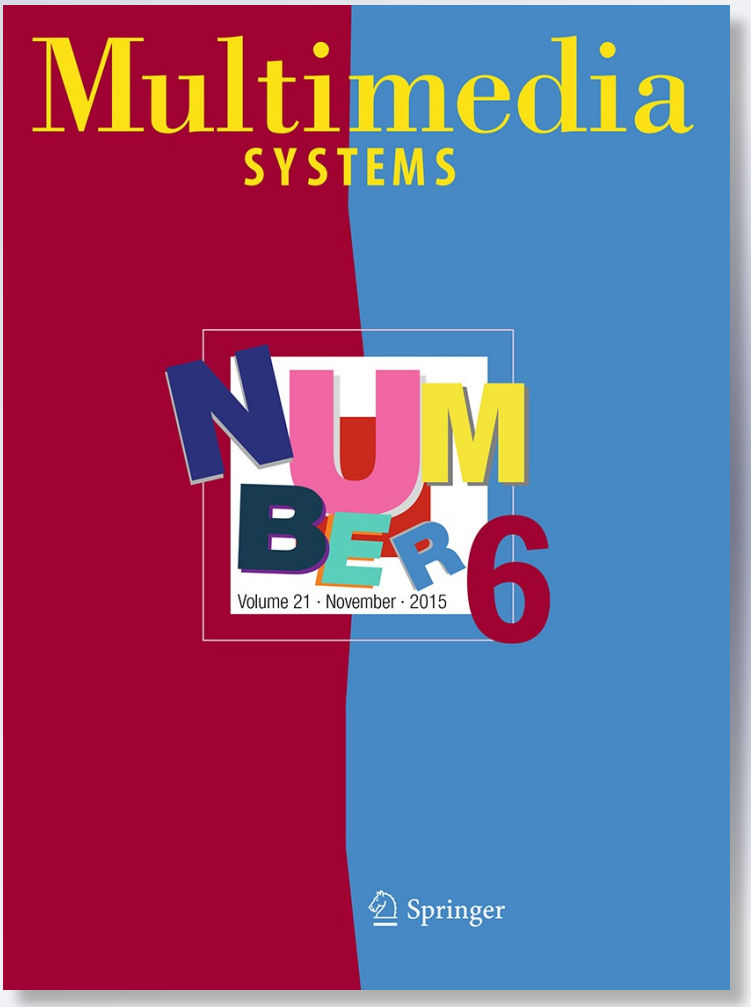
Human gesture recognition using a simplified dynamic Bayesian network

Myung-Cheol Roh & Seong-Whan Lee

Multimedia Systems

ISSN 0942-4962
Volume 21
Number 6

Multimedia Systems (2015) 21:557-568
DOI 10.1007/s00530-014-0414-9



Springer

Multimedia Systems (2015) 21:557–568
DOI 10.1007/s00530-014-0414-9

REGULAR PAPER

Human gesture recognition using a simplified dynamic Bayesian network

Myung-Cheol Roh · Seong-Whan Lee

Received: 11 June 2013 / Accepted: 26 August 2014 / Published online: 9 October 2014

Abstract In video-based human gesture recognition, it is very important to combine useful features and analyze the dynamic structure thereof as efficiently as possible. In this paper, we proposed a dynamic Bayesian network model that is a simplified model of dynamics at the level of hidden variables and employs observation windows of observation time slices for robust modeling and handling of noise and other variabilities. The proposed Simplified dynamic Bayesian network (DBN) was tested on a gesture database and an American sign language database. According to the experiments, the proposed DBN outperformed other methods: Conditional Random Fields (CRFs), conventional Bayesian Networks (BNs), DBNs, and Hidden Markov Models (HMMs). The proposed DBN achieved 98 % recognition accuracy in gesture recognition and 94.6 % in ASL recognition whereas the HMM and the CRF did 80 and 86 % in gesture recognition and 75.4 and 85.4 % in ASL (American Sign Language) recognition, respectively.

Keywords Dynamic Bayesian network · Gesture recognition · Sign language recognition

Communicated by Q. Tian.

M.-C. Roh
S-1 Corporation, Seoul, Korea
e-mail: mcroh@image.korea.ac.kr

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

1 Introduction

The development of computer and robot technologies has resulted in increased attention being given to their role in our daily lives. As computers and robots begin to impact our lives, the technologies of interaction between humans and computers/robots have been highlighted. One of the important features needed for such interactions is a convenient and natural interface. Gesture recognition is one of the intuitive methods for interaction [1]. It plays a very important role in human-computer interaction (HCI), human-robot interaction (HRI), smart environments supporting spontaneous services, automatic sign language interpretation systems, etc. Among the many techniques for gesture recognition, the vision-based method has been actively studied, because of its convenience and intuitiveness. Over the last ten years, the area of vision-based human gesture recognition has received increasing attention as a result of improved hardware technology and mature pattern recognition algorithms. It has found many applications in visual surveillance and human-robot interaction, etc [2–7].

There are several useful tools for human gesture recognition. The most popular one is the hidden Markov model (HMM) [8], which has been successfully applied to hard tasks such as speech, gesture and sign language recognition. Recently, the Conditional Random Field (CRF) has been shown to outperform the HMM [9] in sign language recognition. To build a robust model for gesture recognition, however, we have to consider further requirements: expressing experts' beliefs about dependencies between various features, adapting domain knowledge, etc. These play very important roles in gesture recognition, because a human gesture is a complex movement of various dependent/independent features. Also, an efficient sequence

modeling method is required. Most of the previous works do not provide efficient methods for these factors in combination.

The Bayesian network (BN) is one such powerful tool that can cope with the aforementioned requirements. The BN with its probabilistic characteristics has many advantages such as the following [10–12]:

- Facilitating the inclusion of domain knowledge
- Articulating experts' beliefs (prior knowledge) about the dependencies between different variables
- Providing a natural graphical tool for dealing with the problems that occur throughout applied mathematics and engineering, uncertainty and complexity
- Requiring a small number of parameters, and hence a smaller data set for learning
- Offering an efficient and principled approach for avoiding data over-fitting

- In fact, the BN also provides an efficient method of combining different features and it has been applied to a variety of problems [3, 5, 13–16]. Although the BN is generic to many application fields, it is not suitable for modeling sequential data.

The BN has been extended in the temporal dimension to a variety of dynamic Bayesian networks (DBNs) to deal with sequential signals in object tracking, gesture recognition, and human interaction recognition [3, 4, 7]. Although it is powerful and generic, a conventional DBN becomes structurally and computationally intensive as the number of variables grows. In terms of the DBN structure, it becomes more complicated as more features and characteristics are considered, and the structure depends on the application. One of the most important and hard problems with DBNs is determining the structure. There is no explicit method to find the proper structure. In terms of the computational cost, for the simplest example, we consider a DBN with a fully connected structure. Let's assume that there are n nodes in each time slice and each node can have k discrete states. Then, a node that is not an initial node at time zero requires specification of $(k-1) \times k^n$ conditional probabilities. This means the complexity increases exponentially according to the number of nodes, n. There are many DBN structures that reduce the complexity. However, in this paper, we are not claiming that the proposed DBN has less computational cost than any other. Instead, we are focusing on the DBN structure.

In this paper, we propose a DBN model that is a simplified model of dynamics at the level of hidden variables and employs observation windows of observation time slices for robust modeling and handling of noise and other variabilities. The proposed model achieved recognition rates of 98 % for isolated whole body gesture recognition and 94.6 % for American Sign Language recognition. According to the experiments, the proposed DBN outperformed these existing methods: Naive BNs, Naive DBNs, HMMs, and CRFs (Conditional Random Fields).

The rest of the paper consists of Related work (Sect. 2), the proposed dynamic Bayesian network (Sect. 3), Experimental result and analysis (Sect. 4), and Conclusion (Sect. 5).

2 Related work

There has been a surge of research interest in gesture recognition. Dynamic Programming (DP) is one of the well-known methods for computing the similarity between the input data and a template. DP has been successfully used for speech recognition and gesture recognition [17–19].

The HMM is most widely used method for handling sequential data. It is a statistical model that computes the probability of a trained model given an input data. It has been successfully applied to hand gesture, sign language, speech recognition, etc [8, 20, 21].

Recently, CRF has been shown to outperform the HMM in gesture recognition [9, 22]. The CRF is a discriminative probabilistic model for labeling sequential data.

As aforementioned, the BN and DBN are powerful probabilistic tools for dealing with complex data efficiently. The BN is also called the Directed Acyclic Graph (DAG), which represents random variables and their conditional independences. HMMs are considered the simplest type of DBN. There are many researches on using BN and DBN for dealing with various sequential data: human behavior understanding, object tracking, etc.

Robertson and Reid [5] proposed a probabilistic framework for analyzing human behavior in videos. They used a Naive BN to effect fusion of various data types and used HMMs to encode the time-dependent rules of the scene.

Moënne-Loccoz et al. [23] proposed a recurrent BN for recognizing human behavior from videos. The recurrent BN using temporal features has been applied to the analysis of violence scenes.

Wang et al. [7] proposed a DBN structure to track face based on visual multi-cue. They used the skin color, ellipse shape and an Ada-boost face detector as the multi-cue observations, which were combined by a DBN.

Du et al. [4] proposed another type of DBN for recognizing interaction activities. To overcome the defect of an exponential distribution of long duration activity, a duration node is added with a uniform distribution for modeling the global activity node. They considered five different interacting activities.

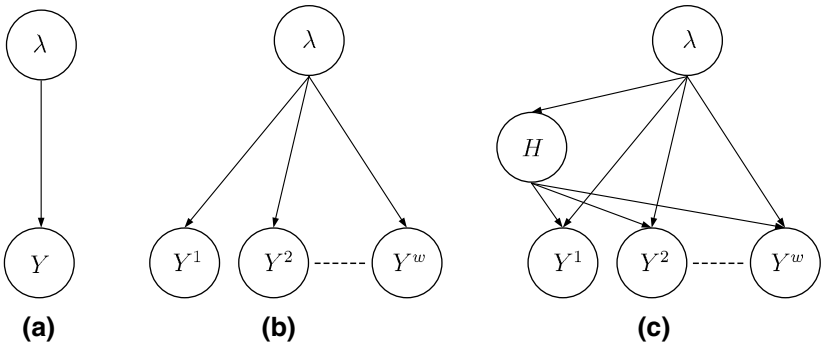


Fig. 1 Graphical representation of **a** a simple Naive BN, **b** a BN with time slice window, and **c** a BN with an additional hidden node (H). Y, and Y^i, $(1 \leq i \leq w)$ are sets of variables. λ represents the class node

Dielmann and Renals [3] proposed a multi-stream DBN with a counter structure for automatic segmentation of a meeting into a sequence of group meeting actions taken from a dictionary of events such as monolog, discussion, and presentation.

Suk et al. [6] proposed a DBN framework for hand gesture recognition. The framework recognized 10 isolated gestures and also performed on continuous gestures. The gestures they defined are quite simple ones.

The BNs and DBNs above are very powerful tools. However, the DBNs become structurally and computationally intensive as the number of variables grows, and there is no remarkable model for whole body gesture and sign language recognition. The above models have been designed for a specific purpose. Therefore, we propose a simplified DBN that can handle complex problem and does not require high computational cost.

3 The dynamic bayesian network modeling

Recognition of human gesture is highly complex problem, due to inter-personal differences and intra-personal dynamic variability as well as the ill-posed problem-characteristics. To understand a gesture, we have to exploit domain knowledge and any constraints among the observation features.

A gesture is described by a sequence of observation vectors

$$\mathbf{V} = \mathbf{v}_1 \mathbf{v}_2 \cdots \mathbf{v}_T, \quad \mathbf{v}_t \in \Re^n, t = 1, \ldots, T \tag{1}$$

where T is number of frames and $\mathbf{v}_t$ is a n-dimensional column vector of an observation, where $1 \leq t < T$.

The BN with its probabilistic characteristics has many advantages, as we mentioned above. In particular, we believe the BN is highly suited to the task of modeling scenes with inter-dependent parts. The use of BN for scene modeling is intuitive, since it can effectively describe a complex scene by factorizing complex features and capturing the dependencies among them.

Optimal classification under the MAP rule is to decide a class given an input feature vector according to

$$\hat{\lambda} = \arg\max_{\lambda} P(\lambda|\mathbf{V}) \tag{2}$$

where λ is the class variable and $\mathbf{V}$ is an observation vector. $\hat{\lambda}$ is a class that shows the highest joint probability of the target model over a set of other class models. Here, we consider a simple BN, as shown in Fig. 1a. It is a type of BN having a root variable and a dependent variable. The root corresponds to a class variable, the value of which is often unknown. $\mathbf{Y}$ is a dependent variable taking real values or multi-dimensional vectors. Robertson and Reid [5] used this model with 3-dimensional features, where the features are independent.

3.1 Bayesian network model with time slice window

As mentioned in the studies using CRFs and recurrent BNs, it is helpful to consider the number of frames at a given time rather than a single frame for reasons of noise tolerance [9, 23, 24]. Thus, we incorporated the concept of an observation window or time slice into the BN. Figure 1b shows a Naive BN with a time slice window (BNwTSW) of width w. There are w variables, each taking a vector within a window.

Given $\mathbf{V}$, the model is evaluated using the joint probability of conditionally independent time slices as follows:

$$P(\mathbf{Y} = \mathbf{V}, \lambda) = P(\lambda) \prod_{t=w}^{T} P(\mathbf{Y} = \mathbf{V}_{t-w+1:t}, \lambda) \tag{3}$$

where

$$\mathbf{V}_{a:b} = \mathbf{v}_a, \mathbf{v}_{a+1}, \ldots, \mathbf{v}_b, \quad a \leq b, \tag{4}$$

$$\mathbf{Y} = Y^1, Y^2, \ldots, Y^w, \quad Y^i \in \mathbf{R}^n. \tag{5}$$

The probabilistic aspect of a BN is given by the Conditional Probability Distribution (CPD) in each node. If the variables are discrete, it can be represented as a Conditional

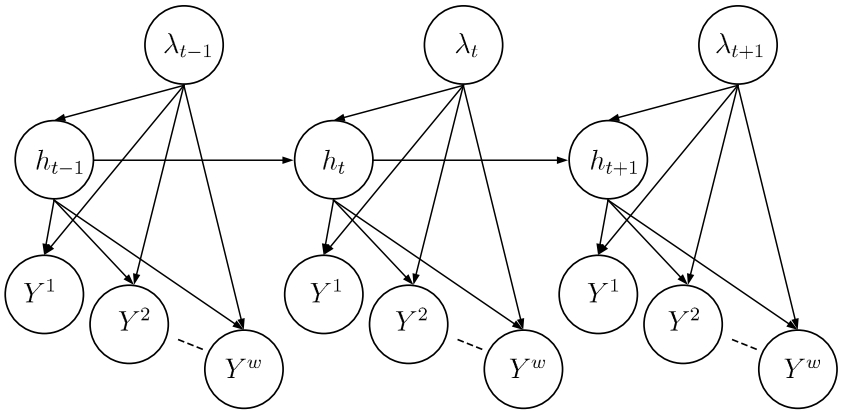


Fig. 2 Graphical representation of the proposed DBN with links among the hidden nodes

Probability Table (CPT). The parameter set Θ of the CPT can often be estimated using an EM algorithm maximizing the likelihood, as given by

$$\hat{\Theta}_{ML} = \arg\max_{\Theta} P(\mathbf{Y}|\Theta^{\lambda}, \lambda). \tag{6}$$

This model is good for a data sequence of a fixed length, w. But gesture signals are sequential and of variable length. If we model the entire sequence $\mathbf{V}$ using the model in Fig. 1b with a w-wide observation window, then the model is essentially a memoryless stationary multinomial process with no characterization of gesture dynamics. In this case, sequence modeling is simply the repeated application of a single model. Therefore, it is not a suitable model for gestures, since it cannot model the sub-patterns in a complete gesture motion.

3.2 The proposed dynamic bayesian network model

To model a gesture as an interesting sequence and increase the modeling power, we introduce a new hidden variable that conditions the observation variables in the window. This in effect defines a number of distinct states, each describing a sub-pattern in a complete gesture motion. Figure 1c shows the basic structure of the model corresponding to a slice of the proposed DBN. The hidden node, H , conditions all the children nodes, which are conditioned by the class node, λ. Let H be in $S = \{1, \ldots, N_h\}$, where N_h is the number of hidden states. With the introduction of H, the capacity of the model increases proportionally over the static model given in Fig. 1b.

As we have aforementioned, there is no explicit method to find a most optimal structure of DBNs. In this paper, we consider that the observations of human gestures are conditionally independent over time and each time slice is connected through the hidden node h. Therefore, the structure remains largely unchanged regardless of which features are used and how many there are. By reducing the connections in this method, we can reduce the complexity of the model. In this model, a node that is neither λs nor hs requires only $(k-1) \times k^2$ conditional probabilities if each node has k discrete states. A complete description of the proposed DBN model is shown in Fig. 2, which shows a repeated time slice of a BN over time. The nodes between time slices are connected via a single link between h_t and h_{t+1}. Note that the links between λs are ignored, since we consider that a single linkage can sufficiently describe the dependency. For simplicity, we assume a uniform prior for $P(\lambda)$.

3.3 Probability evaluation

Given an observation sequence $\mathbf{V}$, the joint probability of $\mathbf{V}$ and the model sequence Λ can be written as

$$P(\mathbf{Y} = \mathbf{V}, \Lambda) = P(\Lambda) \sum_{H} p(\mathbf{Y} = \mathbf{V}|H, \Lambda) P(H|\Lambda) \tag{7}$$

where H is any sequence of hidden states.

The probability evaluation is similar to that of the HMM. For a fixed state sequence $H = h_1 h_2 \ldots h_T$, and $\Lambda = \lambda_1 \lambda_2 \ldots \lambda_T$,

$$P(\mathbf{Y} = \mathbf{V}|H, \Lambda) = \prod_{t=w}^{T} P(\mathbf{Y} = \mathbf{V}_{t-w+1:t}|h_t, \lambda_t) \tag{8}$$

where the sequences of window frames are conditionally independent. The right terms of the equation can be computed by

$$P(\mathbf{Y} = \mathbf{V}_{t-w+1:t}|h_t = j, \lambda_t) = \prod_{s=1}^{t} B_j(\mathbf{V}_{t-w+s}, \lambda_t) \tag{9}$$

by exploiting an additional conditional independence assumption where

$$B_j(k, \lambda) = p(y_t = k|h_t = j, \lambda), \ \forall t \tag{10}$$

in state j.

All possible state sequences of H are required for the probability calculation, but it is computationally unfeasible

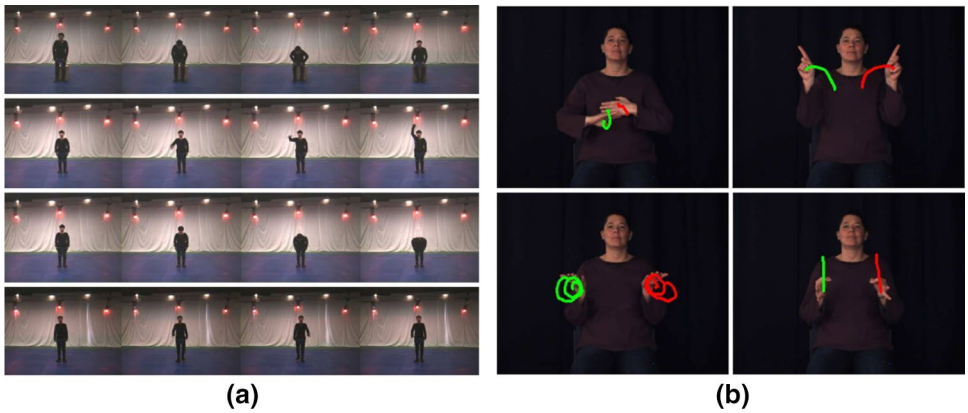


Fig. 3 Example frames of gesture videos in the KUGDB (*left*) and examples of ASL with trajectories of hands in the ASLDB (*right*). **a** *First row* sitting on a chair, *second row* raising a right hand, *third row* bowing, *fourth row* walking at a place]. **b** *Green* and *red* curves represent trajectories of the *left* and the *right* hands, respectively. (*Top left* born, *top right* different, *bottom left* here, *bottom right* decide)

due to its possible combinations are too many in general. However, it can be calculated efficiently by the forward and backward procedures as same as in HMMs. For more details of the forward and backward procedures, please refer to [8].

The probability of the hidden state sequence is

$$P(H|\Lambda) = \pi_{h_1} a_{h_1 h_2} a_{h_2 h_3} \cdots a_{h_{T-1} h_T} \tag{11}$$

where h_t is one of the states in the model λ_t, and π_i and a_{ij} are the initial state transition probability and the state transition probability

$$a_{ij} = P(h_{t+1} = j | h_t = i, \lambda), \ 1 \le i, \ \ j \le N. \tag{12}$$

The transition probabilities can be approximated by counting the frequency of state co-occurrences. We performed experiments on isolated sequence data. Thus, we can set $\lambda_1 = \lambda_2 = \cdots = \lambda_T$.

The final concern involves about the computation of Eq. (7). Due to the window overlap in computing the observation process, the model unduly emphasizes on the observation over the transition process. When $w \gg 1$, the effect of the transition nodes(H) for evaluation of the probability will be insignificant. Therefore, we modify the state transit for evaluation of the ion probability as follows:

$$\hat{a}_{ij} = a_{ij}^{\gamma}, \ \gamma \in \mathbf{R}. \tag{13}$$

In practice, it performs well in the range of 1.5–4 of the γ with the window size of 2–5.

4 Experimental results and analysis

Among the variety of gestures, we have considered five classes of whole body gestures and 48 words of sign language. A whole body gesture involves a whole body movement, while a sign language involves the motion of the two hands and the orientation of the head. The latter is considered to be harder due to the greater vocabulary.

Table 1 Gesture recognition accuracies (%) by HMM, CRF ($w = 7$), BNwTSW ($w = 5$), DBN, and SDBN ($w = 2$, $N_h = 2$) where w and N_h denote window size and the number of states in hidden node, respectively

HMM	CRF	BN	BNwTSW	DBN	SDBN
80	86	84	92	96	98

The SDBN presents the proposed simplified dynamic Bayesian network

We tested the proposed model on the Korea University Gesture Database (KUGDB)[1] and the American Sign Language Database (ASLDB)[2] of Boston University. Each database was divided into training sets and test sets. The test results were compared with those of the existing models: Naive BN, BNwTSW, and Naive DBN, HMMs, and CRF [22].

The quality of feature extraction will affect the performance. However, this paper does not emphasize feature extraction, but rather gesture modeling. One of the reasons we used the databases was to reduce any effect that may arise from the performance of the feature extraction method, enabling us to focus on only the modeling issue. The backgrounds of the videos in the databases are very simple.

We implemented HMM, BN, BNwTSW, DBN and the proposed DBN using the BNT (Bayes Net Toolbox for Matlab) [25]. The CRF was implemented using CRF Toolkit [26]. The platform we used was a MS Windows-based 2.33 GHz PC.

4.1 Gesture recognition

The gesture database KUGDB contains stereo videos of five gestures made by ten subjects. The five gestures include 'bending a waist', 'walking at a place', 'raising a

[1] Korea University Gesture Database, http://gesturedb.korea.ac.kr.

[2] American Sign Language Database, http://www.bu.edu/asllrp/ncslgr.html.

Table 2 ASL recognition accuracies (%) by HMM, CRF ($w = 5$), BN, BNwTSW($w = 3$), DBN, Two-layer CRF, and the SDBN ($w = 4, N_h = 5$) where w and N_h denote window size and the number of states in hidden node, respectively

HMM	CRF [22]	BN	BNwTSW	DBN	Two-layer CRF [22]	SDBN
75.4	85.4	80.2	88.3	73.2	93.5	94.6

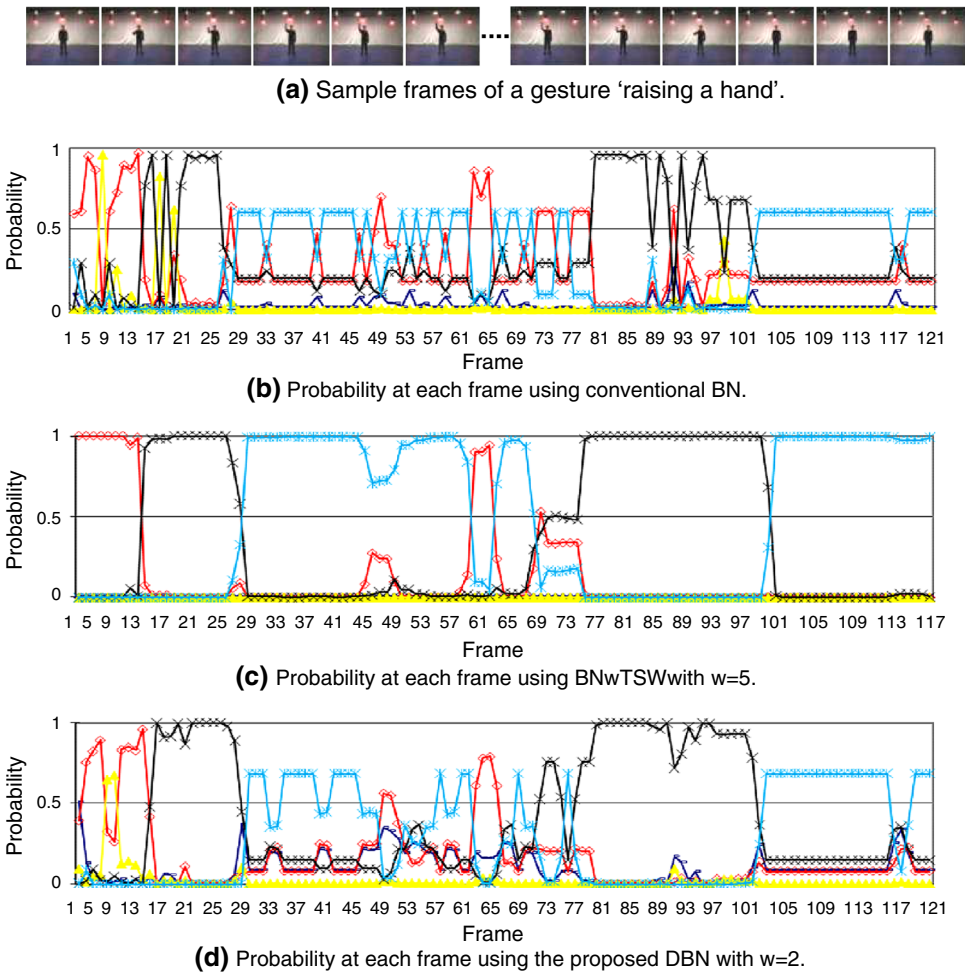


Fig. 4 The sequences of local probabilities over time given a single window of observation over time for the three models, BN, BNwTSW, and the proposed DBN on the experiment of 'raising a hand' gesture recognition. **a** The input video frames. **b–d** The y-axis represents the posterior probability of gesture class. The *black curve* with '×' markers represents the target model, 'raising a hand'. The *blue*, *red*, *yellow* and *cyan* curves represent the models of 'sitting on a chair,' 'standing up from a chair,' 'walking at a place,' and 'bending a waist,' respectively

hand', 'sitting on a chair', and 'standing up from a chair.' Figure 3a shows some sample videos in the KUGDB.

The gestures in the KUGDB were captured in a studio that the background was covered by white fabrics so that the foreground separation can be done easily. This environment facilitates our experiment, because in this paper we do not emphasize the feature extraction part, but rather the recognition model. An example in the KUGDB is shown in Fig. 3a.

We used five-dimensional features as observation vectors, at a given time. The nodes correspond to four mean optical flow vectors of sub-regions and a single variation value of the width–height ratio. The four mean optical flow vectors at time t are calculated using preceding frame at $(t - 1)$. The variation value of width–height ratio is computed by $width_t/height_t - width_{t-1}/height_{t-1}$. The observation vectors are modeled using a 5-Gaussian mixture model which is shared by all of the $Y^1, Y^2, \ldots, Y^w$ variables.

Table 1 shows the gesture recognition results using these six models; HMMs, CRF, Naive BN, BNwTSW, DBN and the proposed DBN. The size w of the window for each method is chosen where it performs best. The number of hidden states in the proposed DBN, N_h is also chosen where the proposed DBN shows the best performance.

In the KUGDB, there are five types of gestures, each with 10 samples. The experiment was carried out using the leave-one-out method. Among the 10 samples, nine were used for training and the remaining one was used for testing. According to Table 1, the proposed DBN performed

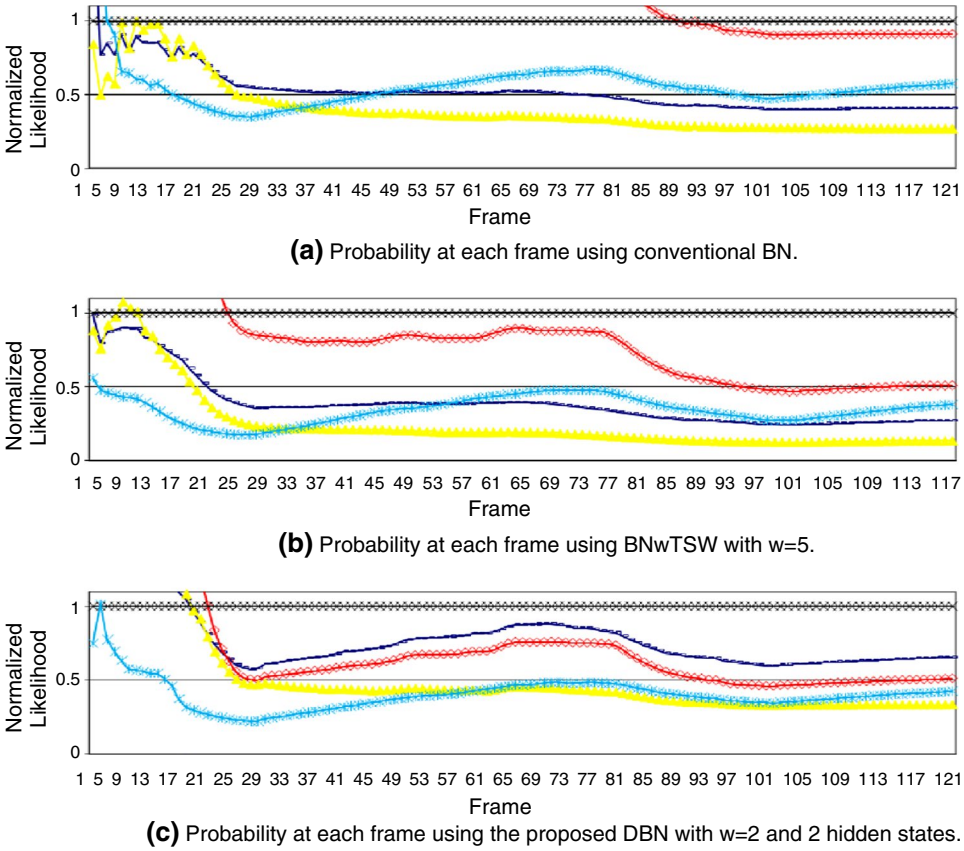


Fig. 5 The accumulated normalized likelihoods for all classes at each frame using conventional BN, BNwTSW, and the proposed DBN on the experiment of 'raising a hand' gesture recognition

well compared to the other methods. However, the performance margin over the conventional DBN is not significant, which can be attributed to the small size of class space. Therefore, we performed an experiment on an ASLDB which has a large class space, vocabularies.

4.2 Sign language recognition

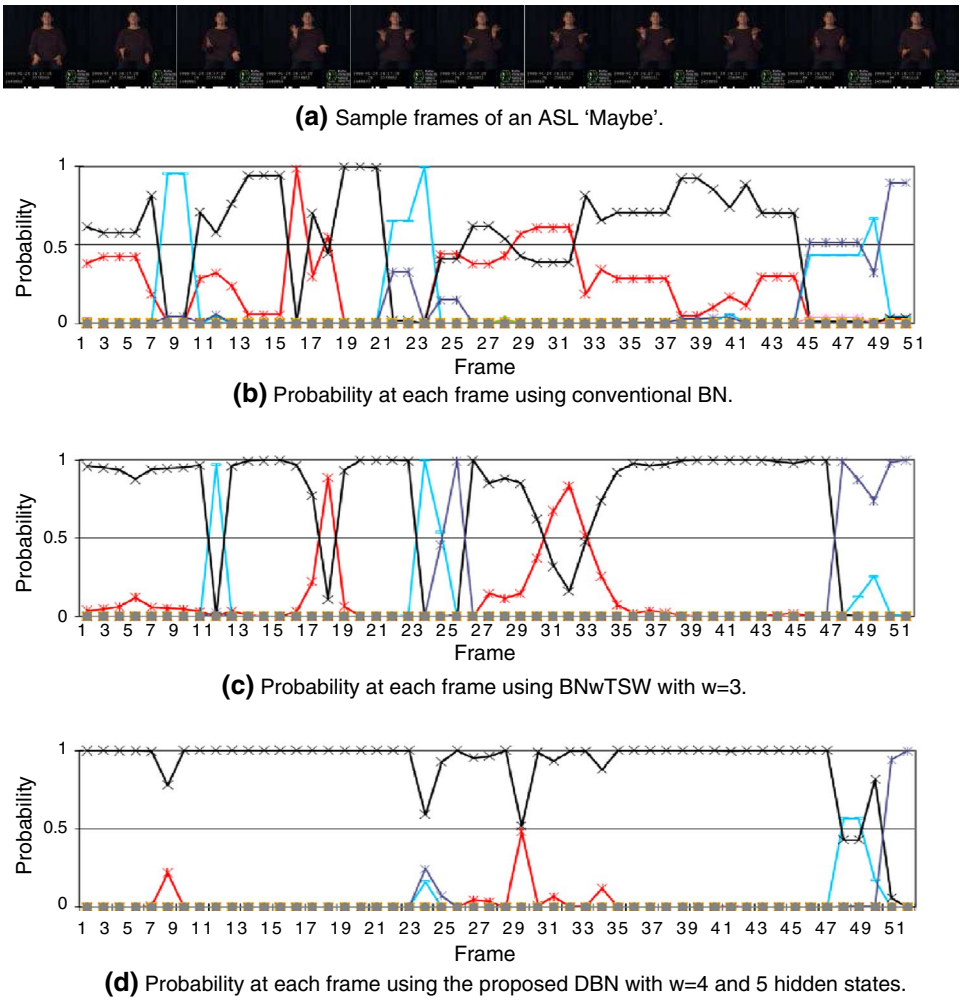
The second test concerns the recognition of isolated sign words. Figure 3b shows some sample videos of the ASL in the database. It contains a set of video data for 48 sign words. Sample words are as follows with their average frame lengths: And (16), Born (33), Arrive (20), Boy (26), Bicycle (23), Butter (28), Big (13), Day (26), Black (12), Car (63), Decide (16), Cold (28), Different (13), Here (67), Farm (24), Interpret (26), Inform (20), Funny (26), Finish (14), Library (52), Good (22), Magazine (40), Hot (20), Know (26), Many (22), Maybe (57), Lie (22), Name (30), Like (24), Night (28), Man (24), Rain (39), Now (12), Read (25), Out (15), Shoes (30), Past (13), Sorry (42), Sit (22), Take-off (25), Strange (20), What (42), Want (22), Work (30), Tell (14), Yesterday (26), Together (11), and Wow (47).

The database contains 10 video sequences per word for training. The data were captured using gloves colored green (left hand) and purple (right hand). Similarly, another 10 sequences per sign were collected with bare hands for testing. The start and end points of each data sequence were manually labeled as provided in the database.

Input to the six recognizers consists of the positions of the left and right hands in relation to the face, the vertical symmetry of the two hands, the state of occlusion between the two hands, and directional codewords of the motion vector of the two hands between two frames. The relative positions of the left and right hands were modeled by a mixture of 35 and 33 Gaussians, respectively. The directional codewords were found by quantizing the directions into eight codes [27] including additional one for 'no movement.' The occlusion state and the symmetry of the two hands have binary values, on and off. The face was detected by [28] and the hands were tracked by the appearance-based tracking method in [29]. The features we used in this experiment were just same as those in [22][3].

Table 2 shows the ASL recognition accuracy evaluated by these methods: HMM, CRF, Naive BN, BNwTSW, DBN, Two-layer CRF and the proposed DBN. The test set includes 480 samples, ten per sign word. The experimental results of the CRF and Two-layer CRF were provided in

[3] We would like to thank H.-D. Yang, the first author of [22] for providing the feature data and the results.



(a) Sample frames of an ASL 'Maybe'.

(b) Probability at each frame using conventional BN.

(c) Probability at each frame using BNwTSW with w=3.

(d) Probability at each frame using the proposed DBN with w=4 and 5 hidden states.

Fig. 6 The sequences of local state probabilities over time given a single window of observation over time for the three models, BN, BNwTSW, and the proposed DBN on the experiment of 'maybe' ASL word recognition. The *black curves* with 'x' markers represent the target model, 'Maybe'. The *red*, *cyan*, and *blue* represent the models of 'Yesterday,' 'Rain,' and 'Here,' respectively

[22] and for fair comparison we used same features in the experiments. The proposed model outperforms the other methods and the performance margin between the conventional and the proposed DBNs is significant. The recognition result of the BNwTSW shows that the method employing the window improved the performance compared with the BN without a window.

4.3 Analysis

Apart from the final performance figures, it is helpful to examine the internal workings of the models. However, since the proposed model involves latent random variables, it is not easy to probe the inner workings and to identify the state and collective dynamics. Here, we are attempting to visualize the probabilistic state of a model that changes over time. Let us consider the sample gesture 'raising a hand' shown in Fig. 4a. The remaining figures of Fig. 4b–d visualize the temporal evolution of the posterior probability of each state given a single window of observations for the three models. For each plot, the black curve knotted with ×s shows the local posterior probabilities of the target model, 'raising a hand' over time. The BN's curve in Fig. 4b shows high variability, presumably due to observation noise, while the BNwTSW shows much smoother curves as a result of the windowed frame. But there has been too much smoothing due to lack of position-independence or stateless smoothing. On the other hand, the result of the proposed DBN is more or less mixed, implying that the proposed DBN captures the local features using state variables, while smoothing out noise using the observation window. The local probability of the target model was maximal for the time frames between 16 and 26, and between 80 and 99, because there were up and down motions while there were still motions of raising hands in the remaining duration. The success of the recognition method approximately depends on the ratio of the duration that the model stayed on top. This can be seen in Fig. 5, which shows the development of the accumulated normalized likelihood for the partial observation sequence $\mathbf{V}_{1:t}$ when the same three models as above were compared.

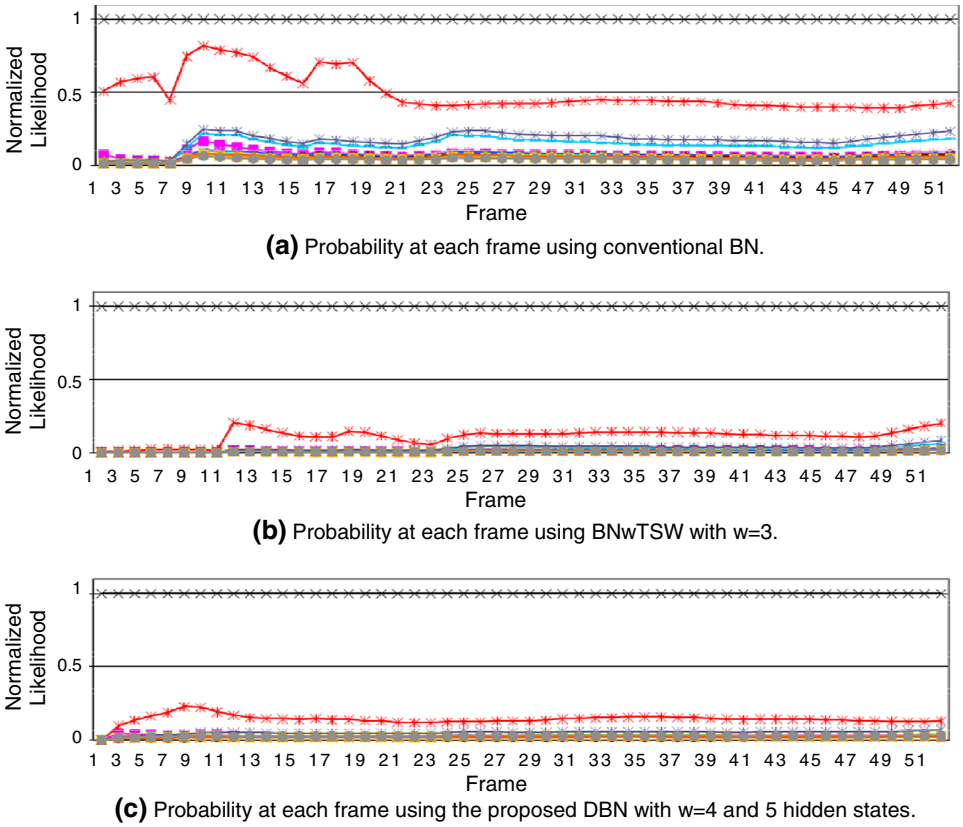


Fig. 7 The accumulated normalized likelihoods at each frame using conventional BN, BNwTSW, and the proposed DBN on the experiment of 'maybe' ASL word recognition

The accumulated normalized likelihood (ANL) for each class is computed by

$$ANL_t = \frac{\text{The likelihood given frames up to time } t}{t,\ t = 1, \ldots, T.} \tag{14}$$

To provide better figures for comparison, the score is scaled to find the relative likelihood to the true class likelihood, as given by

$$\text{Normalized Likelihood}_t = \frac{ANL_t}{ANL_t^{\text{true class}}}. \tag{15}$$

In the early duration, it appears that the likelihoods of some other classes are higher than that of the true class, thus some likelihoods are over 1. However, after a certain time has passed, the normalized likelihood of the true class takes highest one. From Fig. 5, the proposed DBN tends to reach the highest likelihood faster than the others on the time axis.

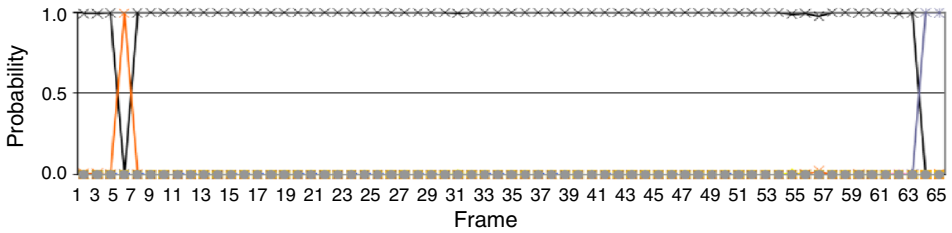
Figure 6 shows an example of ASL word evaluation. The first figure shows the sequence of video frames for the ASL word 'maybe.' Figure 6b–d shows the local frame probabilities using the these three models: BN, BNwTSW with $w = 4$, and the proposed DBN with $w = 5$. For each plot, the black curve knotted with ×s shows the local posterior probabilities of the ASL word 'maybe' over time. Note that, the proposed DBN for the target sign shows the highest values for almost all of the frames, implying a good match with the input sequence. When viewed in terms of the accumulated normalized likelihood in Fig. 7, the same result is obtained, with the target model beating the remaining models by big margins.

According to the experimental results in the Tables 1 and 2, the performance of the proposed DBN is higher than the others. However, we want to confirm that the proposed model works in an intuitively correct manner.

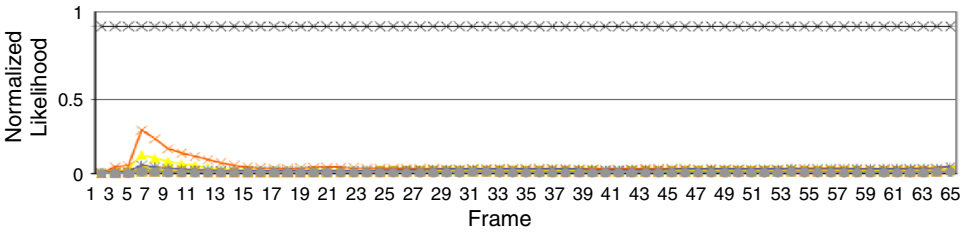
As a method of checking the process, we tried to decode the model given an input sequence, so that we could explain the generation of the sequence by the model. We employed the Viterbi algorithm to decode the best state sequence which is defined as the sequence of hidden states $h_1 h_2 \ldots h_T$ with the maximum joint likelihood.

One example is shown in Fig. 8a, which illustrates the hands' motion trajectories of hands of the ASL word 'here.' Here, the signer repeatedly makes circular patterns with both hands. By noting the repetitive circular patterns made by the hands, it is easy to predict that there will be a corresponding repetitive behavior within the model process. Figure 8b, c represents the probability evolution of model states for each frame using the proposed DBN. Figure 8d shows the evolution of hidden state h_t over time where the vertical axis represents the discrete state space. Although it is short, we can readily identify a repetitive pattern, $2 \rightarrow 1$

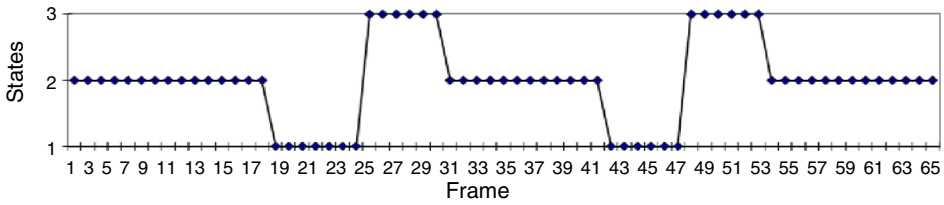
(a) A word 'HERE' with trajectories of both hands in the ASL database.



(b) Probability at each frame using the proposed DBN.



(c) Normalized Likelihood at each frame using the proposed DBN.



(d) The changes of the hidden node's state in the proposed DBN.

Fig. 8 **a** For the ASL word 'here', **b** probabilities for all classes at each frame using the proposed DBN, **c** normalized likelihood at each frame using the proposed DBN, and **d** the transitions of the hidden node's states in the proposed DBN are shown

$\rightarrow$ 3. Note that, this particular order of states is irrelevant since the order itself is determined randomly as given by the initial random condition.

To see modeling power of the time slice window in the BNwTSW and the proposed DBN, we performed an additional experiment using a smaller ASL data set of 24 signs. In the smaller set, the conventional and the proposed DBNs showed 93.75 and 95.8 % recognition accuracies, while we obtained 88.3 and 94.6 % for 48 signs, respectively. The performance margin of the proposed DBN over the BNwTSW is not statistically significant for a smaller set of sign words. This is because the time slice window played a decisive role in modeling ASL, although the hidden node was not applied in the model. However, by comparing the results on the small and large sets of signs, it is clearly shown that BNwTSW has a modeling limitation on a large vocabulary set and complex words because there is no

Table 3 ASL recognition accuracy (%) by the proposed DBN with various window size (w) and number of states in hidden node (N_h)

N_h	w				
	1	2	3	4	5
2	83.3	89.2	90.5	91.3	88.8
3	80.0	87.5	92.9	93.8	90.1
4	70.1	80.3	84.2	90.5	91.2
5	72.1	80.7	92.9	94.6	90.1
6	70.9	85.0	85.0	89.6	90.5

hidden node which can increase the modeling power over time.

Finally, there are two design parameters, namely, N_h, the number of states of h_t, and w, the size of the observation window. It is well known that the larger the value of N_h, the better the model can describe the sequence variability. On the other hand, it is debatable whether using the $w > 1$ does actually improve the performance of the model. Hence, we carried out an experiment to measure the effect of the two variables. The result is summarized in Table 3. The model reached the maximum performance when $N_h = 5$ and $w = 4$. The number of states is not easy to explain in simple terms. But the performance as a function of the window size shows the clear advantage of a moderate windowing size in the temporal dimension.

5 Conclusion

A systematic analysis of human gestures requires an efficient method for articulating expert beliefs about dependencies between various features, adapting domain knowledge. Each of these plays a very important role in gesture recognition, because human gestures are complex movements of various dependent/independent features. It is natural to model gestures using BN-based method which is a good tool to solve complex problem. Therefore, for modeling human gestures, we proposed a DBN model that is a simplified model of dynamics at the level of hidden variables and employs observation windows of observation time slices for robust modeling and handling of noise and other variabilities.

According to the experiments on gesture and ASL recognition, the proposed DBN outperformed the other methods. The proposed DBN achieved 98 % recognition accuracy in gesture recognition and 94.6 % in ASL recognition. The technique of the proposed DBN can be used to solve other complex pattern recognition problems including other modes of human gestures.

Acknowledgments This research was supported by the Implementation of Technologies for Identification, Behavior, and Location of Human based on Sensor Network Fusion Program through the Ministry of Trade, Industry and Energy (Grant No. 10041629) and the 2014 R&D Program for S/W Computing Industrial Core Technology through the Ministry of Science, ICT and Future Planning/Korea Evaluation Institute of Industrial Technology (Project No. 2014-044-023-001), Korea.

References

1. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **37**(3), 311–324 (2007). doi:10.1109/TSMCC.2007.893280
2. Bian, W., Tao, D., Rui, Y.: Cross-domain human action recognition. IEEE Trans. Syst. Man Cybern. Part B Appl. Rev. **42**(2), 298–307 (2012). doi:10.1109/TSMCB.2011.2166761
3. Dielmann, A., Renals, S.: Automatic meeting segmentation using dynamic bayesian networks. IEEE Trans. Multimed. **9**(1), 25–36 (2007)
4. Du, Y., Chen, F., Xu, W., Li, Y.: Recognizing interaction activities using dynamic bayesian network. In: Proceedings of the 17th International Conference on Pattern Recognition, vol. 1, pp. 618–621 (2006)
5. Robertson, N., Reid, I.: Behaviour understanding in video: a combined method. In: Proceedings of The Tenth IEEE International Conference on Computer Vision, vol. 1, pp. 808–815 (2005)
6. Suk, H.I., Shin, B.K., Lee, S.W.: Hand gesture recognition based on dynamic bayesian network framework. Pattern Recognit. **43**(9), 3059–3072 (2010)
7. Wang, T., Diao, Q., Zhang, Y., Song, G., Lai, C., Bradski, G.: A dynamic bayesian network approach to multi-cue based visual tracking. In: Proceedings of the 17th International Conference on Pattern Recognition, vol. 2, pp. 167–170 (2004)
8. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. In: Proceedings of IEEE, vol. 77, pp. 257–286 (1989)
9. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of International Conference on Machine Learning, pp. 282–289, USA (2001)
10. Fenton, N., Neil, M.: Making decisions: using bayesian nets and mcda. Knowl. Based Syst. **14**, 307–325 (2001)
11. Heckerman, D.: A tutorial on learning with Bayesian networks. Technical report msr-tr-95-06, Microsoft Research (1995)
12. Murphy, K.: Dynamic bayesian networks: Representation, inference and learning. Ph.D. thesis, University Of California, Berkeley (2002)
13. Bitmes, J., Bartels, C.: Graphical model architectures for speech recognition. IEEE Signal Process. Mag. **22**(5), 89–100 (2005)
14. Ji, Q., Lan, P., Looney, C.: A probabilistic framework for modeling and real-time monitoring human fatigue. IEEE Trans. Syst. Man Cybern. A **36**(35), 862–875 (2006)
15. Nikolopoulos, S., Papadopoulos, G., Kompatsiaris, I., Patras, I.: Evidence-driven image interpretation by combining implicit and explicit knowledge in a bayesian network. IEEE Trans. Syst. Man Cybern. Part B Appl. Rev. **41**(5), 1366–1381 (2011). doi:10.1109/TSMCB.2011.2147781
16. Park, S., Aggarwal, J.: A hierarchical bayesian network for event recognition of human actions and interactions. Multimed. Syst. **10**(2), 164–179 (2004)
17. Darrell, T., Pentland, A.: Space-time gestures. In: Computer Vision and Pattern Recognition. In: Proceedings of CVPR '93, 1993 IEEE Computer Society Conference on (1993)
18. Li, H., Greenspan, M.: Multi-scale gesture recognition from time-varying contours. Int. Conf. Comput. Vis. **1**, 226–234 (2005)
19. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoustics Speech Signal Proc. **26**(1), 43–49 (1978)
20. Ahmad, M., Lee, S.W.: Human action recognition using shape and CLG-motion flow from multi-view image sequences. Pattern Recognit. **41**(7), 2237–2252 (2008)
21. Starner, T., Weaver, J., Pentland, A.: Real-time american sign language recognition using desk and wearable computer based video. IEEE Trans. Pattern Anal. Mach. Intell. **20**(12), 1371–1375 (1998)
22. Yang, H.D., Sclaroff, S., Lee, S.W.: Sign language spotting with a threshold model based on conditional random fields. IEEE Trans. Pattern Anal. Mach. Intell. **31**(7), 1264–1277 (2009)

23. Moenne-Loccoz, N., Bremond, F., Thonnat, M.: Recurrent bayesian network for the recognition of human behaviors from video. In: Proceedings of 3rd International Conference on Computer Vision Systems, pp. 68–77 (2003)
24. Wang, S., Quattoni, A., Morency, L.P., Demirdjian, D., Darrell, T.: Hidden conditional random fields for gesture recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1521–1527 (2006)
25. Murphy, K.: Bayes net toolbox for Matlab (2014). http://code.google.com/p/bnt/Sept.(2014)
26. Kudo, T.: CRF++: Yet another CRF toolkit (2005). http://code.google.com/p/crfpp/Sept.(2014)
27. Lee, H.K., Kim, J.H.: An hmm-based threshold model approach for gesture recognition. IEEE Trans. Pattern Anal. Mach. Recognit. **21**(10), 961–973 (1999)
28. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE conference on Computer Vision and Patter Recognition, vol. 1, pp. 511–519 (2001)
29. Yang, H.D., Lee, S.W., Lee, S.W.: Multiple human detection and tracking based on weighted temporal texture features. Int. J. Pattern Recognit. Artif. Intell. **20**(3), 377–391 (2006)