Deep Density Clustering of Unconstrained Faces  
(Supplementary Material)  

Wei-An Lin  Jun-Cheng Chen  Carlos D. Castillo  Rama Chellappa  
University of Maryland, College Park  

walin@umd.edu pullpull@cs.umd.edu carlos@cs.umd.edu rama@umiacs.umd.edu  

A. Mathematical Details  

In this section, we first provide the two core mathematical formulations and then present detailed proofs for Lemma 1 and Theorem 1.  

SVDD formulation:  

\[
\begin{align*}
\min_{c, R, \xi} & \quad R + \frac{1}{n_V} \sum_{z \in V(x)} \xi(z) \\
\text{s.t.} & \quad \|\Psi_\theta(z) - c\|^2 \leq R + \xi(z), \\
& \quad \xi \geq 0, \quad \forall z \in V(x), \\
\end{align*}
\]  

(1)  

OC-SVM formulation:  

\[
\begin{align*}
\min_{w, \rho, \xi} & \quad \frac{1}{2} \|w\|^2 + \frac{1}{n_V} \sum_{z \in V(x)} \xi - \rho \\
\text{s.t.} & \quad w^T \Psi_\theta(z) \geq \rho - \xi, \\
& \quad \xi \geq 0, \quad \forall z \in V(x). \\
\end{align*}
\]  

(2)  

A.1. Proof of Lemma 1  

Lemma 1. If $1/n_V < \nu \leq 1$, the SVDD formulation in (1) is equivalent to the OC-SVM formulation in (2) when the evaluation functions for the two are given by  

\[
\begin{align*}
h_{\text{SVDD}}(x) &= R^* - \|\Psi_\theta(x) - c^*\|^2, \\
h_{\text{OC-SVM}}(x) &= w^*^T \Psi_\theta(x) - \rho^*, \\
\end{align*}
\]  

(3)  

with the correspondence $w^* = c^*$, and $\rho^* = c^*^T \Psi_\theta(x_s)$, where $x_s$ is a support vector in (1) that lies on the learned enclosing sphere.  

Proof. The condition corresponds to the case $1/n_V \leq C < 1$ in [1] with $C = 1/(\nu \cdot n_V)$. We introduce the kernel function $K(x_i, x_j) = \Psi_\theta(x_i)^T \Psi_\theta(x_j)$. Since $K(x_i, x_i)$ is constant in our setting, the same dual formulation for (1) and (2) can be written as:  

\[
\begin{align*}
\min_{\alpha} & \quad \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \\
\text{s.t.} & \quad 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n_V} \alpha_i = 1. \\
\end{align*}
\]  

Let $S = \{i \mid 0 < \alpha_i < C\}$. We have the following results:  

\[
\begin{align*}
ce^* &= \sum_{i=1}^{n_V} \alpha_i \Psi_\theta(x_i), \quad R^* = \|\Psi_\theta(x) - c^*\|^2, \\
w^* &= \sum_{i=1}^{n_V} \alpha_i \Psi_\theta(x_i), \quad \rho^* = w^*^T \Psi_\theta(x_s), \\
\end{align*}
\]  

(4)  

where $s \in S$. Substituting into (3) and (4), we obtain  

\[
\begin{align*}
h_{\text{SVDD}}(x) &= 2 \cdot h_{\text{OC-SVM}}(x) = 2 \left[ \sum_{i=1}^{n_V} \alpha_i K(x_i, x) - \rho^* \right]. \\
\end{align*}
\]  

(5)  

A.2. Proof of Theorem 1  

Theorem 1. If $1/n_V < \nu \leq 1$ and $c^*^T \Psi_\theta(x_s) \neq 0$ for some support vector $x_s$, $h_{\text{SVDD}}(x)$ defined in (3) is asymptotically a Parzen window density estimator in the feature space with Epanechnikov kernel.  

Proof. Given the condition, according to Lemma 1, $h_{\text{SVDD}}(x)$ is equivalent to $h_{\text{OC-SVM}}(x)$ with $\rho^* \neq 0$. From the results in [10] and the fact that $\sum \alpha_i = 1$, we obtain:  

\[
\begin{align*}
h_{\text{OC-SVM}}(x) &= \sum_{i=1}^{n_V} \alpha_i \left[ 1 - \frac{1}{2} \|\Psi_\theta(x) - \Psi_\theta(x_i)\|^2 \right] - \rho^* \\
&= \frac{8}{3} \sum_{i=1}^{n_V} \alpha_i K_E \left( \frac{\|\Psi_\theta(x) - \Psi_\theta(x_i)\|}{2} \right) - \rho^* - 1, \\
\end{align*}
\]  

where $K_E(u) = \frac{3}{2}(1 - u^2)$, $|u| \leq 1$ is the Epanechnikov kernel. As a consequence of Proposition 4 in [10] and the proof of Proposition 1 in [11], as $n_V \to \infty$, the fraction of support vector is $\nu$, and the fraction of points with $0 < \alpha_i < 1/(\nu \cdot n_V)$ vanishes. Therefore, either $\alpha_i = 0$ or $\alpha_i = 1/(\nu \cdot n_V)$. We introduce the notation $\hat{S} = \{i \mid \alpha_i = \ldots$
\[ \frac{1}{(\nu \cdot n_V)} \}. \] Then asymptotically,
\[
h_{OC-SVM}(\mathbf{x}) = \frac{8}{3\nu n_V} \sum_{s \in S} K_E \left( \frac{||\Psi_{\theta}(x) - \Psi_{\theta}(x_s)||}{2} \right) - \rho^* - 1,
\]
\[
= \frac{2^{d+3}}{3} \hat{f}(\Psi_{\theta}(\mathbf{x})) - \rho^* - 1, \tag{8}
\]
where \( \hat{f}(z) = \frac{1}{\nu n_V z^2} \sum_{s \in S} K_E \left( \frac{||z - z_s||}{2} \right) \) is a density estimator. As a result, \( h_{SVDD}(\mathbf{x}) \) is equivalent to a Parzen window density estimator with Epanechnikov kernel of bandwidth 2. By scaling properly, Parzen window estimator with different bandwidths can be obtained.

\section*{B. Implementation Details}
We adopt the network architecture presented in [16]. The network is first trained on the CASIA-WebFace dataset [14] using SGD for 750K iterations with a standard batch size 128 and momentum 0.9. The learning rate is set to 0.01 initially and is halved every 100K iterations. The weight decay rates of all the convolutional layers are set to 0, and the weight decay of the final fully connected layer is set to \( 5 \times 10^{-4} \). Then, the model is finetuned with the MSCeleb-1M dataset [6] using the learning rate \( 1 \times 10^{-4} \) for all the convolutional layers, and \( 1 \times 10^{-2} \) for the fully connected layers. The network is then trained with additional 230K iterations. The inputs to the networks are 100×100×3 RGB images. Data augmentation is performed by randomly cropping and horizontally flipping face images. Given a face image, the deep representation is extracted from the pool5 layer with dimension 320.

\section*{C. Baseline Methods}
- Agglomerative Hierarchical Clustering (AHC) [5]: The conventional hierarchical clustering algorithm.
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [2]: A well-known density-based clustering method. We set the parameter MinPts = 5.
- Affinity Propagation (AP) [3]: Affinity Propagation groups data points based on the concept of “message passing”. It automatically finds exemplars and determines the number of clusters.
- Sparse Subspace Clustering using Orthogonal Matching Pursuit (SSC-OMP) [15]: SSC-OMP is a competitive method and runs faster than the classic SSC.
- Joint Unsupervised Learning of deep representations and clusters (JULE) [13]: JULE initializes each image as a cluster. It then iteratively merges images in feature space and updates network parameters.
- Deep Embedded Regularized Clustering (DEPICT) [4]: DEPICT is an efficient image clustering method that runs faster than JULE while attaining comparable performance.
- Proximity-Aware Hierarchical Clustering (PAHC) [7]: PAHC exploits neighborhood similarity based on linear SVMs. This method achieves high clustering performance on several face datasets.
- Approximate Rank-Order Clustering (ARO) [9]: ARO measures pairwise similarity based on shared nearest neighbors. The method is computationally efficient and is highly scalable.
- Conditional Pairwise Clustering (ConPaC) [12]: ConPaC builds a discriminative conditional random field model on the adjacency matrix, and then infers the parameters using the loopy belief propagation. This method outperforms several clustering algorithms on challenging face datasets.

\section*{D. Additional Evaluations on the IJB-B dataset}
Table 1 reports the F-measure and NMI comparisons on the three subtasks in IJB-B. Table 2 summarizes the statistics of these subtasks.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline
Dataset & \textbf{IJB-B-32} & & \textbf{IJB-B-64} & & \textbf{IJB-B-512} & \\
\hline
\textbf{F} & 0.659 & 0.806 & 0.677 & 0.837 & 0.555 & 0.839 \\
\textbf{NMI} & 0.845 & 0.915 & 0.814 & 0.912 & 0.746 & 0.918 \\
\hline
AP [3] & 0.513 & 0.814 & 0.508 & 0.831 & 0.422 & 0.847 \\
DBSCAN [2] & 0.825 & 0.896 & 0.751 & 0.885 & 0.696 & 0.888 \\
SSC-OMP [15] & 0.361 & 0.575 & 0.275 & 0.539 & 0.111 & 0.521 \\
PAHC [7] & 0.798 & 0.891 & 0.786 & 0.898 & 0.650 & 0.882 \\
ARO* [9] & 0.667 & - & 0.574 & - & 0.410 & - \\
ConPaC* [12] & 0.751 & - & 0.656 & - & 0.481 & - \\
\hline
DDC & 0.827 & 0.906 & 0.783 & 0.903 & 0.733 & 0.909 \\
DDC-NEG & 0.851 & 0.919 & 0.818 & 0.915 & 0.761 & 0.918 \\
\hline
\end{tabular}
\caption{BCubed F-measure and NMI performance comparisons. Results reported from the original papers are marked by asterisks (*). The best performance is reported in \textbf{bold.}}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
\textbf{Dataset} & \textbf{# Samples} & \textbf{# Subjects} \\
\hline
\textbf{IJB-B-32} & 1,026 & 32 \\
\textbf{IJB-B-64} & 2,080 & 64 \\
\textbf{IJB-B-512} & 18,251 & 512 \\
\hline
\end{tabular}
\caption{Statistics for the three IJB-B subtasks.}
\end{table}

\section*{References}