Data Fusion through Cross-modality Metric Learning using Similarity-Sensitive Hashing

Michael M. Bronstein, Alexander M. Bronstein
Department of Computer Science, Technion – Israel Institute of Technology, Haifa 32000, Israel
mbron@cs.technion.ac.il, bron@cs.technion.ac.il

Fabrice Michel, Nikos Paragios
Laboratoire de Mathématiques Appliquées aux Systèmes, École Centrale de Paris, France
Equipe GALEN, INRIA Saclay - Île-de-France, Orsay, France
fabrice.michel@ecp.fr, nikos.paragios@ecp.fr

Abstract

Visual understanding is often based on measuring similarity between observations. Learning similarities specific to a certain perception task from a set of examples has been shown advantageous in various computer vision and pattern recognition problems. In many important applications, the data that one needs to compare come from different representations or modalities, and the similarity between such data operates on objects that may have different and often incommensurable structure and dimensionality. In this paper, we propose a framework for supervised similarity learning based on embedding the input data from two arbitrary spaces into the Hamming space. The mapping is expressed as a binary classification problem with positive and negative examples, and can be efficiently learned using boosting algorithms. The utility and efficiency of such a generic approach is demonstrated on several challenging applications including cross-representation shape retrieval and alignment of multi-modal medical images.

1. Introduction

Quantifying similarity or dissimilarity of data is one of the central problems in computer vision and pattern recognition, arising practically in any problem involving comparison, search, or matching. This challenge can be addressed either in unsupervised or supervised manner. Unsupervised methods, often common in classification problems, try separating different data populations based on their statistical properties, using for example the Kulback-Lelbler divergence [8] and mutual information [36]. In a wide range of applications such as content-based retrieval, the data are a high-dimensional representation of complicated concepts which are very difficult or even impossible to model. High-order statistical methods run into computational challenges in high dimensional spaces. One can overcome this limitation through techniques that map the original data into a simpler representation endowed with a metric that correctly represents the data similarity. Problems of this type are usually referred to as *manifold learning* and *nonlinear dimensionality reduction*. Notable algorithms in this family are locally linear embedding (LLE) [23], Isomap [30], Laplacian [1] and diffusion [9] eigenmaps, and multidimensional scaling (MDS) [4, 22].

Supervised methods could handle these cases through the definition of a metric that best separates the observed populations. In the simplest case, one has a set of labeled examples and can employ a naïve Bayes nearest neighbor classifier [3]. More generally, the similarity is given in some parametric form and an optimization problem with respect to these parameters is employed to construct the optimal metric. Methods based on convex programming [37] and support vector machines (SVM) [2, 34] have been proposed. In [12], neighborhood component analysis is considered to learn the Mahalanobis metric. A similar approach is used in [14] in a visual content-based search application.

An important setting of similarity learning is when the desired similarity is *binary* (similar/dissimilar). Such similarities arise, for example, in content-based image retrieval applications, where one wishes to find only images in the database similar to the query image. In [24, 35], this problem was addressed using *similarity-sensitive hashing*. The aim was to find a projection of the data into the space of binary codes such that the Hamming metric between the codes reflects the similarity relations between pairs in the training set. Such a concept is intimately related to the *locality-sensitive hashing* (LSH) [13], where the collision

probability of the hash is inversely related to the distance (typically, Euclidean or L_p) between the hashed data. In the case of similarity-sensitive hashing, this distance is learned from examples. Considering the hash construction as a binary classification problem, [24] proposed an efficient solution for this similarity learning problem using AdaBoost [10]. These ideas were successfully used in [32, 15] for content-based image retrieval as well as other computer vision applications [25].

In many cases, the data that one needs to compare come from different representations or modalities, and often reside in a different spaces. For example, in multimodal medical image alignment, one would like to find similarity between patches in images coming from different imaging modalities, e.g., two MRI contrasts, or a CT and a PET image. The data formation, processing, and representation in these modalities can be completely different and statistically uncorrelated, making the alignment and fusion of such data practically impossible.

In content-based retrieval and copy detection using the *bag of features* paradigm [27, 7], images are represented as histograms of simple visual features taken from a large ($10^3 - 10^6$) vocabulary. Similar approaches have been proposed for indexing and retrieval of 3D shapes [20, 21, 31]. The use of different feature descriptors (e.g. MSER [19] or SIFT [18] in images; spin image descriptors [20] or heat kernel signatures [29] in shapes), vocabularies of different size, or just different versions of the same vocabulary would result in two mutually-incomparable representations. The similarity relation between such multi-modal data is not a metric, hence, does not fall into the standard framework of metric learning or similarity-sensitive hashing. There have been very few attempts to address the problem of *cross-modality similarity learning* like for example in [17] where an SVM-based approach was proposed for medical image alignment.

In this paper, we approach the cross-modality similarity learning problem by means of embedding incommensurable data into a common metric space. The embedding itself is used to parameterize the similarity. In particular, we extend the similarity-sensitive hashing introduced in [24] to the setting in which the input data come from two different spaces. We show that like in the standard similarity-sensitive hashing, cross-modality similarity learning can be efficiently solved using boosting techniques. We evaluate the performance of the method on two difference applications showing its extreme potentials.

The rest of the paper is organized as follows: In Section 2, we review the basics of similarity-sensitive hashing. Section 3 extends this approach to the multi-modal setting. In Section 4, we demonstrate the use of our approach for cross-representation shape retrieval and multi-modal medical image alignment. Section 5 concludes the paper.

2. Similarity-sensitive hashing

Let $X \subseteq \mathbb{R}^m$ be the space of data points, and let $s : X \times X \to \{\pm 1\}$ be an unknown binary similarity function between the data points. The similarity s partitions the set $X \times X$ of all pairs of data points into *positives* $P = \{(x, x') : s(x, x') = +1\}$ and *negatives* $N = \{(x, x') : s(x, x') = -1\}$. The goal of *similarity learning* is to construct another binary similarity function $\hat{s}$ that approximates the unknown s as faithfully as possible. To evaluate the quality of such an approximation, it is common to associate with $\hat{s}$ the expected *false positve* and *negative rates*,

$$\begin{aligned} FP &= E\{\hat{s}(x, x') = +1 | s(x, x') = -1\} \\ FN &= E\{\hat{s}(x, x') = -1 | s(x, x') = +1\}, \end{aligned} \quad (1)$$

and the related *true positive* and *negative rates*, $TP = 1 - FN$ and $TN = 1 - FP$. Here, the expectations are taken with respect to the joint distribution of pairs (in the context of retrieval, where (x, x') are obtained by pairing a query with all the examples in the database, this means the product of marginal distributions).

A popular variant of similarity learning involves embedding of the data points into some metric space $(\mathbb{Z}, d_{\mathbb{Z}})$ by means of a map $\xi : X \to \mathbb{Z}$. The distance $d_{\mathbb{Z}}$ represents the similarity of the embedded points, in the sense that the lower is $d_{\mathbb{Z}}(\xi(x), \xi(x'))$, the higher is the probability that $s(x, x') = +1$. Alternatively, one can find a range of radii R such that with high probability positive pairs have $d_{\mathbb{Z}} \circ (\xi \times \xi) < R$, while negative pairs have $d_{\mathbb{Z}} \circ (\xi \times \xi) > R$. A map ξ satisfying this property is said to be *sensitive* to the similarity s, and it naturally defines a *binary classifier* $\hat{s}(x, x') = \operatorname{sign}(R - d_{\mathbb{Z}}(\xi(x), \xi(x')))$ on the space of pairs of data points. In practice this means that retrieval of a query in a database translates into search of k nearest neighbors or R-neighbors of the query embedded into $\mathbb{Z}$ by ξ.

In [24], the (possibly weighted) n-dimensional Hamming space $\mathbb{H}^n$ was proposed as the embedding space $\mathbb{Z}$. Such a mapping encodes each data point as an n-bit binary string. The correlation between positive similarity of a pair of points and small Hamming distance between their corresponding codes implies that positives are likely to be mapped to the same code. This fact allows to interpret the Hamming embedding as *similarity-sensitive hashing*, under which positive pairs have high collision probability, while negative pairs are unlikely to collide. Such a hashing can be thought of as an optimal LSH, in which sensitivity to the desired similarity is explicitly maximized. The hash also acts as a means of dimensionality reduction when $m \gg n$.

The n-dimensional Hamming embedding can be thought of as a vector $\xi(x) = (\xi_1(x), \ldots, \xi_n(x))$ of binary embed-

dings of the form

$$\xi_i(x) \quad = \quad \begin{cases} 0 & \text{if } f_i(x) \leq 0; \\ 1 & \text{if } f_i(x) > 0, \end{cases} \tag{2}$$

parametrized by a *projection* $f_i : X \to \mathbb{R}$. Each such map ξ_i defines a *weak* binary classifier on pairs of data points,

$$h_i(x, x') \quad = \quad \begin{cases} +1 & \text{if } \xi_i(x) = \xi_i(x'); \\ -1 & \text{otherwise.} \end{cases} \tag{3}$$

Using this terminology, the Hamming metric between the embeddings $\xi(x)$ and $\xi(x')$ of a pair of data points (x, x') can be expressed as a (possibly weighted) superposition of weak classifiers,

$$d_{\mathbb{H}^n}(\xi(x), \xi(x')) \; = \; \frac{1}{2}\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n} \alpha_i h_i(x, x'), \tag{4}$$

where $\alpha_i > 0$ is the weight of the i-th bit ($\alpha_i = 1$ in the unweighted case). Observing the resemblance with cascaded binary classifiers, the idea of constructing the similarity-sensitive embedding using the standard boosting approach was proposed in [24]. Specifically, the use of AdaBoost [10] was made to find a greedy approximation to the minimizer of the exponential loss function $L = E\{e^{-s(x,x')\hat{s}(x,x')}\}$, where in practice the expectation is replaced by an empirical average on the training set. The exponential loss is a reasonable selection of the objective function, as it constitutes an upper bound on the training error. Furthermore, the minimization of L is equivalent to the minimization of the sum of the error rates $FN + FP$ or, alternatively, to the maximization of the gap $TP - FP$. The latter is directly related to the sensitivity of the embedding to the similarity function being learned [24].

3. Cross-modality similarity learning

An important generalization of the metric learning problem is the case in which the similarity is between points from *different* spaces, $X \subseteq \mathbb{R}^m$ and $Y \subseteq \mathbb{R}^{m'}$ (usually, $m \neq m'$). For example, a point in X can be a query image in some representation, while a point in Y can be an image in the database in a different representation. The unknown binary similarity function in this case is $s : X \times Y \to \{\pm 1\}$, comparing data points across modalities. As in the classical similarity learning problem, *cross-modality similarity learning* aims at finding a binary similarity $\hat{s}$ on $X \times Y$ approximating s.

The central contribution of this paper is the extension of the embedding framework to the multimodal case. We propose to construct *two maps* $\xi : X \to \mathbb{H}^n$ and $\eta : Y \to \mathbb{H}^n$ such that $d_{\mathbb{H}^n}(\xi(x), \eta(y))$ is small for $s(x, y) = +1$ and large for $s(x, y) = -1$ with high probability. Following the greedy approach [24], the latter Hamming metric can be constructed sequentially as a superposition of weak binary classifiers, now of the form

$$\begin{aligned} h_i(x, y) \quad &= \quad \begin{cases} +1 & \text{if } \xi_i(x) = \eta_i(y); \\ -1 & \text{otherwise} \end{cases} \\ &= \quad (2\xi_i(x) - 1)(2\eta_i(y) - 1), \end{aligned} \tag{5}$$

where ξ_i and η_i are binary maps parametrized by projections $f_i : X \to \mathbb{R}$ and $g_i : Y \to \mathbb{R}$, respectively. Here, we limit our attention to affine projections of the form $f_i(x) = p_i^{\mathrm{T}} x + a_i$ and $g_i(y) = q_i^{\mathrm{T}} y + b_i$, where p_i and q_i are, respectively, m- and m'-dimensional unit vectors, and a_i and b_i are scalars. Extension to more complex projections is relatively straightforward.

Our boosted cross-modality similarity learning algorithm can be summarized as follows:

Input: K pairs (x_k, y_k) labeled by $s_k = s(x_k, y_k)$.
Output: maps $\xi_i : X \to \{0, 1\}$ and $\eta_i : Y \to \{0, 1\}$, and scalars α_i, $i = 1, \ldots, n$.

1 Initialize weights $w_1(k) = 1/K$.
2 **for** $i = 1, \ldots, n$ **do**
3 Select ξ_i and η_i such that h_i in (5) maximizes

$$r_i \quad = \quad \sum_{k=1}^{K} w_i(k) s_k h_i(x_k, y_k). \tag{6}$$

4 Set $\alpha_i = \frac{1}{2}\log(1 + r_i) - \frac{1}{2}\log(1 - r_i)$.
5 Update weights according to

$$w_{i+1}(k) \quad = \quad w_i(k) e^{-\alpha_i s_k h_i(x_k, y_k)} \tag{7}$$

and normalize by sum.
6 **end**

The algorithm follows very much the standard AdaBoost procedure. It consists of two steps, where first the maximization of the weighted correlation r_i of labels with the outputs of the weak classifier (Step 3) is addressed. This step is followed by the selection of α_i in (Step 4) that minimizes the exponential loss [10]. In case the unweighted version of the Hamming metric is used, Step 4 is skipped, fixing $\alpha_i = 1$.

3.1. Projection selection

Details of projection selection specific to our cross-modality similarity learning problem are concentrated in Step 3. Substituting the affine projection f_i and g_i into (6), we obtain

$$r_i = \sum_{k=1}^{K} w_i(k) s_k \mathrm{sign}\,(p_i^{\mathrm{T}} x_k + a_i) \mathrm{sign}\,(q_i^{\mathrm{T}} y_k + b_i). \tag{8}$$

Maximizing r_i with respect to the projection parameters is difficult because of the sign function. However, this maxi-

mizer is closely related to the maximizer of a simpler function,

$$\hat{r}_i \quad = \quad \sum_{k=1}^{K} v_k (p_i^{\mathrm{T}} \overline{x}_k)(q_i^{\mathrm{T}} \overline{y}_k), \tag{9}$$

where $\overline{x}_k$ and $\overline{y}_k$ are x_k and y_k centered by their weighted means, and $v_k = w_i(k) s_k$. Rewriting the above yields

$$\hat{r}_i \quad = \quad p_i^{\mathrm{T}} \left(\sum_{k=1}^{K} v_k \overline{x}_k \overline{y}_k^{\mathrm{T}} \right) q_i = p_i^{\mathrm{T}} C q_i, \tag{10}$$

where C can be thought of as the difference between weighted covariance matrices of positive and negative pairs of the training data points. This approach to the selection of projection direction can be considered an extension of one introduced in [6]. Unit projection directions p_i and q_i maximizing $\hat{r}_i$ correspond, respectively, to the largest left and right singular vectors of C. In practice, since the minimizers of $\hat{r}_i$ and r_i are not identical, we project x_k and y_k onto the subspaces spanned by M largest left and right singular vectors. Selecting $M \ll m, m'$ allows to greatly reduce the search space complexity. In our experiments, M was empirically set to 5; further increase of M did not bring significant improvement.

The best projection directions p_i and q_i are selected as a linear combination of M largest singular vectors, reducing the search to an M-dimensional space. We generate N pairs of M-dimensional random vectors; each such pair forms a candidate for the pair of projection directions p_i and q_i. For each candidate, we project the training data points obtaining two sets of scalars $x'_k = p_i^{\mathrm{T}} x_k$ and $y'_k = q_i^{\mathrm{T}} y_k$. Next, we search for the scalar parameters a_i and b_i maximizing r_i. For that purpose, for every pair of scalars (a, b), we define the cumulative sum

$$S(a,b) = \sum_{k=1}^{K} \mathbf{1}(x'_k + a \leq 0) \mathbf{1}(y'_k + b \leq 0) v_k, \tag{11}$$

where $\mathbf{1}$ denotes an indicator function. In this notation, r_i can be expressed as $r_i(a,b) = 4S(a,b) + S(-\infty,-\infty) - 2S(a,-\infty) - 2S(-\infty,b)$. In order to find (a,b) maximizing r_i, we quantize the space of candidate pairs (a,b) on a grid of $B \times B$ bins and evaluate $S(a,b)$ and, hence, $r_i(a,b)$ in each bin. This technique, largely resembling the idea of integral images [33], is applied at two resolutions of the grid, thus allowing to control the tradeoff between accuracy and complexity.

4. Applications

4.1. Cross-representation shape retrieval

In this experiment, we tested the proposed approach on a three-dimensional shape retrieval application. In shape retrieval, given a query shape, the goal is to retrieve all the possible transformations of the shape that appear in the database of shapes (e.g. in Figure 1, an ideal response to the query human shape would be all the other human shapes from the database). We used the ShapeGoogle database [21], consisting of a total of 1052 shapes from different classes with rich transformations including isometric deformations, topological changes, missing parts, and different sampling and triangulation. As of today, this is the largest non-rigid shape retrieval benchmark, comprising objects from TOSCA [5], Sumner [28] and Princeton [26] datasets. 583 transformed shapes from ten classes were used as queries against untransformed shapes to which 456 other unrelated shapes were added as negatives. Performance was evaluated in terms of *average precision* (AP), computed as the area below the precision-recall curve for each query, and the *mean average precision* (mAP), computed by averaging AP over all queries.

As shape descriptors, we used bags of geometric words and expressions proposed in [21]. Multiscale heat kernels were used as deformation-invariant local feature descriptors [29]; these descriptors were quantized in a geometric vocabulary. Two types of shape descriptors were used: a standard bag of features counting the frequency of geometric words in a vocabulary of size 32 (BoF 32), and a spatially-sensitive bag of features counting the simultaneous occurrence of pairs of geometric words ("geometric expressions") in a vocabulary of size 8 (SS-BoF 8). The two descriptors were represented as 32- and 64-dimensional vectors, respectively (see Figure 2).

Cross-modality similarity-sensitive hashing with code length up to 96 bits was built to query 64-dimensional SS-BoF 8 shape descriptors against a database of 32-dimensional BoF 32 descriptors. Training was performed on an independent set of shapes containing 10^4 positive and 2×10^5 negative pairs, and took approximately an hour.

Figure 3 shows the mAP achieved by comparing the two different shape descriptors using the learned cross-modality similarity. 48 bits are sufficient to achieve performance not worse than one achieved by comparing each of the descriptors separately using the Euclidean distance (92.68% for SS-BoF 8 and 94.68% for BoF 32). For 96 bits, mAP of the learned distance exceeds 99.2%, slightly inferior to the performance of the standard similarity-sensitive hashing applied to each of the descriptor modalities separately.

4.2. Alignment of multi-modal medical images

In the second experiment, we used the proposed approach as a distance function in multi-modal medical image alignment application. In the problem of non-rigid alignment, we are given a source and target images f and g (for simplicity, scalar-valued), defined on a domain Ω ($\Omega \subset \mathbb{R}^2$ in 2D alignment shown in our experiment here, or $\Omega \subset \mathbb{R}^3$

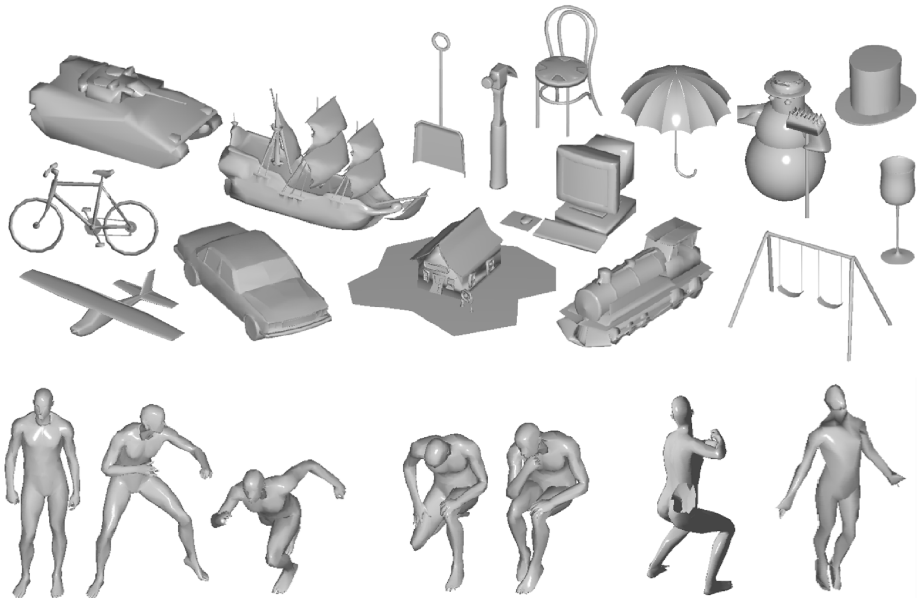


Figure 1. Top: examples of shapes in the ShapeGoogle database used in our cross-modality shape retrieval experiment. Bottom: examples of transformations of the human shape (shown are isometric deformations, topological changes, missing parts, coarse triangulation).

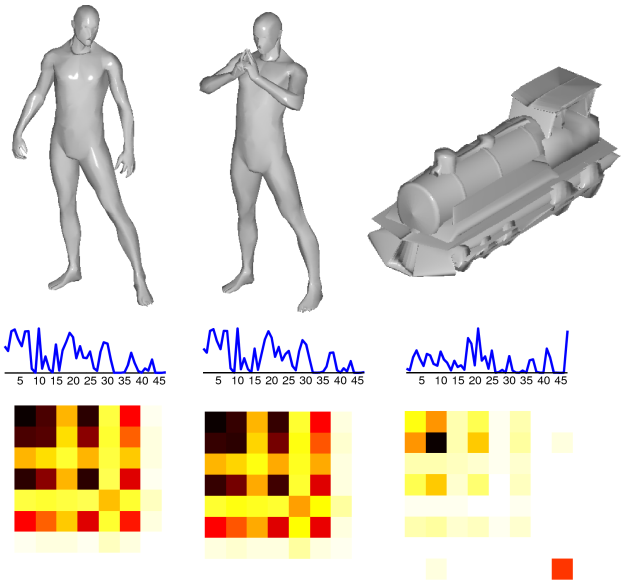


Figure 2. Different shapes (top) and their corresponding descriptors: bag of features descriptor using a vocabulary of 32 geometric words represented as a 32-dimensional vector (middle), and spatially-sensitive bags-of features descriptor using a vocabulary of 8 geometric words represented as a 8×8 matrix (bottom).

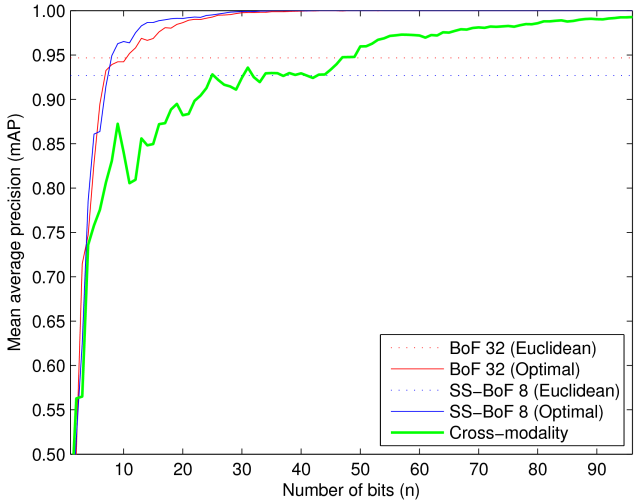


Figure 3. Mean average precision of similarity-sensitive hashing as a function of the code length within modalities: BoF 32 queried against BoF 32 (red), SS-BoF 8 queried agains SS-BoF 8 (blue), and across modalities: SS-BoF 8 queried against BoF 32 (bold green). For reference, the performance of the Euclidean distance within the two modalities is shown (dashed red and blue).

in the case of 3D volume registration). In general, the images are related by a complicated relation,

$$g(x) = h \circ f(\mathcal{T}(x)), \tag{12}$$

for all x on Ω, involving a geometric deformation $\mathcal{T}$ and a non-linearity h explaining the changes of appearance between corresponding points.

State-of-the-art registration methods [11] attempt at estimating the deformation $\mathcal{T}$ on a sparse grid $\Omega' \subset \Omega$

($|\Omega'| \ll |\Omega|$) of *control points*,

$$\mathcal{T}(x) = x + \sum_{p \in \Omega'} \rho(\|x - x_p\|)\Delta_p, \tag{13}$$

where Δ_p is the displacement vector of the control point x_p. Moving a control point results in a local deformation of the image around it; the weighting function ρ measures the contribution of a control point in Ω' to the displacement of point in Ω. The deformation field is found by minimizing the criterion of point-wise similarity between the target and deformed source images,

$$E(\mathcal{T}) = \frac{1}{|\Omega'|} \sum_{p \in \Omega'} \int_{\Omega} \rho^{-1}(\|x - x_p\|) d(g(x), f(\mathcal{T}x)) dx, \tag{14}$$

where d is some similarity function. In order to avoid folding on the deformation grid, a smoothness term on $\mathcal{T}$ is added.

For a practical and efficient numerical solution, problem (14) is posed as an assignment problem in the following way [11]: Let $\mathcal{L} = \{u^1, ..., u^k\}$ be a discrete set of labels corresponding to a quantized version of the deformation space $\Theta = \{\Delta^1, ..., \Delta^k\}$. A label assignment $u_p \in \mathcal{L}$ to a grid node $x_p \in \Omega'$ is associated with displacing the node by the corresponding vector Δ^{u_p}. The deformation field associated with a certain discrete labeling u is $\mathcal{T}_u(x) = x + \sum_{p \in \Omega'} \rho(\|x - x_p\|)\Delta^{u_p}$. Problem 14 can thus be posed as discrete Markov random field (MRF) optimization with respect to the labeling,

$$\begin{aligned} E(u) &= \frac{1}{|\Omega'|} \sum_{p \in \Omega'} \int_{\Omega} \rho^{-1}(\|x - x_p\|) d(g(x), f(\mathcal{T}_u x)) dx \\ &\approx \frac{1}{|\Omega'|} \sum_{p \in \Omega'} V_p(u_p), \end{aligned} \tag{15}$$

where V_p is a *singleton potential function* representing a local dissimilarity measure. Such a formulation allows to plug in any dissimilarity function without modifying the scheme itself.

When the source and the target images arise from different imaging modalities, we land at the problem of computing a cross-modality similarity. Modeling such similarity can be difficult, but learning is possible given examples of aligned images. In our experiment, we used ten T1- and T2-weighted MRI images of the brain (Figure 4) of size 256×256. Each T1 and T2 pair of images was perfectly aligned. We used cross-modality similarity-preserving hashing to learn the distance between 9×9 patches in T1 and T2 MRI images. The training set was created from two pairs of perfectly aligned images and consisted of 78887 positive and 788870 negative pairs of patches. 64-dimensional embedding was trained in about

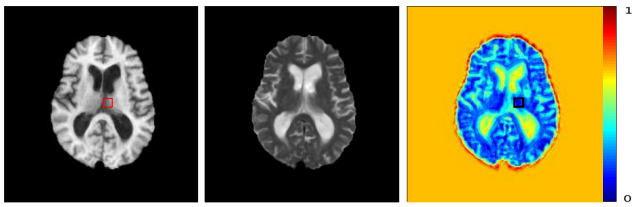


Figure 4. Demonstration of the effectiveness of the learned metric between T1 and T2 images on previously unseen images. The dissimilarity map (right) was generated between the specific patch delineated by the red square (T1-MRI) and the patches around all of the pixels in the center image (T2-MRI). We can see that the blue parts are all situated in regions visually resembling the one around the inside the square.

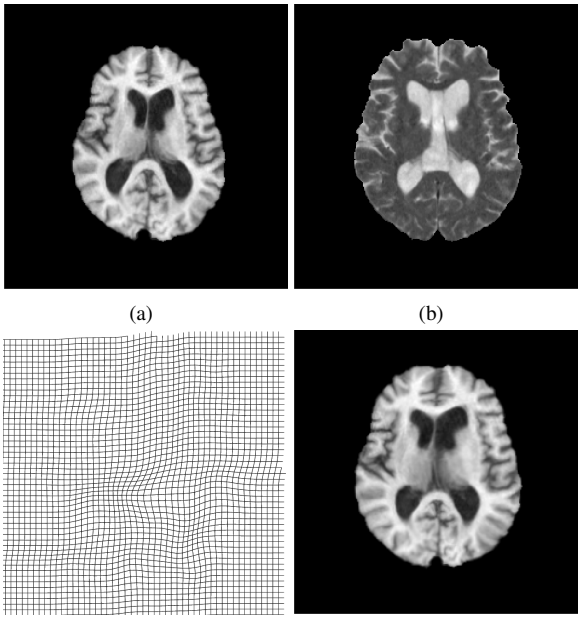
103 minutes. The obtained similarity was smooth and discriminative enough to align images in a non-rigid way (see Figure 4).

The MRF solver from [16] was used to perform the alignment with the learned metric as a point-wise dissimilarity. As a test set, we used the remaining eight images of the brain with groundtruth perfect alignment and manual manual segmentations of ventricles. We selected one image as the reference and aligned the remaining seven images to it. The recovered transformation was then used to warp the segmentations. In order to compare the deformed segmentations to the one of the reference image, we used the DICE coefficient which measures the overlapping proportion of two regions and constitutes an anatomical criterion for conformity of the transformation. For each image, we performed ten alignments with different regularization coefficients in order to take into account the variability of the results with respect to this parameter. Representative alignment results are depicted in Figures 5 and 6.

We compared our method with several metrics commonly used in multi-modal image alignment including mutual information (MI), normalized mutual information (NMI), normalized cross-correlation (NCC) and correlation ratio (CR) [11]. As a reference, the "ideal" case in which each source T1-MRI was swapped with the corresponding T2-MRI image and registered by means of uni-modal alignment using sum of squared differences (U-SSD). As can be see from Figure 7, in terms of the DICE coefficient our method significantly outperforms all other multi-modal alignment approaches and is only slightly inferior to the "ideal" uni-modal case. For a visual evidence of this fact, compare Figure 6 (a) to Figure 6 (b).

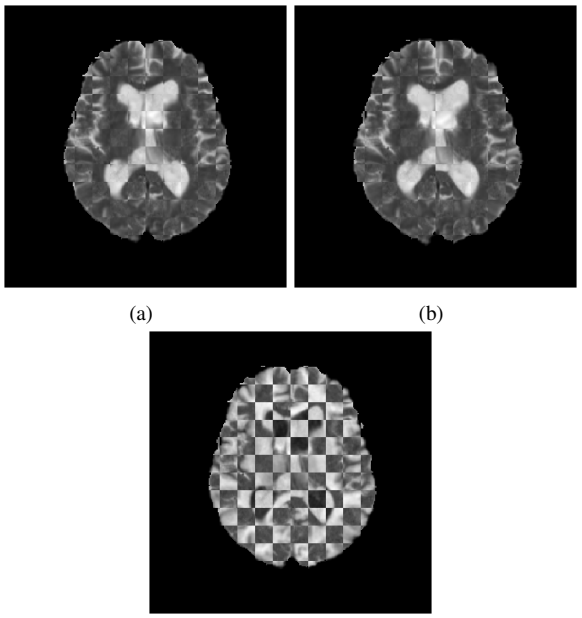
5. Conclusions

We introduced a generalization of similarity-sensitive hashing to multi-modal data. To our knowledge, this is the first attempt to approach the challenging problem of cross-



(a) (b) (c) (d)

Figure 5. Representative result of T1- to T2-weighted MRI registration. (a) Source image. (b) Target image. (c) Transformation applied to a regular grid. (d) Source image deformed with the transformation obtained after alignment.



(a) (b) (c)

Figure 6. Checkerboard visualization of (a) target image and the previously unknown T2-MRI source image deformed with the transformation obtained using our learned metric; (b) target image and the previously unknown T2-MRI source image deformed through uni-modal registration with an SSD metric; (c) target and the deformed source image.

modality similarity learning as an embedding problem. We demonstrated our approach on cross-representation retrieval of non-rigid shapes; in future studies, we intend to show further examples of retrieval and copy detection of images and video across representations. We also showed that using cross-modality similarity learning allows to efficiently perform alignment of medical images acquired with different modalities. Cross-modality similarity can also be interpreted as *model learning*. In future studies, we intend to show the use of our approach for building priors in inverse problems, in particular in image restoration.

While in retrieval applications the Hamming embedding is advantageous due to its low computational and storage complexity and easy integration into existing database managements systems, the Hamming metric is discrete-valued and involves a non-differentiable non-linearity. This fact might complicate some applications. Our approach can be extended to representation of cross-modality similarities by means of embedding into e.g. L_p metrics. In particular, it is straightforward to extend it to the cosine (correlation) similarity by removing the sign function and replacing the AdaBoost iteration by RealBoost.

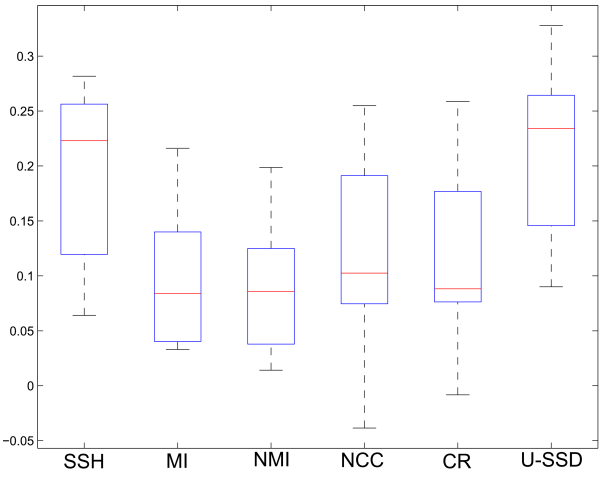


Figure 7. Increase in the DICE coefficient before and after alignment, for our learned metric (SSH) and other metrics commonly used in multi-modal image alignment (MI, NMI, NCC, and CR). DICE coefficient was computed on manually segmentated ventricles. Each diagram averages a total of 70 experiments: 10 alignments with different regularization parameters on 7 patients' images. Uni-modal alignment (U-SSD) is given as a reference.

Acknowledgement

MB is partially supported by the Digiteo research cluster. FM is partially supported by the Stereos+ Grant of the MEDICEN pôle de compétitivité of the Île-de-France region.

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 13:1373–1396, 2003.

[2] S. Bermejo and J. Cabestany. Large margin nearest neighbor classifiers. In *Proc. Artificial and Natural Neural Networks*, pages 669–676, 2001.

[3] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Proc. CVPR*, 2008.

[4] I. Borg and P. Groenen. *Modern multidimensional scaling - theory and applications*. Springer, 1997.

[5] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. *Numerical geometry of non-rigid shapes*. Springer, 2008.

[6] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. The video genome. Technical report, 2010.

[7] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV*, 2007.

[8] A. C. S. Chung, W. M. W. III, A. Norbash, and W. Grimson. Multi-modal image registration by minimizing kullback-leibler distance. In *Proc. MICCAI*, pages 525–532. Springer, 2002.

[9] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS*, 102(21):7426–7431, 2005.

[10] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proc. European Conf. Computational Learning Theory*, 1995.

[11] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios. Dense image registration through mrfs and efficient linear programming. *MedIA*, 12(6):731–741, 2008.

[12] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Proc. NIPS*, 2005.

[13] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. ACM Symp. Theory of Computing*, 1998.

[14] P. Jain, B. Kulis, and K. Grauman. Fast image search for learned metrics. In *Proc. CVPR*, 2008.

[15] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV*, 2008.

[16] N. Komodakis, G. Tziritas, and N. Paragios. Performance vs computational efficiency for optimizing single and dynamic mrfs: Setting the state of the art with primal-dual strategies. *CVIU*, 112(1):14–29, 2008.

[17] D. Lee, M. Hofmann, F. Steinke, Y. Altun, N. D. Cahill, and B. Schölkopf. Learning similarity measure for multi-modal 3d image registration. In *Proc. CVPR*, 2009.

[18] D. Lowe. Distinctive image features from scale-invariant keypoint. *IJCV*, 2004.

[19] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[20] N. J. Mitra, L. Guibas, J. Giesen, and M. Pauly. Probabilistic fingerprints for shapes. In *Proc. SGP*, 2006.

[21] M. Ovsjanikov, A. Bronstein, M. Bronstein, and L. Guibas. Shape google: a computer vision approach to invariant shape retrieval. In *Proc. NORDIA*, 2009.

[22] G. Rosman, A. Bronstein, M. Bronstein, and R. Kimmel. Topologically constrained isometric embedding. *Human Motion: Understanding, Modelling, Capture, and Animation*, 243:243, 2008.

[23] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[24] G. Shakhnarovich. *Learning task-specific similarity*. PhD thesis, MIT, 2005.

[25] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *Proc. CVPR*, page 750, 2003.

[26] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The Princeton shape benchmark. In *Proc. SMI*, pages 167–178, 2004.

[27] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. CVPR*, 2003.

[28] R. Sumner and J. Popović. Deformation transfer for triangle meshes. In *Intl. Conf. Computer Graphics and Interactive Techniques*, pages 399–405, 2004.

[29] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Proc. SGP*, 2009.

[30] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[31] R. Toldo, U. Castellani, and A. Fusiello. Visual vocabulary signature for 3D object retrieval and partial matching. In *Proc. 3DOR*, 2009.

[32] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *Proc. CVPR*, 2008.

[33] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 57(2):137–154, 2002.

[34] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *Proc. NIPS*, 2006.

[35] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. *Proc. NIPS*, 21:1753–1760, 2009.

[36] W. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35–51, 1996.

[37] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In *Proc. NIPS*, 2003.