Is that you?
Metric Learning Approaches for Face Identification

M. Guillaumin, J. Verbeek and C. Schmid
LEAR team, INRIA Rhône-Alpes, France

Supplementary Material

Summary

In this document we present a more complete overview of the clustering using LDML+MkNN.

In the first part, we show all the clusters of the best clustering obtained using LDML+MkNN on our presented experiment on *Labeled Faces in the Wild*. In Figure 1, we reproduce Figure 7 from the submitted paper. The "random" curve is an average curve corresponding to 10 random hierarchical clustering.

The second part of the supplementary material we derive the minimum and maximum labeling cost for all possible clustering sizes.

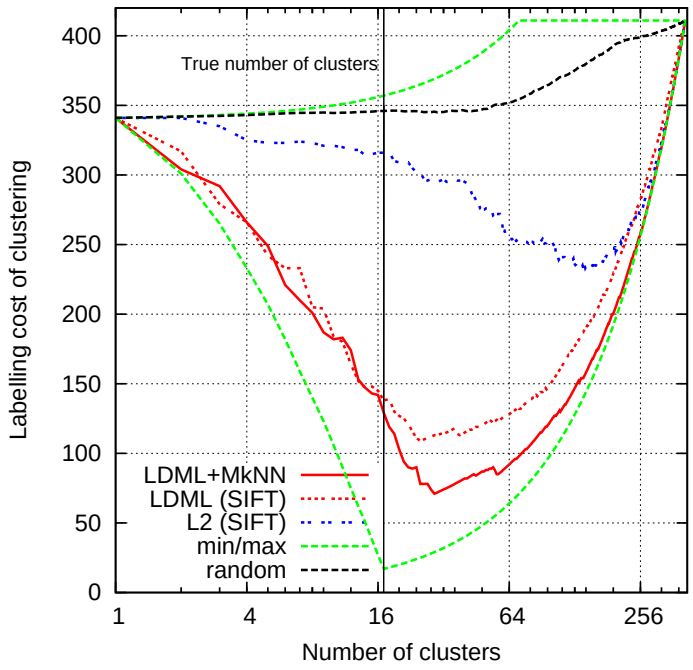


Figure 1: Experimental and theoretical plots of the labelling cost of some clusterings with respect to the number of clusters of a hierarchical clustering. Theoretical bounds correspond to Equation 14 and Equation 16. Relevant constants are $N = 411$, $n_1 = 71$, and $I = 17$.

LDML+MkNN Clusters

About the layout

The clusters are sorted with descending "pureness", *i.e* the frequency of the most frequent person. Ties are broken using the cluster size. The cluster images are displayed from left to right, from top to bottom. Images are grouped by their true identity, and groups are sorted by descending sizes. Hence, the most frequent person's images appear first, then the second most frequent, and so on. Spaces are inserted between groups to enhance clarity. Between the first and the second group, even more spaces are added to obtain a full blank line between the most frequent person and other persons. Clusters 1 to 13 are pure, and only clusters 27 and 29 can be considered arbitrary. Only three clusters have a "pureness" equal or below 0.5, where the most frequent person does not account for at least half of the cluster.

Cluster 1

Frequency of the most frequent person: 1.000

Cluster 2

Frequency of the most frequent person: 1.000

Cluster 3

Frequency of the most frequent person: 1.000

Cluster 4

Frequency of the most frequent person: 1.000

Cluster 5

Frequency of the most frequent person: 1.000

Cluster 6

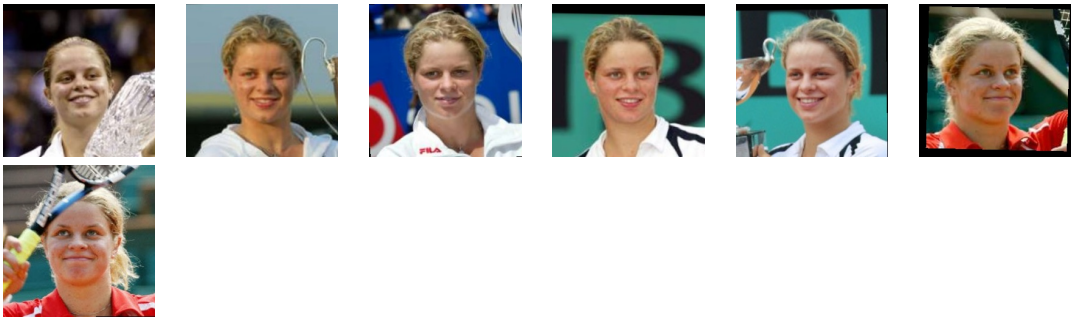
Frequency of the most frequent person: 1.000

Cluster 7

Frequency of the most frequent person: 1.000

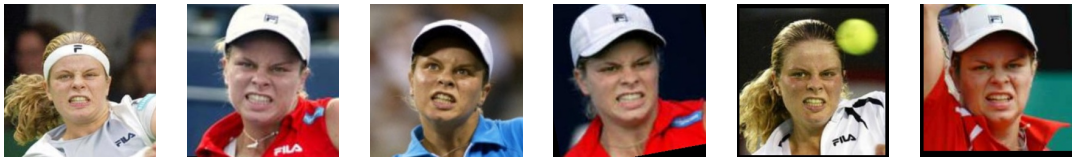
Cluster 8

Frequency of the most frequent person: 1.000



Cluster 9

Frequency of the most frequent person: 1.000



Cluster 10

Frequency of the most frequent person: 1.000

Cluster 11

Frequency of the most frequent person: 1.000

Cluster 12

Frequency of the most frequent person: 1.000


Cluster 13

Frequency of the most frequent person: 1.000

Cluster 14

Frequency of the most frequent person: 0.923

Cluster 15

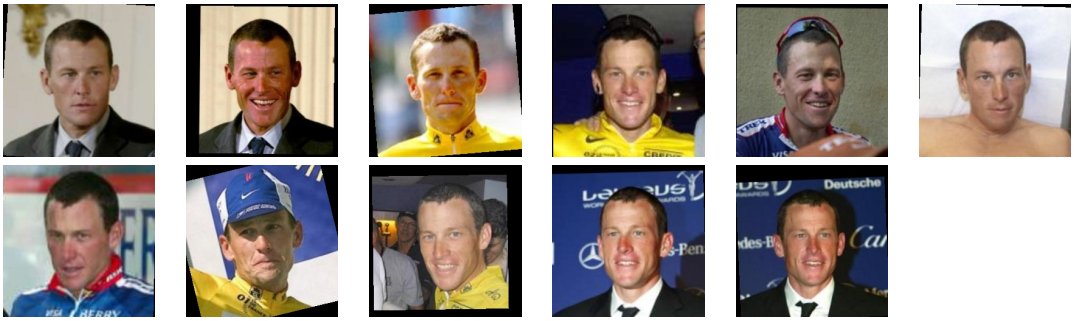
Frequency of the most frequent person: 0.923

Cluster 16

Frequency of the most frequent person: 0.917

Cluster 17

Frequency of the most frequent person: 0.917



Cluster 18

Frequency of the most frequent person: 0.906

Cluster 19

Frequency of the most frequent person: 0.875

Cluster 20

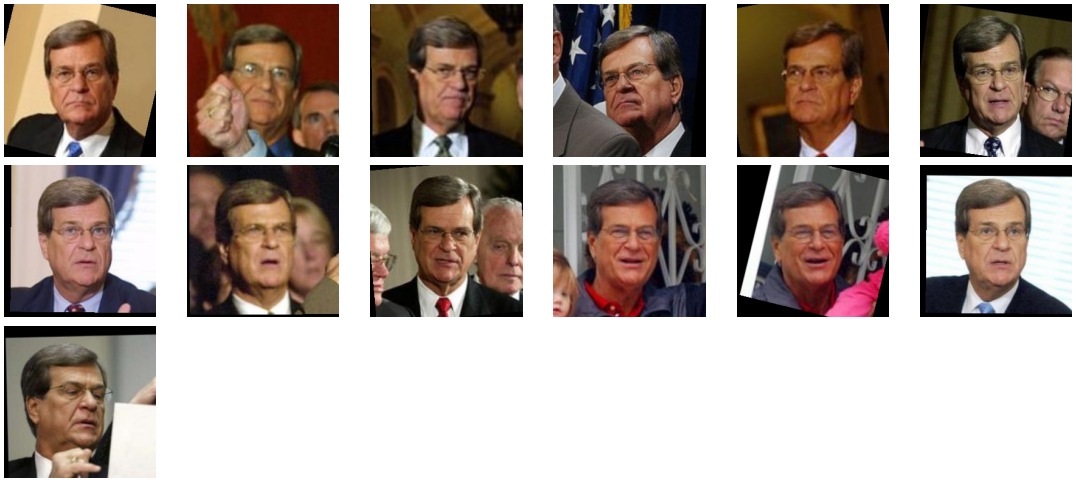
Frequency of the most frequent person: 0.857

Cluster 21

Frequency of the most frequent person: 0.833

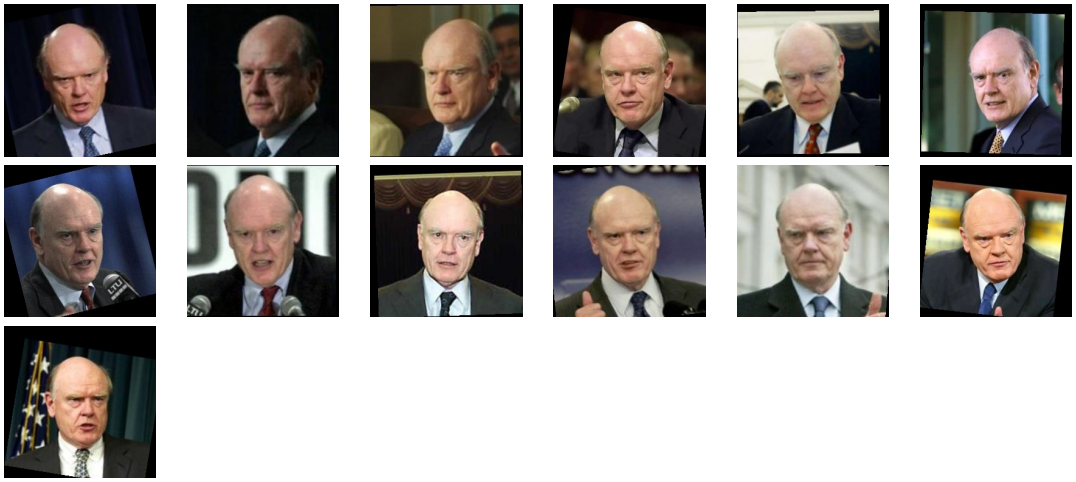
Cluster 22

Frequency of the most frequent person: 0.812



Cluster 23

Frequency of the most frequent person: 0.812

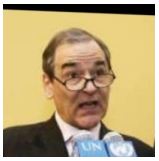


Cluster 24

Frequency of the most frequent person: 0.792

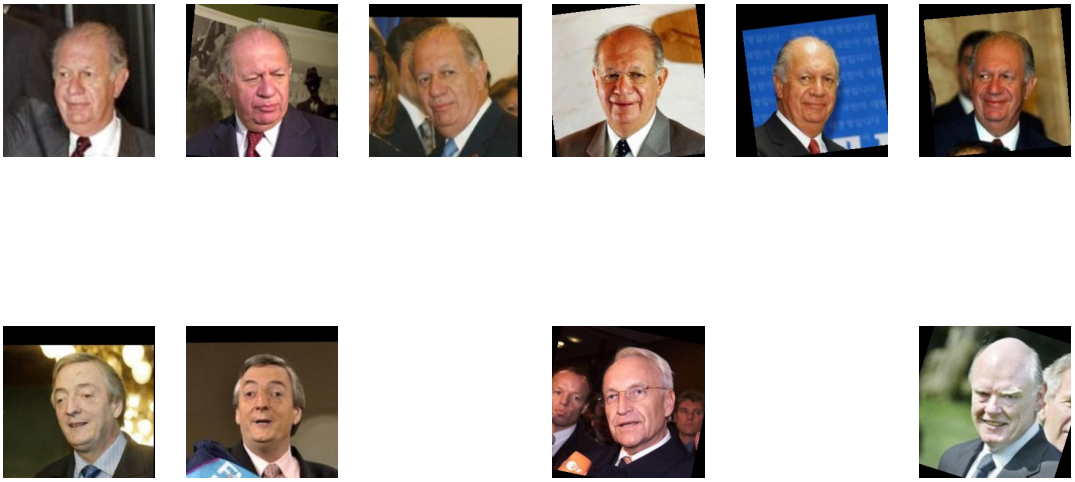
Cluster 25

Frequency of the most frequent person: 0.750

Cluster 26

Frequency of the most frequent person: 0.600



Cluster 27

Frequency of the most frequent person: 0.500

Cluster 28

Frequency of the most frequent person: 0.467

Cluster 29

Frequency of the most frequent person: 0.167

Labelling cost extrema

Here we formally derive the minimal and maximal labeling cost, for arbitrary size of the clustering. As stated in subsection 5.2, we define the labelling cost $\mathcal{L}(c)$ of a cluster c as

$$\mathcal{L}(c) = 1 + \left(n^c - \max_i n_i^c\right), \tag{1}$$

where n_i^c denotes the number of images with label i in cluster c and n^c is the size of cluster c.

Let N be the total number of images and C denote a clustering as a set of clusters, then $|C|$ is the size of C and therefore also the number of clusters. The labelling cost of the clustering C is the sum of labelling costs of its clusters:

$$\begin{aligned} \mathcal{L}(C) &= \sum_{c\in C}\left(1 + (n^c - \max_i n_i^c)\right) && (2)\\ &= |C| + N - \sum_{c\in C}\max_i n_i^c. && (3) \end{aligned}$$

Let us order the class labels by their frequency: label 1 is the most frequent, label 2 the second-most frequent, and so on (up to, say, label I who is the least frequent). Let n_i be the number of images for label i, i.e. $n_1 \geq n_2 \cdots \geq n_I$. Then the following properties straightforwardly hold:

$$\forall i, c, \quad n_i^c \leq n_i \leq n_1 \tag{4}$$

$$\forall i, C, \quad \sum_{c\in C} n_i^c = n_i \tag{5}$$

$$\forall c, \quad \max_i n_i^c \leq n_1 \qquad \text{with equality iff all images with class label 1 are in } c \tag{6}$$

$$\forall C, \quad \sum_{c\in C}\max_i n_i^c \leq |C| n_1 \tag{7}$$

$$\forall C \text{ s.t. } |C| \leq I, \quad \sum_{c\in C}\max_i n_i^c \leq \sum_{i=1}^{|C|} n_i \tag{8}$$

$$\forall C, \quad \sum_{c\in C}\max_i n_i^c \leq N \tag{9}$$

$$\forall C, \quad \mathcal{L}(C) \geq |C| \tag{10}$$

$$\forall C, \quad \mathcal{L}(C) \leq N \tag{11}$$

$$\forall k \leq n_1, \exists C \text{ s.t:} \quad |C| = k \quad \text{and} \quad \forall c, \max_i n_i^c = n_1^c \tag{12}$$

Given a number of clusters $|C|$, the exact lower bound is found by maximizing $\sum_c \max_i n_i^c$. Actually, Equation 8 is an attainable bound. Consider $|C| \leq I$ clusters, and separate the $|C|$ most frequent class labels in individual clusters. Then place the remaining images in any cluster. Then $\sum_{c \in C} \max_i n_i^c = \sum_{i=1}^{|C|} n_i$ and the minimum $\underline{\mathcal{L}}(|C|)$ of $\mathcal{L}(C)$ is

$$\forall C \text{ s.t. } |C| \leq I, \quad \underline{\mathcal{L}}(|C|) = |C| + N - \sum_{i=1}^{|C|} n_i. \tag{13}$$

When $|C| = I$, $\underline{\mathcal{L}}(|C|) = \underline{\mathcal{L}}(I) = I + N - N = I$. The cost is exactly the number of clusters, and the number of different class labels: all clusters are pure. Let us call C^* this clustering. When $|C| > I$, the bound in Equation 10 is reachable by splitting pure clusters of C^* to obtain $|C|$ clusters. Notice that these lower bounds can be seen as a hierachical clustering of a perfect similarity matrix.

Finally, the minimum as a function of the clustering size is given by

$$\underline{\mathcal{L}}(|C|) = \max \left(|C|,\ |C| + N - \sum_{i=1}^{|C|} n_i \right). \tag{14}$$

As for the maximum $\overline{\mathcal{L}}(|C|)$, we now want to minimize $\sum_c \max_i n_i^c$. When $|C| \geq n_1$, we can actually have $\forall c, \max_i n_i^c = 1$, and we can therefore reach the upper bound of Equation 11, by simply ensuring that at most one image of any class label is in a cluster. This is possible because we have enough clusters, even for the most frequent label. Thus

$$\forall C \text{ s.t. } |C| \geq n_1, \quad \overline{\mathcal{L}}(|C|) = N. \tag{15}$$

When $|C| \leq n_1$, we can use Equation 12 to assert that we can achieve a labeling cost of $|C| + N - n_1$, which happens when 1 is the most frequent label in each cluster (there might be other labels equally frequent in those clusters). Next, we show that a bigger cost can not be achieved. For this we look at clusterings with clusters where 1 is not the most frequent label. Let c be such a cluster, and let j be the most frequent label in that cluster. Then there exists a cluster c' with $n_j^{c'} < n_1^{c'}$, otherwise we contradict the fact that label 1 is the most frequent label overall. Now, we can increase the labeling cost by moving an element with label j from cluster c to cluster c', to see this note that all terms in the sum $\sum_c \max_i n_i^c$ stay constant except for c which is reduced by 1.

This finalizes our proof that for a clustering of size $|C|$ the maximum label cost is given by

$$\overline{\mathcal{L}}(|C|) = \min \left(N,\ |C| + N - n_1 \right). \tag{16}$$