# A Bayesian Network Approach to Traffic Flow Forecasting

Shiliang Sun, Changshui Zhang, *Member, IEEE*, and Guoqiang Yu

***Abstract*—A new approach based on Bayesian networks for traffic flow forecasting is proposed. In this paper, traffic flows among adjacent road links in a transportation network are modeled as a Bayesian network. The joint probability distribution between the cause nodes (data utilized for forecasting) and the effect node (data to be forecasted) in a constructed Bayesian network is described as a Gaussian mixture model (GMM) whose parameters are estimated via the competitive expectation maximization (CEM) algorithm. Finally, traffic flow forecasting is performed under the criterion of minimum mean square error (mmse). The approach departs from many existing traffic flow forecasting models in that it explicitly includes information from adjacent road links to analyze the trends of the current link statistically. Furthermore, it also encompasses the issue of traffic flow forecasting when incomplete data exist. Comprehensive experiments on urban vehicular traffic flow data of Beijing and comparisons with several other methods show that the Bayesian network is a very promising and effective approach for traffic flow modeling and forecasting, both for complete data and incomplete data.**

***Index Terms*—Bayesian network, expectation maximization algorithm, Gaussian mixture model, traffic flow forecasting.**

## I. Introduction

URBAN traffic control systems (UTCSs) and freeway management systems around the world are collecting large amount of traffic condition data every day. Typical data include volume, flow rate, occupancy, and speed. Development of systems that put these data to good use for traffic control and management has become an active area of ongoing transportation research, which is usually referred to as Intelligent Transportation Systems (ITS) [1]. In the research area of ITS, traffic flow forecasting is a very important issue. Reliable analysis of historical trends of traffic flows is an important input to many of the traffic management and control systems in operation and under development. Some well-known systems, such as the Split Cycle Offset Optimization Technique (SCOOT) system and the Sydney Coordinated Adaptive Traffic (SCAT) system, integrated the traffic flow forecasting function as fundamental modules. Without an effective forecasting capability, these systems would not operate smoothly.

Manuscript received March 25, 2005; revised July 5, 2005, September 9, 2005, and September 14, 2005. This work was supported by the National Natural Science Foundation of China under Project 60475001. The Associate Editor for this paper was D.-H. Lee.

S. Sun and C. Zhang are with the State Key Laboratory of Intelligent Technology and Systems, Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: sunsl02@mails.tsinghua.edu.cn; zcs@mail.tsinghua.edu.cn).

G. Yu is with NuTech Company Limited, Beijing 100084, China (e-mail: hugo79@gmail.com).

Digital Object Identifier 10.1109/TITS.2006.869623

In this paper, we concentrate on the problem of short-term traffic flow rate forecasting, which is to determine the traffic condition data in the next time interval, usually in the range of 5 min to half an hour. During the past decades, some approaches ranging from simple to complex are proposed for traffic flow forecasting. Simple ones, such as random walk (RW, which is informed only by the current traffic condition), historical average (whose prediction is solely based on the average of all corresponding observed flow rates), and informed historical average (which combines the above two ideas), can only work well in specific situations [1]. One class of the complex approaches, UTCS prediction method, is based on forecasting philosophies similar to the above methods [1]. The first generation of UTCS prediction method relied heavily on historical data, which would bring along similar drawbacks as the historical average method; although the philosophies behind UTCS-2 and UTCS-3 prediction are simple, the forecasting methods are highly complex. There are also other elaborate methods including approaches based on time series models (including ARIMA, seasonal ARIMA) [1]–[3], Kalman filter theory [4], neural network approaches [5], nonparametric methods [6], simulation models [7], local regression models [8], [9], layered models known as the ATHENA model [10] and the KARIMA model [11], fuzzy-neural approach [12], and Markov chain model [13].

Although these methods have alleviated difficulties in traffic modeling and forecasting to some extent, from a careful review we can still find a problem, that is, most of them have not made good use of information from adjacent roads to analyze the trends of the object road. Some approaches did not even use data from adjacent road links at all. Although Chang *et al.* [14] utilized data from other roadways to make judgmental adjustments, the information was still not used to its full potential. Yin *et al.* [12] developed a fuzzy-neural model (FNM) to predict traffic flows in an urban street network that only utilizes upstream flows in the current time interval to forecast the selected downstream flow in the next interval. However, in order to forecast as precisely as possible, information from adjacent (mainly, upstream) roads and the current road should be considered all together, as some researchers advised.

Another drawback of the existing approaches mentioned above is that they hardly work when the data used for forecasting are incomplete (partially missing or unavailable). The incomplete data can be caused by malfunctions or measurement errors in data collection and recording systems, such as failed loop detectors, faulty loop amplifiers, or signal communication and processing errors (the situation of data abnormality caused by incidents/accidents does not belong to the incomplete data

case, since the data are not incomplete, but abnormal). Although the historical average method (filling up the incomplete data with their historical average values) is often adopted to cope with this issue, the forecasting accuracy is quite limited. There has not been a general well-defined approach for incomplete data forecasting yet. Therefore, developing an approach to work in case of incomplete data is of great demand.

The Bayesian network approach, as studied comprehensively in the communities of statistical analysis, artificial intelligence, and machine learning, gives us some inspiration on the theme of traffic flow forecasting. Bayesian networks can fully take into account the causal relationship between random variables statistically, and thus can be employed to model and analyze traffic flows among upstream and downstream road links. From the standpoint of Bayesian networks, the value of an object node can be inferred by its neighbor nodes. Furthermore, forecasting in case of incomplete data is also possible based on the message passing mechanism of Bayesian networks. In this article, we focus on using Bayesian networks to carry out traffic flow modeling and forecasting. Experiments with encouraging results show that this approach is applicable and considerably effective for traffic flow modeling and forecasting both for complete data and incomplete data.

The remainder of the article is organized as follows. After introducing Bayesian networks and some related issues in Section II, we respectively describe the model construction mechanism and experiment results for complete data and incomplete data forecasting in Sections III and IV. Finally, Section V concludes the paper and discusses some directions of future work.

## II. Bayesian Networks

In a transportation network, information from other road links would be helpful to forecast traffic flow at the current link. However, it is very hard to directly describe the influence of traffic flows at all the other links to the traffic flow at the current one, since there would be too many variables to be determined in order to access this relationship. Although there exist direct or indirect relations among different road links, one usually assumes that, given the traffic flows (including the flows at the current and previous intervals) at adjacent links, the traffic flows at the other links are independent of the traffic flow at the current one. In this way, the relation among the studied road links could be simplified. The methodology of Bayesian networks also accords with this postulation of conditional independence. The advantage of this assumption is that the scale of the prediction model can be reduced by cutting down the number of cause nodes in a Bayesian network. As a result, given limited training data, we can estimate the joint probability distribution among all nodes more accurately with a smaller network, and thus get a more precise representation of prediction relations.

A Bayesian network, also known as a causal model, is a directed graphical model for representing conditional independencies between a set of random variables. It is a marriage between probability theory and graph theory, and provides a natural tool for dealing with two problems that occur through applied mathematics and engineering—uncertainty and complexity [15]. In a Bayesian network, an arc from node A to B can be informally interpreted as indicating that A "causes" B [16]. The simplest statement of conditional independence relationships encoded in a Bayesian network can be stated as follows: a node is independent of its ancestors given its parents, where the ancestor/parent relationship is with respect to some fixed topological ordering of the nodes [16]. Therefore, for a Bayesian network consisting of $n$ nodes (random variables) $(x_1, x_2, \ldots, x_n)$, we have the representation for the joint probability distribution

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_{P_i}) \tag{1}$$

where $p(x_i|x_{P_i})$ is the local conditional probability distribution associated with node $i$ and $P_i$ is the set of indices labeling the parents of node $i$ ($P_i$ can be empty if node $i$ has no parents) [15]. The conditional independence relationship allows us to represent the joint probability distribution more compactly and conveniently, especially for large-scale networks. This would both benefit parameter estimation and variable forecasting when used to practical problems. In addition, since our traffic flow forecasting is a problem related to vehicle flows of time series, the Bayesian network model should consider the time factor of traffic flows as well. The intuition that some vehicle flows can cause other vehicle flows in the future temporally and in the downstream spatially indicates the design of Bayesian networks for traffic flows: Directed arcs should flow forward both in time direction and in flow direction.

### A. Representation and Parameter Estimation of Joint Probability Distribution

Before performing traffic flow forecasting in a Bayesian network, one should first derive the joint probability distribution between input and output. In this paper, we adopt Gaussian mixture model (GMM), a weighted combination of several normal distribution functions, to approximate the joint probability distribution in Bayesian networks. The benefits of GMM involve at least three aspects: 1) many events or phenomena in the natural world *per se* obey Gaussian distributions; 2) the Gaussian function has its convenience in mathematical deduction, which can be seen below; 3) one can approximate an arbitrary probability distribution with the combination of a sufficient number of Gaussian distributions. To formulate, let $x$ denote a random variable or a multidimensional random vector, and then the GMM representation of its probability distribution with $M$ mixture components be described as

$$p(x|\Theta) = \sum_{l=1}^{M} \alpha_l p_l(x|\theta_l) \tag{2}$$

where the parameters are $\Theta = (\alpha_1, \ldots, \alpha_M, \theta_1, \ldots, \theta_M)$ and $M$, s.t. $\sum_{l=1}^{M} \alpha_l = 1$ [17]. Each $p_l(\cdot)$ is a Gaussian probability density function parameterized by $\theta_l = (\mu_l, \Sigma_l), l = 1, \ldots, M$.

Usually, maximum likelihood estimation (MLE) can be used to implement parameter estimation when given a training data

set. However, the analytical solutions of the parameters in GMM cannot be obtained by traditional MLE methods. The expectation maximization (EM) algorithm is an iterative method to carry out MLE [18]. It makes up for the disadvantage of traditional MLE methods that usually need to derive the analytical expressions of solutions. The applications of the EM algorithm mainly include two scenarios. The first occurs when the data indeed have hidden parameters, due to problems such as limitations of the observation process. The other occurs when optimizing the likelihood function is analytically intractable, but can be simplified by assuming the existence of values for additional but hidden parameters [18]. The latter one is more common in practice.

Suppose $X$ be the observed data set generated by some distribution. We assume that a complete data set $Z = (X, Y)$ exists and the joint distribution of the corresponding random variable/vector $z = (x, y)$ has the form

$$p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)p(x|\Theta) \tag{3}$$

where $Y$ is the hidden data set and $\Theta$ is the parameter set governing the distribution $p(z|\Theta)$. Via the E-step and M-step of EM algorithm, our objective becomes to seek

$$\Theta^{(i)} = \arg\max_{\Theta} Q\left(\Theta, \Theta^{(i-1)}\right) \tag{4}$$

where $Q(\Theta, \Theta^{(i-1)}) = E_y[\log p(X, Y|\Theta)|X, \Theta^{(i-1)}]$ [18]. EM algorithm is guaranteed to converge at local maximums of the corresponding likelihood function. The iterative equations to obtain the estimate of the new parameters in (2) in terms of the old parameters are given as

$$\begin{aligned}
\alpha_l^{\text{new}} &= \frac{1}{N}\sum_{i=1}^{N} p\left(l|x_i, \Theta^{(i-1)}\right) \\
\mu_l^{\text{new}} &= \frac{\sum_{i=1}^{N} x_i p\left(l|x_i, \Theta^{(i-1)}\right)}{\sum_{i=1}^{N} p\left(l|x_i, \Theta^{(i-1)}\right)} \\
\Sigma_l^{\text{new}} &= \frac{\sum_{i=1}^{N} p\left(l|x_i, \Theta^{(i-1)}\right)\left(x_i - \mu_l^{\text{new}}\right)\left(x_i - \mu_l^{\text{new}}\right)^{\top}}{\sum_{i=1}^{N} p\left(l|x_i, \Theta^{(i-1)}\right)}
\end{aligned} \tag{5}$$

at $l = 1, \ldots, M$, where $N$ is the size of data set $X$ [18].

However, in the EM algorithm, $M$ is a predetermined parameter and the algorithm may converge to a local maximum or the boundary of the parameter space. To find a global maximum is more desirable in most cases. As an extension of the basic EM algorithm, the recently proposed competitive EM (CEM) algorithm overcomes these drawbacks [19]. The CEM algorithm, which includes (5) as a subroutine, is capable of automatically choosing the number of mixing components $M$ and selecting the "split" or "merge" operations efficiently based on some competitive mechanism. Another good characteristic is that it is insensitive to the initial configurations of the number of the mixture components and model parameters. Considering these virtues, in this article, the parameters of a GMM that describe the joint probability distribution of the cause nodes and effect nodes in a Bayesian network are estimated through the CEM algorithm.

### B. Prediction Formulation for GMM

In our work, traffic flow forecasting is regarded as an inference problem in a Bayesian network. The main goal of inference in Bayesian networks is to estimate the values of target nodes given the values of the observed nodes. As can be seen below, the prediction formulation for the GMM is very concise, which is a desirable property for applications.

Let $(E, F)$ be a partitioning of the node indices of a Bayesian network into disjoint subsets and $(x_E, x_F)$ be a partitioning of the corresponding random variables/vectors. Then, the marginal probability of $x_E$ can be formulated as

$$p(x_E) = \sum_{x_F} p(x_E, x_F). \tag{6}$$

Thus, according to Bayesian theory [20], the conditional probability $p(x_F|x_E)$ is equal to

$$p(x_F|x_E) = \frac{p(x_F, x_E)}{p(x_E)} = \frac{p(x_F, x_E)}{\sum_{x_F} p(x_E, x_F)} \tag{7}$$

which can be readily computed for any $x_F$ once the denominator is computed by a marginalization computation. For traffic flow forecasting, we can use observation $x_E$ to forecast $x_F$. Under the rule of minimum mean square error (mmse), the optimal estimation of $x_F$ can be given as [21]

$$\hat{x}_F = E(x_F|x_E). \tag{8}$$

To deduce the specific representation of the optimal forecasting $\hat{x}_F$ under the GMM framework, we first introduce the following lemma given in [22].

*Lemma 1:* Let $G(x; \mu, \Sigma)$ denote a multidimensional normal density function with mean $\mu$ and covariance matrix $\Sigma$. If we rewrite them as $x^{\top} = (x_1^{\top}, x_2^{\top})$, $\mu^{\top} = (\mu_1^{\top}, \mu_2^{\top})$, and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, then $p(x)$ can be described as

$$p(x) = G(x_1; \mu_1, \Sigma_{11})G(x_2; \mu_{x_2|x_1}, \Sigma_{x_2|x_1}) \tag{9}$$

where

$$\begin{aligned}
\mu_{x_2|x_1} &= \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}(\mu_1 - x_1) \\
\Sigma_{x_2|x_1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.
\end{aligned} \tag{10}$$

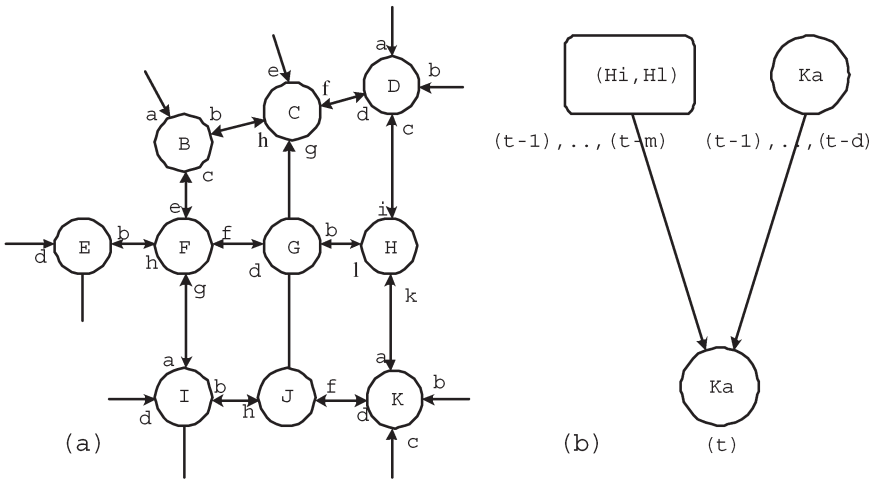


Fig. 1. (a) Patch taken from the East Section of the Third Circle of Beijing City Map where UTC/SCOOT systems are placed. For convenience, roads and flows are coded with English characters. (b) Bayesian network between the current link and its adjacent links for road link $K_a$.

If we define $x^\top = (x_F^\top, x_E^\top)$, $\mu_l^\top = (\mu_{lF}^\top, \mu_{lE}^\top)$, and $\Sigma_l = \begin{pmatrix} \Sigma_{lFF} & \Sigma_{lFE} \\ \Sigma_{lEF} & \Sigma_{lEE} \end{pmatrix}$ with the GMM framework and the above lemma, (2) can be rewritten as

$$\begin{aligned} p(x_F, x_E) &= \sum_{l=1}^{M} \alpha_l G(x; \mu_l, \Sigma_l) \\ &= \sum_{l=1}^{M} \alpha_l G(x_E; \mu_{lE}, \Sigma_{lEE}) G(x_F; \mu_{lF|E}, \Sigma_{lF|E}). \end{aligned} \tag{11}$$

According to (7), the conditional probability density function $p(x_F|x_E)$ can be represented as

$$p(x_F|x_E) = \sum_{l=1}^{M} \beta_l G(x_F; \mu_{lF|E}, \Sigma_{lF|E}) \tag{12}$$

where

$$\begin{aligned} \beta_l &= \frac{\alpha_l G(x_E; \mu_{lE}, \Sigma_{lEE})}{\sum_{j=1}^{M} \alpha_j G(x_E; \mu_{jE}, \Sigma_{jEE})} \\ \mu_{lF|E} &= \mu_{lF} - \Sigma_{lFE}\Sigma_{lEE}^{-1}(\mu_{lE} - x_E) \\ \Sigma_{lF|E} &= \Sigma_{lFF} - \Sigma_{lFE}\Sigma_{lEE}^{-1}\Sigma_{lEF}. \end{aligned} \tag{13}$$

Thus, optimal forecasting $\hat{x}_F$ under the criterion of mmse can be derived as

$$\begin{aligned} \hat{x}_F &= E(x_F|x_E) \\ &= \int x_F p(x_F|x_E)\, dx_F \\ &= \sum_{l=1}^{M} \beta_l \int x_F G(x_F; \mu_{lF|E}, \Sigma_{lF|E})\, dx_F \\ &= \sum_{l=1}^{M} \beta_l \mu_{lF|E} \end{aligned} \tag{14}$$

where $\beta_l$ and $\mu_{lF|E}$ have the same meanings as above. We can see that, with the GMM formulation, the forecasting relationship between the input and the output in a Bayesian network is rather concise, which is very convenient for practical applications.

## III. Model Construction and Experiments for Complete Data

### A. Modeling Mechanism and Flow Chart

Fig. 1(a) shows a patch taken from one urban traffic map of highways. Each circle node in the sketch map denotes a road link. An arrow shows the direction of traffic flow, which reaches the corresponding road link from its upstream link. Paths without arrows are of no traffic flow records. We take vehicle flow data $K_a$ as an instance to show our modeling mechanism. $K_a$ represents the vehicle flow from upstream link $H$ to downstream link $K$. From the view point of Bayesian networks, vehicle flows of $H_i$ and $H_l$ should have causal relations with vehicle flow of $K_a$. Furthermore, considering the time factor, to predict the vehicle flow of $K_a$ at time $t$ [denoted by $K_a(t)$], we should use values $K_a(t-1), K_a(t-2), \ldots, K_a(t-d)$ as well, since these values imply some trend of $K_a(t)$. That is, the historical values of $H_i$, $H_l$, and $K_a$ should all be regarded as the cause nodes of $K_a(t)$ in a Bayesian network. From the aspect of forecasting, the cause nodes serve as the input and the effect node $K_a(t)$ serves as the output. The constructed Bayesian network model is given in Fig. 1(b).

The flow chart of the forecasting procedure can be concluded as follows. 1) Construct a Bayesian network model between the input (cause nodes) and the output (effect node) for a chosen traffic flow on a given road link. 2) Approximate the joint probability distribution of all nodes in the constructed network by GMM using the CEM algorithm explained in Section II-A. 3) Carry out the optimal estimation of traffic flow of the current link in the form of (14).

Usually, the dimension of the joint probability distribution using Bayesian networks is high, and the available data are comparatively not enough, since there are many nodes

to describe. Therefore, this might decrease the accuracy of the sequent parameter estimation. In Section III-B, we will illustrate the relationship of the size of training data and the forecasting error on the same test data for given road links. However, if we carry out forecasting in a data space of much lower dimension, parameter estimation might be more accurate and efficient. Principal component analysis (PCA) or Karhunen–Loeve transform is such an effective tool for linear dimension reduction [23]. When using PCA for dimension reduction, we select some representative components from the input nodes and then estimate the joint probability distribution among these components and the output node. Based on the new relationship, we can carry out traffic flow forecasting more efficiently. We take road $B_b$ as an example to illustrate the PCA procedure for our problem. Assume the original dimension of the joint distribution of traffic flows of $B_b$, $C_e$, $C_f$, $C_g$ with their historical values is 20 when forecasting $B_b(t)$. Thus, the dimension of the input space is 19. We carry our PCA for the 19 input nodes with training data and select a few principal components (e.g., 5, 6, or 7) to represent the input space. Consequently, we can, respectively, reduce the input data to these dimensions and implement forecasting with the training data. From these results, the reduced dimension with the best accuracy could be identified.

### B. Experiments

Simply stated, the problem addressed in this section is to forecast future traffic flow rates at given road links from the historical data of themselves and their neighbor links. The field data for analysis are the vehicle flow rates of discrete time series recorded every 15 min along many road links by the UTC/SCOOT system in the Traffic Management Bureau of Beijing, whose unit is vehicles per hour (veh/h). From the real urban traffic map, we select a representative patch to verify the proposed approach, which is given in Fig. 1(a). The raw data are from March 1 to March 31, 2002, totaling 31 days. Considering the malfunction of detector or transmitter, we screened the days with empty data in view of evaluation. The remaining data for use are of 25 days and totaling 2400 sample points.

To evaluate the performance of our approach, the idea of cross validation is utilized to conduct experiments. Please refer to [20] for the justifications of cross validation in approach evaluations. In doing cross validation, we divided all the samples into two parts. One part serves as a training set, which is utilized to estimate parameters of GMM, and the other serves as a test (validation) set to verify forecasting performance. Every time, we randomly select 2112 samples and treat them as training data. The rest are used for test data. For every road link, the average accuracy across ten times of cross validation is regarded as the accuracy of the corresponding method.

In our experiments, the forecasting orders [parameters $d$ and $m$ as shown in Fig. 1(b)] from the current link and from adjacent links are respectively taken as 4 and 5 ($d = 4$ and $m = 5$), for this configuration could provide enough information for traffic flow forecasting empirically. Then for Fig. 1(b), the joint probability distribution of the Bayesian network is $p(H_i(t-j), H_l(t-j), K_a(t-j+1), j = 1, \ldots, 5)$. Using

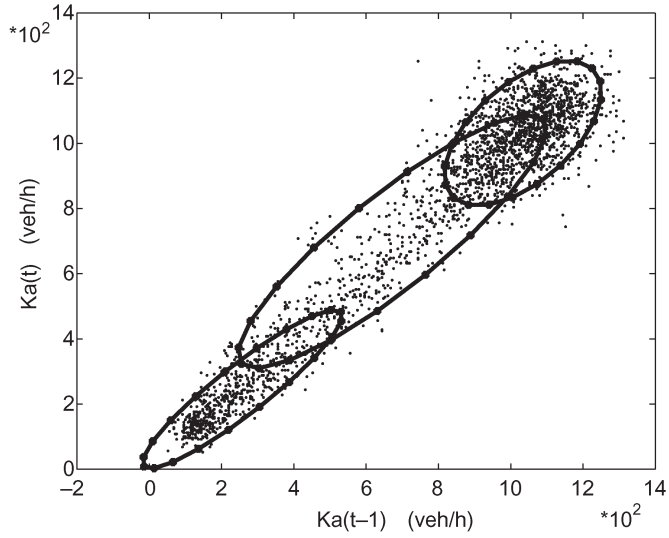


Fig. 2. Training data and estimated GMM for road link $K_a$.

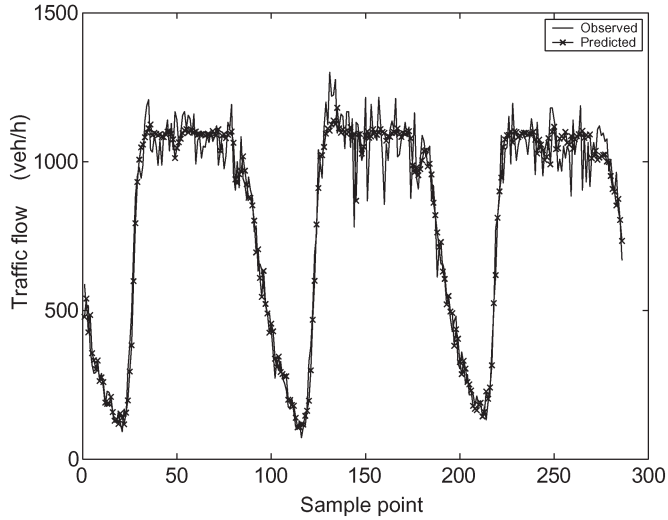


Fig. 3. Traffic flow forecasting result for the last 3 days' data of road link $K_a$.

the CEM algorithm, we can obtain the GMM formulation of this distribution. Fig. 2 shows an observation of this distribution with two dimensions, $K_a(t)$ and $K_a(t-1)$. In Fig. 2, one point corresponds to one sample in the training data set. Each ellipse corresponds to one estimated Gaussian component with the center and shape denoting the estimated parameters. From this figure, we can see that the component number of GMM is selected as three automatically and that the configuration of each ellipse is very regular. This should attribute to the appealing characteristics of the CEM algorithm. It steers clear of the problem of choosing the component number when using the EM algorithm and gets a globally optimal realization of the corresponding joint probability distribution. Sequentially, we can get the optimal estimation of traffic flow of the current link in the form of (14). Fig. 3 gives the final forecasting result of the Bayesian network approach (without PCA for dimension reduction) for the last 3 days' data of link $K_a$.

RW is a classical method for traffic flow forecasting whose core idea is to forecast the current value using the last value [1]. In this paper, RW is adopted as one of the basis lines for

TABLE I
ERROR COMPARISON OF FIVE METHODS FOR SHORT-TERM TRAFFIC FLOW FORECASTING WITH COMPLETE DATA

| | RW | FNM | AR | BN (without PCA) | BN (with PCA) |
|---|---|---|---|---|---|
| $B_b$ | 83.04 | 96.42 | 74.41 | *74.15* | *72.87* |
| $B_c$ | 113.56 | 124.61 | 98.63 | *94.18* | *91.68* |
| $C_f$ | 97.51 | 126.70 | 89.49 | *88.10* | *84.67* |
| $C_h$ | 74.96 | 66.89 | 60.85 | *55.34* | *55.43* |
| $D_c$ | 84.17 | 89.16 | 72.95 | *65.24* | *67.70* |
| $D_d$ | 67.44 | 65.86 | 60.15 | *54.77* | *53.91* |
| $E_b$ | 155.86 | 149.66 | 130.74 | *124.07* | *112.64* |
| $F_e$ | 144.51 | 158.20 | 119.45 | *114.15* | *110.41* |
| $G_b$ | 89.08 | 126.17 | 73.14 | *69.26* | *66.66* |
| $G_d$ | 172.72 | 185.99 | 150.84 | *140.57* | *135.83* |
| $H_i$ | 95.95 | 107.20 | 80.46 | *76.70* | *72.47* |
| $H_k$ | 127.48 | 186.27 | 102.98 | 105.14 | 112.25 |
| $I_a$ | 81.59 | 107.72 | 75.36 | *74.52* | 77.84 |
| $J_f$ | 141.05 | 147.37 | 116.80 | *115.34* | *111.96* |
| $K_a$ | 91.87 | 74.51 | 77.74 | *71.07* | *66.30* |
| SUM | 1620.8 | 1812.7 | 1384.0 | *1322.6* | *1295.6* |

comparisons. Besides, the FNM proposed recently is an effective approach for traffic flow forecasting that utilizes upstream flows in the current time interval to forecast the selected downstream flow in the next interval [12]. It has shown great superiority to the traditional neural network model. In this paper, we also adopt this model for offline analysis as a basis line to carry out comparisons among different approaches. For the parameters involved in FNM, we use the same configuration as the authors suggested in their paper (number of clusters is 10 and training rate is 0.1), since the parameters were found to be enough to produce good data fitting for general urban traffic flow forecasting problems through sensitivity analysis. Experiments with FNM for our forecasting problem also validate that this parameter configuration is among the best through searching from a small set of parameters. In addition, the autoregressive (AR) model, which only uses historical flow rates of the current link to forecast, is also employed as a comparative method. For the convenience of comparison, the order parameter $d$ in the AR model is taken as 4, and the other parameters are obtained through the GMM and CEM algorithm. The second to the fifth columns of Table I give the forecasting results of all road links available through the RW method, FNM approach, AR model, and Bayesian network model (BN without PCA), respectively, with performances evaluated by root mean square error (RMSE). In the same row in Table I, the smaller RMSE corresponds to the better accuracy. Let $\hat{y}$ be the estimate of $N$-dimensional vector $y$, the performance measure RMSE can be expressed as

$$\text{RMSE}(y,\hat{y}) = \left[\frac{1}{N}\sum_{n=1}^{N}\left(y(n)-\hat{y}(n)\right)^2\right]^{\frac{1}{2}}. \qquad (15)$$

The size of training set is an important factor in the accuracy of parameter estimation and thus of the forecasting result. Usually, the larger the training data set is, the better the estimated joint probability distribution will be and the better the forecasting result. Theoretically, however, there is no relationship about the size of training set and the forecasting error that can be adopted here. We have no alternative but to explore the relationship experimentally. We conducted experiments with varying size of training data for parameter estimation and then obtain the corresponding forecasting error on the same test data. Fig. 4 gives the kind of relationship for road links $B_b$ and $K_a$, respectively. For other road links, we obtained quite similar curves. From Fig. 4, we can see the fact with increasing number of training data that the forecasting performance would be improved. The decreasing tendency of the two curves also shows that the training data are a little inadequate for parameter estimation of GMM. Therefore, PCA would be helpful for dimension reduction in order to obtain better forecasting results. As a byproduct, PCA can also reduce the time of parameter estimation and traffic flow forecasting.

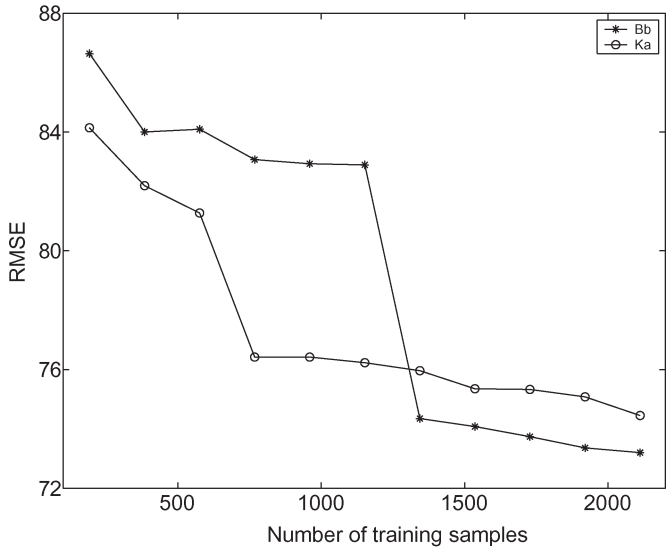


Fig. 4. Size of training data and forecasting error for road links $B_b$ and $K_a$ based on the same test data, respectively.

However, how to choose the reduced dimension number when using PCA is a problem that should be paid attention to. Residual variance is often used to evaluate the fits of PCA and the reduced dimensions. For the definition and usage of residual variance, please refer to [24]. Fig. 5 shows the residual variance of PCA with different dimensions for input nodes of road link $B_b$ on the training data. We see that using dimension 19 for forecasting is quite redundant and we have not focused on the few essential dimensions. To effectively use data and gain good performance, we resort to look for the "elbow" where the curve ceases to decrease significantly with added dimensions. Sequentially, we employ PCA to reduce the input data dimension 19 to these numbers (shown in the rectangular), respectively, and then carry out sequent forecasting. Experiments on the training data show that the best forecasting result is obtained at input data space dimension 6 for road link $B_b$. The RMSE drops from 74.15 to 72.87. For other road links, we also carry out dimension selection by PCA before the parameter estimation of GMM. The final forecasting performances are listed in the last column of Table I. From the results, we can see that incorporating information from both the current link and adjacent links to forecast (Bayesian network method) outperforms using the current link only (AR method). When

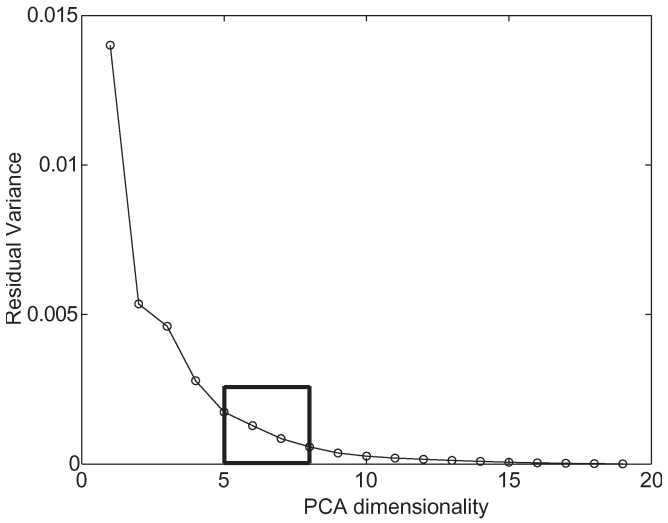


Fig. 5. Residual variance of PCA with different dimensions for input data of road link $B_b$.

using PCA for dimension reduction, the total forecasting error is 1295.6, which is much better than the AR method with forecasting error 1384.0 (13 out of 15 outputs are better than the AR method). Even without the dimension reduction procedure, the forecasting performance is still better than the AR method with RMSE 1322.6 versus 1384.0 (14 out of 15 outputs are better than the AR method). The superior performance should be attributed to Bayesian network utilizing all related information available. In addition, from Table I, one can see that the Bayesian network approach is much better than the RW method and the FNM approach, and even the AR method is better than them. So in the following parts, we use the AR method to compare with our Bayesian network approach.

To quantitatively evaluate our approach, we carry out paired $t$ test between the experimental results of the AR method and the Bayesian network methods. Through paired $t$ test between the AR method and the Bayesian network method without PCA, significant differences are found with $p$-value less than 0.005. Likewise, the $p$-value between the AR method and the Bayesian network method with PCA is obtained and is less than 0.005. These results manifest that the performance of our Bayesian network approach is significantly different from that of the AR method. Combining the comparisons in the last paragraph, we can draw the conclusion that the Bayesian network approach is greatly superior to the AR method.

## IV. Model Construction and Experiments for Incomplete Data

In Section III, we applied the Bayesian network approach to carry out traffic flow modeling and forecasting in case of complete data. However, due to practical limitations, traffic flows recorded can often be incomplete, i.e., partially missing or unavailable. Developing an approach to work in case of incomplete data would be much more important and practical. In this section, we consider the situation where an incomplete data exist. The following parts would demonstrate that a new Bayesian network could be constructed to adapt to this scenario through a slight modification of the former Bayesian network approach.

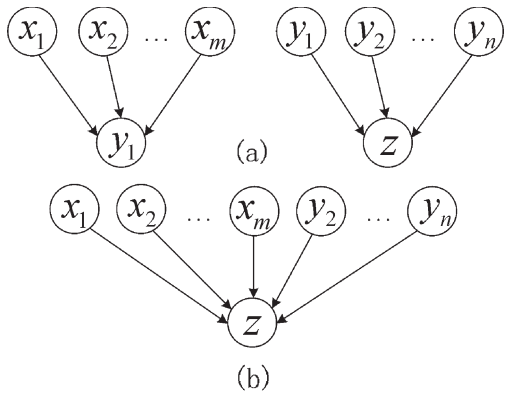


Fig. 6. (a) Two Bayesian networks. (b) Expanded Bayesian network.

### A. Modeling Mechanism and Flow Chart

Suppose we have several random variables denoted by $x_1, x_2, \ldots, x_m$, $y_1, y_2, \ldots, y_n$, and $z$, respectively. $x_1, x_2, \ldots, x_m$ are used to forecast $y_1$ and $y_1, y_2, \ldots, y_n$ are used to forecast $z$ in turn. Then, considering the causal relations in variable forecasting, we can construct two Bayesian networks as shown in Fig. 6(a), where arrows start from the cause nodes and point to the effect nodes. If the data for random variable $y_1$ are missing whereas the data for $x_1, x_2, \ldots, x_m, y_2, \ldots, y_n$ are complete (intact), then how can we forecast $z$? We can construct another Bayesian network to model the new causal relationship, as is given in Fig. 6(b). In the graph, $x_1, x_2, \ldots, x_m$, $y_2$, $\ldots, y_n$ all together serve as the cause nodes of $z$. Since the newly constructed Bayesian network often has more nodes than either of the original graphs, we call it expanded Bayesian network.

The flow chart of our forecasting procedure for incomplete data can be described as follows. 1) Construct an expanded Bayesian network between the input (cause nodes) and output (effect node) for a given road link. 2) Approximate the joint probability density function of all nodes in the expanded Bayesian network by PCA and GMM using the methods explained in Sections II and III. 3) Carry out the optimal estimation of flow rates of the object road link in the form of (14).

### B. Experiments

We take road link $D_d$ as an example to show our Bayesian network approach for incomplete traffic flow forecasting. From the view point of the Bayesian network, vehicle flows of $C_e$, $C_g$, and $C_h$ should have causal relations with vehicle flow of $D_d$. Similarly, vehicle flows of $B_a$ and $B_c$ should have causal relations with vehicle flow of $C_h$. Furthermore, considering the time factor to predict the vehicle flow of $D_d$ at time $t$ (denoted by $D_d(t)$), we should use data $D_d(t-1), \ldots, D_d(t-d)$ as well. That is, some historical values of $C_e$, $C_g$, $C_h$, and $D_d$ could be regarded as the cause nodes of $D_d(t)$ in a Bayesian network. Suppose traffic flow data $C_h(t-m)$ are missing, then we can use the expanded Bayesian network method to forecast

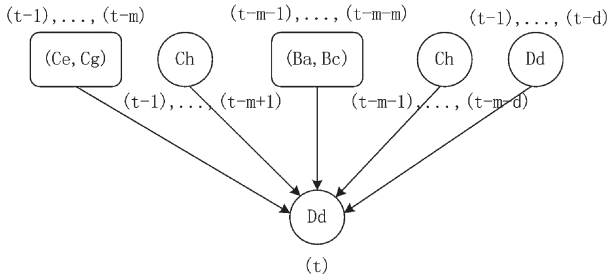


Fig. 7. Expanded Bayesian network for object road link $D_d$.

TABLE II
ERROR COMPARISON OF TWO METHODS FOR SHORT-TERM TRAFFIC FLOW FORECASTING WITH INCOMPLETE DATA

| | AR | Expanded Bayesian network |
|---|---|---|
| $D_d$ | 66.14 | *57.44* |
| $J_f$ | 123.65 | *110.88* |
| $G_d$ | 155.20 | *138.39* |
| $C_f$ | 90.76 | *86.31* |
| SUM | 435.75 | *393.02* |

$D_d(t)$. The expanded Bayesian network for forecasting $D_d(t)$ with missing data $C_h(t-m)$ is shown in Fig. 7.

With the same configuration of the parameters $d$ and $m$ as in Section III, the joint probability density function in Fig. 7 is $p(C_e(t-j), C_g(t-j), B_a(t-j-5), B_c(t-j-5),\ C_h(\cdot), D_d(t-j+1), j=1,\ldots,5)$, where $C_h(\cdot) = (C_h(t-l), C_h(t-l-5), l=1,\ldots,4)$. We can see that the dimension of the joint probability density function is very high (dimension $= 33$). So we use PCA to carry out dimension reduction before the parameter learning of GMM. For other road links, we also employ PCA for dimension reduction before the parameter learning of GMM. The final forecasting performances using the expanded Bayesian network approach evaluated by RMSE are listed in Table II. The forecasting results through the AR method are also reported. From the experimental results, we can find the outstanding improvements of forecasting performance brought by the Bayesian network. For each of the four road links analyzed, the performance of the Bayesian network outperforms the AR method significantly. Paired $t$ test between the results of the AR method and the expanded Bayesian network method also shows that there are significant differences between these two methods with $p$-value less than 0.05. Thus, the effectiveness of our Bayesian network approach for incomplete data forecasting is manifested. Considering the difference of dimensions of the joint probability distribution between these two methods, according to the statistical learning theory [25], we are sure that given enough data for training, forecasting with Bayesian networks would obtain even better results.

## V. DISCUSSIONS AND CONCLUSION

In this paper, we successfully introduce the conception and methodology of Bayesian networks from the statistical analysis, artificial intelligence, and machine learning fields to the community of ITS for the application of traffic flow forecasting. The essence of traffic flow is consistent with the ideology of Bayesian networks. In the traffic flow forecasting theme, as vehicles usually keep travelling from one road to its neighbor roads, we can take it for granted that the traffic volume of the object road link is the result of the flows of its upstream links and its own historical series. Therefore, to construct a Bayesian network for traffic flow of each road link at a given time is reasonable. Besides this intuitive causal relationship, another advantage of the Bayesian network approach is that we can collect the entire cause information to predict the traffic flow of the object link, even in case of incomplete data. Encouraging experimental results with real-world data also manifested the applicability of the Bayesian network approach for traffic flow forecasting.

It is true that traffic flow at the current link is not independent of all the upstream traffic flows. However, to solve a practical problem, some approximation and assumption are usually necessary. In this paper, we adopt the idea of conditional independence, that is, given the adjacent upstream traffic flows at different time delays, traffic flow at the current link is assumed to be independent of other upstream traffic flows. The merits of this assumption have already been exhibited by experimental results.

In the presence of incidents/accidents, our approach can still work in principle. Since it is based on pattern learning from training data, as long as given adequate data that account for these scenarios, the estimated GMM could represent these patterns. And the sequent forecasting procedures would follow similarly. However, to collect a large amount of data in the presence of incidents/accidents is usually difficult. When these situations occur, our approach might not be the optimal choice and some other techniques might be helpful, such as abnormality detection methods, etc.

Besides, concerning computational complexity, one might be suspicious of the applicability of the Bayesian network approach for online traffic flow forecasting. In fact, this should not be worried about at all because the time-consuming computation of parameter estimation could be done in advance and offline. We only need the estimated parameter values and the historical traffic flow data for online forecasting, which can be easily implemented in real time by an ordinary PC today. In addition, our approach can still be improved in the future by considering the seasonal or periodical effect (e.g., daily trends, weekly trends, and monthly trends) of traffic flows. The authors hope to report on such extensions in future publications.

## ACKNOWLEDGMENT

The authors would like to thank Dr. B. Zhang for his sincere help on CEM algorithm and related advises. The authors are also grateful to the anonymous editor and reviewers for giving valuable comments.

## REFERENCES

[1] B. M. William, "Modeling and forecasting vehicular traffic flow as a seasonal stochastic time series process," Ph.D. dissertation, Dept. Civil Eng., Univ. Virginia, Charlottesville, VA, 1999.

[2] C. K. Moorthy and B. G. Ratcliffe, "Short term traffic forecasting using time series methods," *Transp. Plan. Technol.*, vol. 12, no. 1, pp. 45–56, 1988.

[3] S. Lee and D. B. Fambro, "Application of subsets autoregressive integrated moving average model for short-term freeway traffic volume forecasting," *Transp. Res. Rec.*, no. 1678, pp. 179–188, 1999.
[4] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filter theory," *Transp. Res., Part B: Methodol.*, vol. 18 B, no. 1, pp. 1–11, Feb. 1984.
[5] J. Hall and P. Mars, "The limitations of artificial neural networks for traffic prediction," in *Proc. 3rd IEEE Symp. Computers and Communications*, Athens, Greece, 1998, pp. 8–12.
[6] G. A. Davis and N. L. Nihan, "Non-parametric regression and short-term freeway traffic forecasting," *J. Transp. Eng.*, vol. 177, no. 2, pp. 178–188, 1991.
[7] R. Chrobok, J. Wahle, and M. Schreckenberg, "Traffic forecast using simulations of large scale networks," in *Proc. 4th IEEE Int. Conf. Intelligent Transportation Systems*, Oakland, CA, 2001, pp. 434–439.
[8] G. A. Davis, "Adaptive forecasting of freeway traffic congestion," *Transp. Res. Rec.*, no. 1287, pp. 29–33, 1990.
[9] B. L. Smith and M. Demetsky, "Traffic flow forecasting: Comparison of modeling approaches," *J. Transp. Eng.*, vol. 123, no. 4, pp. 261–266, 1997.
[10] M. Danech-Pajouh and M. Aron, "ATHENA, a method for short-term inter-urban traffic forecasting," INRETS, Paris, France, Tech. Rep. 177, 1991.
[11] M. Der Voort, M. Dougherty, and S. Watson, "Combining Kohonen maps with ARIMA time series models to forecast traffic flow," *Transp. Res., Part C Emerg. Technol.*, vol. 4, no. 5, pp. 307–318, Oct. 1996.
[12] H. B. Yin, S. C. Wong, J. M. Xu, and C. K. Wong, "Urban traffic flow prediction using a fuzzy-neural approach," *Transp. Res., Part C Emerg. Technol.*, vol. 10, no. 2, pp. 85–98, Apr. 2002.
[13] G. Q. Yu, J. M. Hu, C. S. Zhang, L. K. Zhuang, and J. Y. Song, "Short-term traffic flow forecasting based on Markov chain model," in *Proc. IEEE Intelligent Vehicles Symp.*, Columbus, OH, 2003, pp. 208–212.
[14] S. C. Chang, R. S. Kim, S. J. Kim, and B. H. Ahn, "Traffic-flow forecasting using a 3-stage model," in *Proc. IEEE Intelligent Vehicle Symp.*, Dearborn, MI, 2000, pp. 451–456.
[15] M. I. Jordan, *Learning in Graphical Models*. Cambridge, MA: MIT Press, 1999.
[16] F. V. Jensen, *Bayesian Networks and Decision Graphs*. New York: Springer-Verlag, 2001.
[17] G. Mclachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.
[18] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," Int. Comput. Sci. Inst., Berkeley, CA, Tech. Rep. 94704, 1998.
[19] B. B. Zhang, C. S. Zhang, and X. Yi, "Competitive EM algorithm for finite mixture models," *Pattern Recognit.*, vol. 37, no. 1, pp. 131–144, 2004.
[20] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
[21] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic, 1970.
[22] C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd ed. New York: Wiley, 1973.
[23] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.
[24] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
[25] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

**Shiliang Sun** was born in Shandong, China, in 1979. He received the B.E. degree in automatic control from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2002 and is currently working toward the Ph.D. degree at the State Key Laboratory of Intelligent Technology and Systems, Department of Automation, Tsinghua University, Beijing.

His research interests include machine learning, pattern recognition, signal processing, and time series analysis.

**Changshui Zhang** (M'01) received the B.S. degree in mathematics from Beijing University, Beijing, China, in 1986 and the Ph.D. degree in automation from Tsinghua University, Beijing, in 1992.

Since July 1992, he has been working as a Teacher at the Department of Automation, Tsinghua University. He is currently a Professor at the Department of Automation, Tsinghua University. His research interests include machine learning, pattern recognition, artificial intelligence, image processing, evolutionary computation, etc.

**Guoqiang Yu** was born in Shandong, China, on 1979. He received the B.S. degree from the Department of Electronic Engineering, Shandong University, Jinan, Shandong, China, in 2001 and the M.E. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2004.

He is currently a Researcher at the Department of Application Science, NuTech Company Limited, Beijing, with focus on the development and research in the algorithms of image processing and pattern recognition.