t^1 = h^1),\end{aligned} \tag{5}$$

where ϕ_t' is a scale. We redefine forward operators as

$$\alpha_t^k(h^1, q^k) = P(O_{1:t}^k, H_t^1 = h^1, Q_t^k = q^k \mid \lambda), \tag{6}$$

and $\alpha_t^H = P(H_t = h \mid \lambda)$. We compute forward operators independently, and how one operator is doing does not affect others. We obtain the likelihood as

$$P(O \mid \lambda) \propto \sum_h \alpha_T^H(h) \sum_q \prod_k \alpha_T^k(h^1, q^k). \tag{7}$$

Via this assumption, we reduce the complexity to $O(T(N^{2K_h} + K_q N^4))$. When approximate inference techniques such as belief propagation are used, the computational complexity is reduced further to $O(T(K_h^2N^2+K_qN^4))$. Usually, we will use 2~3-level hierarchical models, while the number of persons may be large. So the complexity can be simplified as $O(TK_qN^4)$, which is tractable in applications and much smaller than that of traditional multi-modal models that take $O(TN^{2K_q})$ for K_q-person activity recognition. The inference complexity of our model increases linearly with the increase of the numbers of persons and abstract layers, which is an expected characteristic for all multi-channel models. From Eqs.(5)~(7), we can see that the reduction of inference complexity is the result of independence between person chains and our assumption, which proves again that our model is parsimony and less complex in computation.

EXPERIMENT RESULTS

Experiments were conducted to recognize interacting activities between two or three persons in the scenes of a parking lot and footway. Two kinds of individual activities on square were used to demonstrate the flexibility of the model in feature decomposition. Typically, it is easy and unnecessary to do role assignment in applications such as visual-audio activity recognition and two-person interactions, because it is easy to differentiate features (modalities) from each other. We did only role assignment and tested its performance in multi-person activity recognition.

Multi-person activity recognition

We collected our multi-person activity dataset similar to the one in (Liu and Chua, 2006), but our activities were performed in longer time and larger space (Fig.6). The video sequences consist of 352×288-pixel color images at a 15 Hz frame rate. Each of them was acted by three persons for about 50 times, and lasted around 1 min and 900 frames. For each activity, about half of its dataset was used for training. Initially, persons 1 and 2 formed group A, and person 3 formed group B.

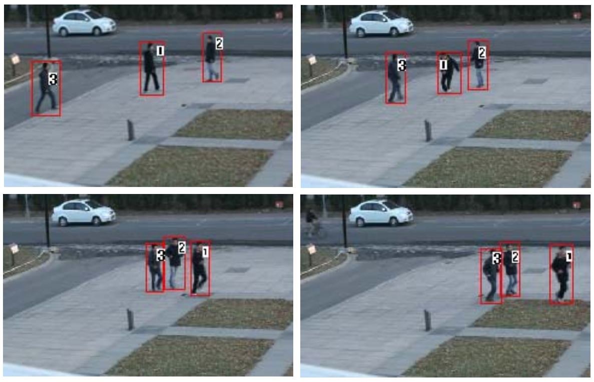

Fig.6 Key frames of three-person activities

Act. 1 Two groups approached oppositely and met, persons 1 and 3 formed a new group and turned to a new direction, and 2 kept his direction.

Act. 2 Similar to Act. 1, but after meeting, persons 1 and 3 turned back, and 2 kept walking and formed a new group with 3.

Act. 3 Initially, persons 1 and 2 approached 3 oppositely and 2 followed 3. Before 1 and 3 met, 3 suspected and turned around, then 3 found 2, and then 3 ran off in another direction with 1 and 2 chasing him.

Act. 4 Persons 1 and 2 followed 3, and after a while 1 sped up and tried to attract 3's attention. 2 approached and snatched 3's belongings and ran off in a new direction with 3 chasing him.

Feature ($\boldsymbol{p}$, $\boldsymbol{v}$, η) is extracted based on the results of people tracking using our previous system (Zhang W. *et al.*, 2006), where $\boldsymbol{p}=(p_x, p_y)$ is the position, $\boldsymbol{v}=(v, \theta)$ is the magnitude and direction of velocity, and $\eta=\bar{v}/\|\bar{\boldsymbol{v}}\|$ is the ratio between the average speed and the norm of the average velocity. There were several track failures because of serious occlusions. The features extracted from one person were used as inputs to the corresponding HMM chain in the production layer of DHMMs.

The DHMM must be designed according to the activity's temporal and spatial characteristics. Generally, the more complex the activity is, the more abstract layers and states in each layer the DHMM has. The initial number of HMM chains at the production layer is usually selected according to the person number, but the number of HMM chains is unknown when there are some chains that start or end. Different from many researchers who learned the structure of the model directly from the dataset, we designed DHMMs by priori information and values of the Bayesian information criterion (BIC) (Schwarz, 1978) to test the performance.

Firstly, the interacting activity is divided into a sequence of sub-activities and, in turn, the sub-activities are refined into a lower layer until primitive actions are reached. Secondly, the number of the actions of each person defines the corresponding chain's state number in the production state layer; the group activity defines the root; each decomposition of the activity defines an abstract layer using a top-down strategy. A tuple $<K_h,\ K_q,\ N_{\{1,2,\ldots,K_h\}},\ M_{\{1,2,\ldots,K_q\}}>$ defines the structure and state space of a DHMM. N_i is the number of states of the ith layer of the abstract layers, and M_i is the number of states of the ith channel in the production layer. In our experiments, for Acts. 3-4 activity recognition, $K_h=2$, $K_q=3$, $N_1=8$, $N_2=6$, $M_i=7$, and the model took state space of 35 and $O(10^3)$ parameters. The inference complexity of the model was $O(12\,544\times T)$. Our activity lasted 1 min, and T was around 900. A total of 10^7 operations were required, which is tractable.

Role assignment is a difficult problem in multi-mode and multi-channel sequence processing problems. Typically, it is done manually in such a way as in (Brand *et al.*, 1997; Gong and Xiang, 2003; Zhang D. *et al.*, 2006; Du *et al.*, 2007; 2008). However, when the number of persons is large, role assignment becomes more difficult. We develop a method similar to that in (Liu and Chua, 2006) to view the roles of agents as unknown parameters and integrate them out by summing all feasible solutions' matching values. We introduce role parameters R as $R=\{R_i^r(O_t^k)\mid R_i^r(O_t^k)=P(O_t^k\mid Q_t^k=i,\ role_k=r)\}$, where $role_k$ means the role of agent k. R^r is viewed as the output probability of role r. When the likelihood is computed, role parameters are multiplied with Eq.(7) and all feasible solutions are summed. Since the estimation of the optimal state path is independent of any agent assignment, we replace the whole right term of Eq.(7) with single likelihood probability under the optimal state path.

We recognized activities using standard DHMMs, MOHHMMs, and ODHMMs. We also compared them with role assignment (+R). The results are shown in Table 1. ODHMMs suffer the same problem as HMMs do when the range of activities becomes more complex. As the sequence length increases, ODHMMs degrade while DHMMs and MOHHMMs keep their performance. Although ODHMMs achieve overall high recognition rates, we can conclude that DHMMs have better results when the activities last longer and have more complex structures, as shown in Table 1. DHMMs will be more accurate for their ability to model individual person's dynamics in theory; however, we learned DHMMs using segments because of their large parameter space, and some local optimization was gained for Act. 3. While MOHHMMs can be learned using the whole sequences, they perform comparably to DHMMs. The results show that none of these models are sensitive to the occasional track failures.

Table 1 Comparison with ODHMMs for multi-person activity recognition

Method	Recognition rate (%)			
	Act. 1	Act. 2	Act. 3	Act. 4
ODHMM	92.0	91.8	83.3	81.2
MOHHMM	96.0	95.9	91.3	91.3
DHMM	96.0	95.9	89.6	91.3
ODHMM+R	89.8	87.9	81.2	81.2
MOHHMM+R	91.3	91.3	86.5	86.5
DHMM+R	91.3	89.6	82.3	89.6

'+R' means with role assignment

We cannot know the exact roles of agents before recognizing their activities. Since the role parameters are summed for all feasible solutions in computation, it is a compensatory method in nature. The method with role parameters works a little worse than that with manual role assignment. However, from the results we can see that our method still achieves overall high recognition rates and is practicable.

Two-person interaction recognition

Five two-person interacting activities were selected as test examples (Zhang *et al.*, 2007; 2008):

Act. 5 Two persons walked on the same path in the same direction with a relatively constant distance between them.

Act. 6 Two persons walked on the same path in the opposite direction.

Act. 7 Two persons ran on the same path in the opposite direction.

Act. 8 Two persons walked in the opposite direction on the same path. They chatted with each other when they met, and then went on separately.

Act. 9 Two persons approached and met, one putting an object on the ground and going away, the other picking up this object and going away.

The activity dataset contains five activities, each having around 30 and a total of 146 video clips lasting 20~30 s. Fig.7 shows some key frames of Act. 9. Although these activities seem simpler than multi-person activities, they are similar to each other; i.e., it is not easy to differentiate them. Unlike above multi-person activities, these kinds of activities also differ a lot from the local characteristic. Besides the feature vector ($\boldsymbol{p}$, $\boldsymbol{v}$, η), we added (γ, α, $\boldsymbol{a}$) into the feature vector for efficiency in representing details, with $\gamma=w/h$ the aspect ratio and w and h the weight and height respectively, α the angle of inclination of the human body, and $\boldsymbol{a}=(a_x, a_y)$ the acceleration.

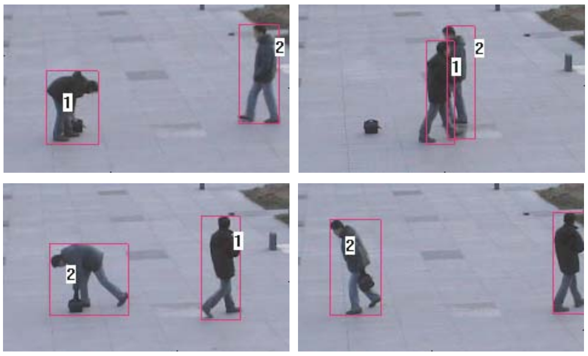

Fig.7 Key frames of Act. 9 (only the interested region of the original images is shown)

Comparisons between the proposed DHMMs and traditional HMMs, CHMMs, and HHMMs were conducted with this dataset. When using HMMs and HHMMs, we merged the feature vectors of two persons into a big one. Table 2 illustrates the recognition results using the proposed method and comparison results with three models. The results show that traditional HMMs can obtain a high recognition rate when the activity structure is relative simple, such as in Acts. 5 and 6. CHMMs can recognize most of these interactions. HHMMs are inferior to CHMMs when this interacting activities dataset is concerned. Since DHMMs are nearly equivalent to CHMMs in a special case, DHMMs obtain comparable results when Acts. 5~8 are concerned. But when Act. 9 is considered, our proposed method is superior to other methods because of the efficient representation of temporal and spatial structures. The overall high errors at Act. 7 were caused by the stronger noises in the observations.

Table 2 Comparison with other three models for interaction activity recognition

Act.	Recognition rate (%)			
	HMMs	CHMMs	HHMMs	DHMMs
5	100.0	100.0	100.0	100.0
6	84.6	100.0	92.3	100.0
7	84.2	84.2	84.2	84.2
8	93.3	100.0	93.3	100.0
9	63.6	81.8	81.8	90.9

Individual activity recognition

Nearly every signal produced by human behaviors can be beneficially decomposed into a group of interacting processes. We also compared our method with HMMs for recognizing individual activities. This dataset includes two activities:

Act. 10 One person walked in the scene.

Act. 11 One person walked in the scene, and picked up an object from the ground and went on.

The same features as in two-person interaction recognition were selected. The recognition rates and comparison are shown in Table 3. We decompose features into ($\boldsymbol{p}$, $\boldsymbol{v}$, γ) and (α, $\boldsymbol{a}$) for DHMMs(1) and decompose features into ($\boldsymbol{p}$, $\boldsymbol{v}$, $\boldsymbol{a}$) and (α, γ) for DHMMs(2) with duration state on the former chain.

Table 3 Comparison with HMMs for single-person activity recognition

Act.	Total activities	Recognized activities		
		HMMs	DHMMs(1)	DHMMs(2)
10	11	10	11	11
11	12	10	10	12

CONCLUSION

This paper presents a novel method to recognize human activities, especially multi-person interacting activities with complex structures. We analyzed the structural characteristics of activities from temporal and spatial points, and presented a decomposed HMM structure to recognize long-term complex interacting activities. DHMMs decreased the dimension of the feature and reduced the network complexity as well as the number of parameters greatly via spatial decomposition and relations uncoupling. DHMMs can not only model the interactions between persons but also represent the details of individual activities. This model worked well in multi-person activity recognition even when the person number is unknown or variable. According to the special structure, we introduced an approximation for model inference and learning. However, it should be pointed out that: (1) the semantic abstraction and partonomic segmentation of activities were mainly based on personal experience, which affects the results greatly; (2) the role assignment was done manually. These will be studied in our future work.

References

Brand, M., Oliver, N., Pentland, A., 1997. Coupled Hidden Markov Models for Complex Action Recognition. Proc. CVPR, p.994-999. [doi:10.1109/CVPR.1997.609450]

Bui, H.H., Venkatesh, S., West, G., 2002. Policy recognition in the abstract hidden Markov model. *J. Artif. Intell. Res.*, **17**:451-499.

Du, Y., Chen, F., Xu, W., 2007. Human interaction representation and recognition through motion decomposition. *IEEE Signal Processing Lett.*, **14**(12):952-955. [doi:10.1109/LSP.2007.908035]

Du, Y., Chen, F., Xu, W., Zhang, W., 2008. Activity recognition through multi-scale motion detail analysis. *Neurocomputing*, **71**:3561-3574. [doi:10.1016/j.neucom.2007.09.012]

Fine, S., Singer, Y., Tishby, N., 1998. The hierarchical hidden Markov model: analysis and applications. *Mach. Learning*, **32**(1):41-62. [doi:10.1023/A:1007469218079]

Forster, M., 2000. Key concepts in model selection performance and generalizability. *J. Math. Psychol.*, **44**:205-231. [doi:10.1006/jmps.1999.1284]

Ghahramani, Z., 2001. An introduction to hidden Markov models and Bayesian networks. *Int. J. Pattern Recogn. Artif. Intell.*, **15**(1):9-42. [doi:10.1142/S0218001401000836]

Gong, S., Xiang, T., 2003. Recognition of Group Activities Using Dynamic Probabilistic Networks. Proc. ICCV, p.742-749. [doi:10.1109/ICCV.2003.1238423]

Intille, S.S., Bobick, A.F., 2001. Recognizing planned, multi-person action. *Comput. Vis. Image Underst.*, **81**(3):414-445. [doi:10.1006/cviu.2000.0896]

Liu, X.H., Chua, C.S., 2006. Multi-agent activity recognition using observation decomposed hidden Markov models. *Image Vis. Comput.*, **24**:166-175. [doi:10.1016/j.imavis.2005.09.024]

Moeslund, T.B., Hilton, A., Krüger, V., 2006. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, **104**(2-3):90-127. [doi:10.1016/j.cviu.2006.08.002]

Murphy, K.P., 2002. Dynamic Bayesian Networks: Representation, Inference and Learning. PhD Thesis, University of California, Berkeley, USA.

Murphy, K.P., Paskin, M., 2001. Linear Time Inference in Hierarchical HMMs. Proc. NIPS, p.833-840.

Nguyen, N., Phung, D., Venkatesh, S., Bui, H.H., 2005. Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Model. Proc. CVPR, p.955-960. [doi:10.1109/CVPR.2005.203]

Oliver, N., Garg, A., Horvitz, E., 2004. Layered representations for learning and inferring office activity from multiple sensory channels. *Comput. Vis. Image Underst.*, **96**(2):163-180. [doi:10.1016/j.cviu.2004.02.004]

Schwarz, G., 1978. Estimating the dimension of a model. *Ann. Statist.*, **6**(2):461-464. [doi:10.1214/aos/1176344136]

Wada, T., Matsuyama, T., 2000. Multiobject behavior recognition by event driven selective attention method. *IEEE Trans. PAMI*, **22**(8):873-887. [doi:10.1109/34.868687]

Zacks, J.Z., Tversky, B., 2001. Event structure in perception conception. *Psychol. Bull.*, **127**(1):3-21. [doi:10.1037/0033-2909.127.1.3]

Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I., 2006. Modeling individual and group actions in meetings with layered HMMs. *IEEE Trans. Multim.*, **8**(3):509-520. [doi:10.1109/TMM.2006.870735]

Zhang, W., Chen, F., Xu, W., Zhang, E., 2006. Real-time Video Intelligent Surveillance System. Proc. ICME, p.1021-1024. [doi:10.1109/ICME.2006.262707]

Zhang, W., Chen, F., Xu, W., Cao, Z., 2007. Decomposition in Hidden Markov Models for Activity Recognition. Proc. MCAM, p.232-241. [doi:10.1007/978-3-540-73417-8_30]

Zhang, W., Chen, F., Xu, W., Du, Y., 2008. Hierarchical group process representation in multi-agent activity recognition. *Signal Processing: Image Commun.*, **23**(10):739-753. [doi:10.1016/j.image.2008.09.001]