Dynamic Bayesian Network for Unconstrained Face Recognition in Surveillance Camera Networks

Le An, Mehran Kafai, *Student Member, IEEE*, and Bir Bhanu, *Fellow, IEEE*

***Abstract*—The demand for robust face recognition in real-world surveillance cameras is increasing due to the needs of practical applications such as security and surveillance. Although face recognition has been studied extensively in the literature, achieving good performance in surveillance videos with unconstrained faces is inherently difficult. During the image acquisition process, the noncooperative subjects appear in arbitrary poses and resolutions in different lighting conditions, together with noise and blurriness of images. In addition, multiple cameras are usually distributed in a camera network and different cameras often capture a subject in different views. In this paper, we aim at tackling this unconstrained face recognition problem and utilizing multiple cameras to improve the recognition accuracy using a probabilistic approach. We propose a dynamic Bayesian network to incorporate the information from different cameras as well as the temporal clues from frames in a video sequence. The proposed method is tested on a public surveillance video dataset with a three-camera setup. We compare our method to different benchmark classifiers with various feature descriptors. The results demonstrate that by modeling the face in a dynamic manner the recognition performance in a multi-camera network is improved over the other classifiers with various feature descriptors and the recognition result is better than using any of the single camera.**

***Index Terms*—Camera networks, dynamic Bayesian network (DBN), face recognition, surveillance.**

I. Introduction

WITH THE broad establishment of surveillance video camera systems in recent years in both public and private venues, the recognition/verification of the subjects is often of interest and importance for purposes such as security monitoring, access control, etc. Some biometric traits such as gait can be used to recognize different subjects [1], however, it is preferred to use more distinct biometric clues such as face to identify a subject. Although face recognition has been studied extensively, face recognition in an unconstrained environment such as in surveillance camera videos remains very challenging and the recognition rate could drop dramatically to less than 10% using standard technologies [2]. The challenges to unconstrained face recognition in surveillance cameras are mainly due to the following reasons.

- Low resolution. In the video captured by surveillance cameras, the pixels that account for the faces are very limited. However, previous studies have shown that faces of size 64 × 64 are required for the existing algorithms to achieve good recognition accuracy [3].
- Arbitrary poses. Usually the subjects are moving freely. Consequently, it is not uncommon that the captured faces have different poses in different cameras.
- Varying lighting conditions. As the lighting is usually not uniform in the coverage area of the surveillance cameras, the illumination on the subject's face could vary significantly as he/she moves (e.g., the subjects walks into the shade from direct sunshine).
- Noise and blurriness. The captured images are often corrupted by noise during transmission and the motion of the subjects usually introduces blurriness.

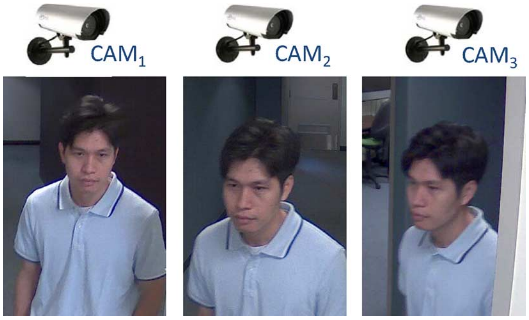

Fig. 1. Subject's face is captured by three cameras from different views in a typical surveillance camera system setup [4]. Pose and resolution of the captured faces vary across different views.

Fig. 1 shows an example of a subject's face captured by three surveillance cameras. The cameras are placed above a portal. The cameras have different viewing angles and none of the cameras captures the full frontal face of the subject. The face images exhibit variations in resolution, lighting condition, and poses. In addition, noise and blurriness are also observed. Under such circumstance, the standard face recognition algorithms such as Eigenfaces [17] would fail to work effectively. Despite the aforementioned difficulty, a multi-camera system provides different views of subjects which are complementary to each other. This enables the potential to improve the recognition performance with low quality input faces from multiple cameras.

Manuscript received December 03, 2012; accepted March 12, 2013. Date of publication April 23, 2013; date of current version June 07, 2013. This work was supported in part by the National Science Foundation (NSF) under Grant 0905671 and in part by the Office of Naval Research (ONR) under Grant N00014-07-C-0311 and Grant N00014-09-C-0388. This paper was recommended by Guest Editor F. Z. Qureshi.

L. An is with the Center for Research in Intelligent Systems and the Department of Electrical Engineering, University of California-Riverside, Riverside, CA 92521 USA (e-mail: lan004@ucr.edu).

M. Kafai is with the Center for Research in Intelligent Systems and the Department of Computer Science, University of California-Riverside, Riverside, CA 92521 USA (e-mail: mkafai@cs.ucr.edu).

B. Bhanu is with the Center for Research in Intelligent Systems, University of California-Riverside, Riverside, CA 92521 USA (e-mail: bhanu@cris.ucr.edu).

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/JETCAS.2013.2256752

In this paper, we propose a dynamic Bayesian network (DBN) based approach to tackle the problem of face recognition in multi-camera systems. Related work and our contributions are listed in Section II. Section III describes the details of the proposed method. In Section IV the experimental results are reported. We conclude this paper in Section V.

II. Related Work and Our Contributions

A. Related Work

To recognize a face in video, different approaches have been proposed. In general there are two principles: using 2D images from video sequences directly, or generating a 3D face model to cope with pose variation.

Within the 2D-based methods, normally the faces are first extracted from the video frames manually or using an automated face detector [18]. Subsequently, either all the face images or only the exemplar face images are used for the recognition task. An appearance manifold was built in [5] to represent each pose by an affine plane to cope with the pose variations in video sequences. In [8] a hidden Markov model (HMM) was used for video-based face recognition. In this model the temporal characteristics were analyzed over time. Stallkamp *et al.* [9] presented a real-time video based face identification system using a local appearance-based model and multiple frame weighting schemes. In [7] the face recognition in video was tackled by exploiting the spatial and temporal information based on Bayesian keyframe learning and nonparametric discriminant embedding. Recently, Biswas *et al.* [10] proposed a learning-based likelihood measurement to match high-resolution frontal view gallery images with probe images from surveillance videos. Wong *et al.* [4] proposed a patch-based image quality assessment method to select a subset of the "best" face images from the video sequences to improve the recognition performance. In [13], the video-based face recognition was converted to the problem of matching two image sets from different video sequences and it needs an independent reference set to align the images sets to be matched.

In an effort to recognize faces using more than one camera, some prior work has been done. Xie *et al.* [11] trained a reliability measure and it was used to select the most reliable camera for recognition. In [12] a cylinder head model was built to track and fuse face recognition results from different cameras. These approaches were tested on videos taken in controlled environment with higher resolution than typical surveillance video data. For application in surveillance cameras, a person re-identification method was proposed in [19] which depends on the robustness of the face tracker. A face recognition framework for mass transport security surveillance was proposed in [20].

In 3D-based approaches, the 3D face models are either computed or captured directly with a 3D scanner. Xu *et al.* [14] developed a framework using 3D face models for pose and illumination invariant face recognition from video sequences by integrating the effects of motion and lighting changes. In [15], the system used the images in the video as probe to compare with the 2D projection of the gallery 3D model. Liao *et al.* [16] used a single image for each individual in the gallery set to construct a 3D model to synthesize various face views. The 3D based methods are in general computationally expensive. Furthermore, a 3D model is difficult to be constructed from low-resolution videos, thus, the application of 3D models in surveillance cameras is limited. The above mentioned methods are summarized and compared in Table I. A recent survey of video based face recognition can be found in [21].

B. Motivation and Contributions of This Paper

Previously Bayesian network has been applied to face recognition. Heusch *et al.* [22] combined intensity and color information for face recognition in a Bayesian network where the observation nodes represented different parts of the face and the hidden nodes described the types of the observations. In [23] an embedded Bayesian network was proposed for efficient face recognition. Beyond the image-based recognition, there has been a growing interest to study the temporal dynamics in video sequences to improve the recognition performance in recent years [6], [7].

We propose a probabilistic approach for video-to-video face recognition using a DBN, utilizing different frames from multiple cameras. DBN has previously been applied to tasks such as speech recognition [24], vehicle classification [25], visual tracking [26], and facial expression recognition [27]. Variant of DBN such as topological DBN has also been proposed to identify human faces across age [28]. In this paper, the DBN is constructed by repeating a Bayesian network over a certain number of time slices with time-dependent variables. In each time slice the observed nodes are from different cameras. During the training, the temporal information is well encoded and the person-specific dynamics are learned. The identity of the testing subject can be inferred using previously trained network structure and parameters. By using DBN we are able to factor the joint probability distribution considering the temporal relationship of the feature evolution process between consecutive frames. Moreover, the DBN is defined and structured in a way that adding more cameras is easy. In addition, if features from one camera were not extracted due to image capture failure, this information can still be inferred by DBN and, therefore, recognition may not fail.

Compared to the previous work [29] in which the DBN structure is manually defined and only two cameras are used for recognition, in this paper, the topological structure in each time slice of DBN is learned automatically in an optimal manner with three cameras involved. In addition, the experimental results are examined thoroughly using multiple performance evaluation criteria using much more data with improved evaluation protocol.

In summary, the contributions of this paper are as follows.

- We propose a probabilistic framework for unconstrained face recognition in a multi-camera surveillance scenario. To the authors' best knowledge, this is the first work using DBN for video-based face recognition in surveillance cameras with more than two cameras. The framework is flexible and can be easily extended to more complicated multi-camera settings. Besides, any feature descriptor is compatible in this framework.
- We test the proposed method on a publicly available multi-camera surveillance video dataset "ChokePoint"

TABLE I
RELATED WORK SUMMARY FOR VIDEO BASED FACE RECOGNITION

Methodology	2D/3D	Multi-Camera?	Advantages and Limitations
Probabilistic-based appearance manifolds [5]	2D	No	Pros: different poses are approximated by an affine plane in the manifold Cons: controlled dataset, images need careful alignment
Manifold and Bayesian inference model [6]	2D	No	Pros: manifold learning and Bayesian inference are combined Cons: view-specific manifold needs to be learned
Spatio-Temporal embedding [7]	2D	No	Pros: the intrinsic temporal structures are preserved Cons: controlled dataset, small pose variation
Hidden Markov Model [8]	2D	No	Pros: effectively model the person specific dynamics in video sequences Cons: controlled dataset, small pose variation
Local appearance based model [9]	2D	No	Pros: real-time system, different weighting schemes in different classification models Cons: the combination of the weighting scheme does not necessarily improve results
Learning-based likelihood measurement [10]	2D	No	Pros: robust high-resolution gallery images and low-resolution video sequences matching Cons: fiducial points are difficult to detect on low-resolution images
Patch-based image selection [4]	2D	Yes (three)	Pros: best images are selected from multiple video sequences to improve recognition Cons: the temporal information in video sequences is not utilized
Camera reliability measurement [11]	2D	Yes (two)	Pros: recognition performance improved over single camera Cons: controlled dataset, high-resolution images, rely on component based face detector
Cylinder head model from cameras [12]	2D	Yes (two)	Pros: transform the face recognition from video to still face recognition Cons: cylinder model is difficult to build with low-resolution uncontrolled videos
Image sets alignment and matching [13]	2D	No	Pros: align two images sets using a reference set to improve the matching performance Cons: additional reference data needed, temporal information from video not utilized
Motion and lighting integration [14]	3D	No	Pros: robust to large changes in facial pose and lighting conditions Cons: controlled dataset, high-resolution images
3D model assisted recognition [15]	3D	No	Pros: better than basic 2D image-based recognition Cons: small pose variation, controlled dataset
3D model from single image [16]	3D	No	Pros: only a single image is required to generate a 3D face model Cons: the high-resolution frontal view face image is required
This paper: Dynamic Bayesian network (DBN)	2D	Yes (three)	Pros: encode the video dynamics, multi-camera based, uncontrolled surveillance videos Cons: computation for training may become expensive with complicated DBN structure

TABLE II
DEFINITION OF THE SYMBOLS USED IN THIS PAPER

Symbol	Definition
K	Number of cameras in the multi-camera setup
k	camera index
T	total number of time slices in the DBN (sequence length)
t	time slice index
CAM_k^t	the random variable representing the feature vector of a face image from the k^{th} camera in time slice t
N	Number of subjects in the gallery
S	the random variable representing the probability distribution over the gallery of subjects

with unconstrained face acquisition [4], in contrast to the other commonly used datasets which were collected in controlled environment. We compare the proposed method with popular benchmark classifiers using different feature descriptors. The superior performance of the proposed DBN approach is verified in different aspects with various evaluation criteria.

- We compare the face recognition performance using all of the three cameras in the ChokePoint dataset against using single camera. Experimental results demonstrate that using multiple cameras improves the recognition performance over any single camera.

Before the detailed algorithms is presented, Table II gives a summary of the symbols used in the following sections for a better understanding.

III. TECHNICAL DETAILS

In the following subsections, we first explain the Bayesian network structure for face recognition from multiple cameras with a single time slice and then the DBN structure with multiple time slices is presented.

A. Bayesian Network

A Bayesian network (BN) is a graphical model, which is defined using a directed acyclic graph. The nodes in the model represent the random variables and the edges define the dependencies between the random variables. Given the value of its parents, each variable is conditionally independent of its nondescendants. A BN can effectively represent and factor the joint probability distributions and it is suitable for the classification tasks. Mathematically, given a set of ordered random variables $X_1, X_2, \ldots, X_n$, the full joint distribution is given by

$$\begin{aligned} p(x_1, x_2, \ldots, x_n) &= p(x_1) \times p(x_2|x_1) \times \cdots \\ &\quad \times p(x_n \,|\, x_1, x_2, \ldots, x_{n-1}) \\ &= \prod_{i=1}^{n} p(x_i \,|\, x_1, \ldots, x_{i-1}). \end{aligned} \tag{1}$$

In the scope of multi-camera face recognition, when several face images of the same subject are captured by different cameras, we construct the corresponding BN using two different kinds of nodes.

- Root node: This is a discrete node on the top of the BN. The node is represented by a random variable S. S is the probability distribution over all the subjects in the gallery

and does not represent the identity of a single subject. The size of the root node indicates the number of the subjects (classes).
- Camera node: This continuous node contains the feature descriptors of the extracted face image from one camera. The number of the camera nodes depends on the number of cameras involved in the surveillance. Different feature descriptors such as local binary patterns (LBP) [30] or local phase quantization (LPQ) [31] can be adopted. The notation CAM is used to represent this random variable.

When a test sequence is provided, the subject's identity s is determined using the *maximum a posterior* (MAP) rule

$$\begin{aligned} s &= \underset{S}{\operatorname{argmax}}\, p(S \mid \mathrm{CAM}_1, \dots, \mathrm{CAM}_K) \\ &= \underset{S}{\operatorname{argmax}}\, \frac{p(\mathrm{CAM}_1, \dots, \mathrm{CAM}_K \mid S) p(S)}{\sum_S p(\mathrm{CAM}_1, \dots, \mathrm{CAM}_K \mid S) p(S)} \end{aligned} \tag{2}$$

where CAM_k is the random variable representing the feature vector from the face image in camera k. $p(S)$ is the prior probability of the presence of each subject and is usually modeled by a uniform distribution. Since the different cameras are capturing the same subject, the camera nodes are not independent. We explain how the BN structure is learned in the next part.

B. Structure Learning

The structure of the BN would greatly impact the accuracy of the model. However, the number of possible structures is super-exponential in the total number of nodes. Therefore, it is desirable to avoid performing exhaustive search for structure learning. In this paper, we use the K2 structure learning algorithm [32] to determine the BN's structure. K2 uses a greedy approach to incrementally add parents to a node according to a chosen scoring function. The search space of K2 algorithm is much smaller than the entire space due to the ordering of the nodes and it guarantees no cycle in the generated structure. We use the completed likelihood Akaike information criterion (CL-AIC) scoring function for this purpose [33]. Fig. 2 shows the K2 learned BN structure. In this case, the subject's identity s is determined by (3) shown at the bottom of the page.

C. Dynamic Bayesian Network for Face Recognition

Compared to the traditional face recognition methods which are typically image based, the video based face recognition is advantageous since the dynamics in different frames for the specific person can be learned to help the recognition of the subject. As suggested in [34], multiple face samples from a video sequence have the potential to boost the performance of the recognition system.

We propose our graphical model as a DBN. DBN differs from HMM in the following aspects: a DBN represents the problem utilizing a set of random variables whereas an HMM uses a single discrete random variable; in a standard first-order HMM modeled as a DBN, the random variables at time slice t depend only on the variables in time slices t and $t-1$ for all $t > 1$; in an HMM all the hidden random variables are combined in a single multi-dimensional node, whereas in a DBN multiple hidden nodes can be present.

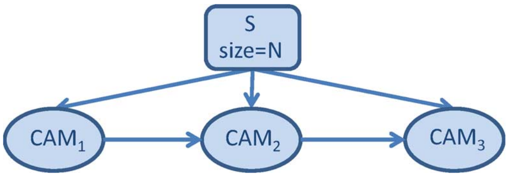

Fig. 2. K2 learned Bayesian network structure [32]. Training data is from the ChokePoint dataset [4].

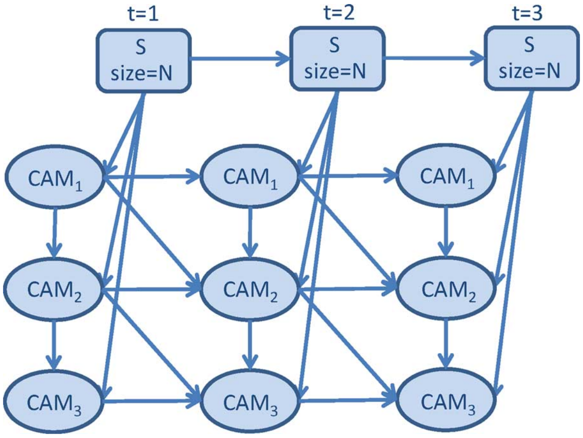

Fig. 3. DBN structure for three time slices with a three-camera setup.

In terms of complexity, an HMM would require $O(T(N^K)^2)$ for inference, $O(N^{2K})$ parameters to specify $P(S^t \mid S^{t-1})$, and $O(TN^K)$ space, where T is the sequence length, N is the number of classes, and K is the number of camera observations. For a DBN, $O(TKN^{K+1})$ is required for inference, and $O(KN^2)$ parameters to specify $P(S^t \mid S^{t-1})$. The DBN has exponentially less parameters and inference is much faster.

Operating a graphic model requires three main steps: defining the structure, learning the parameters, and inference. The structure of the DBN consists of the inter-slice topology and the intra-slice topology. The inter-slice topology is defined as follows. Each time slice $t = 1 \dots T$ has $K+1$ nodes; one root node S, and K camera nodes $\mathrm{CAM}_{k=1\dots K}$. This structure is the same as shown in Fig. 2 for the three-camera setting $(K = 3)$. The intra-slice topology is illustrated in Fig. 3 with three time slices.

After defining the structure, it is required to learn the parameters of the DBN before recognition is performed. Therefore, the probability distribution for each node given its parents should be

$$\begin{aligned} s &= \underset{S}{\operatorname{argmax}}\, p(S \mid \mathrm{CAM}_1, \mathrm{CAM}_2, \mathrm{CAM}_3) = \underset{S}{\operatorname{argmax}}\, \frac{p(\mathrm{CAM}_1, \mathrm{CAM}_2, \mathrm{CAM}_3 \mid S) p(S)}{\sum_S p(\mathrm{CAM}_1, \mathrm{CAM}_2, \mathrm{CAM}_3 \mid S) p(S)} \\ &= \underset{S}{\operatorname{argmax}}\, \frac{p(\mathrm{CAM}_1 \mid S) p(\mathrm{CAM}_2 \mid \mathrm{CAM}_1, S) p(\mathrm{CAM}_3 \mid \mathrm{CAM}_2, S) p(S)}{\sum_S p(\mathrm{CAM}_1, \mathrm{CAM}_2, \mathrm{CAM}_3 \mid S) p(S)}. \end{aligned} \tag{3}$$

determined. In the three-camera setting, for the first time slice this includes

$$p\left(\mathrm{CAM}_1^1\middle|\, S^1\right),\; p\left(\mathrm{CAM}_2^1\middle|\, S^1, CAM_1^1\right),\; p\left(\mathrm{CAM}_3^1\middle|\, S^1, CAM_2^1\right),\; p(S^1). \quad (4)$$

For time slices $t = 2 \ldots T$ it includes

$$\begin{aligned}&p\left(\mathrm{CAM}_1^t\middle|\, S^t, \mathrm{CAM}_1^{t-1}\right)\\ &p\left(\mathrm{CAM}_2^t\middle|\, S^t, \mathrm{CAM}_1^{t-1}, \mathrm{CAM}_2^{t-1}, \mathrm{CAM}_1^t\right)\\ &p\left(\mathrm{CAM}_3^t\middle|\, S^t, \mathrm{CAM}_2^{t-1}, \mathrm{CAM}_3^{t-1}, \mathrm{CAM}_2^t\right),\; p(S^t \mid S^{t-1}).\end{aligned} \quad (5)$$

With new unseen data (evidence), an inference algorithm is applied to compute the marginal probability from the evidence. Specifically, inference determines the subject's identity by $p(S^T \mid \mathrm{CAM}_{k=1,2,3}^{(1:T)})$, where $\mathrm{CAM}_{k=1,2,3}^{(1:T)}$ refers to features from all of the three cameras for time slices 1 to T. In other words, a probability distribution over the set of all the subjects is defined. The goal is to find the marginal probability of each hidden variable. Equation (6) shows how $p(S^t \mid \mathrm{CAM}_{k=1,2,3}^{(1:T)})$ is computed for any $t = 2 \ldots T$

$$\begin{aligned}
&p(S^t \mid \mathrm{CAM}_{k=1,2,3}^{1:T})\\
&\quad = p(S^t, \mathrm{CAM}_1^1 = \mathrm{cam}_1^1, \ldots, \mathrm{CAM}_3^T = \mathrm{cam}_3^T)\\
&\quad\times \underbrace{1/p(\mathrm{CAM}_1^1 = \mathrm{cam}_1^1, \ldots, \mathrm{CAM}_3^T = \mathrm{cam}_3^T)}_{=L(\text{a constant})}\\
&\text{by marginalization}\\
&\quad = \sum_{S^1,\ldots,S^{t-1},S^{t+1},\ldots,S^T} p\left(S^1, \ldots, S^T, \mathrm{CAM}_1^1\right.\\
&\quad \left. = \mathrm{cam}_1^1, \ldots, \mathrm{CAM}_3^T = \mathrm{cam}_3^T\right) \times L\\
&\text{by Bayes net factoring}\\
&\quad = \sum_{S^1,\ldots,S^{t-1},S^{t+1},\ldots,S^T} p(S^1) p\left(\mathrm{CAM}_1^1\middle|\, S^1\right)\\
&\quad\times p\left(\mathrm{CAM}_2^1\middle|\, S^1, \mathrm{CAM}_1^1\right) p\left(\mathrm{CAM}_3^1\middle|\, S^1, \mathrm{CAM}_2^1\right)\\
&\quad\times \prod_{i=2:T} p(S^i \mid S^{i-1}) \prod_{i=2:T} p\left(\mathrm{CAM}_1^i\middle|\, S^i, \mathrm{CAM}_1^{i-1}\right)\\
&\quad\times p\left(\mathrm{CAM}_2^i\middle|\, S^i, \mathrm{CAM}_1^{i-1}, \mathrm{CAM}_2^{i-1}, \mathrm{CAM}_1^i\right)\\
&\quad\times p\left(\mathrm{CAM}_3^i\middle|\, S^i, \mathrm{CAM}_2^{i-1}, \mathrm{CAM}_3^{i-1}, \mathrm{CAM}_2^i\right) \times L\\
&\text{by splitting products}\\
&\quad = \sum_{S^1,\ldots,S^{t-1},S^{t+1},\ldots,S^T} p(S^1) p\left(\mathrm{CAM}_1^1\middle|\, S^1\right)\\
&\quad\times p\left(\mathrm{CAM}_2^1\middle|\, S^1, \mathrm{CAM}_1^1\right) p\left(\mathrm{CAM}_3^1\middle|\, S^1, \mathrm{CAM}_2^1\right)\\
&\quad\times \prod_{i=2:t} p(S^i \mid S^{i-1}) \prod_{i=2:t} p\left(\mathrm{CAM}_1^i\middle|\, S^i, \mathrm{CAM}_1^{i-1}\right)\\
&\quad\times p\left(\mathrm{CAM}_2^i\middle|\, S^i, \mathrm{CAM}_1^{i-1}, \mathrm{CAM}_2^{i-1}, \mathrm{CAM}_1^i\right)\\
&\quad\times p\left(\mathrm{CAM}_3^i\middle|\, S^i, \mathrm{CAM}_2^{i-1}, \mathrm{CAM}_3^{i-1}, \mathrm{CAM}_2^i\right)\\
&\quad\times \prod_{i=t+1:T} p(S^i \mid S^{i-1}) \prod_{i=t+1:T} p\left(\mathrm{CAM}_1^i\middle|\, S^i, \mathrm{CAM}_1^{i-1}\right)\\
&\quad\times p\left(\mathrm{CAM}_2^i\middle|\, S^i, \mathrm{CAM}_1^{i-1}, \mathrm{CAM}_2^{i-1}, \mathrm{CAM}_1^i\right)\\
&\quad\times p\left(\mathrm{CAM}_3^i\middle|\, S^i, \mathrm{CAM}_2^{i-1}, \mathrm{CAM}_3^{i-1}, \mathrm{CAM}_2^i\right) \times L.
\end{aligned} \quad (6)$$

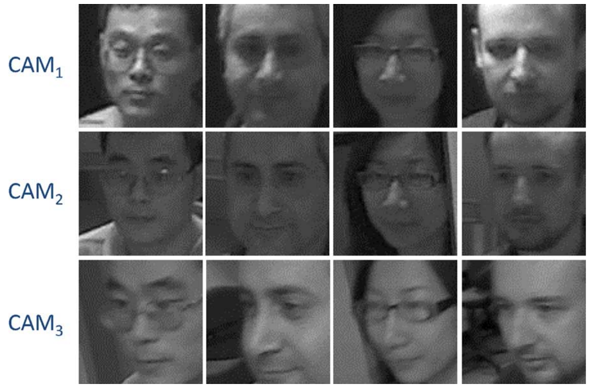

Fig. 4. Sample images from the ChokePoint [4] dataset.

IV. Experiments

A. Experimental Settings

1) Dataset: We use the ChokePoint dataset [4] which is designed for evaluating face recognition algorithms under real-world surveillance conditions. This dataset is challenging for face recognition task as the captured faces are unconstrained in terms of pose, lighting, and image quality. Although many face datasets exist, to the authors' best knowledge, the ChokePoint dataset is the only available open surveillance video dataset with multiple cameras. Fig. 4 shows some sample images from this dataset. The setting for the network involves three cameras mounted above two portals (P1 and P2) that captured the video sequences of the moving subjects while the subjects were either entering (E) or leaving (L) the portals in a natural manner. In total four data subsets are available (P1E, P1L, P2E, and P2L). In each subset, four sequences are provided (S1, S2, S3, and S4) and each sequence contains the recorded videos from three cameras (C1, C2, and C3). In P1 25 subjects were involved and in P2 there were 29 participants. The resolution of the captured frames are 800 × 600 at a frame rate of 30 fps and the cropped faces with size 96 × 96 from the original video frames are provided.

2) DBN Structure: The DBN is constructed with five time slices. The size of the DBN is determined empirically to offset the complexity of the network and to ensure sufficient dynamics to be encoded as a temporal clue. In each time slice we use the learned structure in Fig. 2. For parameter learning, the EM algorithm is used and the junction tree algorithm is chosen for inference. With nonoptimized Matlab implementation, the training takes about 42 s on a PC with 3 GHz CPU and 8 GB RAM. For testing, the inference takes about 60 s.

In our experiments, we use faces from all of the four subsets (P1E, P1L, P2E, and P2L). S1, S2, S3, and S4 are all used except for P2E in which we use only S3 and S4 due to incomplete data in P2E_S1 and P2E_S2. In each sequence 40 instances are used for training or testing. Each instance consists of 15 face images (three cameras in each time slice, five time slices in total). In each run, we perform cross-validation on the same subset (i.e., train on sequence P1E_S1 and test on P1E_S2, P1E_S3, and P1E_S4). The averaged results are reported for each subset separately.

3) Feature Descriptors: For face recognition, various feature descriptors have been proposed and applied. Local binary

TABLE III
RANK-1 RECOGNITION RATES ON DIFFERENT TESTING SEQUENCES (IN %)

Data→	P1E				P1L				P2E				P2L			
Method→	NN	LDA	SVM	DBN	NN	LDA	SVM	DBN	NN	LDA	SVM	DBN	NN	LDA	SVM	DBN
LBP	43.7	40	78.2	**89.7**	40.9	39	55.2	**85.8**	55.8	70.6	84.5	**95.3**	88.1	85.2	81.3	**94.1**
LPQ	40.1	38.5	55.36	**91.7**	54.5	42.3	60	**86.3**	64.2	64.1	83.2	**89**	88.2	89.9	85.5	**97.2**
HOG	63	45.5	70.3	**84.3**	50.9	36.8	71.5	**89**	49.7	33.7	65.8	**75.5**	43.2	19.1	43.2	**73.6**

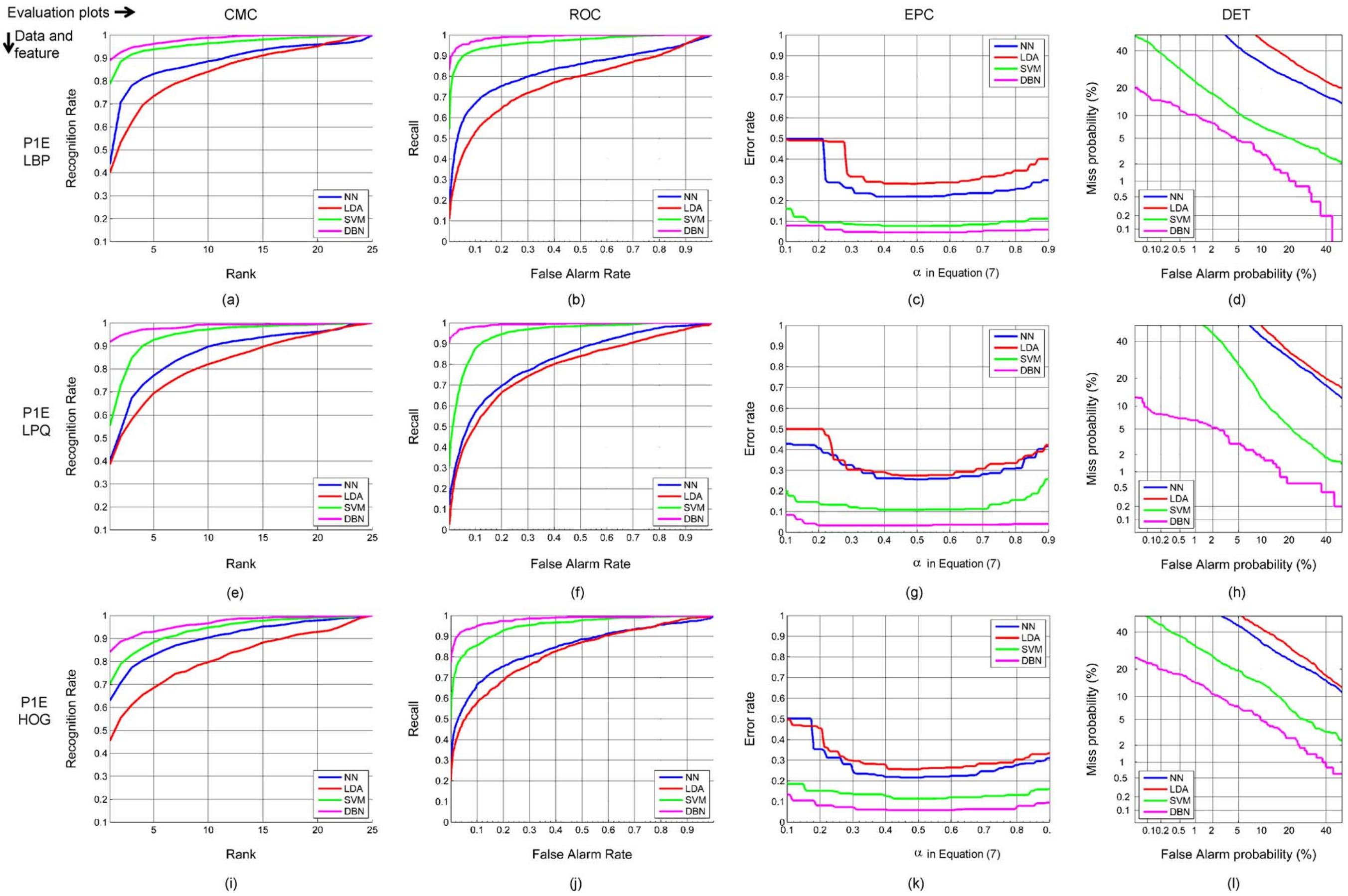

Fig. 5. Evaluation results for P1E. From left to right: different plots (CMC, ROC, EPC, and DET). From top to bottom: different feature descriptors (LBP, LPQ, and HOG). Best viewed in color.

pattern (LBP) and its derivatives are among the most popular choices [30], [35]. To tackle with blurred face recognition, local phase quantization (LPQ) has been adopted [31]. Recently, inspired by the success in object recognition tasks, histogram of oriented gradients (HOG) has been applied to face recognition [36]. In our experiments, we choose to use these three popular feature descriptors: LPQ, LBP, and HOG. For LBP and LPQ operation, the image is divided into the blocks of size 16 × 16. In LBP, $\mathrm{LBP}_{8,2}^{u2}$ is used as suggested in [30]. The parameters for LPQ are set to $M = 7, \alpha = 1/7$, and $\rho = 0.9$. For HOG, the image is divided into nine blocks and the number of orientation bins is set to 15. Note that any feature descriptors can be applied in the proposed framework. The dimensionality of the extracted feature vectors is reduced to 50 using PCA to enforce the efficiency during computation.

4) Classifiers Compared: The DBN is compared with three benchmark classifiers: nearest neighbor (NN), linear discriminant analysis (LDA) and support vector machine (SVM). These classifiers are commonly used in recognition tasks. In the SVM classifier, its linear version is used. For these classifiers, the same training and testing data are used as for DBN. After multiple testing samples are classified, we adopt the majority voting scheme to decide the final class label (identity) of each subject.

B. Experimental Results

1) Comparison With Different Classifiers: To compare with other classifiers, the rank-1 recognition rates for the four groups of sequences P1E, P1L, P2E, and P2L are reported in Table III. In most cases, NN and LDA are less able to discriminate the faces from the unconstrained video sequences due to the challenging dataset used. SVM improves the results by seeking for the maximum separation between the features of distinct subjects. Regardless of the choice of the feature descriptor, the proposed DBN classifier, compared to NN, LDA, and SVM, performs best in different sequences as a result of the encoding of the person-specific dynamics in the video and the fusion of multi-camera inputs.

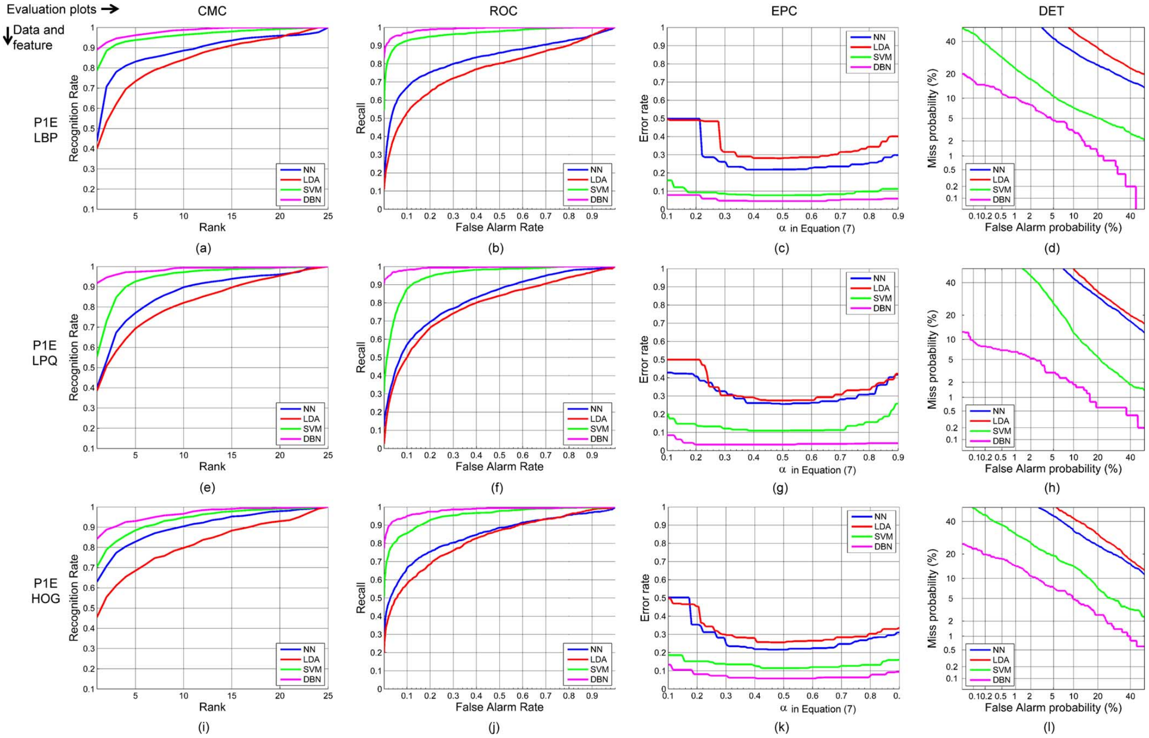

Fig. 6. The evaluation results for P1L. From left to right: different plots (CMC, ROC, EPC, and DET). From top to bottom: different feature descriptors (LBP, LPQ, and HOG). Best viewed in color.

To carefully investigate the performance of the classifiers, for each sequence four evaluation plots are presented: cumulative match characteristic (CMC) curve, receiver operating characteristic (ROC) plot, expected performance curve (EPC) [37], and detection error tradeoff (DET) plot. Fig. 5 shows the results for the P1E sequence. Fig. 5(a), (e), and (i) presents the CMC curves for LBP, LPQ, and HOG, respectively. The recognition rates for the top 25 ranks are reported as the gallery includes 25 subjects in P1. Compared to the other classifiers, the recognition results are more accurate using the proposed DBN at different ranks. The comparison of the results among different feature descriptors confirms the superiority of the proposed method over the other classifiers.

Fig. 5(b), (f), and (j) presents the ROC plots using the three different feature descriptors. The recognition using LPQ is better than LBP and HOG in terms of ROC performance. The reason is that LPQ is inherently designed as a blur invariant feature descriptor while the captured faces by the surveillance cameras show explicit blurriness due to subject's motion. Note that with different feature descriptors, the performance of the DBN is constantly better than the other classifiers in most cases. This indicates that the performance gain of the proposed method is not entirely feature dependent.

The EPCs in Fig. 5(c), (g), and (k) compare DBN with the other classifiers from the viewpoint of the tradeoff between false alarm and false reject probabilities. The x-axis represents $\alpha \in \mathbb{R}$ where $\alpha \in [0, 1]$ and the y-axis corresponds to the error rate β defined as

$$\beta = \alpha \times \mathrm{FAR} + (1 - \alpha \times \mathrm{FRR}) \qquad (7)$$

where FAR is the false alarm ratio and FRR represents the false rejection ratio. For all of the three feature descriptors, DBN reports lower error rate compared to the other classifiers. More importantly, the error rate is almost constant for all values of α.

Fig. 5(d), (h), and (l) presents the DET plots comparing the decision error rate of DBN versus the other classifiers. The performance is characterized by the miss and false alarm probabilities. Both x and y axes are scaled nonlinearly by their standard normal deviates such that a normal Gaussian distribution will plot as a straight line. The results show that the DBN reports less miss probability with equal false alarm probability compared to NN, LDA, and SVM. It is important to point out that not only the DBN outperforms the other classifiers as shown in the DET plots, but even in cases where DBN and SVM seem to have similar performance [e.g., Fig. 7(e)], the DET plot shows that DBN achieves significantly less miss probability [Fig. 7(h)].

Figs. 6–8 show results for sequences P1E, P1L, P2E, and P2L, respectively. The observations of Figs. 6–8 are similar to that of Fig. 5. Overall, compared to NN, LDA, and SVM, DBN is more robust in recognition and less prone to error.

2) Multiple Camera Versus Single Camera: We compare the recognition performance using three cameras against using only one camera in the proposed DBN framework. The DBN structure for a single camera is derived from Fig. 3 by removing the other two camera nodes. Table IV show the rank-1 recognition rates comparisons using three cameras together (ALL) against using only a single camera on sequences from P1E, P1L, P2E, and P2L. As can be seen, regardless of the specific choice of the feature descriptor, the recognition rates with three cameras

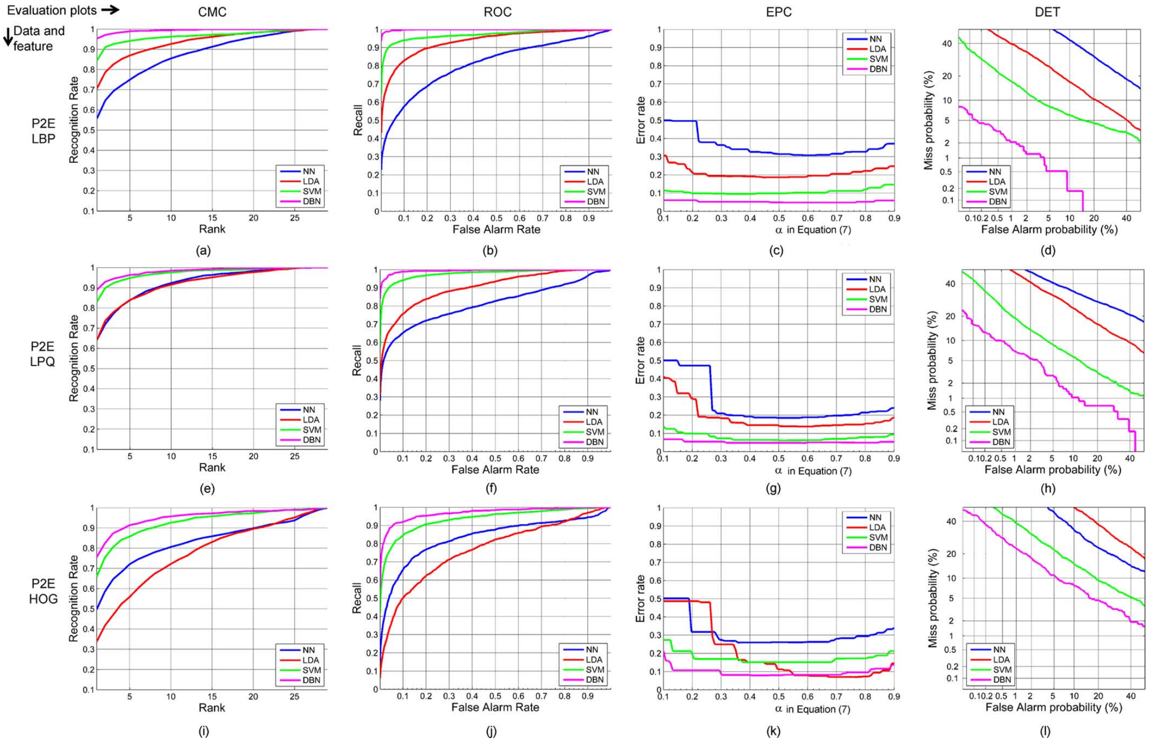

Fig. 7. The evaluation results for P2E. From left to right: different plots (CMC, ROC, EPC, and DET). From top to bottom: different feature descriptors (LBP, LPQ, and HOG). Best viewed in color.

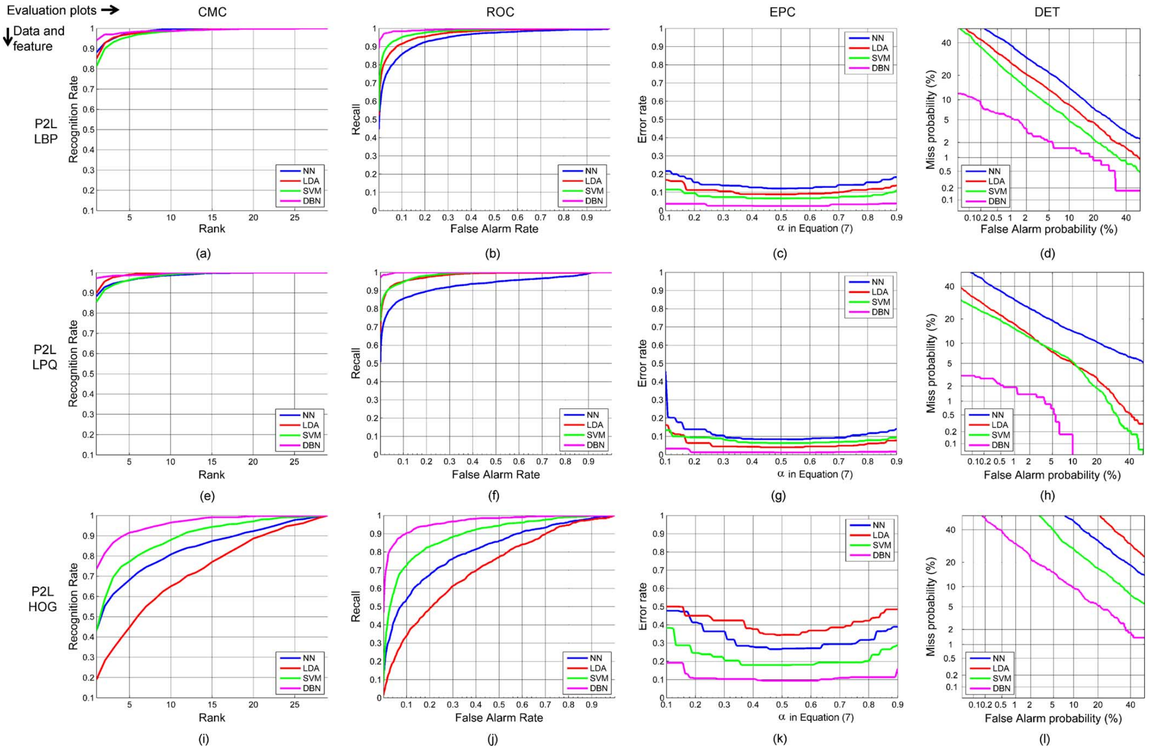

Fig. 8. The evaluation results for P2L. From left to right: different plots (CMC, ROC, EPC, and DET). From top to bottom: different feature descriptors (LBP, LPQ, and HOG). Best viewed in color.

are higher than using any of a single camera. The reason is that DBN takes into account the relationship of the three cameras through the dependencies, thus the complementary information from each camera is utilized to help improve the recognition performance. Also note that in most cases CAM_2 (C_2) provides higher recognition rates compared to CAM_1 (C_1) and CAM_3 (C_3) due to the near frontal faces it captured with relatively higher video quality. Although the performance using HOG is,

TABLE IV
RANK-1 RECOGNITION RATES WITH DIFFERENT CAMERAS ON DIFFERENT TESTING SEQUENCES (IN %)

Data→	P1E				P1L				P2E				P2L			
Camera→	C_1	C_2	C_3	ALL	C_1	C_2	C_3	ALL	C_1	C_2	C_3	ALL	C_1	C_2	C_3	ALL
LBP	77.3	81.3	75.3	**89.7**	80	81.3	75.3	**85.8**	74.1	86.2	84.5	**95.3**	55.8	65.2	63.2	**94.1**
LPQ	78.3	84	73.3	**91.7**	75.3	86	60	**86.3**	79.9	84.8	77.9	**89**	65.8	70.4	67.2	**97.2**
HOG	71.3	73.7	73.7	**84.3**	73.7	77	76.7	**89**	67.6	65.9	71	**75.5**	68.4	69.3	69	**73.6**
Average	75.7	79.7	74.1	**88.6**	76.3	81.4	70.7	**87**	73.9	79	77.8	**86.6**	63.3	68.3	66.5	**73.6**

in general, inferior to LBP and LPQ, for different camera, HOG gives similar recognition rates. This is due to the tolerance of the HOG descriptor for small pose variations.

V. CONCLUSION

We proposed a multi-camera face recognition system using DBN and this framework is suitable for applications such as surveillance monitoring in camera networks. In the proposed method, videos from multiple cameras are effectively utilized to provide the complementary information for robust recognition results. In addition, the temporal information among different frames are encoded by DBN to establish the person-specific dynamics to help improve the recognition performance. Experiments on a surveillance video dataset with a three-camera setup show that the proposed method performs better than the other benchmark classifiers using different feature descriptors by different evaluation criteria. Regarding the generality of our method, the feature nodes in the DBN can be replaced with any choice of informative feature descriptors and the proposed framework can be extended to the camera systems with different number of cameras.

ACKNOWLEDGMENT

The authors would like to thank National ICT Australia Limited (NICTA) for providing the ChokePoint dataset.

REFERENCES

[1] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[2] M. Grgic, K. Delac, and S. Grgic, "SCface—Surveillance cameras face database," *Multimedia Tools Appl.*, vol. 51, no. 3, pp. 863–879, Feb. 2011.

[3] Y. M. Lui, D. Bolme, B. Draper, J. Beveridge, G. Givens, and P. Phillips, "A meta-analysis of face recognition covariates," in *Proc. IEEE 3rd Int. Conf. Biometrics: Theory, Appl., Syst.*, Sep. 2009, pp. 1–8.

[4] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2011, pp. 74–81.

[5] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2003, vol. 1, pp. 313–320.

[6] W. Fan, Y. Wang, and T. Tan, , T. Kanade, A. Jain, and N. Ratha, Eds., "Video-based face recognition using Bayesian inference model," in *Audio- and Video-Based Biometric Person Authentication*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2005, vol. 3546, pp. 122–130.

[7] W. Liu, Z. Li, and X. Tang, "Spatio-temporal embedding for statistical face recognition from video," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 374–388.

[8] X. Liu and T. Cheng, "Video-based face recognition using adaptive hidden Markov models," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2003, vol. 1, pp. 340–345.

[9] J. Stallkamp, H. Ekenel, and R. Stiefelhagen, "Video-based face recognition on real-world data," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[10] S. Biswas, G. Aggarwal, and P. Flynn, "Face recognition in low-resolution videos using learning-based likelihood measurement model," in *Int. Joint Conf. Biometr.*, Oct. 2011, pp. 1–7.

[11] B. Xie, V. Ramesh, Y. Zhu, and T. Boult, "On channel reliability measure training for multi-camera face recognition," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Feb. 2007, p. 41.

[12] J. Harguess, C. Hu, and J. Aggarwal, "Fusing face recognition from multiple cameras," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Dec. 2009, pp. 1–7.

[13] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen, "Image sets alignment for video-based face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2626–2633.

[14] Y. Xu, A. Roy-Chowdhury, and K. Patel, "Pose and illumination invariant face recognition in video," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshop Biometr.*, Jun. 2007, pp. 1–7.

[15] U. Park, H. Chen, and A. Jain, "3D model-assisted face recognition in video," in *Proc. 2nd Can. Conf. Comput. Robot Vis.*, May 2005, pp. 322–329.

[16] C.-T. Liao, S.-F. Wang, Y.-J. Lu, and S.-H. Lai, "Video-based face recognition based on view synthesis from 3D face model reconstructed from a single image," in *Proc. IEEE Int. Conf. Multimedia Expo*, Apr. 2008, pp. 1589–1592.

[17] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.* vol. 3, no. 1, pp. 71–86, Jan. 1991.

[18] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision* vol. 57, no. 2, pp. 137–154, May 2004.

[19] M. Bäuml, K. Bernardin, M. Fischer, H. Ekenel, and R. Stiefelhagen, "Multi-pose face recognition for person retrieval in camera networks," in *Proc. 7th IEEE Int. Conf. Adv. Video Signal Based Surveill*, Sep. 2010, pp. 441–447.

[20] B. Lovell, S. Chen, A. Bigdeli, E. Berglund, and C. Sanderson, "On intelligent surveillance systems and face recognition for mass transport security," in *Proc. Int. Conf. Control, Automat., Robot. Vis.*, Dec. 2008, pp. 713–718.

[21] C. Shan, "Face recognition and retrieval in video," in *Video Search and Mining*, D. Schonfeld, C. Shan, D. Tao, and L. Wang, Eds. Berlin, Germany: Springer, 2010.

[22] G. Heusch and S. Marcel, "Bayesian networks to combine intensity and color information in face recognition," in *Proc. 3rd Int. Conf. Adv. Biometr.*, Berlin, Germany, 2009, pp. 414–423.

[23] A. Nefian, "Embedded Bayesian networks for face recognition," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2002, vol. 2, pp. 133–136.

[24] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Process.* vol. 2002, no. 1, pp. 1274–1288, Jan. 2002.

[25] M. Kafai and B. Bhanu, "Dynamic Bayesian networks for vehicle classification in video," *IEEE Trans. Ind. Informat.*, vol. 8, no. 1, pp. 100–109, Feb. 2012.

[26] T. Wang, Q. Diao, Y. Zhang, G. Song, C. Lai, and G. Bradski, "A dynamic Bayesian network approach to multi-cue based visual tracking," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2004, vol. 2, pp. 167–170, vol. 2.

[27] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 699–714, May 2005.

[28] D. Bouchaffra, "Topological dynamic Bayesian networks: Application to human face identification across ages," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 1–8.

[29] L. An, M. Kafai, and B. Bhanu, "Face recognition in multi-camera surveillance videos using dynamic Bayesian network," in *Proc. 6th ACM/IEEE Int. Conf. Distrib. Smart Cameras*, Oct. 2012, pp. 1–6.

[30] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Eur. Conf. Comput. Vis.*, 2004, pp. 469–481.

[31] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using local phase quantization," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[32] G. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Mach. Learn.*, vol. 9, no. 4, pp. 309–347, 1992.

[33] T. Xiang and S. Gong, "Visual learning given sparse data of unknown complexity," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, vol. 1, pp. 701–708.

[34] N. Poh, C. H. Chan, J. Kittler, S. Marcel, C. McCool, E. Rúa, J. Castro, M. Villegas, R. Paredes, V. Štruc, N. Pavešić, A. Salah, H. Fang, and N. Costen, "An evaluation of video-to-video face verification," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 4, pp. 781–801, Dec. 2010.

[35] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, vol. 1, pp. 786–791.

[36] O. Dèniz, G. Bueno, J. Salido, and F. D. la Torre, "Face recognition using histograms of oriented gradients," *Pattern Recognit. Lett.* vol. 32, no. 12, pp. 1598–1603, 2011.

[37] S. B. Bengio, J. Marithoz, and M. Keller, "The expected performance curve," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 9–16.

Le An received the B.Eng. degree in telecommunications engineering from Zhejiang University, Hangzhou, China, in 2006, and the M.S. degree in electrical engineering from Eindhoven University of Technology, Eindhoven, The Netherlands, in 2008. He is currently a Ph.D. student in electrical engineering at the Center for Research in Intelligent Systems at the University of California, Riverside, CA, USA.

His research interests include image processing, computer vision, pattern recognition, and machine learning. His current research focuses on image enhancement and unconstrained face recognition in camera networks.

Mehran Kafai (S'11) received the B.S. degree in computer engineering from the Bahonar University, Kerman, Iran, in 2002, the M.S. degree in computer engineering from the Sharif University of Technology, Tehran, Iran, in 2005, and the M.S. degree in computer science from the San Francisco State University, San Francisco, CA, USA, in 2009. He is currently working toward the Ph.D. degree in computer science at the Center for Research in Intelligent Systems at the University of California, Riverside, CA, USA.

His research interests are in computer vision, pattern recognition, machine learning, and data mining. His recent research has been concerned with robust object recognition algorithms.

Bir Bhanu (S'72–M'82–SM'87–F'95) received the S.M. and E.E. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, and the M.B.A. degree from the University of California, Irvine, CA, USA.

He is the Distinguished Professor of electrical engineering; a Cooperative Professor of computer science and engineering, mechanical engineering, and bioengineering; and the Director of the Center for Research in Intelligent Systems and the Visualization and Intelligent Systems Laboratory with the University of California, Riverside (UCR). In addition, he serves as the Director of National Science Foundation (NSF) The Integrative Graduate Education and Research Traineeship program on Video Bioinformatics at UCR. He has been the Principal Investigator of various programs for NSF, Defense Advanced Research Project Agency, National Aeronautics and Space Administration, Air Force Ofce of Scientic Research, Office of Naval Research, Army Research Office, and other agencies and industries in the areas of video networks, video understanding, video bioinformatics, learning and vision, image understanding, pattern recognition, target recognition, biometrics, autonomous navigation, image databases, and machine and vision applications. He is the author or coauthor of 400 reviewed technical publications, including over 100 journal papers and 40 book chapters; seven published books; and three edited books. He is the holder of 18 (three pending) patents. His current research interests include computer vision, pattern recognition and data mining, machine learning, artificial intelligence, image processing, image and video database, graphics and visualization, robotics, human and computer interactions, and biological, medical, military and intelligence applications.

Dr. Bhanu is Fellow of American Association for the Advancement of Science, International Association of Pattern Recognition, and The International Society for Optical Engineering.