MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan
3-20

Multivariate Sparse Bayesian Regression and Its Application for Facial Feature Detection

Yoshio Iwai[†] and Roberto Cipolla[‡]
[†]Graduate School of Engineering Science, Osaka University
[‡]Department of Engineering, University of Cambridge

Abstract

The processing of facial images has received considerable attention by computer vision researchers because of the broad range of potential applications for systems that are able to encode and interpret facial images. Especially, reliable facial feature detection and tracking in an image sequence are still challenging problems. In this paper, we propose an extension of the RVM (relevance vector machine) for multivariate Bayesian regression and its application for automatically locating facial features after initial training.

1 Introduction

The processing of facial images has received considerable attention by computer vision researchers because of the broad range of potential applications for systems that are able to encode and interpret facial images. Some applications, such as those reviewed in [1], include personal identification and access control, video phone and teleconferencing, forensic applications, human-computer interaction, and automated surveillance.

Reliable facial feature detection and tracking in an image sequence are still challenging problems. Much work has been done on extracting facial features using methods such as template matching[2, 3], edge detection[4], and deformable models[5]. It is important to set the initial positions near the correct feature points.

To estimate initial positions of facial features, another approach exists that maps an input image to initial positions of facial features. The advantage of this approach is fast computation. Neural networks and linear regression can be used for such a purpose[6]. In the case of using a neural network, it requires to determine the number of nodes in a hidden layer, and to iterate optimization for learning. On the other hand, linear regression does not require any optimization step because the optimal solution can be solved analytically, but its approximation accuracy is worse than that of neural networks. The kernel-based linear regression[7], therefore, has been proposed for addressing this problem, but it arose another problem that many training data must be stored for calculation of kernel functions that correspond the hidden nodes of a neural network.

Tipping had proposed the RVM (relevance vector machine) for obtaining sparse solutions to regression tasks[8]. Sparseness of solution addresses the memory problem, and we can estimate facial feature positions directly from an image after training. If the target function is, however, multivariate, the performance of the RVM is worse because the RVM does not consider the multivariate regression case. The multivariate Bayesian regression is powerful for modeling various target functions. In this paper, we propose an extension of RVM to multivariate Bayesian regression and its application for automatically locating facial features.

2 Multivariate Relevance Vector Machine

2.1 Kernel-Based Multivariate Linear Regression

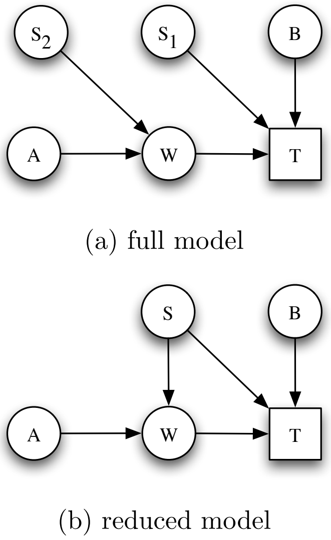
The training examples are a set of input - output pair $\{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^N$, where $\mathbf{x}_n = (x_{1n}, x_{2n}, \cdots, x_{pn})^T$ and $\mathbf{t}_n = (t_{1n}, t_{2n}, \cdots, t_{qn})^T$. Here, p and q are the dimensions of the feature vector and the target function, respectively. Our aim is to obtain a vector function of the form:

$$\mathbf{y}\left(\mathbf{x}; W\right) = \phi(\mathbf{x}) W^T, \tag{1}$$

where W is the q-by-(N+1) weight matrix and $\phi(\mathbf{x})$ is a N+1 dimensional vector of non-linear basis functions. We are only interested in the data centric basis functions of the form: $\phi(\mathbf{x}) = (1, K(\mathbf{x}_1, \mathbf{x}), K(\mathbf{x}_2, \mathbf{x}), \cdots, K(\mathbf{x}_N, \mathbf{x}))^T$.

Let $T = (\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_n)^T \in R^{N\times q}$ and $\Phi(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) = (\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \cdots, \phi(\mathbf{x}_n))^T \in R^{N\times(N+1)}$. Assuming a Gaussian noise to the model we obtain

$$T = \Phi(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) W^T + E, \tag{2}$$



(a) full model

(b) reduced model

Figure 1: Probabilistic models

with a mean-zero Gaussian noise process,$E \in R^{N\times q}$, whose covariance matrix is $B^{-1}\otimes S \in R^{N\times N}\otimes R^{q\times q}$ where $\otimes$ is the Kronecker product. S is a covariance matrix of parameters and B is a covariance matrix of sampling data. We assume $B = \mathrm{diag}(\beta_i) \in R^{N\times N}$ because each sampling is performed independently.

The joint probability of all parameters is given by

$$p(W,S,T,A,B) = p(T|W,S,B)p(W|S,A)p(S)p(A)p(B), \tag{3}$$

where $A = \mathrm{diag}(\alpha_i) \in R^{N\times N}$.

We assume the covariance matrices of T and W are the same, i.e. S, for simplicity of calculation. Of course, more complicated case can be considered that the covariance matrices are different like $p(T|W,S_1,B)p(W|A,S_2)p(S_1)p(S_2)p(A)p(B)$. Figure 1 shows probabilistic models of each case.

2.2 Inference

From equation (2),

$$T|W,S,B \sim N(\Phi W^T, B^{-1}\otimes S), \tag{4}$$

and we obtain

$$\begin{aligned} p(T|W,S,B) &\propto \\ &|S|^{-\frac{N}{2}}|B^{-1}|^{-\frac{q}{2}}e^{-\frac{1}{2}trS^{-1}(T-\Phi W^T)^T B(T-\Phi W^T)}. \end{aligned} \tag{5}$$

The prior, $p(W|A,S)$, can be chosen arbitrary, and then we choose

$$W|A,S \sim N(0, A^{-1}\otimes S), \tag{6}$$

and we get

$$p(W|A,S) \propto |S|^{-\frac{N}{2}}|A^{-1}|^{-\frac{q}{2}}e^{-\frac{1}{2}trS^{-1}WAW^T}. \tag{7}$$

The joint posterior distribution of the weight matrix, W, and covariance matrix, S, can be rewritten as follows:

$$\begin{aligned} p(W,S|T,A,B) &\propto p(T|W,S,B)p(W|A,S) \\ &\propto |S|^{-\frac{2N}{2}}|A^{-1}|^{-\frac{q}{2}}|B^{-1}|^{-\frac{q}{2}} \\ &\quad \times e^{-\frac{1}{2}trS^{-1}[(T-\Phi W^T)^T B(T-\Phi W^T)+WAW^T]} \\ &\propto |S|^{-\frac{2N}{2}}|A^{-1}|^{-\frac{q}{2}}|B^{-1}|^{-\frac{q}{2}} \\ &\quad \times e^{-\frac{1}{2}trS^{-1}[(W-\hat{W})\Sigma^{-1}(W-\hat{W})^T+G]} \end{aligned} \tag{8}$$

where

$$\begin{aligned} \hat{W}^T &= (\Phi^T B\Phi + A)^{-1}\Phi^T BT = \Sigma\Phi^T BT \\ \Sigma &= (\Phi^T B\Phi + A)^{-1} \\ G &= (T-\Phi\hat{W}^T)^T B(T-\Phi\hat{W}^T) + \hat{W}A\hat{W}^T. \end{aligned}$$

To find the marginal posterior distribution of the matrix, W, the joint posterior distribution are integrated with respect to S. This can be performed easily by recognizing that the posterior distribution is exactly of the same form as an Inverse Wishart distribution except for a proportionality constant.

$$\begin{aligned} p(W|T,A,B) &= \int p(W,S|T,A,B)dS \\ &\propto \frac{1}{|G + (W-\hat{W})\Sigma^{-1}(W-\hat{W})^T|^{\frac{N-q-1}{2}}}. \end{aligned} \tag{9}$$

Therefore, $W = \hat{W}$ maximizes $p(W|T,A,B)$.

2.3 Optimizing the Hyperparameters

We derive with the marginal likelihood

$$\begin{aligned} p(T|S,A,B) &= \int p(T|W,S,B)p(W|S,A)dW \\ &\propto |S|^{-\frac{N}{2}}|A^{-1}|^{-\frac{q}{2}}|B^{-1}|^{-\frac{q}{2}}|\Sigma|^{\frac{q}{2}} \\ &\quad \times e^{-\frac{1}{2}trS^{-1}[(T-\Phi\hat{W})^T B(T-\Phi\hat{W})+\hat{W}A\hat{W}^T]} \end{aligned} \tag{10}$$

The marginal likelihood, $\log p(T|S,A,B)$, must be maximized over the hyperparameters, A, B , and the noise covariance, S.

$$\frac{\partial}{\partial \log \alpha_i}\log p(T|S,A,B) = 0, \tag{11}$$

$$\frac{\partial}{\partial \log \beta_i}\log p(T|S,A,B) = 0, \tag{12}$$

$$\frac{\partial}{\partial S}\log p(T|S,A,B) = 0. \tag{13}$$

The derivatives are

$$\frac{\partial}{\partial \log \alpha_i}\log p(T|S,A,B) = \frac{q}{2} - \frac{1}{2}\left[q\Sigma_{ii} + (\hat{W}^T S^{-1}\hat{W})_{ii}\right]\alpha_i, \tag{14}$$

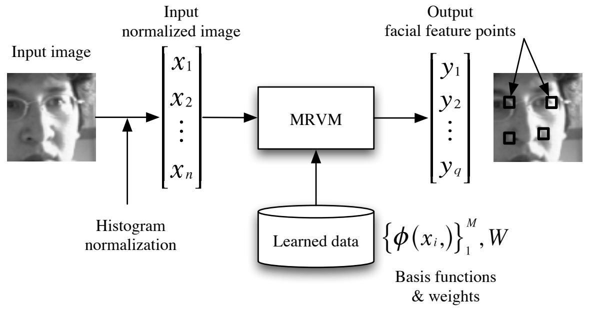


Figure 2: The process flow of facial feature detection

$$\frac{\partial}{\partial \log \beta_i} \log p(T|S, A, B) =$$
$$\frac{q}{2} - \frac{1}{2}\left[q(\Phi\Sigma\Phi^T)_{ii} + ((T - \Phi\hat{W}^T)S^{-1}(T - \Phi\hat{W}^T)^T)_{ii}\right]\beta_i,$$

and

$$\frac{\partial}{\partial S} \log p(T|S, A, B) =$$
$$-NS^{-1} + S^{-1}G^T S^{-1} - \frac{1}{2}\mathrm{diag}(-NS^{-1} + S^{-1}G^T S^{-1})$$

Setting this to zero and solving for them gives a re-estimation rule:

$$\alpha_i = \frac{q}{q\Sigma_{ii} + (\hat{W}^T S^{-1}\hat{W})_{ii}}, \tag{15}$$

$$\beta_i = \frac{q}{q(\Phi\Sigma\Phi^T)_{ii} + ((T - \Phi\hat{W}^T)S^{-1}(T - \Phi\hat{W}^T)^T)_{ii}} \tag{16}$$

and

$$S = \frac{G^T}{N} = \frac{G}{N}. \tag{17}$$

Following MacKay[9] in defining quantities $\gamma_i \equiv q(1 - \alpha_i\Sigma_{ii})$, leads to the following re-estimation rule:

$$\alpha_i = \frac{\gamma_i}{(\hat{W}^T S^{-1}\hat{W})_{ii}}, \tag{18}$$

which was observed to lead to much faster convergence although it does not guarantee the local maximization.

When $S = I_q$ and $B = \beta^{-1}I_N$, the proposed method is equal to the RVM proposed by Tipping. Therefore, the method extends the RVM naturally.

3 Application — Facial Feature Detection

In this section, we explain an application of MRVM (multivariate RVM) for facial feature detection. We assume that face regions are already segmented from an input image by an appropriate method such as boosted cascade detector[10]. The application outputs facial feature points (x_i, y_i) in the face region.

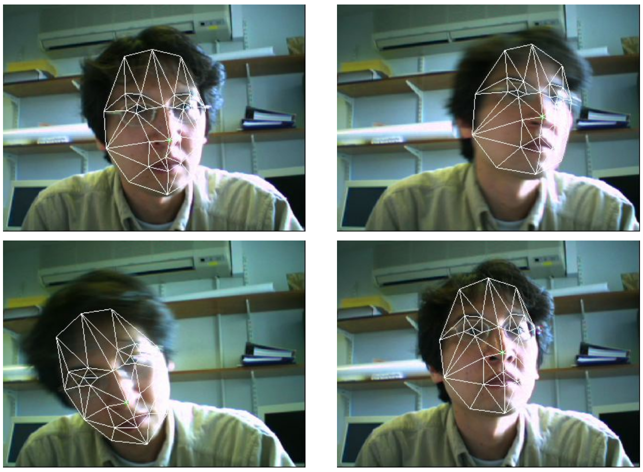


Figure 3: training data: images and facial feature points

The input face regions are normalized to fixed size determined in advance, and then pixel values of them are normalized to $[0, 1]$ by using the histogram normalization. The normalized input face region is used as an input, $\mathbf{x}$, of MRVM. By using a MRVM learned already, we can estimate facial feature positions, $\mathbf{y}$, from equation (1). The flow of these processes is illustrated in Fig. 2.

4 Experimental Results

We conduct an experiment to verify the accuracy of regression. We use an MPEG-2 image sequence (150 frames, 640 × 480 pixels, 8-bit gray scale) as training data. Facial feature points of training data are collected by using the flexible feature matching[11] and incorrect detection is adjusted manually. Figure 3 shows examples of training data.

4.1 Regression Performance

We conduct an experiment to verify the regression performance of the MRVM, and use the whole input image sequence as training data (400 data). Figure 4 shows examples of face regions used for the MRVM and RVM as inputs, and Fig. 5 shows examples of prediction results of nose position by the MRVM. White rectangles are training data used as true value, and black rectangles are prediction of nose position by the MRVM.

Regression results of the nose position show in Fig. 6. Thick lines are target values (nose position, x and y) and thin lines are regression values predicted by the MRVM and RVM. The horizontal axis is frame no, and the vertical axis is the posi-

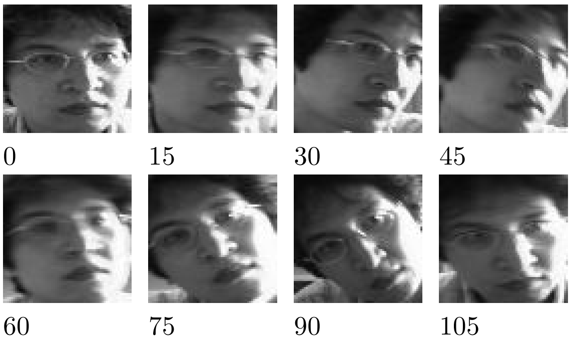


Figure 4: Examples of input face regions for MRVM and RVM: the numbers below figures shows frame no.

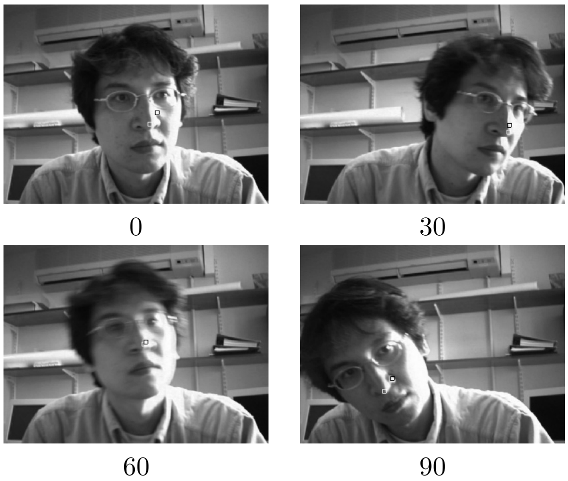


Figure 5: Prediction results of the nose position: white rectangles are true value, and black rectangles are prediction.

tion normalized by the image size. In this case, the numbers of relevance vectors of MRVM and RVM are 33 and 19+27=46, respectively. The training data are greatly reduced and the average regression errors of MRVM and RVM are 0.0300 and 0.0288, respectively. The regression accuracy of MRVM is slightly worse than that of RVM because the number of relevance vectors of MRVM is less than that of RVM.

5 Conclusion

In this paper, we proposed an extension of the RVM for multivariate Bayesian regression and its application for facial feature detection. We conducted experiments to verify the regression and prediction accuracy. In future work, we will conduct experiments to check the performance of the MRVM in detail.

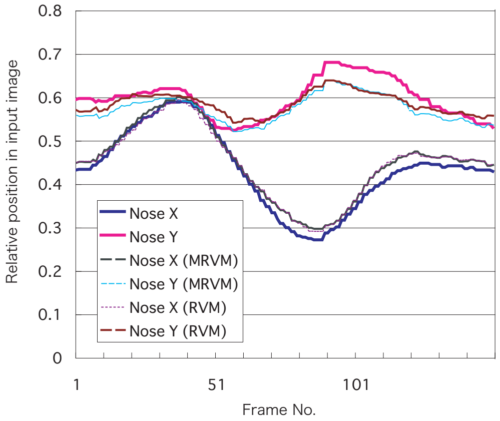


Figure 6: regression performance of nose position, (x, y)

References

[1] John Daugman. Face and gesture recognition: Overview. *IEEE Trans. on PAMI*, 19(7):675–676, July 1997.

[2] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. on PAMI*, 19(7):775–779, July 1997.

[3] H. Wu, Q. Chen, and M. Yachida. Face detection from color images by fuzzy pattern matching. *IEICE Trans. on Information and System*, J80-D-II(7):1774–1785, 1997.

[4] C. Kotropoulos, A. Tefas, and I. Pitas. Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions. *Pattern Recognition*, 33:1935–1947, 2000.

[5] D. Pramadihant, Y. Iwai, and M. Yachida. Integrated person identification and expression recognition from facial images. *IEICE Trans. on Inform. & Sys.*, E84-D(7):850–866, 2001.

[6] D.B. Rowe. *Multivariate Bayesian Statistics*. ChapMan & Hall/CRC, London, 2003.

[7] N. Cristianini and J. S.-Tayler. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. CUP, Cambridge, 2000.

[8] M.E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learing Research*, 1(211-244), 2001.

[9] D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[10] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings of Computer Vision and Pattern Recognition*, pp. 511–518, 2001.

[11] T. Hirayama, Y. Iwai, and M. Yachida. Face recognition based on efficient facial scale estimation. In *AMDO*, pp. 201–212, Palma de Mallorca, Spain, 2002.