PEOPLE MONITORING USING FACE RECOGNITION WITH OBSERVATION CONSTRAINTS

Ji Tao and Yap-Peng Tan

School of Electrical & Electronic Engineering
Nanyang Technological University, Singapore

ABSTRACT

We propose a people monitoring system to recognize people by probabilistic inference methods exploiting low-level facial feature and high-level domain knowledge. In particular, the faces of people in the view of a monitoring camera are first detected and modeled. Optimal recognition of people leaving and entering a closed-room is accomplished by exploiting temporal correlation and constraints among the observed face sequence. The optimality is achieved in the sense of maximizing a joint posterior probability of multiple observations. Experimental results of real and synthetic data suggest the efficacy of the proposed system.

1. INTRODUCTION

With the increased concern for physical security in the face of global terrorism and outbreaks of infectious viruses, automated video surveillance has received unprecedented attention over the past few years. One main task of a surveillance system is to associate each person with an identity or to correspond a same person observed at different time instances. The results of the system allow the derivation of such useful information as how long a person has stayed in the room, how many people are in the room during a certain period, and who they are. Potential applications thus include, for example, understanding the human activities in a monitored work place, keeping aware of the user identities in an intelligent room, and even identifying who could possibly be infected by a newly identified Severe Acute Respiratory Syndrome (SARS) victim.

A number of possible solutions to the problems of concern have been proposed in the literature. For example, biometrics have been successfully used for person recognition and satisfactory results have been achieved. Traditional biometric-based recognition methods range from fingerprint identification to iris/retina scan [1]. However, these methods involve intrusive data collection, i.e., requiring human proactive action and collaboration in the course of identification or authentication, and thus work mainly in well-controlled environments. Face recognition has also received extensive attention from researchers for it can address to some extent the above drawback while still maintaining a rather high recognition rate, reaching 90% for the best cases according to the study of Fromherz *et al.* [2]. However, this technique, in general, may suffer difficulties from changes in face orientation, camera view angle, and illumination condition, as well as different facial expressions and wears, and thus is often used to provide auxiliary information in many applications. Many approaches have been proposed to improve the robustness of face recognition by various means of feature modeling, which can be categorized into geometric feature-based methods, template-based methods, and more recently HMM-based methods. These methods are designed to recognize each observed face from an existing database based on some rule of maximum likelihood, and mainly rely on features observed at a single time instance/duration.

In comparison, we propose in this paper a probabilistic framework to combat this problem in the context of closed-room monitoring. The system consists of two modules: a feature extraction module to detect/model the faces of people entering or leaving the only entrance/exit of a closed room (e.g., a lab, class room, or meeting room) for the purpose of recognition in an unintrusive manner, and a people recognition module to correspond each observed person with a person previously entering the room or to identify him/her as a new person unseen before. Rather than using only a single face observation, we perform recognition by exploiting the temporal correlation and constraints among multiple observations acquired at different time instances. Consequently, it can effectively remedy the problems caused by low-resolution and/or arbitrary pose of faces extracted from real-time surveillance video. Experimental results demonstrate that the proposed approach can obtain promising recognition rates, as compared with the existing maximum likelihood approach.

2. FEATURE EXTRACTION

Having reviewed many existing methods, we choose to modify and combine two functions provided in Intel Open Source Computer Vision Library (OpenCV) [3], HarrFaceDetection and HMMFaceRecognition, to establish a people monitoring system that is capable of automatically detecting, extracting and modeling human faces from video sequences.

The face detector was initially proposed by Viola [4] and improved by Lienhart [5], and it works effectively by employing the Ada-boost algorithm. The embedded HMM

0-7803-8554-3/04/$20.00

(EHMM) face recognizer was developed by Nefian *et al.* [6] to exploit the natural structure of frontal faces, showing outstanding performance. With a number of face images of a same person, we can train his/her EHMM using a set of observation vectors obtained from the corresponding 2D-DCT coefficients. The likelihood of an unknown face observation $\boldsymbol{O}$ with respect to an established face model Λ can be calculated by the doubly embedded Viterbi algorithm, and a likeness measure $P[\boldsymbol{O} \sim \Lambda]$ can then be obtained. The system interface is shown in Fig. 1 with some sample faces extracted from a test video sequence.

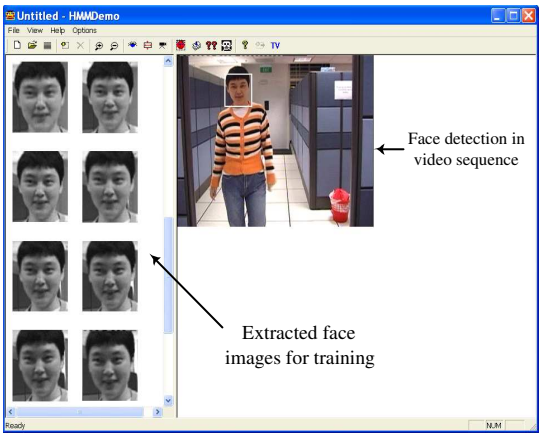


Fig. 1. Interface for face detection and modeling.

It should be noted that the people monitoring framework to be proposed below can employ many human features/attributes for which a likelihood measure is defined. To make the system more accurate and less intrusive, we have only made use of facial feature in the work reported in this paper.

3. PROBABILISTIC REASONING FRAMEWORK

3.1. Problem formulation

Since hidden Markov models (HMMs) offer a simple graphical way to perform probabilistic inference by combining prior knowledge and incomplete observation data [7], our first attempt to the problem is to establish an HMM for its recognition task. In general, an HMM can be fully characterized by a parameter set $\lambda = \{A, B, \pi\}$, including the probabilities to measure state transition distribution (A), observation symbol distribution (B), and initial state distribution (π), which are usually pre-learned and defined based on a fixed number of states. Unfortunately, this does not apply to our case, in which the number of people as well as their behavior patterns vary from place to place and are generally hard to learn from prior data.

To construct a framework well suited for the problem of our concern, we make use of the lattice structure and parameter definition of HMMs and formulate the problem of people recognition as follows. Assume that the room is empty when the system is initially activated. Each time a person is entering, we append a new state (identity) S_i to the state set (database); a new observation O_t is recorded when a person leaves at time t. Either a state or an observation is represented by the facial feature of the corresponding person. At time t, the state set consists of the current identities in the database (i.e., people that possibly remain in the room at time t), denoted as $\mathbf{S}(t) = \{S_1 \cdots S_{N_t}\}$. After observing a number of people leaving, an observation sequence $O = \{O_1 \cdots O_T\}$ can be obtained. Once given the time-variant model $\lambda(t)$, we can use the *Viterbi* algorithm to find the optimal state sequence $Q = \{q_1 \cdots q_T\}$ associated with the observation sequence by maximizing a joint posterior probability $P(Q, O|\lambda(t))$, so that each leaving person can be recognized from those who entered the closed-room before.

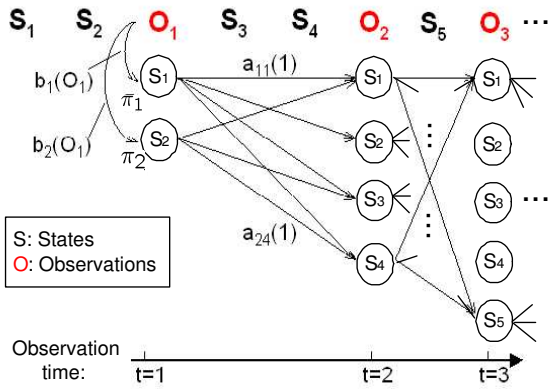


Fig. 2. Illustration of how the proposed probabilistic reasoning framework is developed.

An example of the proposed framework is shown in Fig. 2. It can be seen that the framework has a lattice structure similar to HMMs; however, there exist several key differences that distinguish our framework from HMMs. First, the time-variant parameter set $\lambda(t)$ is derived at each observation time instance based on the current observations and previous states rather than some training data. Second, the number of states in our model is not stationary but increasing over time before a decision is made. Third, the states are non-deterministic because more than one states could represent the same person.

3.2. Framework construction

The critical problem in constructing the proposed framework is to estimate the time-variant parameter set $\lambda(t)$ at each observation time, with the knowledge of which one can easily find the optimal path by using the *Viterbi* algorithm. Our solutions are given as follows.

• *Initial state distribution π:*

Since N_1 people entered before the first exit and, without any other prior knowledge, intuitively each of them has an equal probability to leave, we have

$$\pi_i = \frac{1}{N_1},\ 1 \le i \le N_1. \tag{1}$$

• *Output probability of state i at time t, $b_i(O_t)$:*

This probability represents how likely O_t is from S_i. It is simply estimated by the likeness measure of observation O_t with respect to the face model S_i as defined in Sec. 2 as

$$b_i(O_t) = P[O_t \sim S_i],\ 1 \le i \le N_t. \tag{2}$$

• *State transition probability $a_{ij}(t)$:*

The probability $a_{ij}(t)$ measures the odds for person S_j to leave the room at time $t+1$ given that S_i leaves at time t. To compute it, we first define a set of probabilities within time pair $\{t, t+1\}$ as shown in Fig. 3. $P[S_{i,t^+} = 1]$ is the probability of S_i remaining in the room at time t^+, and $t^{+/-}$ denotes the time instance right after/before the observation O_t is made. The likelihood matrix M characterizes the similarities between people entering from O_t to O_{t+1} (new appended states) and people observed so far (existing states at time t), defined by their likeness measures as

$$\boldsymbol{M} = \{m_{uv} | m_{uv} = P[S_{N_t+v} \sim S_u]\}. \tag{3}$$

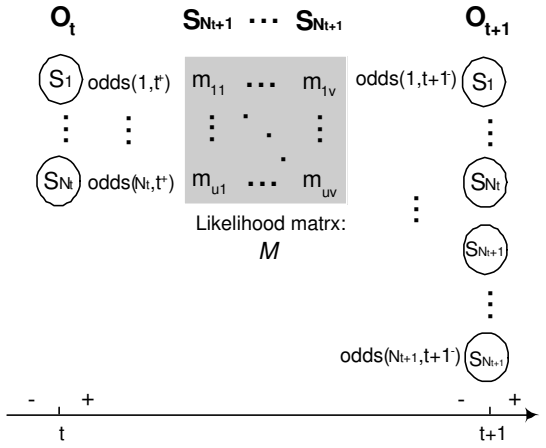


Fig. 3. The likelihood matrix M and the people existence odds at times t^+ and $t+1^-$, where $odds(i, t^+) = P[S_{i,t^+} = 1]$, $odds(j, t+1^-) = P[S_{j,t+1^-} = 1]$.

Lacking other prior knowledge, we assume that the probability of one's exit is proportional to that of his/her existence, e.g., if a person does not reside in the room at all, he/she could not leave the room. Therefore the transition probability can be computed as

$$a_{ij}(t) = \frac{P[S_{j,t+1^-} = 1 | q_t = S_i]}{\sum_j P[S_{j,t+1^-} = 1 | q_t = S_i]}, \tag{4}$$

where the denominator is a normalization factor. By assuming the status of each person is independent of the others, the numerator in Eq. (4) can be expanded as

$$\begin{aligned} &P[S_{j,t+1^-} = 1 | q_t = S_i] = \\ &\textstyle\sum_{\text{all } cond} P[S_{j,t+1^-} = 1 | cond] \cdot P[cond | q_t = S_i], \end{aligned} \tag{5}$$

where $P[cond|q_t = S_i] = \prod_{i=1}^{N_t} P[S_{i,t^+} = \theta_i | q_t = S_i]$. $cond = \{S_{1,t^+} = \theta_1 \cdots S_{N_t,t^+} = \theta_{N_t}\}$ represents one of the possible realizations of $\{S_{1,t^+} \cdots S_{N_t,t^+}\}$ over all $\boldsymbol{\theta}$, and $\theta_i = 1$ *or* 0 designates that the *status* of person S_i is in or out of the room.

To estimate the probability $P[S_{j,t+1^-} = 1 | cond]$, we take into consideration the people entering between t and $t+1$ and update the likelihood matrix to incorporate the domain knowledge imposed by a specific *cond* as

$$\widetilde{m}_{uv} = \eta[\rho(u)\varepsilon(u)m_{uv}]. \tag{6}$$

$\varepsilon(u)$ is equal to zero if $S_{u,t^+} = 1$ in the *cond* and equal to one otherwise, reflecting the knowledge that one could not enter the room if he/she is already inside. Let $PT(S_{\hat{i}}, t)$ be the partial best path ending in $S_{\hat{i}}$ at time t (retrieved by the tracing back array ψ in the *Viterbi* algorithm). Then $\rho(u)$ is set to zero if $S_u \notin PT(S_{\hat{i}}, t)$ and to one otherwise, accounting for the fact that one could not enter again if he/she has not left the room. The normalization operation η is in place to ensure that the summations of the likelihoods corresponding to all possible situations are equal to one. Viewing $\widetilde{m}_{uv}$ as the probability of S_{N_t+v} to be S_u returning, we can derive the conditional probability in Eq. (5) as

$$\begin{aligned} &P[S_{j,t+1^-} = 1 | cond] = \\ &\begin{cases} P[S_{j,t^+} = 1 | cond] + \sum_v \widetilde{m}_{jv} & 1 \le j \le N_t \\ 1 - \sum_u \widetilde{m}_{u(j-N_t)} & \text{otherwise.} \end{cases} \end{aligned} \tag{7}$$

If we also know the probability $P[S_{i,t^+} = \theta | q_t = S_i]$, we can calculate $a_{ij}(t)$ based on Eqs. (4), (5) and (7). To estimate this unknown, we initially set $P[S_{i,1^+} = 1 | q_1 = S_{\hat{i}}] = 1$ for all $i \neq \hat{i}$ and $P[S_{\hat{i},1^+} = 1 | q_1 = S_{\hat{i}}] = 0$, since all the people enter before $t = 1$ should be inside at time 1^+ except for the one who just left. For recursion, we need to estimate the probability of $P[S_{j,t+1^+} = 1 | q_{t+1} = S_{\hat{j}}]$ when proceeding to the next observation time. Again, we use the array ψ to retrieve the previous state of $S_{\hat{j}}$, say $S_{\hat{i}}$, and thus the probabilities of $P[S_{i,t^+} = 1 | q_t = S_{\hat{i}}]$ and $P[S_{j,t+1^-} = 1 | q_t = S_{\hat{i}}]$. Accordingly, we can obtain

$$\begin{aligned} &P[S_{j,t+1^+} = 1 | q_{t+1} = S_{\hat{j}}] = \\ &\begin{cases} 0 & j = \hat{j} \\ P[S_{j,t+1^-} = 1 | q_t = S_{\hat{i}}] - \varphi\gamma(j) & \text{otherwise,} \end{cases} \end{aligned} \tag{8}$$

where $\gamma(j) = P[S_{j,t+1^-} = 1 | q_t = S_{\hat{i}}] - P[S_{j,t^+} = 1 | q_t = S_{\hat{i}}]$ and $\varphi = (1 - P[S_{k',t+1^-} = 1 | q_t = S_{\hat{i}}]) / \sum_{j \neq \hat{j}} \gamma(j)$. Furthermore, the summation of the existence odds of all states at any time instance $t^{+/-}$ (i.e., $\sum_i P[\mathbf{s}_{i,t^{+/-}} = 1 | q_t = \mathbf{s}_j]$) for any t and j is always equal to the number of people who really reside in the room at that time.

To this end, we have estimated the time variant parameter set $\lambda(t)$ of the proposed framework. Batch recognition of people exiting (observations) can be performed at

any time by selecting the path with the highest score, i.e., maximum joint posterior probability. In addition, people re-entering the room can be identified through a local maximum likelihood scheme: if a state S_i appears twice in the selected path at time t_1 and t_2, respectively, the state $S_{\hat{j}} = \arg\max_{S_j \in \mathbf{S}(t_2)/\mathbf{S}(t_1)} P[S_j \sim S_i]$ can be viewed as S_i re-entering and merged with S_i.

4. EXPERIMENTAL RESULTS

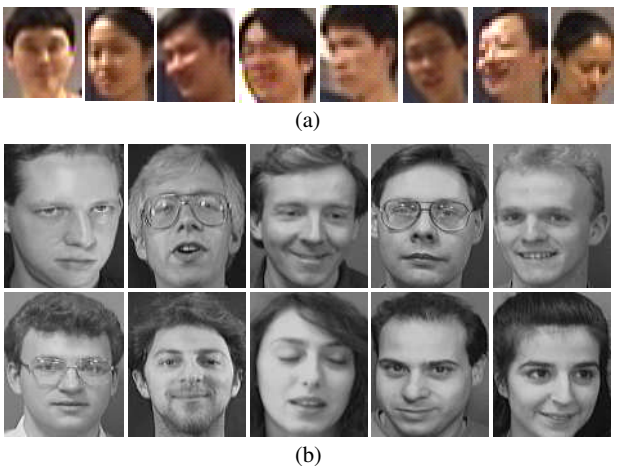
We tested the proposed monitoring system using two test sequences: one was a real video captured in a research laboratory by a dual-camera system monitoring the lab's only entrance, and the other one was synthesized from the Olivetti Research Ltd. database (400 images of 40 individuals, 10 images per individual at the resolution of 92×112 pixels). The real video recorded eight different people, each entering and leaving the lab for three times. Ten face images were extracted for each person to train the face model when he/she was observed for the first time. For the synthetic data, we only used two images of each individual for training to limit the recognition rate since they were acquired in relatively high resolution and well-controlled pose/illumination, a condition that is difficult to come by in practice. Fig. 4 presents some sample face images from the two test sequences.

To simulate the real entry-exit observation sequences over time, a synthetic process generator was used to randomly rearrange faces in the database to represent people entering/leaving the room based on the rule that one cannot enter unless he/she is outside the room, and vice versa. In our experiments, every 40 images of 10 different individuals (4 images per person) were grouped, yielding a process with 20 entries and 20 exits (each individual entered and left twice). Totally, we obtained eight such synthetic processes from this face database.

For comparison, we also implemented a recognition approach based on maximum likelihood classification [8]. Each person entering the lab was classified as either *re-entering* or *new* against a likelihood threshold T_m, while identities of people leaving were revealed by selecting the face model with the highest likelihood from the database. The ground truth was obtained manually. It should be noted that the results of the maximum likelihood approach were rather sensitive to the threshold T_m, inappropriate selection of which could cause false recognition of classifying a person re-entering as new or vice versa. In contrast, the proposed approach did not require any hard threshold for making decision, and obtains notably improved recognition rates, as shown in Table 1.

Table 1. Recognition rates comparison of the maximum likelihood (ML) approach and the proposed approach.

Data	ML approach	Proposed approach
Real Video	82.3%	98.5%
Synthetic Data	85.0%	100%



(a)

(b)

Fig. 4. (a) The sample face images of eight people extracted from the first test sequence. (b) Ten individuals in face database used to compose the synthetic sequence.

5. CONCLUSION

We have presented a novel probabilistic framework for closed-room people monitoring using face recognition. Rather than identifying each single face observation from a database, the framework is devised to recognize people based on multiple face observations by exploiting their temporal correlations. In addition, the proposed framework combines low-level facial features and domain-specific knowledge to estimate and update its parameters at each observation instance. Experimental results demonstrate that the proposed model outperforms the existing maximum-likelihood approach when using the same face model.

6. REFERENCES

[1] W. Shen, M. Surette, and R. Khanna, "Evaluation of automated biometrics-based identification and verification systems", *Proc. of the IEEE*, Vol. 85, 1997.

[2] T. Fromherz, P. Stucki, and M. Bichsel, "A survey of face recognition", *MML Technical Report*, No 97.01, Dept. of Computer Science, Univ. of Zurich, 1997.

[3] Open source computer vision library reference manual, *Intel Corporation*.

[4] P. Viola and M. Jones, "Robust real-time object detection", *Technical Report 2001/01, Compaq CRL*, Feb. 2001.

[5] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection", *IEEE ICIP*, 2002.

[6] A.V. Nefian and M.H. Hayes III. "An embedded HMM based approach for face detection and recognition", *IEEE ICASSP*, 1999.

[7] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. of the IEEE,* Vol. 77, 1989.

[8] J.M. Siskind and Q. Morris, "A maximum-likelihood approach to visual event classification", in *the Proc. of ECCV*, 1996.