Supplementary Material for the paper: Higher-order Co-occurrence Features based on Discriminative Co-clusters for Image Classification

Takumi Kobayashi
takumi.kobayashi@aist.go.jp

National Institute of Advanced Industrial Science and Technology
1-1-1, Umezono, Tsukuba, Japan

In this material, we show the EM method [3] which is basically used to produce co-clusters as described in Sec.2.2, and its modified versions to cope with various domains used in the experiments (Sec.4).

A EM method

As described in Sec.2.2, suppose we obtain the (classifier) weight $\boldsymbol{W}$ on $\mathcal{R} \times \mathcal{R}$, actually $\boldsymbol{W} \in \Re_{+}^{L\times L}$ by finely partitioning the domain space $\mathcal{R}$ into L bins. Let $x_i \in \Re^d$ represent the quantitative data at the i-th bin, and it is augmented as $\tilde{x}_{ij} = \left[\begin{smallmatrix} x_i \\ x_j \end{smallmatrix}\right] \in \Re^{2d}$ to represent the i,j-th joint bin corresponding to the element W_{ij} on $\mathcal{R} \times \mathcal{R}$. We normalize the weight by $\boldsymbol{W} \leftarrow \boldsymbol{W} / \sum_{ij}^{L} W_{ij}$ such that it is regarded as the probability distribution, $p(\tilde{x}_{ij}) \triangleq W_{ij}$, to which the EM method is applied.

Given the number of co-clusters D, the EM method [3] is applied on the probability distribution $\boldsymbol{W}$ as follows.

Algorithm 1 : EM method

```
Input: $\boldsymbol{W} \in \Re_{+}^{L\times L}$, $D$
Initialize $\alpha_k^{(0)} \in \Re,\ \mu_k^{(0)} \in \Re^{2d},\ \Sigma_k^{(0)} \in \Re^{2d\times 2d}\ (k=1,..,D),\ \ t=0$
repeat
    $\gamma_{ijk} = \frac{\alpha_k^{(t)} \mathcal{N}(\tilde{x}_{ij}; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_k^D \alpha_k^{(t)} \mathcal{N}(\tilde{x}_{ij}; \mu_k^{(t)}, \Sigma_k^{(t)})},\ \forall i,j,k$
    $\alpha_k^{(t+1)} = \sum_{ij}^{L} W_{ij} \gamma_{ijk},\ \forall k$
    $\mu_k^{(t+1)} = \frac{1}{\alpha_k^{(t+1)}} \sum_{ij}^{L} W_{ij} \gamma_{ijk} \tilde{x}_{ij},\ \forall k$
    $\Sigma_k^{(t+1)} = \frac{1}{\alpha_k^{(t+1)}} \sum_{ij}^{L} W_{ij} \gamma_{ijk} \{\tilde{x}_{ij} - \mu_k^{(t+1)}\}\{\tilde{x}_{ij} - \mu_k^{(t+1)}\}^{\top},\ \forall k$
    $t \leftarrow t+1$
until convergence
Output: $\mathcal{N}_k(x_1,x_2) = \mathcal{N}\left(\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right];\ \mu_k,\ \Sigma_k\right)$ with $\alpha_k\ \ (k=1,..,D)$
```

The EM method produces the cluster component functions $\mathcal{N}_k(x_1,x_2)$ with their prior weights α_k which are subsequently used to determine the discriminative co-clusters g_k as shown in (3) of the paper. Note that $\mathcal{N}(\tilde{x};\, \mu,\, \Sigma)$ indicates the Gaussian distribution function with the mean μ and the covariance matrix Σ.

B EM method on symmetric domain

In the experiment on cancer detection (Sec.4.1), we consider the symmetric co-occurrences to render the rotation invariance, which results in the symmetric (classifier) weight $W_{ij} = W_{ji}, \forall i, j$.

The EM method is modified for such a symmetric weight (probability distribution) $\boldsymbol{W}$ as follows.

Algorithm 2 : EM method on symmetric domain

```
Input: $\boldsymbol{W} \in \Re_{+}^{L\times L}$, $D$
 1: Initialize $\alpha_k^{(0)} \in \Re$, $\mu_k^{(0)} \in \Re^{2d}$, $\Sigma_k^{(0)} \in \Re^{2d\times 2d}$ $(k = 1,..,D)$, $t = 0$
 2:     $\alpha_{D+k}^{(0)} = \alpha_k^{(0)}$, $\mu_{D+k}^{(0)} = \mathsf{flip}(\mu_k^{(0)})$, $\Sigma_{D+k}^{(0)} = \mathsf{flip}(\Sigma_k^{(0)})$ $(k = 1,..,D)$
 3: repeat
 4:     $\gamma_{ijk} = \frac{\alpha_k^{(t)} \mathcal{N}(\tilde{x}_{ij}; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_k^{2D} \alpha_k^{(t)} \mathcal{N}(\tilde{x}_{ij}; \mu_k^{(t)}, \Sigma_k^{(t)})}$, $\forall i, j, k$
 5:     $\bar{\alpha}_k = \sum_{ij}^{L} W_{ij}\gamma_{ijk}$, $\forall k$
 6:     $\bar{\mu}_k = \frac{1}{\bar{\alpha}_k} \sum_{ij}^{L} W_{ij}\gamma_{ijk}\tilde{x}_{ij}$, $\forall k$
 7:     $\bar{\Sigma}_k = \frac{1}{\bar{\alpha}_k} \sum_{ij}^{L} W_{ij}\gamma_{ijk}\{\tilde{x}_{ij} - \bar{\mu}_k\}\{\tilde{x}_{ij} - \bar{\mu}_k\}^{\top}$, $\forall k$
 8:     $\alpha_k^{(t+1)} = \frac{1}{2}\{\bar{\alpha}_k + \bar{\alpha}_{D+k}\}$, $k = 1, \cdots, D$
 9:     $\mu_k^{(t+1)} = \frac{1}{2}\{\bar{\mu}_k + \mathsf{flip}(\bar{\mu}_{D+k})\}$, $k = 1, \cdots, D$
10:     $\Sigma_k^{(t+1)} = \frac{1}{2}\{\bar{\Sigma}_k + \mathsf{flip}(\bar{\Sigma}_{D+k})\}$, $k = 1, \cdots, D$
11:     $\alpha_{D+k}^{(t+1)} = \alpha_k^{(t+1)}$, $\mu_{D+k}^{(t+1)} = \mathsf{flip}(\mu_k^{(t+1)})$, $\Sigma_{D+k}^{(t+1)} = \mathsf{flip}(\Sigma_k^{(t+1)})$, $k = 1, \cdots, D$
12:     $t \leftarrow t + 1$
13: until convergence
Output: $\mathcal{N}_k(x_1, x_2) = \mathcal{N}\left(\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right];\, \mu_k,\, \Sigma_k\right) + \mathcal{N}\left(\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right];\, \mathsf{flip}(\mu_k),\, \mathsf{flip}(\Sigma_k)\right)$ with $\alpha_k$ $(k = 1,..,D)$
```

The function flip flips a vector or a matrix symmetrically: for $\boldsymbol{a}, \boldsymbol{b} \in \Re^d, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \Re^{d\times d}$,

$$\mathsf{flip}\left(\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}\right) = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{a} \end{bmatrix}, \quad \mathsf{flip}\left(\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^{\top} & \boldsymbol{C} \end{bmatrix}\right) = \begin{bmatrix} \boldsymbol{C} & \boldsymbol{B}^{\top} \\ \boldsymbol{B} & \boldsymbol{A} \end{bmatrix}.$$

To cope with the symmetric weight, the cluster components are (temporarily) doubled in symmetric forms using the flip (lines 2, 11 and **Output**). The lines 8 $\sim$ 10 are applied for numerical stability.

C EM method on circular domain

In the case that the primitive quantitative data are gradient orientations (Sec.4.2), the (classifier) weight $\boldsymbol{W}$ is obtained on the circular domains (orientations) of 2π cycle. In order to cope with the circularity, we modify the EM method as follows.

Algorithm 3 : EM method on circular domain

```
Input: $\boldsymbol{W} \in \Re_{+}^{L\times L}$, $D$
1: Initialize $\alpha_k^{(0)} \in \Re$, $\mu_k^{(0)} \in \Re^{2d}$, $\Sigma_k^{(0)} \in \Re^{2d\times 2d}$ $(k=1,..,D)$, $t=0$
2: repeat
3:     $\gamma_{ijk} = \frac{\alpha_k^{(t)}\mathcal{N}_{\mathsf{circ}}(\tilde{x}_{ij};\mu_k^{(t)},\Sigma_k^{(t)})}{\sum_k^D \alpha_k^{(t)}\mathcal{N}_{\mathsf{circ}}(\tilde{x}_{ij};\mu_k^{(t)},\Sigma_k^{(t)})}$, $\forall i,j,k$
4:     $\alpha_k^{(t+1)} = \sum_{ij}^L W_{ij}\gamma_{ijk}$, $\forall k$
5:     $\mu_k^{(t+1)} = \mathsf{arg}\left\{\frac{1}{\alpha_k^{(t+1)}}\sum_{ij}^L W_{ij}\gamma_{ijk}\mathsf{exp}(\mathtt{i}\tilde{x}_{ij})\right\}$, $\forall k$
6:     $\Sigma_k^{(t+1)} = \frac{1}{\alpha_k^{(t+1)}}\sum_{ij}^L W_{ij}\gamma_{ijk}\{\tilde{x}_{ij} \overset{\mathsf{circ}}{-} \mu_k^{(t+1)}\}\{\tilde{x}_{ij} \overset{\mathsf{circ}}{-} \mu_k^{(t+1)}\}^\top$
7:     $t \leftarrow t+1$
8: until convergence
Output: $\mathcal{N}_k(x_1,x_2) = \mathcal{N}_{\mathsf{circ}}\left(\left[\begin{smallmatrix}x_1\\x_2\end{smallmatrix}\right];\ \mu_k,\ \Sigma_k\right)$ with $\alpha_k$ $(k=1,..,D)$
```

We assume that the clusters are localized, not spreading over whole of the circular domain. On that assumption, the clusters can be approximated by Gaussian distribution locally around the mean (mode). It should be noted that such pairs of circular data form the distribution on a torus [2], not a sphere, and the clustering using von Mises-Fisher distribution [1] can not be directly applied. The main differences from Algorithm 1 are lines 5 and 6 according to the circularity. The functions exp and arg operate on each element:

$$\mathsf{exp}(\mathtt{i}\boldsymbol{a}) = \begin{bmatrix}\mathsf{exp}(\mathtt{i}a_1)\\ \vdots \\ \mathsf{exp}(\mathtt{i}a_d)\end{bmatrix},\ \mathsf{arg}\{\mathsf{exp}(\mathtt{i}\boldsymbol{a})\} = \begin{bmatrix}\mathsf{arg}\{\mathsf{exp}(\mathtt{i}a_1)\}\\ \vdots \\ \mathsf{arg}\{\mathsf{exp}(\mathtt{i}a_d)\}\end{bmatrix} = \begin{bmatrix}a_1\\ \vdots \\ a_d\end{bmatrix},$$

where $\mathtt{i}$ indicates the imaginary unit. The operator $\overset{\mathsf{circ}}{-}$ is modified *minus* so as to fit the circular domain:

$$a \overset{\mathsf{circ}}{-} b = \begin{cases} a-b-2\pi & \text{if} \quad a-b>\pi \\ a-b & \text{if} \quad |a-b|<\pi \\ a-b+2\pi & \text{if} \quad a-b<-\pi \end{cases}.$$

Thus, $\mathcal{N}_{\mathsf{circ}}$ is also a modified Gaussian distribution function using $\overset{\mathsf{circ}}{-}$[1]:

$$\mathcal{N}_{\mathsf{circ}}(\tilde{x};\ \mu,\ \Sigma) = \frac{1}{\sqrt{(2\pi)^{2d}|\Sigma|}}\exp\{-\frac{1}{2}(x \overset{\mathsf{circ}}{-} \mu)^\top\Sigma^{-1}(x \overset{\mathsf{circ}}{-} \mu)\}.$$

References

[1] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(12): 1345–1382, 2005.

[2] T. Kobayashi and N. Otsu. Image feature extraction using gradient local autocorrelations. In *European Conference on Computer Vision*, pages 346–358, 2008.

[3] Y. Xiao and G. Xuan. Fast em algorithm of multi-dimensional histogram in medical images. In *International Conference on Diagnostic Imaging and Analysis*, pages 328–333, 2002.

[1] Strictly speaking, $\mathcal{N}_{\mathsf{circ}}$ is not probability density function since $\int\mathcal{N}_{\mathsf{circ}}(\tilde{x};\mu,\Sigma)d\tilde{x} \neq 1$ on this circular domain, but we adopt this simply modified form due to computational convenience.