FACE RECOGNITION BASED ON SEPARABLE LATTICE 2-D HMMS USING VARIATIONAL BAYESIAN METHOD

Kei Sawada, Akira Tamamori, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda (Nagoya Institute of Technology, Japan)

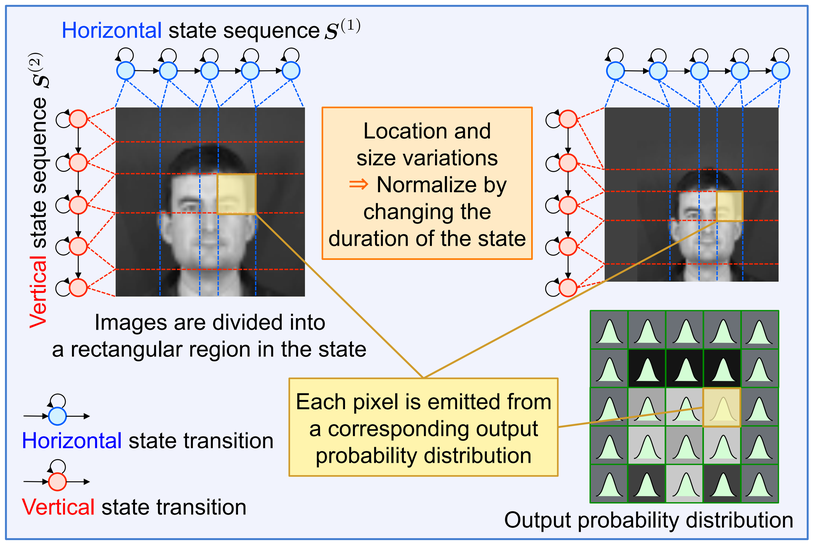
1. Introduction

- Image recognition based on statistical approaches
  - Eigen-image and subspace methods based on PCA
  - Heuristic normalization techniques for each task are required
- Separable lattice 2-D HMMs (SL2D-HMMs) [Kurata, et al.; '06]
  - Training and normalization are integrated
  - ML criterion produces point estimation of model parameters
  - ⇒ Estimation accuracy may be decreased due to the over-fitting
- Bayesian criterion
  - Use of prior distribution and marginalization of model parameters

Apply Bayesian criterion to separable lattice 2-D HMMs

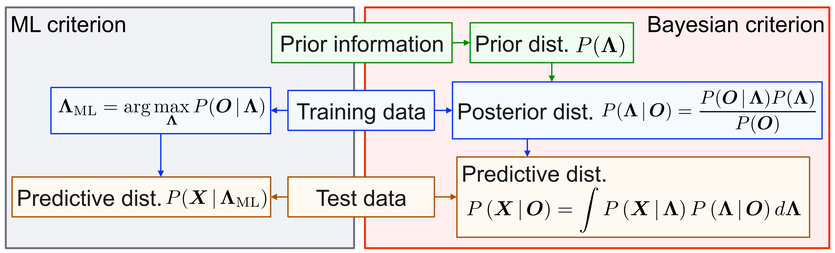
2. Separable lattice 2-D HMMs

- Separable lattice 2-D hidden Markov models
  - SL2D-HMMs with horizontal and vertical Markov chains
  - ⇒ An elastic matching in both horizontal and vertical directions



3. Bayesian criterion

- Maximum likelihood (ML) criterion
  - ML criterion produces point estimation ⇒ Over-fitting problem
- Bayesian criterion
  - Use of prior distribution and marginalization of model parameters
  - Complex integral and expectation calculations
  - ⇒ Effective approximation techniques are required



Λ : Set of model parameters O : Training data X : Test data

4. Separable lattice 2-D HMMs using variational Bayesian method

- Maximum a posteriori (MAP) method [Gauvain, et al.; '94]
  - Estimation of model parameters by maximizing posterior probability

$$\Lambda_{MAP} = \arg\max_{\Lambda} P(O|\Lambda)P(\Lambda)$$

  - Use of prior distribution
  - Over-fitting problem because of point estimates
- Variational Bayesian (VB) method [Attias; '99]
  - Estimation of approximated posterior distribution
  - Define a low bound of log marginal likelihood

$$\ln P(O) = \ln \sum_{S} \int P(O,S|\Lambda)P(\Lambda)d\Lambda \geq \sum_{S} \int Q(S,\Lambda) \ln \frac{P(O,S|\Lambda)P(\Lambda)}{Q(S,\Lambda)} d\Lambda = \mathcal{F}(Q)$$

Jensen's inequality

S : State sequence
$Q(S,\Lambda)$: Arbitrary dist.

  - Relation between the log marginal likelihood and the lower bound

$$\mathcal{F} = \ln P(O) - \mathrm{KL}(Q(S,\Lambda) \,||\, P(S,\Lambda|O)) \Rightarrow P(S,\Lambda|O) \approx Q(S,\Lambda)$$

  - Assume that random variables are conditionally independent

$$Q(S,\Lambda) = Q(S)Q(\Lambda) = Q(S^{(1)})Q(S^{(2)})Q(\Lambda)$$

$Q(\cdot)$: Variational posterior dist.

  - Estimation of posterior distribution based on maximizing $\mathcal{F}$

VB E-step	$Q'(S^{(1)}) = \arg\max_{Q(S^{(1)})} \mathcal{F}$ $Q'(S^{(2)}) = \arg\max_{Q(S^{(2)})} \mathcal{F}$
VB M-step	$Q'(\Lambda) = \arg\max_{Q(\Lambda)} \mathcal{F}$

Alternately update

  - Derive variational posterior distribution

$$Q(S^{(1)}) \propto \exp\left[\sum_{S^{(2)}} \int Q(S^{(2)})Q(\Lambda) \ln P(O,S^{(1)},S^{(2)}|\Lambda)d\Lambda\right]$$

$$Q(S^{(2)}) \propto \exp\left[\sum_{S^{(1)}} \int Q(S^{(1)})Q(\Lambda) \ln P(O,S^{(1)},S^{(2)}|\Lambda)d\Lambda\right]$$

$$Q(\Lambda) \propto P(\Lambda) \exp\left[\sum_{S^{(1)}} \sum_{S^{(2)}} Q(S^{(1)})Q(S^{(2)}) \ln P(O,S^{(1)},S^{(2)}|\Lambda)\right]$$

  - Use of prior distribution and marginalization of model parameters
- Prior distribution
  - Conjugate prior distribution
    - Posterior dist. belongs to the same dist. family as the prior dist.

Initial state probability	Dirichlet distribution
State transition probability	Dirichlet distribution
Output probability distribution	Gauss-Wishart distribution

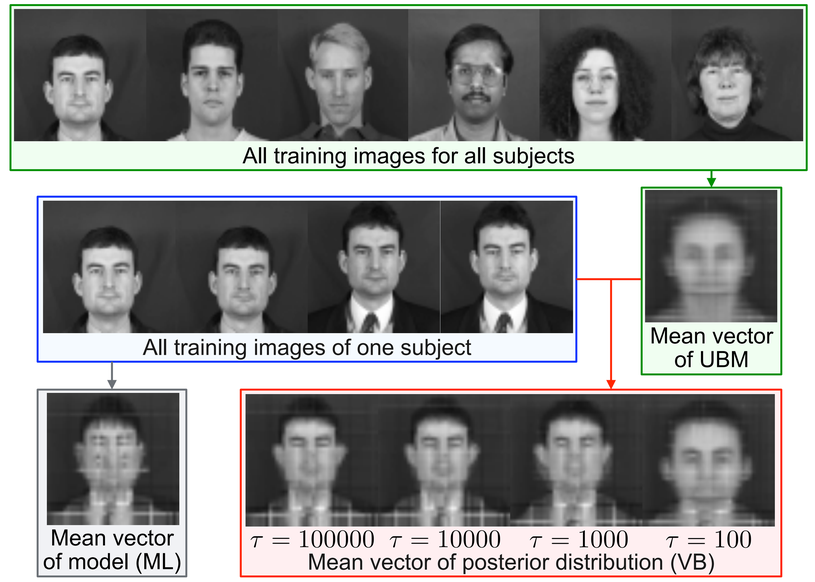
  - Universal background model (UBM)
    - UBM is trained from all training data for all subjects
    - ⇒ UBM roughly represents a training data
  - Tuning parameter τ
    - Representation of the reliability of the UBM
    - τ is small ⇒ Prior distribution has a larger impact on posterior distribution
    - τ is large ⇒ Prior distribution has a smaller impact on posterior distribution

5. Experiments

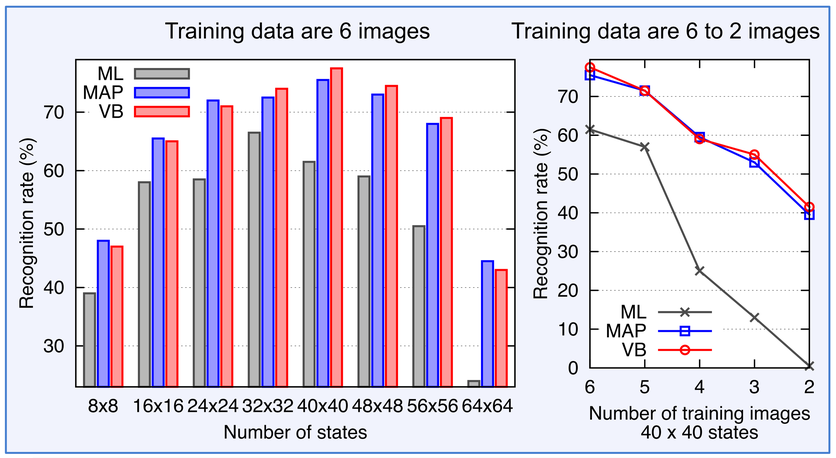
- Experimental conditions

Database	XM2VTS
Image size	64×64, grayscale
Training data	6, 5, 4, 3, 2 images per person × 100 subjects
Test data	2 images per person × 100 subjects
Number of states	8×8, 16×16, 24×24, 32×32, 40×40, 48×48, 56×56, 64×64
Tuning parameter τ	50, 100, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, 100000

- Examples of training images and mean vectors



- Results



ML : ML criterion (conventional) MAP and VB : Bayesian criterion (proposed)

- Bayesian criterion achieved significantly higher recognition rates than ML criterion
- The difference between ML criterion and Bayesian criterion became larger when small numbers of training images were used
- The use of a prior distribution was more effective than the marginalization of model parameters