An Accurate Evaluation of MSD Log-likelihood and its Application in Human Action Recognition

Nuha Zamzami and Nizar Bouguila
Concordia University, Montreal, QC, Canada

Objectives

1. To propose a new parameterization of the Multinomial Scaled Dirichlet (MSD) [1] log-likelihood function based on a truncated series consisting of Bernoulli polynomials.
2. To adopt the mesh algorithm for computing this log-likelihood used for parameters estimation within the mixture model framework.
3. To propose a model selection approach which is seamlessly integrated with the parameters estimation and avoids several drawbacks of the standard Expectation-Maximization algorithm.

Approximating the Paired Log-Gamma Difference

- The MSD likelihood function is given by:

$$\mathcal{L}(p,\psi,\beta;\mathbf{X}) = -\Big(\ln\Gamma(1/\psi+N)-\ln\Gamma(1/\psi)\Big) - \sum_{w=1}^{W} x_w \ln(\beta_w) + \sum_{w=1}^{W}\left(\ln\Gamma\Big(1/\frac{\psi}{p_w}+x_w\Big)-\ln\Gamma\Big(1/\frac{\psi}{p_w}\Big)\right) \quad (1)$$

where $N=\sum_{w=1}^{W} x_w$, $\psi = 1/A$ is the overdispersion parameter, and $p_w = \psi\alpha_w$.

- We use the approximation of the paired log-gamma difference method

$$\ln\Gamma(1/x+y)-\ln\Gamma(1/x) \approx -y\ln x + D_m(x,y) \quad (2)$$

when y is an integer, $|x|, \min(|y-1|,|y|)<1$, $xy\le\delta$ and:

$$D_m(x,y) = \sum_{n=2}^{m}\frac{(-1)^n\phi_n(y)}{n(n-1)}x^{(n-1)}$$

where $\phi_n(y) = B_n(y) - B_n$ is the old type Bernoulli polynomial, $B_n(y)$ and B_n indicate the nth Bernoulli polynomial, and nth Bernoulli number ($B_n = B_n(0)$), respectively.

The Mesh Algorithm for Evaluating the Log-likelihood

- First, generate the mesh using:

$$x_w^{(l)} = \lfloor \alpha_w^{(l-1)}\delta \rfloor \quad (3)$$

- Then, we select the level of the mesh L, so it would be the smallest integer satisfying:

$$\sum_{l=1}^{L} x_w^{(l)} \ge x_w \quad \text{for all } w = 1,\ldots,W$$

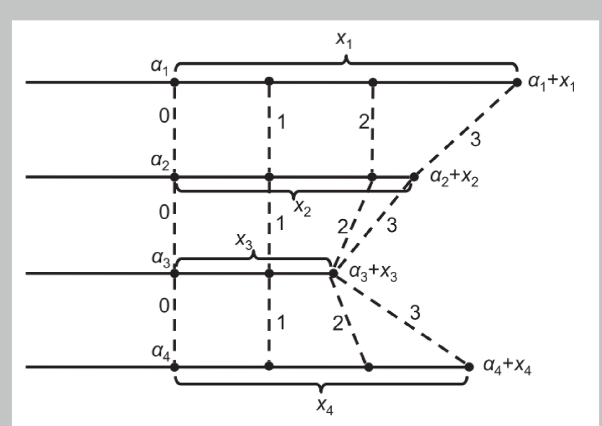

Figure 1: A graphical depiction of the mesh algorithm [2].

- Afterwards, we adjust $x_w^{L'_w}$ such that $\sum_{l=1}^{L'_w} x_w^{(l)} = x_w$, and all the remaining $x_w^{(l)}(l > L'_w)$ will be set to zero.
- With this adjusted mesh, we can use the approximation in (Eq. 2) to compute the MSD log-likelihood as the sum of each $X^{(l)}$ log-likelihood, as:

$$\mathcal{L}(p,\psi,\beta;X^{+}) = \sum_{l=1}^{L}\mathcal{L}(p^{(l-1)+},\psi^{(l-1)},\beta^{(l-1)+};X^{(l)+}) \quad (4)$$

where X^{+} is the vector of non-zero elements in X.

The Estimation and Selection Framework

- The algorithm starts with a large number of components and iteratively deletes components as they become irrelevant.
- The model selection is based on the minimum message length (MML) criterion, defined as:

$$\Theta_{MML} = \arg\min_{\Theta}\Big\{-\ln P(\Theta) - \mathcal{L}(\mathcal{X},\mathcal{Z}|\Theta) + \frac{1}{2}\ln|\mathbf{I}(\Theta)| + \frac{\mathcal{D}(\Theta)}{2}\Big(1+\ln\frac{1}{12}\Big)\Big\} \quad (5)$$

where $\mathcal{L}(\mathcal{X},\mathcal{Z}|\Theta)$ the complete-data log-likelihood, $P(\Theta)$ is the prior distribution, $\mathbf{I}(\Theta)$ is the expected Fisher information matrix, and $\mathcal{D}(\Theta)$ denotes the model dimensionality.

- We use the component-wise EM procedure (CEM), and any weak component will be annihilated.
- MML criterion is re-evaluated for non-zero components only until the message length difference becomes insignificant.

Results-Frame Level

- Frames extracted from the UCF sports dataset.
- Each extracted frame is treated as an image from which a set of interest points are detected and described using Scale-Invariant Feature Transform (SIFT), then represented as count vectors using Bag of Features (BoF) approach.

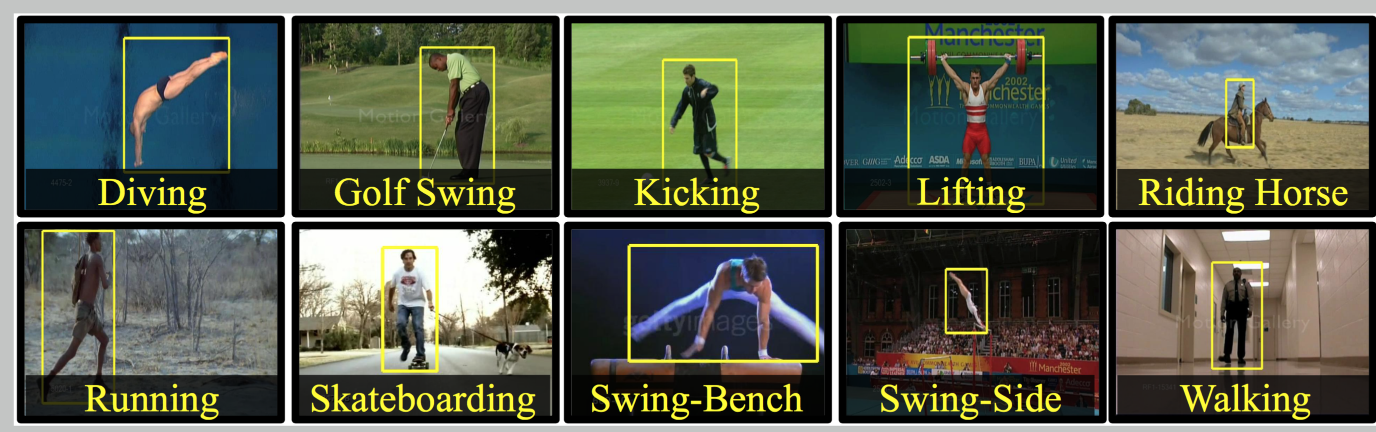

Figure 2: Sample frames from UCF sports dataset.

MM	DMN	MSD	MSD_Mesh
61.72%	66.97%	66.97%	**78.60%**

Table 1: The average recognition accuracy for UCF sports dataset.

*MM: Mixture of Multinomials, DMN: Mixture of Dirichlet-Multinomial, MSD: Mixture of MSD as in [1], MSD_Mesh: the proposed framework.

Results-Video Sequences

- Each video is represented as a vector of count data using the extension of Bag of words paradigm to videos.
- Detection of the local neighborhood with a significant variations is via Spatio Temporal Interest Points (STIP), then 3D SIFT descriptor is used.

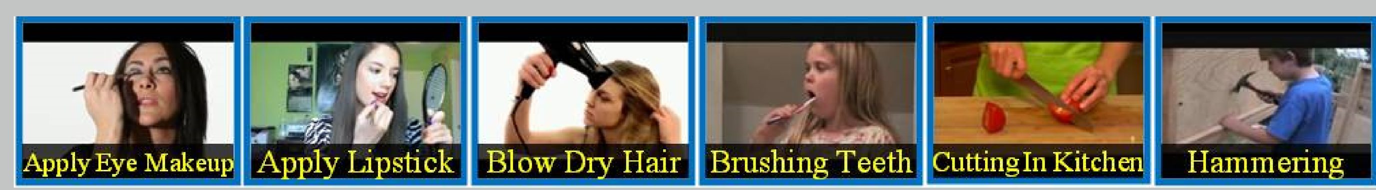

Figure 3: Sample Human-Object Interaction frames from UCF101.

MM	DMN	MSD	MSD_Mesh
81.77%	82.29%	84.90%	**87.76%**

Table 2: The average recognition accuracy for Human-Object Interaction subset.

Figure 4: Sample Playing Musical Instruments frames from UCF101.

MM	DMN	MSD	MSD_Mesh
81.89%	82.65%	89.54%	**92.60%**

Table 3: The average recognition accuracy for Playing Musical Instruments subset.

Conclusion

- The mesh method is generally more stable and provides an accurate computation of the log-likelihood function leads to a significant improvement in the clustering accuracy.
- The proposed algorithm successfully selected the optimal number of components, that agrees with the prespecified ones for different datasets.

References

[1] Nuha Zamzami and Nizar Bouguila.
Text modeling using multinomial scaled dirichlet distributions.
In *Recent Trends and Future Technology in Applied Intelligence. IEA/AIE 2018. Lecture Notes in Computer Science, vol 10868*, pages 69–80. Springer, 2018.

[2] Peng Yu and Chad A Shaw.
An efficient algorithm for accurate computation of the dirichlet-multinomial log-likelihood function.
Bioinformatics, 30(11):1547–1554, 2014.

[3] Mario A. T. Figueiredo and Anil K. Jain.
Unsupervised learning of finite mixture models.
IEEE Transactions on Pattern Analysis & Machine Intelligence, (3):381–396, 2002.

Contact Information

- Nuha Zamzami n_zamz@encs.concordia.ca