# Supplementary Material for Paper ID 2012

November 8, 2013

This article describes an interpretation of our Gradient Vector $\mathcal{G}$ with respect to Fisher Vector, as well as the technical details for the EM algorithm of M-PCCA.

## 1 Gradient Vector $\mathcal{G}$ vs Fisher Vector

In this section, we show that, compared to Fisher Vector, our Gradient Vector representation encodes only information private to $x$ and $y$. Thus our representation, a concatenation of gradient vector $\mathcal{G}$ and latent vector $\mathcal{Z}$, incorporate less redundant information compared to Fisher Vector representation. Here we only discuss gradient vector with respect to $x$. Interpretation for $y$ can be similarly obtained. Note that $\{\Psi_x^k\}_{k=1}^K$ are all constrained to be diagonal. We reserve the derivatives in a matrix form only for the sake of notation convenience.

In Gaussian mixture model for $v = (x, y)$, the gradient of the model likelihood with respect to $x$ is

$$\frac{\partial E(\mathcal{L}_{Gauss})}{\partial \mu_x^k} = 2w_k(\Psi_x^k)^{-1}\left\{\mu_x^k - \frac{\sum_i \gamma_{i,k} x_i}{\sum_i \hat{\gamma}_{i,k}}\right\}, \tag{1}$$

$$\frac{\partial E(\mathcal{L}_{Gauss})}{\partial \Psi_x^k} = w_k(\Psi_x^k)^{-1}\left\{\Psi_x^k - \frac{\sum_i \gamma_{i,k}(x_i - \mu_x^k)(x_i - \mu_x^k)^\top}{\sum_i \hat{\gamma}_{i,k}}\right\}(\Psi_x^k)^{-1}. \tag{2}$$

The resulting Fisher Vector representation is

$$\left\{\frac{\partial E(\mathcal{L}_{Gauss})}{\partial \mu_x^k}, \frac{\partial E(\mathcal{L}_{Gauss})}{\partial \Psi_x^k}, \frac{\partial E(\mathcal{L}_{Gauss})}{\partial \mu_y^k}, \frac{\partial E(\mathcal{L}_{Gauss})}{\partial \Psi_y^k}\right\} \tag{3}$$

Compared with the formulas used in [2], the above representations are exactly the same except for a constant coefficient, which is cancelled out in our intra-normalization. Let $\widetilde{x}_{i,k}$ denote

$$\widetilde{x}_{i,k} = x_i - W_x^k z_{i,k}. \tag{4}$$

We rewrite the gradient vector $\mathcal{G}$ of $x$ for comparison.

$$\frac{\partial E(\mathcal{L}_{\mathcal{CCA}})}{\partial \mu_x^k} = 2w_k(\Psi_x^k)^{-1}\left\{\mu_x^k - \frac{\sum_i \gamma_{i,k} \widetilde{x}_{i,k}}{\sum_i \hat{\gamma}_{i,k}}\right\}, \tag{5}$$

$$\frac{\partial E(\mathcal{L}_{\mathcal{CCA}})}{\partial \Psi_x^k} = w_k(\Psi_x^k)^{-1}\left\{\widetilde{\Psi}_x^k - \frac{\sum_i \gamma_{i,k}(\widetilde{x}_{i,k} - \mu_x^k)(\widetilde{x}_{i,k} - \mu_x^k)^\top}{\sum_i \hat{\gamma}_{i,k}}\right\}(\Psi_x^k)^{-1}. \tag{6}$$

where $\widetilde{x}_{i,k} = x_i - W_x^k z_{i,k}$. This originates from our model and assumption on $x$, $y$ and $z$ as their shared information, where

$$x = W_x z + \mu_x + \epsilon_x, \tag{7}$$

$$y = W_y z + \mu_y + \epsilon_y. \tag{8}$$

For each submodel $k$, denote

$$\widetilde{x}_k = x - W_x^k z_k, \tag{9}$$

$$\widetilde{y}_k = y - W_y^k z_k. \tag{10}$$

$\widetilde{x}_k$ and $\widetilde{y}_k$ can be considered to be variables encoding information private to $x$ and $y$ individually in each submodel, with $\widetilde{x}_{i,k}$ and $\widetilde{y}_{i,k}$ being their samples. Var$(\widetilde{x}_k)$ and Var$(\widetilde{y}_k)$ are then $\widetilde{\Psi}_x^k$ and $\widetilde{\Psi}_y^k$, respectively. Our gradient vector representation is thus different from Fisher Vector only in that we subtract the shared part between $x$ and $y$ in every component, before aggregating descriptors. The resulting Fisher Vector representation is

$$\left\{ \frac{\partial E(\mathcal{L}_{CCA})}{\partial \mu_x^k}, \frac{\partial E(\mathcal{L}_{CCA})}{\partial \Psi_x^k}, \frac{\partial E(\mathcal{L}_{CCA})}{\partial \mu_y^k}, \frac{\partial E(\mathcal{L}_{CCA})}{\partial \Psi_y^k} \right\} \tag{11}$$

# 2 Derivation of the EM algorithm for M-PCCA

Assume we have a set of training data $D = \{v_i\}$, $v_i = (x_i, y_i)$. In this section, we discuss how to use EM algorithm to estimate parameters $\Theta = \{w_k, W_x^k, W_y^k, \mu_x^k, \mu_y^k, \Psi_x^k, \Psi_y^k\}$ for M-PCCA. The likelihood function we need to maximize is defined by

$$\mathcal{L}(\Theta; D) = \sum_i \log \left\{ \sum_k w_k p(v_i|k) \right\}. \tag{12}$$

M-PCCA model includes two types of latent variables, $Z = \{z_{i,k}\}$ and $\Gamma = \{\gamma_{i,k}\}$. In the **E-step**, we update $Z = \{z_{i,k}\}$ and $\Gamma = \{\gamma_{i,k}\}$ by calculating their posterior distributions given old M-PCCA model parameters $\Theta$. $\gamma_{i,k}$ has a 0-1 posterior distribution, where $\gamma_{i,k} = 1$ indicates that sample data $v_i$ is generated by the $k$-th submodel. Given an M-PCCA model, the expectation $\gamma_{i,k} = 1$ is given by,

$$\begin{aligned} \hat{\gamma}_{i,k} &= E(\gamma_{i,k}) \\ &= p(k|v_i). \end{aligned} \tag{13}$$

$z_{i,k}$ has a Guassian posterior distribution $\mathcal{N}(\hat{z}_{i,k}, \Sigma_z^{i,k})$, given by

$$\begin{aligned} p(z_{i,k}|v_i, k) &= p(z_{i,k}|x_i, y_i, k) \\ &= \frac{p(x_i, y_i|z_{i,k}, k)p(z_{i,k})}{p(x_i, y_i|k)} \\ &= \frac{p(x_i|z_{i,k}, k)p(y_i|z_{i,k}, k)p(z_{i,k})}{p(x_i, y_i|k)}. \end{aligned} \tag{14}$$

Following the notations in our paper, we have

$$p(x_i|z_{i,k},k) = \mathcal{N}(x_i - W_x^k z_{i,k}|\mu_x^k, \Psi_x^k), \tag{15}$$

$$p(y_i|z_{i,k},k) = \mathcal{N}(y_i - W_y^k z_{i,k}|\mu_y^k, \Psi_y^k), \tag{16}$$

$$p(z_{i,k}|k) = \mathcal{N}(z_{i,k}|0, I). \tag{17}$$

The mean and covariance matrix of $z_{i,k}$ can then be estimated by

$$\begin{aligned}\hat{z}_{i,k} &= E(z_{i,k}) \\ &= [{W_x^k}^\top, {W_y^k}^\top]\mathbf{\Sigma}_k^{-1}\begin{bmatrix} x_i - \mu_x^k \\ y_i - \mu_y^k \end{bmatrix},\end{aligned} \tag{18}$$

$$\begin{aligned}\Sigma_z^{i,k} &= \text{Var}(z_{i,k}) \\ &= I - [{W_x^k}^\top, {W_y^k}^\top]\mathbf{\Sigma}_k^{-1}\begin{bmatrix} W_x^k \\ W_y^k \end{bmatrix},\end{aligned} \tag{19}$$

$$\begin{aligned}\langle z_{i,k} z_{i,k}^\top \rangle &= E(z_{i,k} z_{i,k}^\top) \\ &= \Sigma_z^{i,k} + \hat{z}_{i,k}\hat{z}_{i,k}^\top.\end{aligned} \tag{20}$$

In the **M-step**, we use the estimations of those latent variables to optimize $\Theta$. To begin with, we introduce the complete-data log-likelihood as in [1]

$$\mathcal{L}(\Theta; D, Z, \Gamma) = \sum_i \sum_k \gamma_{i,k} \log\left\{w_k p(v_i, z_{i,k}|k)\right\}, \tag{21}$$

where

$$p(v_i, z_{i,k}|k) = p(v_i|z_{i,k},k)p(z_{i,k}|k) \tag{22}$$

$$= p(x_i|z_{i,k},k)p(y_i|z_{i,k},k)p(z_{i,k}|k). \tag{23}$$

Denote $\widetilde{x}_{i,k}$ and $\widetilde{y}_{i,k}$ by

$$\widetilde{x}_{i,k} = x_i - W_x^k z_{i,k}, \tag{24}$$

$$\widetilde{y}_{i,k} = y_i - W_y^k z_{i,k}. \tag{25}$$

Then we have

$$\begin{aligned}\log p(v_i, z_{i,k}|k) =& \log p(x_i|z_{i,k},k) + \log p(y_i|z_{i,k},k) + \log p(z_{i,k}|k) \\ =& -\frac{d}{2}\log 2\pi - \frac{1}{2}z_{i,k}^\top z_{i,k} \\ & - \frac{n}{2}\log 2\pi - \frac{1}{2}\log|\Psi_x^k| - \frac{1}{2}(\widetilde{x}_{i,k} - \mu_x^k)^\top (\Psi_x^k)^{-1}(\widetilde{x}_{i,k} - \mu_x^k) \\ & - \frac{m}{2}\log 2\pi - \frac{1}{2}\log|\Psi_y^k| - \frac{1}{2}(\widetilde{y}_{i,k} - \mu_y^k)^\top (\Psi_y^k)^{-1}(\widetilde{y}_{i,k} - \mu_y^k) \\ =& -\frac{d+m+n}{2}\log 2\pi - \frac{1}{2}(\log|\Psi_x^k| + \log|\Psi_y^k|) - \frac{1}{2}z_{i,k}^\top z_{i,k} \\ & - \frac{1}{2}\left\{(\widetilde{x}_{i,k} - \mu_x^k)^\top (\Psi_x^k)^{-1}(\widetilde{x}_{i,k} - \mu_x^k) + (\widetilde{y}_{i,k} - \mu_y^k)^\top (\Psi_y^k)^{-1}(\widetilde{y}_{i,k} - \mu_y^k)\right\}.\end{aligned} \tag{26}$$

Given hidden parameters $Z$ and $\Gamma$, we need to maximize the expectation of the complete data log likelihood $E(\mathcal{L})$ with respect to $\Theta$,

$$E(\mathcal{L}) = -\frac{1}{2}\sum_i \sum_k \hat{\gamma}_{i,k}\{\log w_k + \log p(v_i, z_{i,k}|k)\} \tag{27}$$

subject to

$$\sum_k w_k = 1. \tag{28}$$

We use Lagrange method to optimize the above problem, as in [3]

$$\max_\Theta E(\mathcal{L}) + \lambda(\sum_k w_k - 1), \tag{29}$$

where $\lambda$ is Lagrange multiplier. Then we have the final result

$$w_k = \frac{1}{N}\sum_i \hat{\gamma}_{i,k}, \tag{30}$$

$$\mu_x^k = \frac{\sum_i \hat{\gamma}_{i,k}(x_i - W_x^k \hat{z}_{i,k})}{\sum_i \hat{\gamma}_{i,k}}, \tag{31}$$

$$\mu_y^k = \frac{\sum_i \hat{\gamma}_{i,k}(y_i - W_y^k \hat{z}_{i,k})}{\sum_i \hat{\gamma}_{i,k}}, \tag{32}$$

$$W_x^k = \left\{\sum_i \hat{\gamma}_{i,k}(x_i - \mu_x^k)\hat{z}_{i,k}^\top\right\}\left\{\sum_i \hat{\gamma}_{i,k}\langle z_{i,k} z_{i,k}^\top\rangle\right\}^{-1}, \tag{33}$$

$$W_y^k = \left\{\sum_i \hat{\gamma}_{i,k}(y_i - \mu_y^k)\hat{z}_{i,k}^\top\right\}\left\{\sum_i \hat{\gamma}_{i,k}\langle z_{i,k} z_{i,k}^\top\rangle\right\}^{-1}, \tag{34}$$

$$\Psi_x^k = \frac{\sum_i \hat{\gamma}_{i,k}(x_i - W_x^k \hat{z}_{i,k} - \mu_x^k)(x_i - W_x^k \hat{z}_{i,k} - \mu_x^k)^\top}{\sum_i \hat{\gamma}_{i,k}} + W_x^k \Sigma_z^k W_x^{k\top}, \tag{35}$$

$$\Psi_y^k = \frac{\sum_i \hat{\gamma}_{i,k}(y_i - W_y^k \hat{z}_{i,k} - \mu_y^k)(y_i - W_y^k \hat{z}_{i,k} - \mu_y^k)^\top}{\sum_i \hat{\gamma}_{i,k}} + W_y^k \Sigma_z^k W_y^{k\top}. \tag{36}$$

In particular, the most relevant part to those used in our paper can be written as

$$w_k = \frac{1}{N}\sum_i \hat{\gamma}_{i,k}, \tag{37}$$

$$\mu_x^k = \frac{\sum_i \hat{\gamma}_{i,k}\widetilde{x}_{i,k}}{\sum_i \hat{\gamma}_{i,k}}, \tag{38}$$

$$\mu_y^k = \frac{\sum_i \hat{\gamma}_{i,k}\widetilde{y}_{i,k}}{\sum_i \hat{\gamma}_{i,k}}, \tag{39}$$

$$\Psi_x^k = \frac{\sum_i \hat{\gamma}_{i,k}(\widetilde{x}_{i,k} - \mu_x^k)(\widetilde{x}_{i,k} - \mu_x^k)^\top}{\sum_i \hat{\gamma}_{i,k}} + W_x^k \Sigma_z^k W_x^{k\top}, \tag{40}$$

$$\Psi_y^k = \frac{\sum_i \hat{\gamma}_{i,k}(\widetilde{y}_{i,k} - \mu_y^k)(\widetilde{y}_{i,k} - \mu_y^k)^\top}{\sum_i \hat{\gamma}_{i,k}} + W_y^k \Sigma_z^k W_y^{k\top}. \tag{41}$$

Denote $\Psi_x^k - W_x^k \Sigma_z^k W_x^{k\top}$ by $\widetilde{\Psi}_x^k$. The corresponding gradient vectors for $x$ are thus

$$\frac{\partial E(\mathcal{L})}{\partial \mu_x^k} = 2w_k (\Psi_x^k)^{-1} \left\{ \mu_x^k - \frac{\sum_i \gamma_{i,k} \widetilde{x}_{i,k}}{\sum_i \hat{\gamma}_{i,k}} \right\}, \tag{42}$$

$$\frac{\partial E(\mathcal{L})}{\partial \Psi_x^k} = w_k (\Psi_x^k)^{-1} \left\{ \widetilde{\Psi}_x^k - \frac{\sum_i \gamma_{i,k} (\widetilde{x}_{i,k} - \mu_x^k)(\widetilde{x}_{i,k} - \mu_x^k)^\top}{\sum_i \hat{\gamma}_{i,k}} \right\} (\Psi_x^k)^{-1}. \tag{43}$$

Similarly, gradient vectors for $y$ are

$$\frac{\partial E(\mathcal{L})}{\partial \mu_y^k} = 2w_k (\Psi_y^k)^{-1} \left\{ \mu_y^k - \frac{\sum_i \gamma_{i,k} \widetilde{y}_{i,k}}{\sum_i \hat{\gamma}_{i,k}} \right\}, \tag{44}$$

$$\frac{\partial E(\mathcal{L})}{\partial \Psi_y^k} = w_k (\Psi_y^k)^{-1} \left\{ \widetilde{\Psi}_y^k - \frac{\sum_i \gamma_{i,k} (\widetilde{y}_{i,k} - \mu_y^k)(\widetilde{y}_{i,k} - \mu_y^k)^\top}{\sum_i \hat{\gamma}_{i,k}} \right\} (\Psi_y^k)^{-1}. \tag{45}$$

where $\widetilde{\Psi}_y^k = \Psi_y^k - W_y^k \Sigma_z^k W_y^{k\top}$.

# References


[1] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006. 3

[2] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–8. IEEE, 2007. 1

[3] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999. 4