Supplementary Materials for Bayesian Hierarchical Dynamic Model for Human Action Recognition

Rui Zhao[1], Wanru Xu[2], Hui Su[1,3], Qiang Ji[1]
[1]RPI, [2]Beijing Jiaotong University, [3]IBM Research
{zhaorui.zju,bjtuxuwanru}@gmail.com, huisuibmres@us.ibm.com, qji@ecse.rpi.edu

1. Parameterization of HDM

We use multinomial distributions for initial and transition distributions. We use Poisson distribution for duration distribution. We assume the duration value is drawn only when entering a state. And then the state chain remains at the same state for the drawn duration. A regular transition happens at the end of the duration. Using the notation introduced in the main paper. The conditional probability are defined as follows.

$$P(Z_1 = j) = \pi_j \tag{1}$$

$$P(Z_t = j|D_{t-1} = d', Z_{t-1} = i) = \begin{cases} A_{ij}, & \text{if } d' = 1 \\ \delta(i,j), & \text{otherwise} \end{cases} \tag{2}$$

$$P(D_1 = d|Z_1 = i) \triangleq C_{id} = \frac{\tau_i^d e^{-\tau_i}}{d!} \tag{3}$$

$$P(D_t = d|D_{t-1} = d', Z_t = i) = \begin{cases} C_{id} & \text{if } d' = 1 \\ \delta(d, d'-1), & \text{otherwise} \end{cases} \tag{4}$$

where $\pi_j \geq 0, A_{ij} \geq 0, \tau_i > 0$ and $\sum_{j=1}^{Q} \pi_j = 1, \sum_{j=1, j\neq i}^{Q} A_{ij} = 1, A_{ii} = 0, \forall i = 1, ..., Q$. $\delta(i,j) = 1$ if $i = j$ and 0 otherwise. We forbid self-transition *i.e.* $A_{ii} = 0$ to disambiguate the duration count used during inference [8]. For emission distribution, we use a mixture of Gaussian.

$$P(X_t = \mathbf{o}|Z_t = i) = \sum_{k=1}^{M} W_{ik} \mathcal{N}(\mathbf{o}|\mu_{ik}, \Sigma_{ik}) \tag{5}$$

where M is the number of mixtures and W_{ik} is the weight of k^{th} mixture under i^{th} state with $\sum_{k=1}^{M} W_{ik} = 1, W_{ik} > 0, \forall i = 1, ..., Q$. $\mu_{ik} \in \mathbb{R}^O$, $\Sigma_{ik} \in \mathbb{R}^{O\times O}$ are mean and covariance matrix of k^{th} mixture under i^{th} state respectively. We assume the same number of mixtures under different states.

We place a conjugate prior for each parameter. For the multinomial parameters $\pi, \mathbf{A}$ and $\mathbf{W}$, Dirichlet priors are used with hyperparameters respectively $\eta_0 \in \mathbb{R}_+^Q, \eta \in \mathbb{R}_+^{Q\times Q}$ and $v \in \mathbb{R}_+^Q$. For the Poisson parameters τ, we use Gamma prior $\xi = \{a \in \mathbb{R}_+^Q, b \in \mathbb{R}_+^Q\}$. For emission mean μ and covariance Σ, Normal-Inverse-Wishart priors are used with hyperparameters $\{\mu_0, \kappa_0, \Lambda_0, \nu_0\}$, where $\mu_0 \in \mathbb{R}^O$, $\kappa_0 > 0$, $\Lambda_0 \in \mathbb{R}^{O\times O}$ positive definite and $\nu_0 > O - 1$.

Specifically, we have

$$P(\pi) = Dir(\pi|\eta_0) \propto \prod_{j=1}^{Q} \pi_j^{\eta_{0j}-1} \tag{6}$$

$$P(\mathbf{A}_{i:}) = Dir(\mathbf{A}_{i:}|\eta_i) \propto \prod_{j=1}^{Q} A_{ij}^{\eta_{ij}-1},\ i = 1, ..., Q \tag{7}$$

$$P(\tau_i) = Gam(\tau_i|a_i, b_i) \propto \tau_i^{a_i-1} e^{-b_i\tau_i},\ i = 1, ..., Q \tag{8}$$

$$P(\mathbf{W}_{i:}) = Dir(\mathbf{W}_{i:}|v) \propto \prod_{k=1}^{M} W_{ik}^{v_k-1},\ i = 1, ..., Q \tag{9}$$

$$\begin{aligned} P(\mu_{ik}, \Sigma_{ik}) &= NIW(\mu_{ik}, \Sigma_{ik}|\mu_0, \kappa_0, \Lambda_0, \nu_0) \\ &\propto |\Sigma_{ik}|^{-(\nu_0+O)/2-1} \exp\big(-\frac{1}{2}tr(\Lambda_0\Sigma_{ik}^{-1}) - \frac{\kappa_0}{2}(\mu_{ik}-\mu_0)^T\Sigma_{ik}^{-1}(\mu_{ik}-\mu_0)\big), \\ i &= 1, ..., Q, k = 1, ..., M \end{aligned} \tag{10}$$

2. MAP-EM algorithm

The MAP-EM algorithm solves the following optimization problem on parameters θ given hyperparameters ϕ and a set of observations $\{\mathbf{X}_n\}$.

$$\begin{aligned} \theta^* &= \arg\max_\theta \sum_n \log \sum_{\mathbf{Z}_n, \mathbf{D}_n} P(\mathbf{X}_n, \mathbf{Z}_n, \mathbf{D}_n|\theta) + \log P(\theta|\phi) \\ &= \arg\max_\theta \sum_n \log P(\mathbf{X}_n|\theta) + \log P(\theta|\phi) \end{aligned} \tag{11}$$

Due to the presence of hidden variables, the marginal likelihood must be evaluated by summing over all the hidden variables $\{\mathbf{Z}_n, \mathbf{D}_n\}$. We adopt Expectation-Maximization (EM) algorithm to handle the learning with hidden variables, where we iterate between E-step, which computes a tight lower bound to the marginal log-likelihood and M-step, which maximizes the lower bound with respect to model parameters. In our case, M-step needs to be modified to incorporate the effect of prior distributions in a similar way to [3].

Specifically, for E-step, we compute $Q(\theta, \hat{\theta}) = E_{P(\mathbf{Z},\mathbf{D}|\mathbf{X},\hat{\theta})}[\log P(\mathbf{X}, \mathbf{Z}, \mathbf{D}|\theta)]$, where $\hat{\theta}$ is current estimate of parameters. Given the parameterization as described in Section 1, the joint distribution $P(\mathbf{X}, \mathbf{Z}, \mathbf{D}|\theta)$ belongs to the exponential family. Then it can be shown that $Q(\theta, \hat{\theta})$ can be decomposed into summation of expected sufficient statistics over individual parameter. Leveraging on the chain structure and the explicit duration assumption, we can extend the forward-backward algorithm used in HMM to efficiently compute the inquired expected sufficient statistics. Following the notation of [8], we define the following quantities.

$$\alpha_t(i, d) = P(Z_t = i, D_t = d, X_{1:t}) \tag{12}$$

$$\beta_t(i, d) = P(X_{t+1:T}|Z_t = i, D_t = d) \tag{13}$$

For compactness, we define notation $b_i(o_t) = P(X_t = o_t|Z_t = i)$. The forward messages α and backward messages β can be computed using the following recursions.

$$\alpha_t(i, d) = \alpha_{t-1}(i, d+1)b_i(o_t) + \big(\sum_{j\neq i} \alpha_{t-1}(j, 1)A_{ji}\big)b_i(o_t)C_{id},\ \forall t > 1 \tag{14}$$

$$\beta_t(i, d) = \begin{cases} b_i(o_{t+1})\beta_{t+1}(i, d-1), & \text{if } d > 1 \\ \sum_{j\neq i} A_{ij}b_j(o_{t+1})\big(\sum_{d\geq 1} C_{jd}\beta_{t+1}(j, d)\big), & \text{if } d = 1 \end{cases},\ \forall t < T \tag{15}$$

with initial condition $\alpha_1(i, d) = \pi_i b_i(o_1)C_{id}, \beta_T(i, d) = 1$. After computing the messages, we can compute the following

probabilities, which are used to compute expected sufficient statistics involved in Q.

$$\gamma_t(i) := P(Z_t = i|\mathbf{X}) = P(Z_t = i, \mathbf{X})/P(\mathbf{X}) \tag{16}$$

$$\zeta_t(i,j) := P(Z_{t-1} = i, Z_t = j|\mathbf{X}) = \alpha_{t-1}(i,1)A_{ij}b_j(o_t)\big(\sum_{d\geq 1} C_{jd}\beta(j,d)\big)/P(\mathbf{X}) \tag{17}$$

$$\omega_t(i,d) := P(Z_t = i, Z_{t-1} \neq i, D_t = d|\mathbf{X}) = \big(\sum_{j\neq i} \alpha_{t-1}(j,1)A_{ji}\big)b_i(o_t)C_{id}\beta_t(i,d)/P(\mathbf{X}) \tag{18}$$

where $P(\mathbf{X}) = \sum_i \alpha_t(i,d)\beta_t(i,d), \forall t$. To compute $\gamma_t(i)$, we use the following recursion.

$$\gamma_t(i) = \gamma_{t+1}(i) + \sum_{j\neq i} \big(\zeta_{t+1}(i,j) - \zeta_{t+1}(j,i)\big),\ \forall t < T \tag{19}$$

The recursion is result of the following equality

$$P(Z_t = i|\mathbf{X}) - P(Z_t = i, Z_{t+1} \neq i|\mathbf{X}) = P(Z_{t+1} = i|\mathbf{X}) - P(Z_t \neq i, Z_{t+1} = i|\mathbf{X}) \tag{20}$$

The initial condition is $\gamma_T(i) = \sum_{d\geq 1} \alpha_T(i,d)$. To avoid numerical underflow. The forward-backward inference is performed in log domain as suggested in [7].

For M-step, we compute the updates of parameters by solving the following problem.

$$\hat{\theta} = \arg\max_{\theta} R(\theta, \hat{\theta}) \tag{21}$$

where $R(\theta, \hat{\theta}) = Q(\theta, \hat{\theta}) + \log P(\theta|\phi)$. Due to the hierarchical structure, initial state parameter π, transition parameter $\mathbf{A}$, duration parameter τ are different for different sequences. Thus they are updated for individual sequence. While emission parameters $\{\mathbf{W}, \mu, \Sigma\}$ are shared across sequences and updated once for all sequences.

Provided that we can compute expected sufficient statistics using Eq. (16)-(18) and we choose conjugate prior, the solution for $\pi_n, \mathbf{A}_n$ have closed-form solution similar to the results in HMM derived in [3]. τ_n can also be computed using a similar derivation. The updates are as follows. [1]

$$\pi^*_{ni} = \frac{P(Z^n_1 = i|\mathbf{X}_n) + \eta_{0i}}{\sum_j \eta_{0j}} \tag{22}$$

$$A^*_{nij} = \frac{\sum_t P(Z^n_t = i, Z^n_{t+1} = j|\mathbf{X}_n) + \eta_{ij}}{\sum_t \sum_j (P(Z^n_t = i, Z^n_{t+1} = j|\mathbf{X}_n) + \eta_{ij})} \tag{23}$$

$$\tau^*_{ni} = \frac{\sum_t \sum_{1\leq d\leq t} P(Z^n_t = i, Z^n_{t-1} \neq i, D^n_t = d|\mathbf{X}_n)d + a_i}{\sum_t \sum_{1\leq d\leq t} P(Z^n_t = i, Z^n_{t-1} \neq i, D^n_t = d|\mathbf{X}_n) + b_i} \tag{24}$$

We now consider the updates for emission parameters $\mathbf{W}, \mu, \Sigma$. We introduce another variable M^n_t to indicate the mixture component index for n^{th} sequence t^{th} frame. The update of W_{ik} can be done in a similar way to temporal parameters.

$$Q(W_{ik}, \hat{\theta}) = \sum_{n=1}^{N}\sum_{t=1}^{T_n} P(Z^n_t = i, M^n_t = k|\mathbf{X}_n) \tag{25}$$

$$\begin{aligned} R(W_{ik}, \hat{\theta}) &= Q(W_{ik}, \hat{\theta}) + \log Dir(\mathbf{W}_i|v) \\ &= Q(W_{ik}, \hat{\theta}) + \sum_{k=1}^{M}(v_k - 1)\log W_{ik} + s \end{aligned} \tag{26}$$

where s is a constant independent of W_{ik}. Maximize $R(W_{ik}, \hat{\theta})$ with respect to W_{ik} subject to $\sum_{k=1}^{M} W_{ik} = 1$ yield [2]

$$W^*_{ik} = \frac{\sum_n \sum_t P(Z^n_t = i, M^n_t = k|\mathbf{X}_n) + v_k}{\sum_n \sum_t \sum_k (P(Z^n_t = i, M^n_t = k|\mathbf{X}_n) + v_k)} \tag{27}$$

[1] We use posterior mean as estimate instead of the exact MAP estimate to ensure positive estimated values on parameters in case the expected sufficient statistics are less than 1 due to data scarcity. For exact MAP estimate, we need to use substitution in Eq. (22)-(23) with $\eta_{0i} \leftarrow \eta_{0i} - 1, \eta_{ij} \leftarrow \eta_{ij} - 1$.

[2] Similar to Eq. (22)-(23), for exact MAP estimate, we need to use substitution in Eq. (27) with $v_k \leftarrow v_k - 1$.

For MoG parameters, we have

$$Q(\mu_{ik}, \Sigma_{ik}, \hat{\theta}) = \sum_{n=1}^{N}\sum_{t=1}^{T_n} P(Z_t^n = i, M_t^n = k|\mathbf{X}_n) \log P(X_t^n, D_t^n = 1, Z_t^n = i, M_t^n = k|\mu_{ik}, \Sigma_{ik}) \tag{28}$$

$$= \sum_{n=1}^{N}\sum_{t=1}^{T_n} P(Z_t^n = i, M_t^n = k|\mathbf{X}_n)\big[-\frac{1}{2}\log|\Sigma_{ik}| - \frac{1}{2}tr(X_t^n(X_t^n)^T\Sigma_{ik}^{-1}) + \mu_{ik}^T\Sigma_{ik}^{-1}X_t^n - \frac{1}{2}\mu_{ik}^T\Sigma_{ik}^{-1}\mu_{ik}\big] + s$$

where $tr(A)$ is the trace of matrix A. s is a constant that does not depend on μ_{ik}, Σ_{ik}. Then

$$R(\mu_{ik}, \Sigma_{ik}, \hat{\theta}) = Q(\mu_{ik}, \Sigma_{ik}, \hat{\theta}) + \log NIW(\mu_{ik}, \Sigma_{ik}|\kappa_0, \mu_0, \nu_0, \Lambda_0) \tag{29}$$

$$= Q(\mu_{ik}, \Sigma_{ik}, \hat{\theta}) - \frac{\kappa_0}{2}(\mu_{ik} - \mu_0)^T\Sigma_{ik}^{-1}(\mu_{ik} - \mu_0) - \frac{\nu_0 + O + 2}{2}\log|\Sigma_{ik}| - \frac{1}{2}tr(\Lambda_0\Sigma_{ik}^{-1}) + s$$

where s is a constant that does not depend on μ_{ik}, Σ_{ik}. Set the gradient of $R(\mu_{ik}, \Sigma_{ik}, \hat{\theta})$ with respect to μ_{ik} and Σ_{ik} to zero, we can obtain the updates for μ_{ik} and Σ_{ik} as follows.

$$\mu_{ik}^* = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n} P(Z_t^n = i, M_t^n = k|\mathbf{X}_n)X_t^n + \kappa_0\mu_0}{\sum_{n=1}^{N}\sum_{t=1}^{T_n} P(Z_t^n = i, M_t^n = k|\mathbf{X}_n) + \kappa_0} = \frac{\tilde{m}_{ik}}{\tilde{N}_{ik}} \tag{30}$$

$$\Sigma_{ik}^* = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n} P(Z_t^n = i, M_t^n = k|\mathbf{X}_n)(X_t^n - \mu_{ik}^*)(X_t^n - \mu_{ik}^*)^T + \kappa_0(\mu_0 - \mu_{ik}^*)(\mu_0 - \mu_{ik}^*)^T + \Lambda_0}{\tilde{N}_{ik} - \kappa_0 + \nu_0 + O + 2} \tag{31}$$

$$= \frac{\tilde{N}_{ik}\bar{S}_{ik} - \tilde{m}_{ik}\tilde{m}_{ik}^T + \tilde{N}_{ik}(\kappa_0\mu_0\mu_0^T + \Lambda_0)}{\tilde{N}_{ik}(\tilde{N}_{ik} - \kappa_0 + \nu_0 + O + 2)}$$

where

$$\tilde{N}_{ik} = \sum_{n=1}^{N}\sum_{t=1}^{T_n} P(Z_t^n = i, M_t^n = k|\mathbf{X}_n) + \kappa_0 \tag{32}$$

$$\tilde{m}_{ik} = \sum_{n=1}^{N}\sum_{t=1}^{T_n} X_t^n P(Z_t^n = i, M_t^n = k|\mathbf{X}_n) + \kappa_0\mu_0 \tag{33}$$

$$\bar{S}_{ik} = \sum_{n=1}^{N}\sum_{t=1}^{T_n} X_t^n(X_t^n)^T P(Z_t^n = i, M_t^n = k|\mathbf{X}_n) \tag{34}$$

3. MLE estimate of hyperparameters

For η_0, η, ξ, v, they are solved by maximizing the likelihood of corresponding Dirichlet distribution. We use a fixed-point update proposed in [5]. For ξ, maximum likelihood estimate of corresponding Gamma distribution is computed using gradient based update proposed in [6]. For emission hyperparameters, we set the $\kappa_0 = 1$ and $\nu_0 = O + 2$ as fixed and solve for μ_0, Λ_0 by maximizing the corresponding Normal-Inverse-Wishart distribution, where closed-form solution exists with fixed κ_0 and ν_0.

4. Computing the total covariance

Here we prove the first equality of Eq. (9) in the main paper.

$$\begin{aligned}
V[y|\mathbf{X}] &= E[yy^T|\mathbf{X}] - E[y|\mathbf{X}]E[y|\mathbf{X}]^T \\
&= E_\theta[E[yy^T|\mathbf{X}, \theta]] - E_\theta[E[y|\mathbf{X}, \theta]]E_\theta[E[y|\mathbf{X}, \theta]]^T \\
&= E_\theta[E[yy^T|\mathbf{X}, \theta]] - E_\theta[E[y|\mathbf{X}, \theta]E[y|\mathbf{X}, \theta]^T] \\
&+ E_\theta[E[y|\mathbf{X}, \theta]E[y|\mathbf{X}, \theta]^T] - E_\theta[E[y|\mathbf{X}, \theta]]E_\theta[E[y|\mathbf{X}, \theta]]^T \\
&= E_\theta[V[y|\mathbf{X}, \theta]] + V_\theta[E[y|\mathbf{X}, \theta]]
\end{aligned}$$

5. More results of uncertainty analysis

Here we compare the confusion matrix of classification with the corresponding covariance matrix C of the categorical distribution of label vector. The diagonal entries of the covariance matrix reflect the within-class uncertainty level. The higher the value, the more uncertainty. The off-diagonal entries of the covariance matrix reflect the pair-wise between-class uncertainty. The value should be close to 0 if the between-class uncertainty is low. Here we report the average covariance over all testing data. For example, as shown in Figure 1, the four actions *draw x, draw circle, draw circle counter-clockwise*, and *draw triangle* have both high within-class uncertainty and high between-class uncertainty. This is consistent with the confusion matrix where *draw circle counter-clockwise* are mostly confused with *draw triangle*. Similarly in Figure 2, we observe *hand catch, high throw*, and *draw x* are likely to be confused with each other. In Figure 3, one action *aim and fire gun* has high within-class uncertainty and high between-class uncertainty with a few other actions. The classification results also show confusion of *aim and fire gun* with these actions. Similarly in Figure 4, the action *tennis forehand* has the highest within class uncertainty and tend to be confused with actions like *golf swing, baseball pitch*. Based on these results, we argue it is important to consider the uncertainty level before making a classification prediction.

References

[1] V. Bloom, D. Makris, and V. Argyriou. G3d: A gaming action dataset and real time action recognition evaluation framework. In *CVPR Workshop*, 2012.

[2] C. Chen, R. Jafari, and N. Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *ICIP*, 2015.

[3] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *TSAP*, 1994.

[4] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *CVPR Workshop*, 2010.

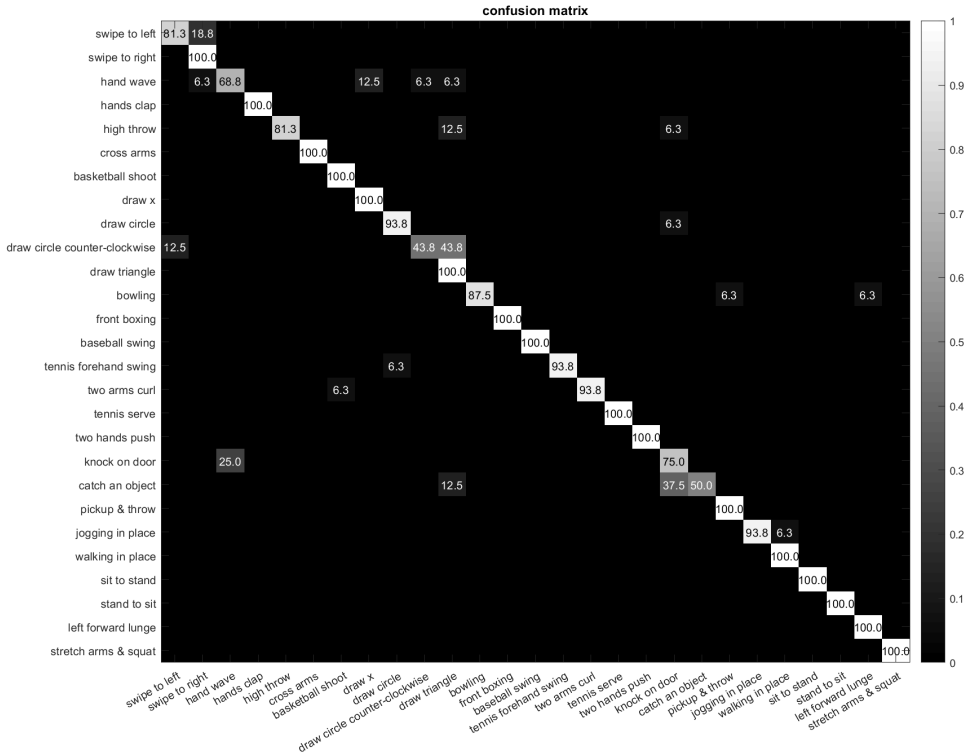
[5] T. Minka. Estimating a dirichlet distribution, 2000.

[6] T. P. Minka. Estimating a gamma distribution. *Microsoft Research, Cambridge, UK, Tech. Rep*, 2002.

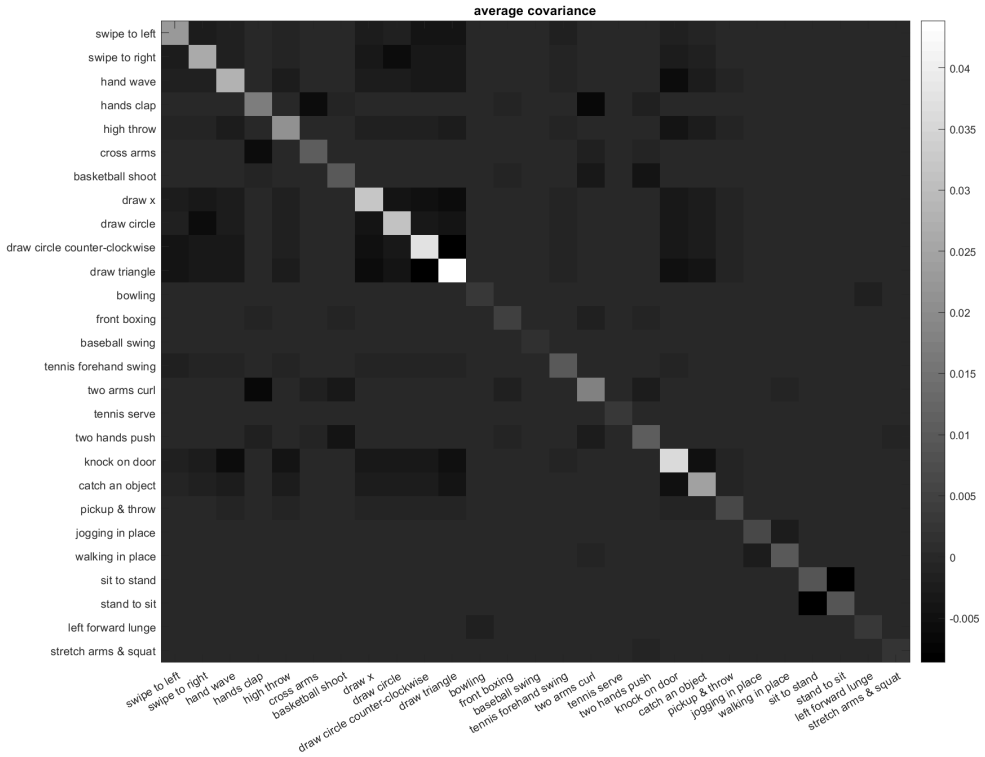
[7] K. P. Murphy. Hidden semi-markov models (hsmms). 2002.

[8] S.-Z. Yu and H. Kobayashi. An efficient forward-backward algorithm for an explicit-duration hidden markov model. *Signal processing letters*, 2003.

[9] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013.

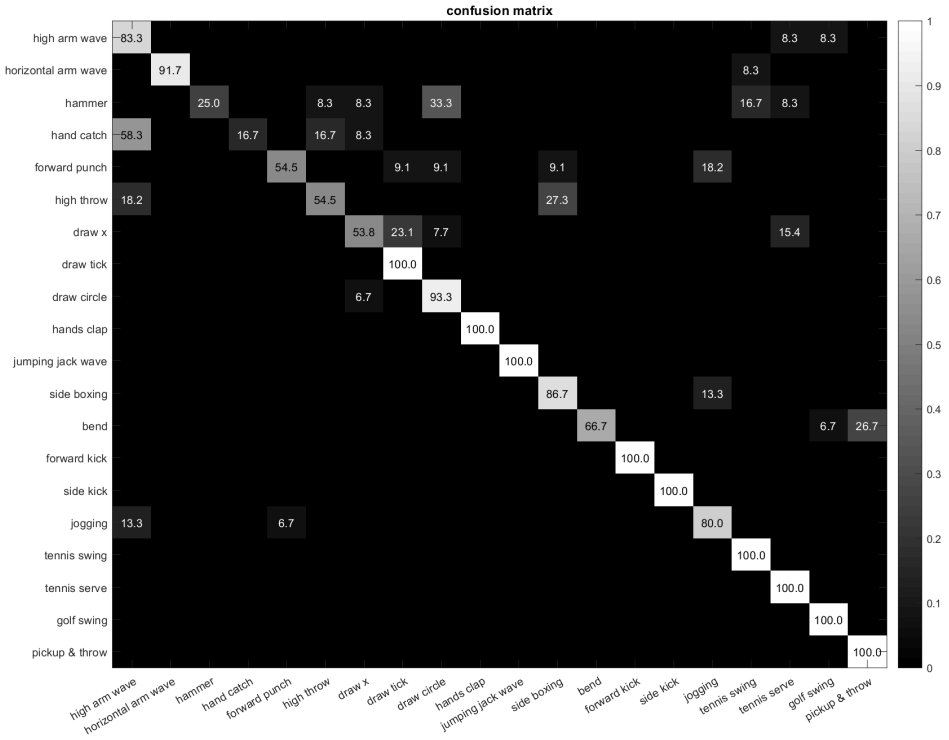


(a) Confusion matrix

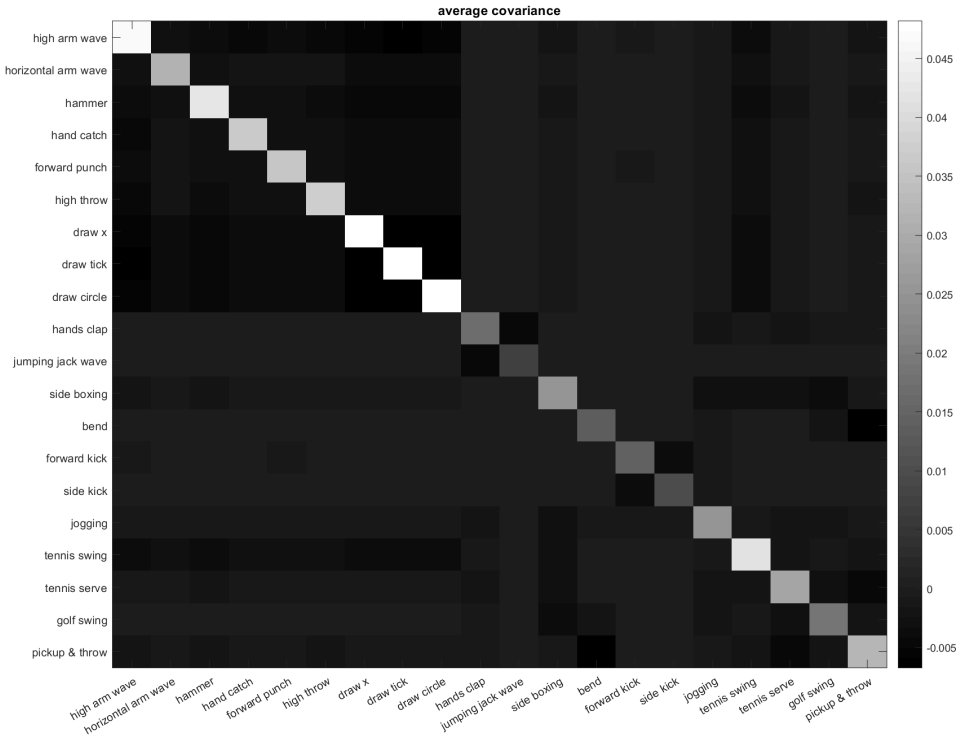


(b) Label distribution covariance

Figure 1. More results on UTD dataset [2].

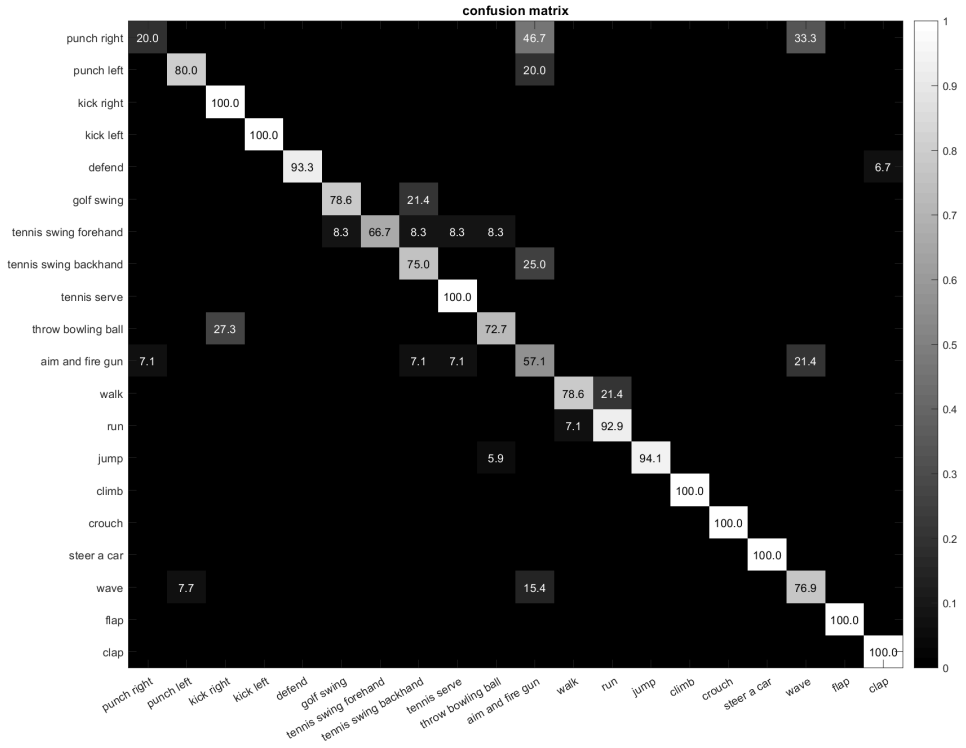


(a) Confusion matrix

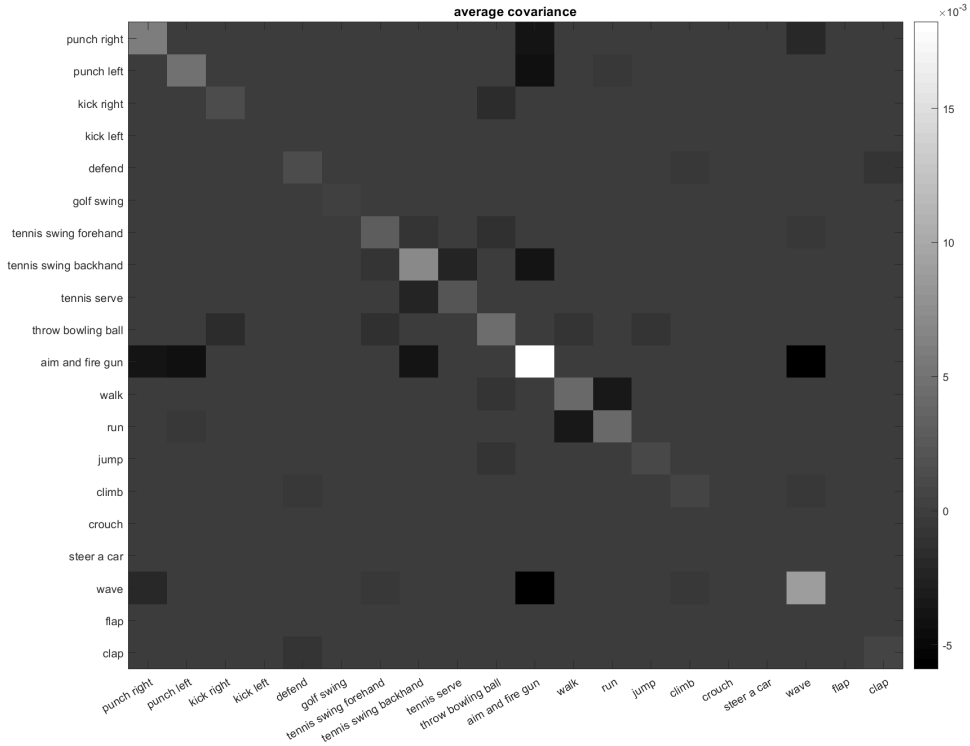


(b) Label distribution covariance

Figure 2. More results on MSRAction3D dataset [4].

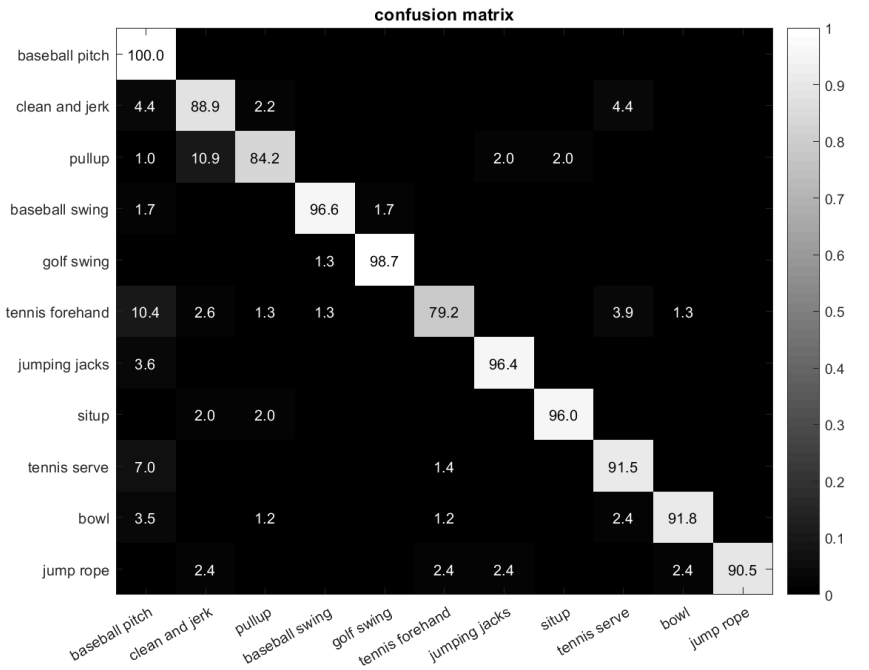


(a) Confusion matrix

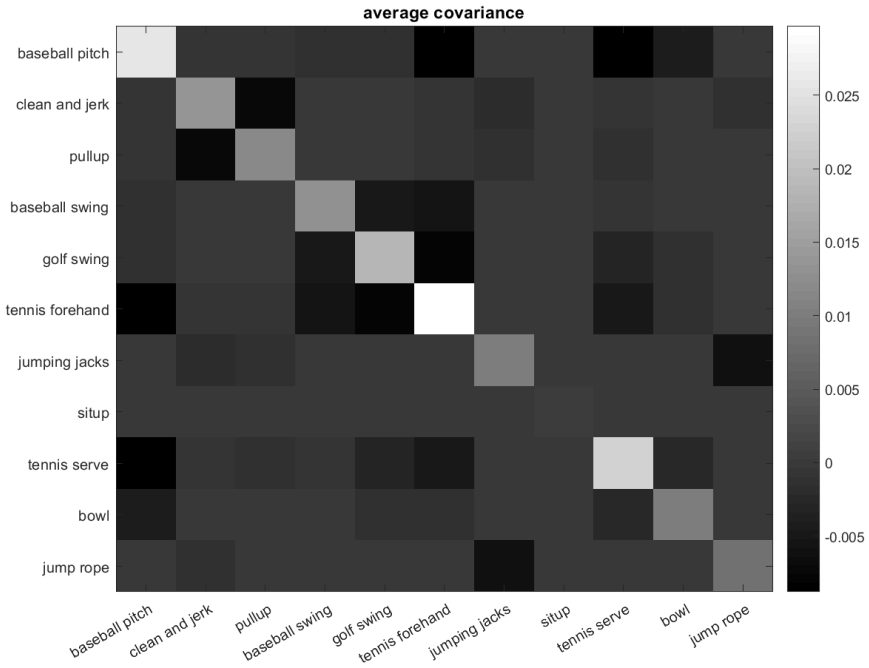


(b) Label distribution covariance

Figure 3. More results on G3D dataset [1].



(a) Confusion matrix



(b) Label distribution covariance

Figure 4. More results on Penn dataset [9].