Face Recognition in Multi-Camera Surveillance Videos using Dynamic Bayesian Network

Le An, Mehran Kafai, Bir Bhanu
Center for Research in Intelligent Systems, University of California, Riverside
lan004@ucr.edu, mkafai@cs.ucr.edu, bhanu@cris.ucr.edu

***Abstract*—Face recognition in surveillance videos is inherently difficult due to the limitation of the camera hardware as well as the image acquisition process in which non-cooperative subjects are recorded in arbitrary poses and resolutions in different lighting conditions with noise and blurriness. Furthermore, as multiple cameras are usually distributed in a camera network and the subjects are moving, different cameras often capture the subject in different views. In this paper, we propose a probabilistic approach for face recognition suitable for a multi-camera video surveillance network. A Dynamic Bayesian Network (DBN) is used to incorporate the information from different cameras as well as the temporal clues from consecutive frames. The proposed method is tested on a public surveillance video dataset. We compare our method to different well-known 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 can be improved.**

I. Introduction

In recent years, surveillance video cameras have been widely established in both public and private venues for the purpose of security monitoring, access control, etc. Individual recognition in such camera networks is often desirable especially for law enforcement purposes. Although other biometric traits such as gait can be used to recognize different subjects [1], it is preferred to use more distinct trait such as face to identify a subject. Face recognition has been studied extensively. However, face recognition in an uncontrolled environment such as in surveillance camera videos remains very challenging and the recognition rate could drop dramatically to less than 10% [2]. The challenges to face recognition in surveillance cameras are mainly due to the following reasons:

- Low resolution. The surveillance cameras normally captures images at low resolution (e.g. 800×600) compared to consumer cameras. The pixels that account for the faces are very limited. Previous studies have shown that faces of size 64×64 are required for the existing algorithms to achieve good recognition accuracy [3].
- Arbitrary poses. The image acquisition process in surveillance cameras is non-intrusive. The subjects are moving freely in the field-of-view of the cameras. 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 surveillance cameras, the illumination on the subject's face could vary significantly (i.e., the subjects walks into the shade from direct sunshine.)
- Noise and blurriness. The captured images are often corrupted by noise and the motion of the subjects usually introduces blurriness.

Fig. 1 shows an example of a subject's face captured by three surveillance cameras. The cameras have different viewing angles and none of the cameras captures the frontal face of the subject. The face images exhibit variations in resolution, lighting condition and poses. In addition, noise, blurriness and occlusion are also observable. Under such circumstance, the traditional face recognition algorithm such as Eigenface, which was developed for face recognition at fixed poses using still images, would fail to work effectively. Despite the aforementioned difficulty, a multi-camera system provides different views of the subjects which are complementary to each other and enable the potential to improve the recognition performance.

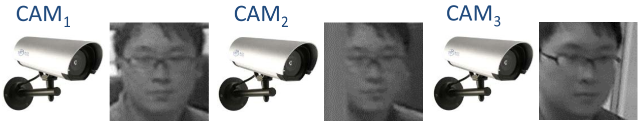

Fig. 1. The subject's face are captured by 3 cameras from different views in a typical surveillance camera system.

A. Related Work

To recognize a face in videos, different approaches have been proposed. In [4] a Hidden Markov Model (HMM) was used for video-based face recognition. In this model the temporal characteristics were analyzed over time. Xu *et al.* [5] developed a framework for pose and illumination invariant face recognition from video sequences by integrating the effects of motion and lighting changes. In [6] the face recognition in video was tackled by exploiting the spatial and temporal information based on Bayesian keyframe learning and nonparametric discriminant embedding. These methods were tested on controlled datasets that do not address the difficulties in real surveillance video data. Recently, Biswas *et al.* [7] proposed a learning-based likelihood measurement to match high-resolution frontal view gallery images with probe images from surveillance videos. Wong *et al.* [8] 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. These methods essentially tried to recognize faces in frontal view, which might not be available in surveillance records of a subject.

Another solution for unconstrained face recognition is to build up 3D face models. In [9] a 3D morphable model was generated as a linear combination of basis exemplars. The model was fit to an input image by changing the shape and albedo parameters of the model. Barreto and Li [10] proposed a framework for 3D face recognition system with variation of expression. The 3D based methods are in general computationally expensive. Furthermore, a 3D model is difficult to be constructed from low-resolution inputs.

In an effort to recognize faces using more than one cameras, 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 have been 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 [13] which depends on the robustness of the face tracker.

B. Contributions of This Paper

There has been a growing interest to study the temporal dynamics in video sequences to improve the recognition performance in recent years [14]. In this paper, we propose a video-to-video face recognition approach using a Dynamic Bayesian Network (DBN), utilizing different frames from multiple cameras. DBN has previously been applied to tasks such as speech recognition [15] and vehicle classification [16]. 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 contributions of this paper are listed below:

- 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 that introduces the DBN for face recognition in surveillance camera systems with multiple cameras. The framework is flexible and can be easily adapted to any specific camera system settings with different number of cameras. Any feature descriptors can be used in this framework.
- We test the proposed method on a publicly available multi-camera surveillance video dataset [8] with unconstrained face acquisition, in contrast to the other datasets which were taken in a controlled environment. We compare the proposed method with other representative classifiers using different feature descriptors.

The remainder of this paper is organized as follows. Section 2 describes the details of the proposed method. In Section 3 the experimental results are reported. We conclude this paper in Section 4. For better understanding of the symbols used in this paper, we define the symbols in Table I.

TABLE I
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

II. Technical Details

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

A. Bayesian Networks

A Bayesian network (BN) is a graphic model that is defined as a directed acyclic graph. The nodes in the model represent the random variables and the edges define the dependencies between the random variables. Each variable is conditionally independent of its non-descendants, given the value of its parents. A BN can effectively represent and factor the joint probability distributions and it is suitable for the classification tasks. In the scope of multi-camera face recognition, when several face images of the same subject are captured by different cameras at a certain time, we construct the 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 classes/subjects.
- 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) [17] or local phase quantization (LPQ) [18] can be adopted. The notation CAM is used to represent this random variable.

Fig. 2 shows structure of the BN. CAM_k is the random variable representing the feature vector from the face image in camera k. Given the identity of the subject, the camera nodes are assumed to be conditionally independent.

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

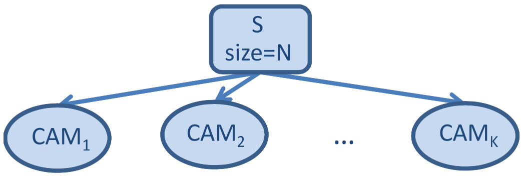

Fig. 2. The BN structure with root node and camera nodes.

$$\begin{aligned} s &= \underset{S}{\operatorname{argmax}}\, p(S|CAM_1,\ldots,CAM_K) \\ &= \underset{S}{\operatorname{argmax}}\, \frac{p(CAM_1,\ldots,CAM_K|S)p(S)}{\sum_S p(CAM_1,\ldots,CAM_K|S)p(S)} \qquad (1) \\ &= \underset{S}{\operatorname{argmax}} \prod_{i=1}^{K} \frac{p(CAM_k|S)p(S)}{p(CAM_k)} \end{aligned}$$

where $p(S)$ is the prior probability of the presence of each subject and is usually modeled by a uniform distribution.

B. Dynamic Bayesian Networks 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 [19], 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. 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.

In terms of complexity, an HMM would require $O(T(N^K)^2)$ for inference, $O(N^{2K})$ parameters to specify $P(S^t|S^{t-1})$, 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 $P(S^t|S^{t-1})$. The DBN we model from the HMM variant has exponentially less parameters and inference is much faster.

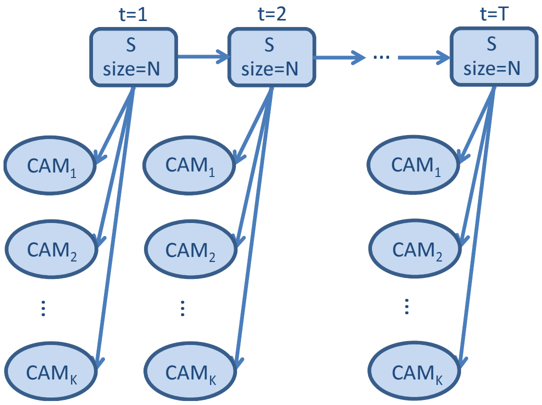

Fig. 3. The DBN structure for T time slices.

Operating a graphic model requires three main steps: defining 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 $t = 1 \ldots T$ has $K+1$ nodes; one root node S, and K camera nodes $CAM_{k=1\ldots K}$. This structure is the same as shown in Fig. 2.

The number of possible structures is super-exponential in the total number of nodes; therefore, it is best to avoid performing exhaustive search for structure learning. In this paper, we use the K2 structure learning algorithm to determine the structure. K2 uses a greedy approach to incrementally add parents to a node according to a chosen scoring function. We use the Completed Likelihood Akaike Information Criterion (CL-AIC) scoring function for this purpose. CL-AIC for video analysis optimizes the prediction and explanation capabilities of the specified model simultaneously. The intra-slice topology is illustrated in Fig. 3 with T 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 determined. For the first time slice this includes:

$$p(CAM_1|S),\ p(CAM_2|S), \ldots,\ p(CAM_K|S),\ p(S). \qquad (2)$$

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

$$\begin{gathered} p(CAM_1^t|S^t),\ p(CAM_2^t|S^t), \ldots,\ p(CAM_K^t|S^t) \\ p(S^t|S^{t-1}) \qquad (3) \end{gathered}$$

Given a dataset with N subjects, N distributions with different parameters are required, one for each value of S^t to determine $p(CAM_k^t|S^t)$. An inference algorithm is applied to compute the marginal probability from the evidence (testing data). Specifically, inference determines the subject's identity by $p(S^T|CAM_{k=1\ldots K}^{(1:T)})$, where $CAM_{k=1\ldots K}^{(1:T)}$ refers to features from all cameras for time slices 1 to T. In other words, a probability distribution over the set of all the subjects is defined. Equation 4 shows how $p(S^t|CAM_{k=1\ldots K}^{(1:T)})$ is computed for any $t = 2 \ldots T$.

III. Experiments

A. Dataset and Parameter Settings

Dataset: We use the ChokePoint dataset [8] which is designed for evaluating face recognition algorithms under real-world surveillance conditions. Two subsets of the video sequences from portal 1 (P1) are used. In total 25 subjects are involved. The first subset contains 4 sequences recorded when the subjects were entering the portal (P1E_S1, P1E_S2, P1E_S3, P1E_S4). The second subset was recorded in the opposite direction when the subjects were leaving the portal (P1L_S1, P1L_S2, P1L_S3, P1L_S4). In each sequence, videos of the moving subjects captured by three cameras that are mounted in an array above the portal.

$$
\begin{aligned}
p(S^t|CAM^{1:T}_{k=1:K}) &= p(S^t, CAM^1_1 = cam^1_1, \ldots, CAM^T_K = cam^T_K) \times \underbrace{1/p(CAM^1_1 = cam^1_1, \ldots, CAM^T_K = cam^T_K)}_{=L \text{ (a constant)}} \\
\text{by marginalization} &= \sum_{S^1,\ldots,S^{t-1},S^{t+1},\ldots,S^T} p(S^1, \ldots, S^T, CAM^1_1 = cam^1_1, \ldots, CAM^T_K = cam^T_K) \times L \\
\text{by Bayes net factoring} &= \sum_{S^1,\ldots,S^{t-1},S^{t+1},\ldots,S^T} p(S^1) \prod_{i=2:T} p(S^i|S^{i-1}) \prod_{i=1:T} p(CAM^i_{k=1:K}|S^i) \times L \\
\text{by splitting products} &= \sum_{S^1,\ldots,S^{t-1},S^{t+1},\ldots,S^T} p(S^1) \prod_{i=2:t} p(S^i|S^{i-1}) \prod_{i=1:t} p(CAM^i_{k=1:K}|S^i) \times \\
&\qquad \prod_{i=t+1:T} p(S^i|S^{i-1}) \prod_{i=t+1:T} p(CAM^i_{k=1:K}|S^i) \times L \\
&= \sum_{S^1,\ldots,S^{t-1},S^{t+1},\ldots,S^T} p(S^1, \ldots, S^t, CAM^1_{k=1:K}, \ldots, CAM^t_{k=1:K}) \times \\
&\qquad p(S^{t+1}, \ldots, S^T, CAM^{t+1}_{k=1:K}, \ldots, CAM^T_{k=1:K}|S^t) \times L \\
&= \sum_{S^1,\ldots,S^{t-1}} p(S^1, \ldots, S^t, CAM^1_{k=1:K}, \ldots, CAM^t_{k=1:K}) \times \\
&\qquad \sum_{S^{t+1},\ldots,S^T} p(S^{t+1}, \ldots, S^T, CAM^{t+1}_{k=1:K}, \ldots, CAM^T_{k=1:K}|S^t) \times L
\end{aligned} \tag{4}
$$

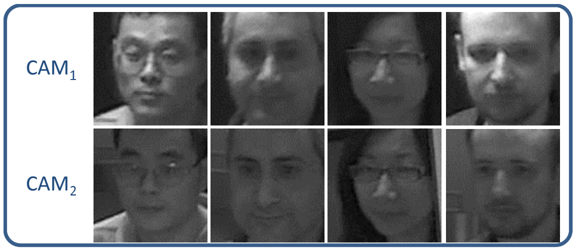

Fig. 4. Sample images from the ChokePoint [8] dataset.

In the experiments, videos from two cameras (CAM_1 and CAM_2) are used. The resolution of the captured frames are 800×600 at a frame rate of 30 fps. Based on the availability of the data, 20 frames from each camera in each sequence for every subject are used. The detected and aligned faces are provided with the dataset. The faces are resized to 96×96 and histogram normalization is applied as a preprocessing step to reduce the non-uniform lighting effects. The experiments are conducted in two rounds: when training is performed on sequences in P1L, P1E is used for testing and vice versa. This dataset is challenging for face recognition task as the faces captured are unconstrained and of low quality. Fig. 4 shows some sample images.

DBN construction: The DBN is constructed with 5 time slices. From the 20 frames in the training sequence, every 5 frames are used together as one training sample. The number of time slices is determined empirically to alleviate overfitting and underfitting. During testing each sample is also constructed with 5 time slices.

Feature descriptors: With the focus to examine the performance of the proposed classifier with different feature descriptors, we use raw pixel intensity values along with two popular face descriptors LBP and LPQ. For LBP and LPQ operation, the image is divided into the blocks of size 16×16. In LBP, $LBP^{u2}_{8,2}$ is used as suggested in [17]. The parameters for LPQ are set to $M = 7$, $\alpha = 1/7$ and $\rho = 0.9$. Note that any feature descriptors can be applied in the proposed framework. The dimensionality of the extracted features is reduced to 50 using PCA to enforce the efficiency during computation.

Methods compared: The proposed method is compared with three commonly used classifiers: nearest neighbor (NN), linear discriminant analysis (LDA) and support vector machine (SVM). In the SVM classifier, linear kernel is used. After multiple testing samples are classified, we adopt majority voting to decide the final class label for each subject.

B. *Experimental Results*

We first examine the effect of using multiple cameras for recognition. Table II and Table III shows the rank-1 recognition rates comparisions using two cameras against using only one camera on P1E and P1L. It can be seen that in general the performance by using two cameras is better than any of the single camera using different feature descriptors. This is expected since the complementary information from two cameras helps to improve the recognition performance. The recognition rate in CAM_1 is consistently higher than CAM_2 because the view of CAM_1 is more frontal (see Fig. 4), thus providing more faithful clue to identify the subjects.

To compare with other classifiers, the rank-1 recognition rates for P1E and P1L are reported here in Table IV and Table V respectively. In this challenging surveillance dataset, NN and LDA are not able to discriminate the faces in unconstrained situations and can not yield competitive results. SVM improves the results by seeking for the maximum separation between the features from distinct subjects. Compared to

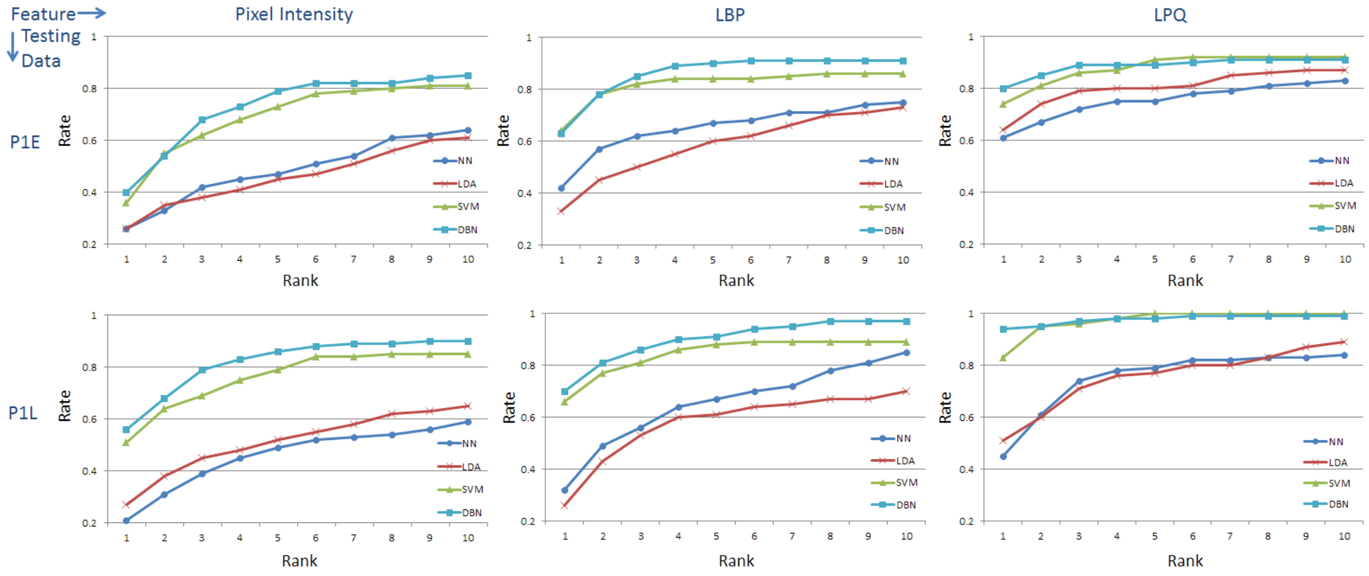

Fig. 5. Cumulative match characteristic (CMC) curves for the testing sequences. From top to bottom: results from testing sets P1E and P1L. From left to right: results obtained using different feature descriptors (pixel intensity, LBP and LPQ).

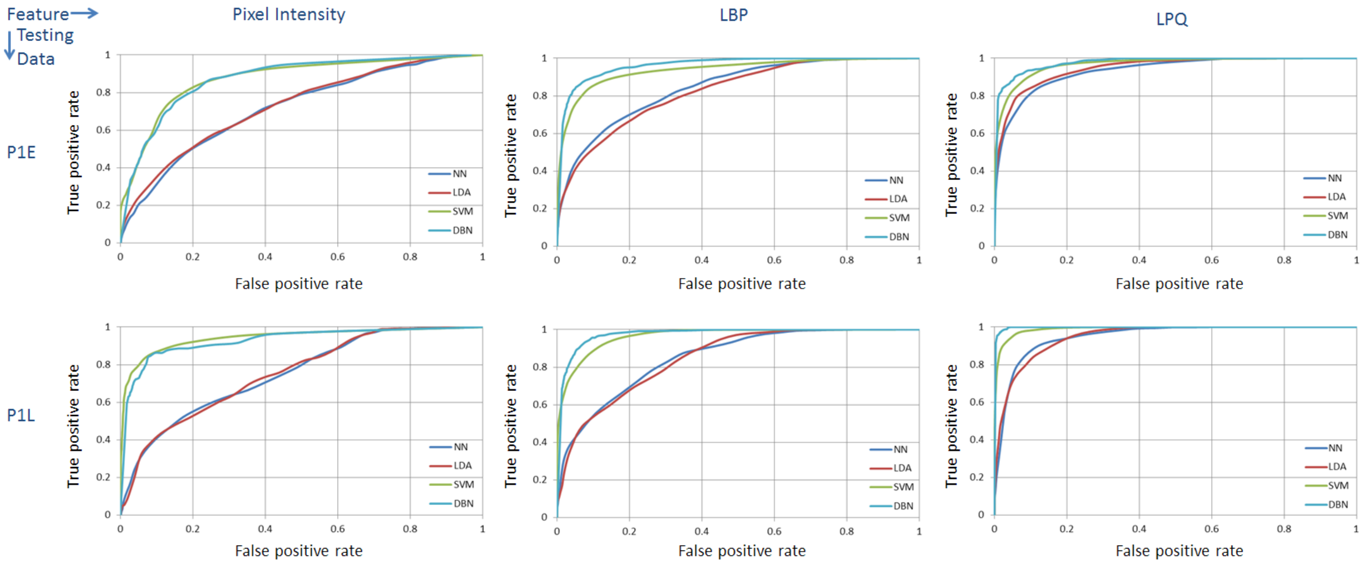

Fig. 6. Receiver operating characteristic (ROC) curves for the testing sequences. From top to bottom: results from testing sets P1E and P1L. From left to right: results obtained using different feature descriptors (pixel intensity, LBP and LPQ).

SVM, DBN performs better in most of the sequences due to the encoding of the person-specific dynamics in the video.

On average, compared to the NN, LDA and SVM classifiers, the average improvements of DBN on P1E and P1L are 29.3%, 29.3% and 4.8% respectively. The recognition rates using LPQ is remarkably better than LBP and pixel intensity values. 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 difference 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 feature dependent.

The cumulative match characteristic (CMC) curves are shown in Fig. 5. The recognition rates in top 10 ranks are reported as further ranks are not useful in practice for a close dataset with limited number of subjects. Compared to the other classifiers, the recognition results are more accurate using the proposed DBN classifier at different ranks. The comparison of the results among different feature descriptors confirms the superiority of the proposed method over the other classifiers.

The receiver operating characteristic (ROC) curves are presented in Fig. 6. The classification performance of the proposed DBN is better than the other classifiers with more

TABLE II
RANK-1 RECOGNITION RATES WITH DIFFERENT CAMERAS ON P1E

Camera→ Feature↓	CAM_1	CAM_2	$CAM_1 + CAM_2$
Intensity	28%	27%	**40%**
LBP	**64%**	33%	63%
LPQ	59%	47%	**80%**
Average	50.3%	35.7%	**61%**

TABLE III
RANK-1 RECOGNITION RATES WITH DIFFERENT CAMERAS ON P1L

Camera→ Feature↓	CAM_1	CAM_2	$CAM_1 + CAM_2$
Intensity	41%	16%	**56%**
LBP	59%	38%	**70%**
LPQ	70%	55%	**94%**
Average	56.7%	36.3%	**73.3%**

TABLE IV
RANK-1 RECOGNITION RATES WITH DIFFERENT CLASSIFIERS ON P1E

Classifier→ Feature↓	NN	LDA	SVM	DBN
Intensity	26%	26%	36%	**40%**
LBP	42%	33%	**64%**	63%
LPQ	61%	64%	74%	**80%**
Average	43%	41%	58%	**61%**

TABLE V
RANK-1 RECOGNITION RATES WITH DIFFERENT CLASSIFIERS ON P1L

Classifier→ Feature↓	NN	LDA	SVM	DBN
Intensity	21%	27%	51%	**56%**
LBP	32%	26%	66%	**70%**
LPQ	45%	51%	83%	**94%**
Average	32.7%	34.7%	66.7%	**73.3%**

descriptive features (LBP and LBP) especially at low false positive rates. When simple features (pixel intensity values) are used, SVM and DBN perform similarly.

IV. CONCLUSIONS

In this paper, a DBN based multi-camera face recognition algorithm suitable for surveillance camera systems is proposed. Videos from multiple cameras are effectively utilized in this model. Besides the facial information in individual frames, the temporal information among adjacent frames are learned by DBN to establish the person-specific dynamics to help improve the recognition performance. Experiments on a surveillance video dataset show that the proposed method performs better compared to the other classifiers using different feature descriptors. The feature nodes in DBN can be replaced with any informative feature descriptors since the advantageous of the DBN is not feature dependent. In addition, this framework is flexible and can be easily adapted to surveillance systems with different number of cameras.

ACKNOWLEDGMENTS

This work was supported in part by NSF grant 0905671 and ONR grants N00014-07-C-0311 and N00014-09-C-0388.

REFERENCES

[1] J. Han and B. Bhanu, "Individual recognition using gait energy image," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 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 *Biometrics: Theory, Applications, and Systems, 2009. BTAS '09. IEEE 3rd International Conference on*, Sept. 2009, pp. 1 –8.

[4] X. Liu and T. Cheng, "Video-based face recognition using adaptive hidden Markov models," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1, June 2003, pp. I–340 – I–345 vol.1.

[5] Y. Xu, A. Roy-Chowdhury, and K. Patel, "Pose and illumination invariant face recognition in video," in *Computer Vision and Pattern Recognition Workshop on Biometrics (CVPRW), 2011 IEEE Computer Society Conference on*, June 2007, pp. 1 –7.

[6] W. Liu, Z. Li, and X. Tang, "Spatio-temporal embedding for statistical face recognition from video," in *Proceedings of the 9th European conference on Computer Vision - Volume Part II*, ser. ECCV'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 374–388.

[7] S. Biswas, G. Aggarwal, and P. Flynn, "Face recognition in low-resolution videos using learning-based likelihood measurement model," in *Biometrics (IJCB), 2011 International Joint Conference on*, Oct. 2011, pp. 1 –7.

[8] 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 *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, June 2011, pp. 74 –81.

[9] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 9, pp. 1063 – 1074, Sept. 2003.

[10] C. Li and A. Barreto, "An integrated 3D face-expression recognition approach," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 3, May 2006, p. III.

[11] B. Xie, V. Ramesh, Y. Zhu, and T. Boult, "On channel reliability measure training for multi-camera face recognition," in *Applications of Computer Vision, 2007. WACV '07. IEEE Workshop on*, Feb. 2007, p. 41.

[12] J. Harguess, C. Hu, and J. Aggarwal, "Fusing face recognition from multiple cameras," in *Applications of Computer Vision (WACV), 2009 Workshop on*, Dec. 2009, pp. 1 –7.

[13] M. Bäuml, K. Bernardin, M. Fischer, H. Ekenel, and R. Stiefelhagen, "Multi-pose face recognition for person retrieval in camera networks," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, Sept. 2010, pp. 441 –447.

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

[15] 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. [Online]. Available: http://dx.doi.org/10.1155/S1110865702206083

[16] M. Kafai and B. Bhanu, "Dynamic Bayesian networks for vehicle classification in video," *Industrial Informatics, IEEE Transactions on*, vol. 8, no. 1, pp. 100 –109, Feb. 2012.

[17] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with Local Binary Patterns," in *European Conference on Computer Vision*, 2004, pp. 469–481.

[18] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using Local Phase Quantization," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, Dec. 2008, pp. 1 –4.

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