Fantope Regularization in Metric Learning

Marc T Law, Nicolas Thome, Matthieu Cord

Fantope Regularization in Metric Learning

Marc T. Law Nicolas Thome Matthieu Cord

Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France

Abstract

This paper introduces a regularization method to explicitly control the rank of a learned symmetric positive semidefinite distance matrix in distance metric learning. To this end, we propose to incorporate in the objective function a linear regularization term that minimizes the k smallest eigenvalues of the distance matrix. It is equivalent to minimizing the trace of the product of the distance matrix with a matrix in the convex hull of rank-k projection matrices, called a Fantope. Based on this new regularization method, we derive an optimization scheme to efficiently learn the distance matrix. We demonstrate the effectiveness of the method on synthetic and challenging real datasets of face verification and image classification with relative attributes, on which our method outperforms state-of-the-art metric learning algorithms.

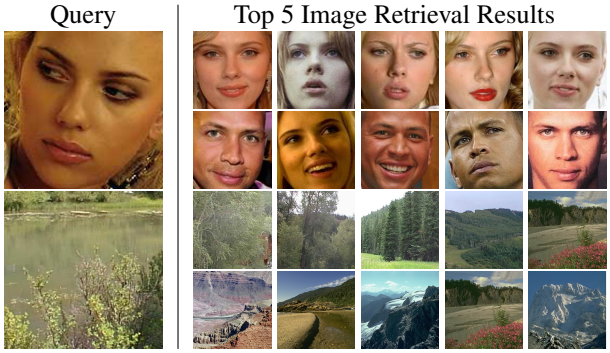

Figure 1. Top 5 similarity search for two queries from the *Public Figure Face* and *Outdoor Scene Recognition* datasets. We show for each query the 5 most similar images using our metric learning approach (first row), and the well-known metric learning approach LMNN (second row). On these examples, our scheme performs better and succeeds to return semantically relevant images. This shows the importance of the proposed regularization scheme to learn a meaningful distance matrix and limit overfitting.

1. Introduction

Distance metric learning is useful for many Computer Vision tasks, such as image classification [14, 17, 26], retrieval [3, 8] or face verification [10, 18]. It emerges as a promising learning paradigm, in particular because of its ability to learn with attributes [20], further offering the appealing possibility to perform zero-shot learning, or to generalize to new classes at near zero cost [17].

Metric learning algorithms produce a linear transformation of data which is optimized to fit semantical relationships between training samples. Different aspects of the learning procedure have recently been investigated: how the dataset is annotated and used in the learning process, *e.g.* using pairs [18], triplets [21] or quadruplets [13] of samples; design choices for the distance parameterization; extensions to large scale context [17], *etc*. Surprisingly, few attempts have been made for deriving a proper regularization scheme, especially in the Computer Vision literature. Regularization in metric learning is however a critical issue, as it often limits model complexity, the number of independent parameters to learn, and thus overfitting. Models learned with regularization usually better exploit correlations between features and often have improved predictive accuracy [14].

In this paper, we propose a novel regularization approach for metric learning that explicitly controls the rank of the learned distance matrix. Figure 1 illustrates the relevance of our approach. We present retrieval results after metric learning with the proposed method, and provide an illustrative comparison with LMNN [26], which is one of the most popular non-regularized metric learning algorithms. The regularization scheme introduced in this paper significantly improves the performance of the semantical visual search.

The remainder of the paper is organized as follows. Section 2 positions the paper with respect to related works. Our regularization framework is introduced in Section 3 and the resulting optimization scheme in Section 4. Section 5 presents toy experiments to grasp the meaning of the proposed regularization. Section 6 demonstrates the effectiveness of our metric learning scheme in two challenging computer vision applications. Finally, Section 7 concludes the paper and gives directions for future work.

Notations: let $\mathbb{S}^d$ and $\mathbb{S}^d_+$ denote the sets of $d \times d$ real-valued symmetric and symmetric positive semidefinite (PSD) matrices, respectively. For matrices $\mathbf{A} \in \mathbb{S}^d$ and $\mathbf{B} \in \mathbb{S}^d$, denote the Frobenius inner product by $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^\top \mathbf{B})$ where tr denotes the trace of a matrix. $\Pi_{\mathbb{S}^d_+}(\mathbf{A})$ is the orthogonal projection of the matrix $\mathbf{A} \in \mathbb{S}^d$ onto the positive semidefinite cone $\mathbb{S}^d_+$. For a given vector $\mathbf{a} = (a_1, \ldots, a_d)^\top \in \mathbb{R}^d$, $\text{Diag}(\mathbf{a}) = \mathbf{A} \in \mathbb{S}^d$ corresponds to a square diagonal matrix such that $\forall i, A_{i,i} = a_i$. $\lambda(\mathbf{A})$ is the vector of eigenvalues of matrix $\mathbf{A}$ arranged in non-increasing order. $\lambda(\mathbf{A})_i$ is the i-th largest eigenvalue of $\mathbf{A}$. $\mathbf{x}_i \in \mathbb{R}^d$ (resp. $\mathbf{x}_j \in \mathbb{R}^d$) is the vector representation of image p_i (resp. p_j) and we note $\mathbf{x}_{ij} = (\mathbf{x}_i - \mathbf{x}_j)$. Finally, for $x \in \mathbb{R}$, let $[x]_+ = \max(0, x)$.

2. Related work

Image representation for classification has been deeply investigated in recent years [4, 19]. The traditional Bag-of-Words representation [24] has been extended for the coding step [9, 28] as well as for the pooling [1], or with bio-inspired models [22, 25]. Nonetheless, similarity metrics are also crucial to compare, classify and retrieve images.

We focus in this work on supervised distance metric learning methods. Some of them consider sets of similar and dissimilar pairs of images for training [6, 18, 27]. They learn a distance metric that preserves distance relations among the training data. Other methods consider triplets [3, 8, 21, 26] of images, which are easy to generate in classification. For instance, LMNN [26] learns a distance metric for k-Nearest Neighbors (k-NN) approach using those triplet-wise training sets.

In this paper, we consider the widely used Mahalanobis distance metric $D_\mathbf{M}$ that is parameterized by the PSD matrix $\mathbf{M} \in \mathbb{S}^d_+$ such that $D^2_\mathbf{M}(p_i, p_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}_{ij})^\top \mathbf{M} \mathbf{x}_{ij}$. It can also be rewritten:

$$D^2_\mathbf{M}(p_i, p_j) = \langle \mathbf{M}, \mathbf{x}_{ij} \mathbf{x}_{ij}^\top \rangle \tag{1}$$

In Computer Vision, many approaches do not learn the Mahalanobis distance matrix $\mathbf{M}$ explicitly, but prefer working on a specific matrix decomposition: *i.e.* $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$ where $\mathbf{L} \in \mathbb{R}^{e \times d}$ and d is the data dimension. An objective function to minimize over $\mathbf{L}$ is defined using a loss function expressed over the different constraints of the training set [17, 18]. Although the resulting optimization is very fast, it is not convex w.r.t. $\mathbf{L}$, leading to many local minima with different objective values that depend on the initialization of $\mathbf{L}$. In addition, an explicit regularization term is rarely introduced in the learning scheme. For instance, that lack of regularization makes LMNN prone to overfitting [3]. To limit this shortcoming, many approaches [17, 18, 26] perform *early stopping* which stops an iterative optimization process before convergence. However, this method needs to be carefully tuned for each dataset.

Different types of regularization in the objective function defined over $\mathbf{M} \in \mathbb{S}^d_+$ have been proposed in the machine learning literature. Schultz and Joachims [21] use the squared Frobenius norm $\|\mathbf{M}\|^2_F$, following the SVM framework to learn a diagonal PSD distance matrix. However, the diagonal form of their model does not benefit from correlations between data. The ITML method (*Information-Theoretic Metric Learning* [6]) uses a LogDet regularizer that constrains the distance matrix to be strictly positive definite, which in practice often results in high-rank solutions that are subject to overfitting. Another powerful way to regularize, is to control the rank of $\mathbf{M}$. Imposing a low-rank solution limits the number of free parameters in the metric, and hence prevents overfitting. To that end, some methods [14, 16, 23] add the trace $\text{tr}(\mathbf{M})$ as a regularization term, because it is a convex surrogate for $\text{rank}(\mathbf{M})$. However, it does not allow an explicit control over the rank of $\mathbf{M}$: the trace of the distance matrix reaches its minimum possible value iff the distance matrix is a zero matrix. In practice, this trivial solution is never obtained because of the associated constraints.

In this paper, we investigate a new optimization scheme with a regularization term that explicitly controls the rank of $\mathbf{M}$. Such a scheme allows to avoid overfitting without any trick such as *early stopping*. The main contributions of this paper are: 1) We introduce a new regularization strategy based on the convex hull of rank-k projection matrices, called Fantope, which allows to explicitly control the rank of distance matrices. 2) We propose an efficient algorithm to solve the new optimization scheme. 3) Our framework outperforms state-of-the-art metric learning methods on synthetic and challenging real Computer Vision datasets.

3. Metric learning Fantope regularization

Objective function: a metric learning algorithm aims at determining $\mathbf{M}$ such that the metric satisfies most of the constraints defined by the training information. It is generally formulated as an optimization problem of the form:

$$\min_\mathbf{M} \mu R(\mathbf{M}) + \ell(\mathbf{M}, \mathcal{A}) \tag{2}$$

where $\ell(\mathbf{M}, \mathcal{A})$ is a loss function that penalizes constraints that are not satisfied, $R(\mathbf{M})$ is a regularization term on the parameter $\mathbf{M}$ of the metric, and $\mu \geq 0$ is the regularization parameter. $\ell(\mathbf{M}, \mathcal{A})$ measures the ability of the matrix $\mathbf{M}$ to satisfy some distance constraints given in the training set. The type of constraints depends on the way relationships between training samples are provided, *e.g.* relations between pairs, triplets, quadruplets [13] *etc*. The details on the design of the set $\mathcal{A}$ and the loss $\ell(\mathbf{M}, \mathcal{A})$ are specified in Section 4.1. In this paper, we focus on defining an effective regularization term $R(\mathbf{M})$.

3.1. Motivation for the proposed regularization

As mentioned in Section 2, controlling the rank of the PSD distance matrix $\mathbf{M}$ is a powerful way to limit overfitting and to better exploit correlations between features. A standard way to promote low-rank solutions is to use the nuclear norm $\|\mathbf{M}\|_*$ as a regularization term. In the case of PSD matrices, the nuclear norm corresponds to the trace: $\forall \mathbf{M} \in \mathbb{S}_+^d, \|\mathbf{M}\|_* = \text{tr}(\mathbf{M})$. However, trace(-norm) regularization is somewhat limited as it seeks a rank-0 matrix (*i.e.* $\mathbf{M} = \mathbf{0}$). Alternatively, we propose a regularization term that reaches its minimum when the rank of the learned PSD matrix is smaller or equal to a fixed target rank. We then formulate the regularization term $R(\mathbf{M})$ as the sum of the k smallest eigenvalues of $\mathbf{M} \in \mathbb{S}_+^d$:

$$R(\mathbf{M}) = \sum_{i=d-k+1}^{d} \lambda(\mathbf{M})_i \tag{3}$$

Such a minimization of $R(\mathbf{M})$ will naturally converge to a subspace corresponding to the $(d-k)$ most significant eigenvalues. As the rank of the PSD matrix $\mathbf{M} \in \mathbb{S}_+^d$ is the number of its non-zero eigenvalues and all the eigenvalues of $\mathbf{M} \in \mathbb{S}_+^d$ are non-negative, the proposed regularization term $R(\mathbf{M})$ allows an explicit control over the rank of $\mathbf{M}$:

$$R(\mathbf{M}) \text{ equals 0 iff } \text{rank}(\mathbf{M}) \leq d-k \tag{4}$$

We explain in the following how to express $R(\mathbf{M})$ in a convenient way.

3.2. Explicit rank control regularization

Using Ky Fan's theorem [7], we can rewrite the sum of the k smallest eigenvalues of any symmetric matrix $\mathbf{M}$ as the trace $\text{tr}(\mathbf{WM})$ where $\mathbf{W}$ is in the convex hull of the set comprising outer product of orthonormal matrices (rank-k projection matrices). This convex hull is called a Fantope [5]. Our regularization term (Eq. (3)) may be expressed as:

$$R(\mathbf{M}) = \text{tr}(\mathbf{WM}) = \langle \mathbf{M}, \mathbf{W} \rangle \tag{5}$$

where the matrix $\mathbf{W} \in \mathbb{S}_+^d$ (in a Fantope) allows to project the matrix $\mathbf{M}$ onto the target k-dimensional subspace.

A simple way to construct such a matrix $\mathbf{W} \in \mathbb{S}_+^d$ is to use the eigendecomposition of $\mathbf{M} \in \mathbb{S}_+^d$: $\mathbf{M} = \mathbf{V_M}\text{Diag}(\lambda(\mathbf{M}))\mathbf{V_M}^\top$ where $\mathbf{V_M}$ is an orthogonal matrix. As $\lambda(\mathbf{M})$ is arranged in non-increasing order, a simple threshold allows to project data onto the subspace generated by the k eigenvectors corresponding to the k smallest eigenvalues. Let us construct $\mathbf{w} = (w_1, \ldots, w_d)^\top \in \mathbb{R}^d$ such that:

$$w_i = \begin{cases} 0 \text{ if } 1 \leq i \leq d-k \text{ (the first } d-k \text{ elements)} \\ 1 \text{ if } d-k+1 \leq i \leq d \text{ (the last } k \text{ elements)} \end{cases} \tag{6}$$

We then express $\mathbf{W}$ as:

$$\mathbf{W} = \mathbf{V_M}\text{Diag}(\mathbf{w})\mathbf{V_M}^\top \tag{7}$$

From Eq. (7), it is simple to verify that the definition of $R(\mathbf{M})$ in Eq. (5) matches with the one in Eq. (3):

$$R(\mathbf{M}) = \text{tr}(\mathbf{WM}) = \text{tr}(\mathbf{V_M}\text{Diag}(\mathbf{w})\mathbf{V_M}^\top\mathbf{V_M}\text{Diag}(\lambda(\mathbf{M}))\mathbf{V_M}^\top)$$
$$= \text{tr}(\text{Diag}(\mathbf{w})\text{Diag}(\lambda(\mathbf{M}))) = \mathbf{w}^\top\lambda(\mathbf{M}) = \sum_{i=d-k+1}^{d} \lambda(\mathbf{M})_i$$

As the last k elements of $\lambda(\mathbf{M})$ (the k smallest eigenvalues of $\mathbf{M}$) equal 0 iff $\text{rank}(\mathbf{M}) \leq d-k$, one can deduce the expected property given in Eq. (4) that $R(\mathbf{M}) = 0$ iff the rank of $\mathbf{M}$ is smaller or equal to $d-k$.

Fantope regularization is a generalization of trace regularization. Indeed, for every matrix $\mathbf{M} \in \mathbb{S}_+^d$, $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_d\mathbf{M})$. Trace regularization is equivalent to a Fantope regularization where $\text{tr}(\mathbf{WM})$ is the sum of the d smallest eigenvalues of $\mathbf{M}$ ($\mathbf{W} = \mathbf{V_M}\text{Diag}(\mathbf{1})\mathbf{V_M}^\top = \mathbf{I}_d$).

It is also worth noting that $\mathbf{W}$ could be fixed in the convex hull of rank-k projection matrices without exploiting the eigendecomposition of $\mathbf{M}$ (as constructed in Eq. (7)). In this case, a (strictly) positive value of $R(\mathbf{M}) = \text{tr}(\mathbf{WM})$ is not necessarily the sum of the k smallest eigenvalues of $\mathbf{M}$. However, if $\text{tr}(\mathbf{WM})$ equals 0, then $R(\mathbf{M})$ includes the sum of the k smallest eigenvalues of $\mathbf{M}$ and the rank of $\mathbf{M}$ is then smaller or equal to $d-k$ [5].

4. Metric learning optimization algorithm

4.1. Optimization problem

Constraints: we focus on quadruplet-wise constraints [13] that encompass pairwise and triplet-wise constraints. They involve distance comparisons of the form $D(p_k, p_l) > D(p_i, p_j)$ for any quadruplet of images $q = (p_i, p_j, p_k, p_l)$. Our goal is to learn a metric $D_\mathbf{M}$ parameterized by $\mathbf{M}$ that satisfies the following constraint for all q in a training set $\mathcal{A}$:

$$\forall q \in \mathcal{A}, D^2_\mathbf{M}(p_k, p_l) \geq \delta_q + D^2_\mathbf{M}(p_i, p_j) \tag{8}$$

where δ_q is a safety margin specific to each quadruplet q. The triplet constraint $D^2_\mathbf{M}(p_i, p_k) \geq 1 + D^2_\mathbf{M}(p_i, p_j)$ can be trivially obtained from Eq. (8) with $q = (p_i, p_j, p_i, p_k)$ and $\delta_q = 1$. The formulation in Eq. (8) is also able to express relationships between a set of similar pairs $\mathcal{S}$ or dissimilar pairs $\mathcal{D}$, as used for example in [6, 18]. The dissimilar pair $(p_i, p_j) \in \mathcal{D}$ can be integrated with $q = (p_i, p_i, p_i, p_j)$ and $\delta_q = l$ leading to the constraint $D^2_\mathbf{M}(p_i, p_j) \geq l$ where l is the minimum value to consider images p_i and p_j as dissimilar. In the same way, the similar pair $(p_i, p_j) \in \mathcal{S}$ can be integrated with $q = (p_i, p_j, p_i, p_i), \delta_q = -u$, leading to the constraint $u \geq D^2_\mathbf{M}(p_i, p_j)$ where u is a given upper bound that enforces the distance between two similar images p_i

and p_j to be smaller than the given threshold u. We specify in the experiments (Section 6) how l and u are defined.

Using Eq. (1), our quadruplet-wise constraints in Eq. (8) using $q = (p_i, p_j, p_k, p_l) \in \mathcal{A}$ can be rewritten:

$$\forall q \in \mathcal{A}, \langle \mathbf{M}, \mathbf{x}_{kl}\mathbf{x}_{kl}^\top - \mathbf{x}_{ij}\mathbf{x}_{ij}^\top \rangle \geq \delta_q \tag{9}$$

Optimization: in order to learn a metric $D_\mathbf{M}$ that obeys the constraints in Eq. (9), we define a global loss $\ell(\mathbf{M}, \mathcal{A}) = \sum_{q\in\mathcal{A}} \ell_\mathbf{M}(q)$ that accumulates losses over all the quadruplets in the training set $\mathcal{A}$. We design the loss for a single quadruplet: $\ell_\mathbf{M}(q) = \max\left(0, \delta_q + \langle \mathbf{M}, \mathbf{x}_{ij}\mathbf{x}_{ij}^\top - \mathbf{x}_{kl}\mathbf{x}_{kl}^\top \rangle\right)$. By including our regularization term and $\ell(\mathbf{M}, \mathcal{A})$, our optimization problem becomes:

$$\min_{\mathbf{M}\in\mathbb{S}_+^d} f_\mathbf{W}(\mathbf{M}) = \mu R(\mathbf{M}) + \ell(\mathbf{M}, \mathcal{A}) \tag{10}$$

where

$$\begin{aligned} f_\mathbf{W}(\mathbf{M}) =& \mu\langle \mathbf{M}, \mathbf{W}\rangle \\ &+ \sum_{q\in\mathcal{A}} \left[\delta_q + \langle \mathbf{M}, \mathbf{x}_{ij}\mathbf{x}_{ij}^\top - \mathbf{x}_{kl}\mathbf{x}_{kl}^\top \rangle\right]_+ \end{aligned} \tag{11}$$

where $\mu \geq 0$ is a regularization parameter and $\langle \mathbf{M}, \mathbf{W}\rangle$ is the sum of the k smallest eigenvalues of $\mathbf{M}$.

4.2. Solving the optimization problem

Although the function defined in Eq. (11) is not globally convex due to the constraint $\langle \mathbf{M}, \mathbf{W}\rangle = \sum_{i=d-k+1}^{d} \lambda(\mathbf{M})_i$, it is convex w.r.t. $\mathbf{M}$ when $\mathbf{W}$ is fixed. We then first propose to perform a subgradient descent over $\mathbf{M}$. We alternate the update of $\mathbf{M}$ and $\mathbf{W}$ by fixing one of these matrices and updating the other. $\mathbf{M}$ is updated by performing a subgradient descent: the subgradient of Eq. (11) w.r.t. $\mathbf{M}$ is:

$$\bigtriangledown_\mathbf{M} = \mu\mathbf{W} + \sum_{q\in\mathcal{A}^+} \left(\mathbf{x}_{ij}\mathbf{x}_{ij}^\top - \mathbf{x}_{kl}\mathbf{x}_{kl}^\top\right) \tag{12}$$

where $\mathcal{A}^+$ is the subset of constraints in $\mathcal{A}$ that are not satisfied (Eq. (9)). The obtained value after subgradient descent over $\mathbf{M}$ is projected onto the cone of PSD matrices at each iteration. $\mathbf{W}$ is updated by construction as explained in Section 3.2 so that $\langle \mathbf{M}, \mathbf{W}\rangle$ is the sum of the k smallest eigenvalues of $\mathbf{M}$. That process stops when the objective value (Eq. (10)) stops decreasing. The global learning scheme is described in Algorithm 1.

4.3. Efficiency discussion

An alternative method to solve the problem in Eq. (11) is to switch the update between $\mathbf{M}$ and $\mathbf{W}$ after a full subgradient descent over $\mathbf{M}$ (*i.e.* fix $\mathbf{W}$ and optimize over $\mathbf{M}$ until convergence, then construct $\mathbf{W}$ (Eq. (7)), and iterate). Note that this option is computationally demanding since the outer loop that alternates between $\mathbf{M}$ and $\mathbf{W}$ has to be performed several times until convergence, requiring several full subgradient optimizations for which the projection onto the cone of PSD matrices is performed at each iteration. In addition, we experimentally noticed that this optimization strategy did not improve accuracy.

Algorithm 1 Metric Learning with Fantope Regularization

input : Training constraints $\mathcal{A}$, hyper-parameter μ and step size $\eta > 0$.
output : $\mathbf{M} \in \mathbb{S}_+^d$
  Initialize $\mathbf{M} \in \mathbb{S}_+^d$, $\mathbf{W} \leftarrow \mathbf{V}_\mathbf{M}\text{Diag}(\mathbf{w})\mathbf{V}_\mathbf{M}^\top$ (Eq. (7))
  repeat
    Compute $\bigtriangledown_\mathbf{M}$ (Eq. (12))
    $\mathbf{M} \leftarrow \Pi_{\mathbb{S}_+^d}(\mathbf{M} - \eta\bigtriangledown_\mathbf{M})$
    $\mathbf{W} \leftarrow \mathbf{V}_\mathbf{M}\text{Diag}(\mathbf{w})\mathbf{V}_\mathbf{M}^\top$ (Eq. (7))
  until stopping criterion (*e.g.* convergence)

When the input space dimension d is large, the eigendecomposition required at each iteration of the subgradient descent (Algorithm 1) also becomes computationally expensive. As in [14], we propose an adaptation of the *Alternating Direction Method of Multipliers* (ADMM) [2] to learn a metric. We then adapt Eq. (10) in this way:

$$\min_{\mathbf{M}\in\mathbb{S}^d, \mathbf{Z}\in\mathbb{S}^d} f_\mathbf{W}(\mathbf{M}) + g(\mathbf{Z}) \text{ s.t. } \mathbf{M} = \mathbf{Z} \tag{13}$$

where

$$g(\mathbf{Z}) = \begin{cases} 0 & \text{if } \mathbf{Z} \in \mathbb{S}_+^d \\ +\infty & \text{if } \mathbf{Z} \notin \mathbb{S}_+^d \end{cases} \tag{14}$$

and $f_\mathbf{W}(\mathbf{M})$ is given in Eq. (11). Introducing a Lagrange multiplier $\Lambda \in \mathbb{S}^d$, we obtain the augmented Lagrangian:

$$\mathcal{L}_\rho(\mathbf{M}, \mathbf{Z}, \Lambda) = f_\mathbf{W}(\mathbf{M}) + g(\mathbf{Z}) + \langle \Lambda, \mathbf{M} - \mathbf{Z}\rangle + \frac{\rho}{2}\|\mathbf{M} - \mathbf{Z}\|_F^2 \tag{15}$$

where $\rho > 0$ is a scaling parameter. The ADMM algorithm written in scaled form follows the successive updates described in Algorithm 2, where $\mathbf{U} = \frac{1}{\rho}\Lambda$. Algorithm 2 finds the optimal $\mathbf{M}$ before updating $\mathbf{W}$, as previously proposed. However, the approximation and speed up in Algorithm 2 comes from the constraint $\mathbf{M} \in \mathbb{S}_+^d$ which has been replaced by the constraint $\mathbf{M} \in \mathbb{S}^d$, whereas $g(\mathbf{Z})$ promotes a PSD solution matrix.

5. Synthetic example

We propose to start exploring the behavior of our Fantope regularization method using a synthetic dataset with a target metric $D_\mathbf{T}$ parameterized by a known low-rank distance matrix $\mathbf{T} \in \mathbb{S}_+^d$. For this purpose, we create a random symmetric positive definite matrix $\mathbf{A} \in \mathbb{S}_+^e$ with $\text{rank}(\mathbf{A}) = e$ and $e < d$, and define the target PSD distance matrix $\mathbf{T} \in \mathbb{S}_+^d$: $\mathbf{T} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ with $\text{rank}(\mathbf{T}) = \text{rank}(\mathbf{A}) = e$.

Algorithm 2 Metric Learning with Fantope Regularization (ADMM version)

input : Constraints $\mathcal{A}$ and hyper-parameters μ, ρ

Initialize $t = 1$, $\mathbf{M}^t = \mathbf{Z}^t \in \mathbb{S}_+^d$, $\mathbf{U}^t \leftarrow \mathbf{0}$, $\mathbf{W}^t \leftarrow \mathbf{V}_{\mathbf{M}^t}\text{Diag}(\mathbf{w})\mathbf{V}_{\mathbf{M}^t}^\top$ (Eq. (7))

repeat

$\mathbf{M}^{t+1} \leftarrow argmin_{\mathbf{M}\in\mathbb{S}^d} f_{\mathbf{W}^t}(\mathbf{M}) + \frac{\rho}{2}\|\mathbf{M} - (\mathbf{Z}^t - \mathbf{U}^t)\|_F^2$

$\mathbf{Z}^{t+1} \leftarrow \Pi_{\mathbb{S}_+^d}(\mathbf{M}^{t+1} + \mathbf{U}^t)$

$\mathbf{U}^{t+1} \leftarrow \mathbf{U}^t + \mathbf{M}^{t+1} - \mathbf{Z}^{t+1}$

$\mathbf{W}^{t+1} \leftarrow \mathbf{V}_{\mathbf{M}^{t+1}}\text{Diag}(\mathbf{w})\mathbf{V}_{\mathbf{M}^{t+1}}^\top$ (Eq. (7))

$t \leftarrow t + 1$

until stopping criterion

return $\Pi_{\mathbb{S}_+^d}(\mathbf{M}^t)$

We generate a set $\mathcal{X}$ of feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ from a uniform distribution in $[0, 1[$ for each component. The distance between two feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is given by: $D^2_{\mathbf{T}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)\mathbf{T}(\mathbf{x}_i - \mathbf{x}_j)$. In order to build a training set $\mathcal{A}$, we randomly sample pairs of distances using quadruplets in $\mathcal{X}^4$ and get the ground-truth using $D^2_{\mathbf{T}}$, so that: $\forall(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l) \in \mathcal{A}, D^2_{\mathbf{T}}(\mathbf{x}_k, \mathbf{x}_l) > D^2_{\mathbf{T}}(\mathbf{x}_i, \mathbf{x}_j)$. The set $\mathcal{A}$ is used to learn our matrix $\mathbf{M}$ by solving Eq. (10) where $\delta_q = 1$ and $\mathbf{W} \in \mathbb{S}_+^d$ such that $\text{rank}(\mathbf{W}) = (d - e)$ as defined in Eq. (7).

A test set $\mathcal{T}$ and a validation set $\mathcal{V}$ are generated in the same way as $\mathcal{A}$. To illustrate the relevance of the proposed method, we focus on having a small e and large d: we set $e = 10$, $d = 50$, $|\mathcal{A}| = 10^4$, $|\mathcal{V}| = |\mathcal{T}| = 10^6$ and $|\mathcal{X}| = 8000$. In this setting, 80% of the features are noisy.

Evaluation Metrics: we compute the number of satisfied constraints on the test set $\mathcal{T}$, the accuracy being measured as the percentage of satisfied constraints on $\mathcal{T}$. We also compare the similarity between the learned PSD matrix $\mathbf{M} \in \mathbb{S}_+^d$ and the target matrix $\mathbf{T} \in \mathbb{S}_+^d$. The similarity between $\mathbf{M}$ and $\mathbf{T}$ is measured as the distance $\|\mathbf{M} - \mathbf{T}\|_F^2 = \sum_{ij}(M_{i,j} - T_{i,j})^2$. $\mathbf{M}$ and $\mathbf{T}$ are rescaled so that their largest element is 1.

Results: to evaluate the impact of Fantope regularization, we compare the following metric learning schemes:

–*No regularization*: setting $\mu = 0$ in Eq. (11), and applying a subgradient descent over $\mathbf{M} \in \mathbb{S}_+^d$ [1].

–*Subgradient Descent over* $\boldsymbol{L}$: setting $\mu = 0$ in Eq. (11), Eq. (10) is solved using a subgradient descent over $\mathbf{L} \in \mathbb{R}^{e\times d}$ where $\mathbf{M} = \mathbf{L}^\top\mathbf{L}$ [2].

–*Trace(-norm) Regularization*: setting $\mu > 0$ and $\mathbf{W} = \mathbf{I}_d$.

–*Fantope Regularization*: setting $\mu > 0$.

–*Fantope and Trace Regularization*: replacing the regularization term $\mu\text{tr}(\mathbf{WM})$ by $R(M) = \gamma\text{tr}(\mathbf{M}) + \mu\text{tr}(\mathbf{WM})$.

For each method, the hyper-parameters $\gamma > 0$ and $\mu > 0$ are determined based on the validation set $\mathcal{V}$.

[1] This scheme usually leads to high-rank solutions prone to overfitting.

[2] This method is often used in the Computer Vision literature [17, 18]. Although the problem is not convex w.r.t. $\mathbf{L}$, this method controls the rank of $\mathbf{M}$ and avoids overfitting as $\text{rank}(\mathbf{M}) = \text{rank}(\mathbf{L}) \leq e$ with $e < d$.

Regularization	Acc.	rank($\mathbf{M}$)	$\|\mathbf{M} - \mathbf{T}\|_F^2$
No Regularization	89.3%	31	1.07
SD over $\mathbf{L}$	92.7%	10	0.44
Trace	95.1%	4	0.38
Fantope	**97.5%**	10	0.04
Fantope and Trace	**98.0%**	10	0.03

Table 1. Toy experiment results. Fantope regularization allows to approximate the target matrix $\mathbf{T}$ better than other methods.

Table 1 reports the accuracies and distances between $\mathbf{T}$ and the learned matrices $\mathbf{M}$. Methods without explicit regularization ($\mu = 0$ in Eq. (11)) obtain the worst results (89.3% and 92.7% accuracy). Trace regularization ignores most of the noisy features but learns a matrix whose rank is a lot smaller than the target rank $e = 10$. That leads to an accuracy of 95.1% and illustrates the fact that trace regularization cannot fine-control the rank of the solution matrix, although it promotes low-rank solutions. Finally, Fantope regularization outperforms the other methods by reaching 97.5% accuracy (and 98% when combined with trace regularization). In addition, the rank of the learned matrix corresponds exactly to the target rank.

We also ran the Fantope regularization with ADMM (Algorithm 2) and got 96.6% accuracy. It performs slightly worse than Algorithm 1 because there is no projection onto the cone of PSD matrices at each iteration. Nonetheless, it performs better than the methods that do not use Fantope regularization. We will use only the Algorithm 1 in the following experiments.

6. Experiments

We evaluate the proposed metric learning regularization method in two different Computer Vision applications. The first experiment is a face verification task, for which the similarity constraints come from relations between pairs of face images that are either similar or dissimilar. In the second experiment, we evaluate recognition performance on image classification with relative attributes [20]. In this context, we work with features defined in attribute space.

6.1. Face verification: LFW

In the face verification task, we are provided with pairs of face images. The goal is to learn a classifier that determines whether image pairs are similar (represent the same person) or dissimilar (represent two different persons).

6.1.1 Experiment setup

Dataset and evaluation metric: we use the publicly available *Labeled Faces in the Wild* (LFW) dataset [11]. It contains more than 13,000 images of faces collected from the Web and can be considered as the current state-of-the-art face recognition benchmark. We focus in this paper on the "restricted" paradigm where we are only provided with two sets of pairs of images: set $\mathcal{S}$ of similar pairs (same person) and set $\mathcal{D}$ of dissimilar images (different person). We follow the standard evaluation protocol that uses *View 2* data for training and testing (10 predefined folds of 600 image pairs each), and *View 1* for validation.

To generate our constraints, we use $\mathcal{S}$ and $\mathcal{D}$ and we set the upper bound $u = 0.5$ and the lower bound $l = 1.5$ following the scheme explained in Section 4.1 . The distance of a test pair is compared to the threshold $\frac{l+u}{2} = 1$ to determine whether the pair is similar or dissimilar.

Image representation: we use the same input features and setup as popular metric learning methods [6, 10, 18] that were already tested on this dataset. We strictly follow the setup described in [18]. We use the SIFT descriptors [15] computed by [10] available on their website. Each face image is represented by 27 SIFT descriptors. Those 27 descriptors are concatenated in a single histogram, and a element-wise square-root is performed on this histogram to return face image representations $\mathbf{x}_i$.

Initialization of the distance matrix $\mathbf{M} \in \mathbb{S}_+^d$: let e be the target rank of the learned matrix $\mathbf{M} \in \mathbb{S}_+^d$. To initialize the PSD matrix $\mathbf{M}$, we first compute the matrix $\mathbf{L} \in \mathbb{R}^{e \times d}$ that is composed of the coefficients for the e most dominant principal components of the training data. $\mathbf{M}$ is then constructed by computing $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$.

6.1.2 Results

We now provide a quantitative evaluation of our method in the described setup. The target rank e of our regularization term is fixed to $e = 40$, as in [18].

Impact of regularization: we compare here the impact of Fantope regularization over trace regularization. Table 2 shows classification accuracies when solving Eq. (10) with both regularization methods. Fantope regularization outperforms trace regularization by a large margin (82.3% *vs.* 77.6%). This illustrates the importance of having an explicit control on the rank of the distance matrix. In the following, we combine trace and Fantope regularization by replacing the regularization term $R(\mathbf{M}) = \mu \text{tr}(\mathbf{WM})$ by $R(\mathbf{M}) = \gamma \text{tr}(\mathbf{M}) + \mu \text{tr}(\mathbf{WM})$, with $\gamma \ll \mu$.

State-of-the-art results: we now compare Fantope Regularization to other popular metric learning algorithms. Table 3 shows performances of ITML [6], LDML [10] and PCCA [18] reported in [10] and [18] in the linear metric learning setup. These methods are the most popular metric learning methods when the task is to decide whether a pair is similar or dissimilar. Fantope regularization, which reaches $82.3 \pm 0.5\%$ accuracy, outperforms ITML and LDML and is comparable to PCCA on LFW in this setup. We explain in the following how our method can reach 83.5 ± 0.5 %.

Regularization Method	Accuracy (in %)
Trace-norm Regularization	77.6 ± 0.7
Fantope Regularization	82.3 ± 0.5

Table 2. Accuracies (mean and standard error) obtained on LFW in the "restricted" setup with our learning framework in different regularization settings.

Method	Accuracy (in %)
ITML [10]	76.2 ± 0.5
LDML [10]	77.5 ± 0.5
PCCA [18]	82.2 ± 0.4
Proposed Method	$\mathbf{83.5 \pm 0.5}$

Table 3. Results (mean and standard error) on LFW in the "restricted" setup of state-of-the-art linear metric learning algorithms and of our method with *early stopping*.

Number of iterations	10	100	1000	10^4
Accuracy (in %)	79.2 $\pm$ 0.5	79.3 $\pm$ 0.5	75.8 $\pm$ 0.5	63.2 $\pm$ 0.5

Table 4. Accuracy of Mignon's code [18] on LFW as a function of the number of iterations of gradient descent. The performance of PCCA [18] greatly depends upon the *early stopping* criterion.

Impact of early stopping: it is worth mentioning that accuracy of 82.2% obtained with PCCA [18] is obtained by performing *early stopping*. Table 4 reports the accuracies we obtained on LFW by testing the code of PCCA [18] provided by its authors, as a function of the number of iterations of gradient descent. 82.2% is the accuracy obtained with 30 iterations. We can notice that the PCCA performance decreases for larger numbers of iterations (*e.g.* 75.8% and 63.2% with 1000 and 10000 iterations, respectively). As in [18], we integrated this *early stopping* criterion in our method and determined the maximum number of iterations of subgradient descent from the validation set *View 1*. We reach an accuracy of $83.5 \pm 0.5\%$. To the best of our knowledge, this is the best result obtained for linear metric learning methods in the same setup (same input features). As a conclusion, our regularization scheme makes our method much more robust than PCCA [18] to *early stopping*.

Impact of the hyper-parameter μ: Fig. 2 illustrates the impact of the Fantope regularization on the rank of the solution matrix $\mathbf{M} \in \mathbb{S}_+^d$ and on the accuracy on LFW as we modify the value of μ (Eq. (11)) when we perform *early*

stopping. We observe that μ has a real impact on the rank of the solution matrix: the rank of $\mathbf{M}$ decreases as μ increases and reaches the expected rank $e = 40$ for high values of μ. On the other hand, the accuracy of the method first increases and eventually decreases as μ increases. Nonetheless, the recognition performed with high values of μ (82.3%) is still better than without regularization (81.2% with $\mu = 0$).

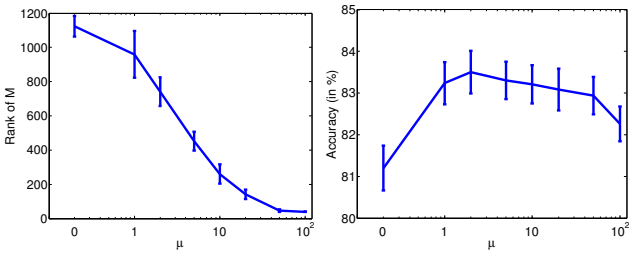

Figure 2. (left) rank and (right) accuracy of the learned metric on LFW in the "restricted" setup as a function of the hyper-parameter μ with *early stopping*. The expected rank is $e = 40$. The proposed regularization controls rank($\mathbf{M}$) while improving accuracy when compared to the absence of regularization ($\mu = 0$).

6.2. Metric learning in attribute space

In this subsection, we focus the image classification task where the goal is to assign an image to a predefined class. Particularly, we focus on the case where classes are described with attributes. Attributes are human-nameable (high-level) concepts used to describe images. For instance, in the context of scene recognition, they can describe the degree of presence of openness or perspective in images. In the image classification task with attributes, we are provided images described with attributes. Each image p_i is described by a vector $\mathbf{x}_i \in \mathbb{R}^d$ where d is the number of attributes. The j-th element of $\mathbf{x}_i$ represents the score (degree) of presence of the j-th attribute in $\mathbf{x}_i$.

6.2.1 Experiment setup

To evaluate and compare our Fantope regularization approach, we follow a classification framework inspired from [20] for scene and face recognition on the OSR [19] and PubFig [12] datasets.

Datasets: we experiment with the two datasets used in [20]: *Outdoor Scene Recognition* (OSR) [19] containing 2688 images from 8 scene categories and a subset of *Public Figure Face* (PubFig) [12] containing 771 images from 8 face categories. We use the image features made publicly available by [20]: a 512-dimensional GIST [19] descriptor for OSR and a concatenation of the GIST descriptor and a 45-dimensional Lab color histogram for PubFig. We also use relative ordering of classes according to some semantic attributes (*e.g.* images in face class (a) have a stronger presence of attribute "smiling" than images in class (b)).

Classification model	OSR	PubFig
Gaussian Distribution [20]	69.7 ± 1.5	70.6 ± 1.8
LMNN	71.7 ± 1.7	74.3 ± 1.9
LMNN + Trace	72.4 ± 2.0	75.0 ± 1.6
LMNN + Fantope (ours)	**73.7 ± 1.8**	**77.5 ± 1.6**

Table 5. Test accuracies (mean and standard deviation in %) obtained on OSR and Pubfig. Fantope regularization improves recognition in the classification task.

Baselines: we use two baselines: (1) The relative attribute learning problem described in [20] that uses relative attribute annotations on classes to compute high-level representations of images $\mathbf{x}_i \in \mathbb{R}^d$, a Gaussian distribution is learned for each class. (2) the *Large Margin Nearest Neighbor* (LMNN) [26] that is a popular metric learning method used for image classification. For each image, LMNN tries to satisfy the condition that members of a predefined set of target neighbors (of the same class) are closer than samples from other classes. High-level representations $\mathbf{x}_i \in \mathbb{R}^d$ are used as input features of the LMNN classifier. We use the publicly available codes of [20] and [26].

Integration of regularization: we modify the code of [26] to integrate trace and Fantope regularization, the stopping criterion is the convergence of the algorithm (*i.e.* the objective function stops decreasing).

Learning setup: we use the same experimental setup as [20]. $N = 30$ training images are used per class to learn the representations $\mathbf{x}_i$ and classifiers, the rest is for testing. The performance is measured as the average classification accuracy across all classes over 30 random train/test splits.

6.2.2 Results

Table 5 reports accuracies of baselines and our proposed regularization method on both OSR and PubFig datasets.

Fantope regularization applied to LMNN significantly improves recognition over baselines, particularly on PubFig. It outperforms the classic LMNN algorithm (without regularization) with a margin of 2 and 3% on OSR and PubFig, respectively. Trace-norm regularization also outperforms the absence of regularization. These results validate the importance of a proper regularization for predictive accuracy. Fantope regularization finds a low e-dimensional subspace where distances can be computed with $e < d$ (*e.g.* $e = 8$ with $d = 11$ on PubFig) and allows to exploit correlations between features better than methods that learn a high-rank distance matrix. In this case, each feature corresponds to the score of presence of an attribute in images. Notably, by considering the learned matrix $\mathbf{M} \in \mathbb{S}_+^d$ as a covariance matrix, the most correlated attributes w.r.t. the Pearson product-moment correlation coefficient are "*smiling*", "*chubby*" and "*male-looking*" on the PubFig dataset.

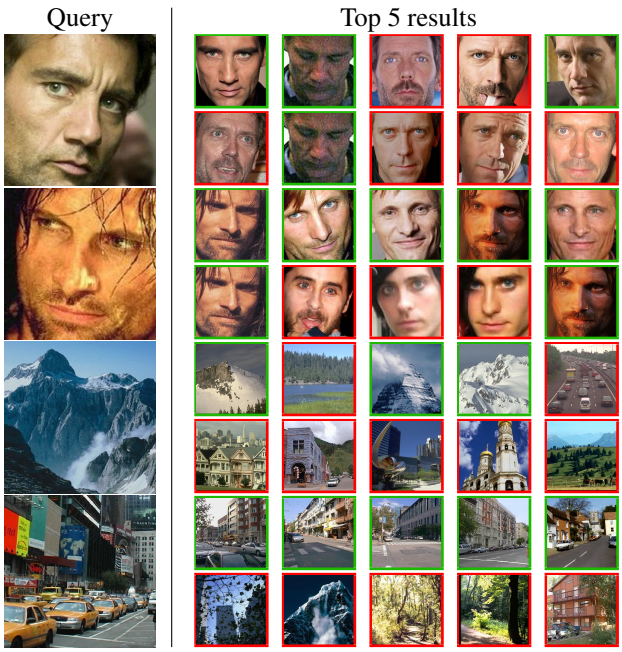

Figure 3. Some results of similarity search on the PubFig and OSR datasets. We show for each query the 5 nearest neighbors returned by our method (first row) and by LMNN (second row). Results in green correspond to images in the same class as the query whereas results in red are images from different classes.

This result is expected as the women of the PubFig dataset (Scarlett Johansson and Miley Cyrus) are annotated in [20] as more chubby and smiling more than most men of the dataset. On the OSR dataset, the attributes "*close depth*", "*open*" and "*perspective*", which are all related to the notion of depth, are also strongly correlated.

Fig. illustrate on some examples how our scheme is effective to learn semantics. Particularly on PubFig, the learned metric gives priority to semantical similarity rather than visual similarity: the images retrieved by the classic LMNN are more visually similar than the images returned by our Fantope regularization. However, they are more often in different categories than the category of the query.

7. Conclusion

We proposed a new regularization scheme for metric learning that explicitly controls the rank of the learned distance matrix. Our method generalizes the trace regularization, and can be applied to various optimization frameworks to impose a meaningful structure on the learned PSD matrix. We also derived an efficient metric learning algorithm that combines the regularization term with a loss function that can incorporate constraints between pairs or triplets of images. We also demonstrate that regularization greatly improves recognition on both controlled and real datasets, showing the relevance of this new regularization to limit overfitting. Future work includes the learning of a better designed ADMM formulation scheme that takes into account the fact that the objective function is not convex.

References

[1] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Arajo. Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding (CVIU)*, 117(5):453 – 465, 2013. 2

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 4

[3] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *JMLR*, 11:1109–1135, 2010. 1, 2

[4] M. Cord and P. Cunningham. *Machine learning techniques for multimedia*. Springer, 2008. 2

[5] J. Dattorro. *Convex optimization and Euclidean distance geometry*. Meboo Publishing USA, 2005. 3

[6] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007. 2, 3, 6

[7] K. Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences of the United States of America*, 35(11):652, 1949. 3

[8] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*, 2007. 1, 2

[9] H. Goh, N. Thome, M. Cord, and J. Lim. Unsupervised and supervised visual codes with restricted boltzmann machines. In *ECCV*, 2012. 2

[10] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009. 1, 6

[11] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 6

[12] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 7

[13] M. T. Law, N. Thome, and M. Cord. Quadruplet-wise image similarity learning. In *ICCV*, 2013. 1, 2, 3

[14] D. Lim, B. McFee, and G. Lanckriet. Robust structural metric learning. In *ICML*, 2013. 1, 2, 4

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 6

[16] B. McFee and G. Lanckriet. Metric learning to rank. In *ICML*, 2010. 2

[17] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2624–2637, 2013. 1, 2, 5

[18] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 1, 2, 3, 5, 6

[19] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 2, 7

[20] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011. 1, 5, 7, 8

[21] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *NIPS*, 2003. 1, 2

[22] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007. 2

[23] C. Shen, J. Kim, L. Wang, and A. van den Hengel. Positive semidefinite metric learning with boosting. In *NIPS*, 2009. 2

[24] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 2

[25] C. Theriault, N. Thome, and M. Cord. Extended coding and pooling in the hmax model. *IEEE Transactions on Image Processing*, 22(2):764–777, 2013. 2

[26] K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009. 1, 2, 7

[27] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *NIPS*, 2002. 2

[28] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009. 2