SIAM J. IMAGING SCIENCES
Vol. 6, No. 4, pp. 2370–2403

Image Registration with Sparse Approximations in Parametric Dictionaries*

Alhussein Fawzi† and Pascal Frossard†

Abstract. We examine in this paper the problem of image registration from the new perspective where images are given by sparse approximations in parametric dictionaries of geometric functions. We propose a registration algorithm that looks for an estimate of the global transformation between sparse images by examining the set of relative geometrical transformations between the respective features. We propose a theoretical analysis of our registration algorithm, and we derive performance guarantees based on two novel important properties of redundant dictionaries, namely the robust linear independence and the transformation inconsistency. We propose several illustrations and insights about the importance of these dictionary properties and show that common properties such as coherence or the restricted isometry property fail to provide sufficient information in registration problems. We finally show with illustrative experiments on simple visual objects and handwritten digit images that our algorithm outperforms baseline competing methods in terms of transformation-invariant distance computation and classification.

Key words. image alignment, sparse approximation, parametric dictionary, dictionary properties

AMS subject classifications. 68U10, 94A08, 68T10, 90C59

DOI. 10.1137/130907872

1. Introduction.

With the ever-increasing quantity of information produced by sensors, efficient processing techniques for identifying meaningful information in high-dimensional data sets become crucial. One of the key challenges is to be able to identify relevant objects captured at different times, from various viewpoints, or by different sensors. Sparse signal representations, which linearly decompose signals into key features, have recently been shown to be a powerful tool in image analysis tasks [35, 20, 8]. In general, it is, however, necessary to align signals a priori in order to derive meaningful comparisons or distances in the analysis. Image alignment or *registration* thus represents a crucial yet nontrivial task in many image processing and computer vision applications, such as object detection, localization, and classification, to name a few.

In this paper, we propose a registration algorithm for sparse images that are given as a linear combination of geometric features drawn from a parametric *dictionary*. The estimation of the global geometric transformation between images is performed first by building a set of candidate transformation solutions with all the relative transformations between features in each image. The transformation that leads to the smallest transformation-invariant distance is finally selected as the global transformation estimate. While image registration is generally a complex optimization problem, our algorithm offers a low complexity solution when the im-

*Received by the editors January 30, 2013; accepted for publication (in revised form) August 21, 2013; published electronically November 26, 2013.
http://www.siam.org/journals/siims/6-4/90787.html
†Signal Processing Laboratory (LTS4), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland (alhussein.fawzi@epfl.ch, pascal.frossard@epfl.ch).

ages have a small number of constitutive components. We analyze its theoretical performance, which mainly depends on the construction of the dictionary that supports the sparse image representations. We introduce two novel properties for redundant dictionaries, namely the robust linear independence (RLI) and transformation inconsistency, which permit us to characterize the performance of the registration algorithm. The benefits of these properties are studied in detail and compared to common properties such as the coherence or the restricted isometry property (RIP). We finally provide illustrative registration and classification experiments, where our algorithm outperforms baseline solutions from the literature, particularly when relative transformations between images are large.

The image registration problem has been widely investigated from different perspectives in the literature, but not from the point of view of sparse image approximations as studied in this paper. Image registration algorithms are usually classified into direct (pixel-based) methods and feature-based methods [30]. We review these two classes of methods, and refer the reader to [37, 30] for a general survey on image alignment.

Direct pixel-based methods simply consist in trying all candidate transformations and seeing how much the pixels agree when the images are transformed relative to each other. A major drawback of these methods is their inefficiency when the number of candidate transformations becomes large. Therefore, hierarchical coarse-to-fine techniques based on image pyramids have been developed [1, 34] to offer a compromise between accuracy and computational complexity. In a different approach, the authors of [27, 36] formulate the registration problem as a low-rank matrix recovery problem with sparse noise and leverage the recent advances in convex optimization to find the optimal transformation that best aligns the images. The approaches developed in [28, 11] map the images to a canonical space where deformations take a simple form and thus allows easier registration.

The popular *feature-based* approaches [31] represent a more efficient class of methods for image registration. They are usually built on several steps: (i) *feature detection*, which searches for stable distinctive locations in the images, (ii) *feature description*, which provides a description of each detected location with an invariant descriptor, (iii) *feature matching* between the images, and (iv) *transformation estimation*, which estimates the global transformation by looking at matched features. Note that it is crucial in this class of methods to describe the features in a transformation-invariant way for easier matching. We refer the reader to [23] for a comparison of the main different methods. A popular example of the feature-based approach relies on the scale invariant feature transform (SIFT) [19] that combines the difference-of-Gaussian (DoG) detector with a descriptor based on image gradient orientations around the keypoint. The SIFT method is invariant to rotation, scaling, and translation, and some of its extensions achieve invariance to affine transformations [25]. Moreover, the SIFT descriptors are often used in combination with affine-invariant detectors such as those proposed in [22, 24, 16, 2] for affine image registration. Even though SIFT has been very successful in many computer vision applications, it is built mostly on empirical results, and several parameters need to be set manually. Feature-based methods in general are not well suited for estimating large transformations between target images, as the matching accuracy and keypoint localization degrade for large transformations.

Finally, we mention some recent advances in transformation-invariant distance estimation, which is closely related to image registration. The transformation-invariant distance is defined

as the minimum distance between the possible transformations of two patterns. In general, the signals generated by the possible transformations of a pattern can be represented by a nonlinear manifold. Computing the transformation-invariant distance between two patterns or, equivalently, the manifold distance is thus a difficult problem in general. The authors in [29] locally approximate the transformation-invariant distance with the distance between the linear spaces that are tangent to both manifolds. Vasconcelos and Lippman [33] go beyond the limitations of local invariance in tangent distance methods by embedding the tangent distance computation in a multiresolution framework. Kokiopoulou and Frossard in [17] achieve global invariance by approximating the original pattern with a linear combination of atoms from a parametric dictionary. Thanks to this approximation, the manifold is given in a closed form, and the objective function becomes equal to a difference of convex functions that can be globally minimized using cutting plane methods. Unfortunately, this class of optimization methods has a slow convergence rate with complexity limitations in practical settings.

In this paper, we propose to examine the image registration problem from a novel perspective by building on our earlier work [9] where we consider that images are given in the form of sparse approximations. Unlike the existing methods, this approach guarantees invariance to transformations of arbitrary magnitude and is generic with respect to the transformation group considered in the registration problem. The detailed analysis of our new framework further provides useful insights on the connections between image registration problems and sparse signal processing.

The rest of this paper is organized as follows. In section 2, we formulate the problem of registration of sparse images and present our registration algorithm. Section 3 proposes a theoretical performance analysis of our algorithm and introduces two new dictionary properties. We finally present illustrative experiments in section 4.

2. Registration of sparse images.

2.1. Preliminaries. We first define the notation and conventions used in this paper. We denote, respectively, by $\mathbb{R}$, $\mathbb{R}^+$, $\mathbb{R}_*^+$ the set of real numbers, the set of nonnegative real numbers, and the set of positive real numbers. We consider images to be continuous functions in $L^2 = \{f : \mathbb{R}^2 \to \mathbb{R} : \int_{-\infty}^{+\infty} |f(x)|^2 dx < \infty\}$. We denote the scalar product associated with L^2 as $\langle f, g\rangle = \int_{-\infty}^{+\infty} f(x)g(x)dx$, and the norm by $\|f\|_2 = \sqrt{\int_{-\infty}^{+\infty} |f(x)|^2 dx}$. Then, we define $\mathcal{T}$ to be a transformation group and denote by $\circ$ its associated composition rule. We consider that the group $\mathcal{T}$ includes the transformations between pairs of images in our registration problem. We represent any transformation $\eta \in \mathcal{T}$ by a vector in $\mathbb{R}^P$ (where P denotes the dimension of $\mathcal{T}$) containing the parameters of the transformation.

Alternatively, we represent a transformation $\eta \in \mathcal{T}$ with its unitary representation $U(\eta)$ in L^2. Therefore, for any $\eta \in \mathcal{T}$, $U(\eta)$ is the function that maps an image f to its transformed image $U(\eta)f \in L^2$ by η. Moreover, as $U(\eta)$ is a unitary operator, we have $\|U(\eta)f\|_2 = \|f\|_2$. In order to avoid heavy notation, we also use f_η to denote $U(\eta)f$. We give in Table 1 some examples of transformation groups and their unitary representation in L^2.

The group $\mathbb{R}^2$ is the group of translations in the plane. The special Euclidean group $SE(2)$ is the group of translations and rotations in the plane. Its dimension is equal to 3 (two degrees of freedom are associated with the translation, and one is associated with rotation). The similarity group $SIM(2)$ of the plane is the set of transformations consisting

Table 1

Examples of transformation groups and their unitary representation in L^2. Parameters with a prime are associated with a secondary transformation η', and R_θ denotes the rotation matrix with angle θ.

Group	Parameters	Composition	Unitary representation
	η	$\eta \circ \eta'$	$U(\eta)f = f_\eta$
$\mathbb{R}^2$	b	$b + b'$	$f(x_1 - b_1, x_2 - b_2)$
Special Euclidean group $SE(2)$	(b, θ)	$(b + R_\theta b', \theta + \theta')$	$f\left(R_{-\theta}(x - b)\right)$
Similarity group $SIM(2)$	(b, a, θ)	$(b + aR_\theta b', aa', \theta + \theta')$	$a^{-1} f(\frac{R_{-\theta}}{a}(x - b))$

of translations, isotropic dilations, and rotations. This group plays a particular importance in transformation-invariant image processing since it contains the basic transformations to which we usually want to be invariant.

Finally, if $c \in \mathbb{R}^n$ and $1 \leq p < \infty$, we denote by $\|c\|_p$ the ℓ_p norm of c defined by $\|c\|_p = \left(\sum_{i=1}^n |c_i|^p\right)^{1/p}$. Note that the notation $\|\cdot\|_2$ is overloaded since it denotes either the continuous L^2 norm or the discrete ℓ_2 norm. However, the distinction between both cases will be clear from the context.

2.2. Problem formulation. We formulate now the registration problem that we consider in the paper. Let I_1 and I_2 be two images in L^2. We are interested in computing the optimal transformation between images I_1 and I_2. Hence, we formulate the original alignment problem as follows:

$$(\mathrm{P}'): \text{ Find } \eta_0' = \operatorname*{argmin}_{\eta \in \mathcal{T}} \|U(\eta)I_1 - I_2\|_2 .$$

We denote by $d(I_1, I_2) = \|U(\eta_0')I_1 - I_2\|_2$ the *transformation-invariant distance* between I_1 and I_2. It corresponds to the regular Euclidean distance when the images are aligned optimally in the L^2 sense. Unfortunately, computing the transformation η_0' and the transformation-invariant distance $d(I_1, I_2)$ is a hard problem since the objective function is typically nonconvex and exhibits many local minima.

In order to circumvent this problem, we consider that the images are well approximated by their sparse expansion in a series of geometric functions. Specifically, let $\mathcal{D}$ be a *parametric dictionary of geometric features* constructed by transforming a generating function $\phi \in L^2$ as follows:

$$\mathcal{D} = \{\phi_\gamma : \gamma \in \mathcal{T}_d\} \subset L^2, \tag{2.1}$$

where $\mathcal{T}_d \subset \mathcal{T}$ is a finite discretization of the transformation group $\mathcal{T}$ and $\phi_\gamma = U(\gamma)\phi$ denotes the transformation of the generating function ϕ by γ. We denote by p and q the respective K-sparse approximations of I_1 and I_2 in the dictionary $\mathcal{D}$:

$$\begin{aligned} p &= \sum_{i=1}^K c_i \phi_{\gamma_i}, \\ q &= \sum_{i=1}^K d_i \phi_{\delta_i}. \end{aligned} \tag{2.2}$$

Since the dictionary $\mathcal{D}$ contains features that represent potential parts of the image, we assume that coefficients c_i and d_i are all nonnegative so that the different features do not cancel each other.

We refer to any element ϕ_γ in $\mathcal{D}$ as a *feature* or *atom*. We suppose in this paper that the generating function ϕ is nonnegative. In addition, we suppose for simplicity that $\gamma \mapsto \phi_\gamma$ defines a one-to-one mapping. This assumption means that the generating function does not have any symmetries in $\mathcal{T}$.[1] Finally, we suppose without loss of generality that the mother function ϕ is normalized so that $\|\phi\|_2 = 1$.

We can now reformulate the registration problem as the problem of finding the optimal relative transformation between sparse patterns. In particular, we reformulate our registration problem as follows:

$$\text{(P): Find } \eta_0 = \operatorname*{argmin}_{\eta \in \mathcal{T}} \|U(\eta)p - q\|_2 .$$

The smallest distance $d(p,q) = \|U(\eta_0)p - q\|_2$ is the transformation-invariant distance computed between the sparse image approximations p and q. Compared to the original problem, the images I_1 and I_2 are replaced by their respective sparse approximations p and q. This presents some potential advantages in applications where users do not have access to the original images; more importantly, the prior information on the support of p and q effectively guides the registration process, as we will see in the next paragraph. We should note that if the images are not well approximated by their sparse expansions, the solution of (P) may substantially differ from the true transformation obtained by solving (P′).

2.3. Registration algorithm. Now we propose a novel and simple algorithm to solve the registration problem for images given by their sparse approximations. The core idea of our registration algorithm lies in the *covariance* property of the dictionary $\mathcal{D}$: a global transformation applied on the image induces an equivalent transformation on the corresponding features.[2] Thanks to this covariance property, it is possible to infer the global transformation between the images by a simple computation of the relative transformations between the features in both images.

Specifically, let $\mathcal{T}_a^{p,q}$ be the set of relative transformations between pairs of features taken, respectively, in p and q: $\mathcal{T}_a^{p,q} = \{\delta_i \circ \gamma_j^{-1} : 1 \le i,j \le K\}$. We can thus estimate the relative transformation between the images by solving the following relaxed problem of (P):

$$\text{(}\hat{\text{P}}\text{): Find } \hat{\eta} = \operatorname*{argmin}_{\eta \in \mathcal{T}_a^{p,q}} \|U(\eta)p - q\|_2 .$$

The minimum of the objective function $d_a(p,q) = \|U(\hat{\eta})p - q\|_2$ is defined as the *approximate transformation-invariant distance* between I_1 and I_2.

Even though problems (P) and ($\hat{\text{P}}$) share some similarities, they differ in an important aspect, the search space. It is reduced from $\mathcal{T}$ to the finite set $\mathcal{T}_a^{p,q}$. This constrains the

[1] We extend this assumption to the more general setting where the stabilizer of ϕ defined by $\mathcal{S}_\phi = \{\gamma \in \mathcal{T} : U(\gamma)\phi = \phi\}$ is a finite set in Appendix B.

[2] The meaning of covariance that is used in this paper is not to be confused with that of covariance used in statistics.

estimated transformation to be equal to a transformation that exactly maps two features taken, respectively, from p and q. The assumption that $\mathcal{T}$ can be replaced by $\mathcal{T}_a^{p,q}$ originates from the observation that features are covariant to the global transformation applied on the original image. Even though this assumption is not necessarily true for all features when innovation exists between the images (other than a global transformation), we expect to have at least one feature whose transformation is consistent with the optimal transformation η_0. We analyze in detail the error due to this assumption in section 3. The advantage of replacing $\mathcal{T}$ by $\mathcal{T}_a^{p,q}$ is, however, immediate: we have reduced an intractable problem to a problem whose search space is of cardinality at most K^2. Since K is generally chosen to be small enough, the problem ($\hat{\mathrm{P}}$) can be efficiently solved by a full search over all the elements of $\mathcal{T}_a^{p,q}$. The registration algorithm is summarized in Algorithm 1.

Algorithm 1. Image registration algorithm.

Input: sparse approximations $p = \sum_{i=1}^{K} c_i \phi_{\gamma_i}$ and $q = \sum_{i=1}^{K} d_i \phi_{\delta_i}$.

1. Construct the set $\mathcal{T}_a^{p,q}$:

$$\mathcal{T}_a^{p,q} = \{\delta_i \circ \gamma_j^{-1} : 1 \leq i, j \leq K\}.$$

2. Estimate the transformation $\hat{\eta}$ and $d_a(p,q)$:

$$\hat{\eta} \leftarrow \operatorname*{argmin}_{\eta \in \mathcal{T}_a^{p,q}} \|U(\eta)p - q\|_2 ,$$
$$d_a(p,q) \leftarrow \|U(\hat{\eta})p - q\|_2 .$$

3. Return $(\hat{\eta}, d_a(p,q))$.

The value of K controls the computational complexity of Algorithm 1: a large value of K results in a large cardinality of the search space $\mathcal{T}_a^{p,q}$. Furthermore, the value of K also generally controls the error in the approximation of the original images by their sparse expansions p and q. We discuss in more detail the influence of K on our registration algorithm in section 4. Note finally that we have supposed for simplicity that both images I_1 and I_2 are approximated by the same number of features. However, it is easy to see that one can generalize it to the case where the numbers of features are different in the two images. In this case, we have $|\mathcal{T}_a^{p,q}| = K_1 K_2$ instead of K^2, where K_1 and K_2 are the number of features in I_1 and I_2, respectively.

In the next section, we analyze the performance of the proposed registration algorithm in different settings and focus in particular on the influence of the dictionary $\mathcal{D}$ on the registration performance.

3. Theoretical analysis.

In this section, we examine the penalty of relaxing the original problem (P') into ($\hat{\mathrm{P}}$) in terms of registration performance. We first discuss the framework and the assumptions used in our analysis. Then, we study a simple case where the image patterns are exactly related by a (possibly very large) geometrical transformation. We show that under a mild assumption on the dictionary, our algorithm achieves perfect registration. We

then extend the analysis to the general case and introduce two key properties of the dictionary (namely, *RLI* and *transformation inconsistency*). We show that under some conditions on these properties, our algorithm succeeds in recovering the correct relative transformation with a bounded error in the general case, as long as the innovation between the images (other than the global geometrical transformation) is controlled. We give at each step of the analysis the main intuitions and several examples to illustrate the novel notions introduced in our analysis.

3.1. Analysis framework. First we define a performance metric to measure the image registration accuracy. As we want to capture the performance of our registration algorithm with respect to the optimal image alignment obtained by solving (P′), a natural metric consists in computing the difference between the transformation-invariant distance and its approximate version, i.e., $E'(p,q,I_1,I_2) = |d_a(p,q) - d(I_1,I_2)|$. We assume in this paper, however, that the images are given by their sparse expansions. Therefore, we use an alternative registration performance given by $E(p,q) = d_a(p,q) - d(p,q)$, where we use the transformation-invariant distance computed between the sparse image approximations p and q instead of the original images. Note that $E(p,q) \geq 0$ since $\mathcal{T}_a^{p,q} \subset \mathcal{T}$.

We relate in the following proposition the two registration metrics $E(p,q)$ and $E'(p,q,I_1,I_2)$ to the sparse approximation errors $\|I_1 - p\|_2$ and $\|I_2 - q\|_2$.

Proposition 3.1. $E'(p,q,I_1,I_2) \leq E(p,q) + \|I_1 - p\|_2 + \|I_2 - q\|_2$.

Proof. We have

$$\begin{aligned} E'(p,q,I_1,I_2) &= |d_a(p,q) - d(I_1,I_2)| \\ &= |d_a(p,q) - d(p,q) + d(p,q) - d(I_1,I_2)| \\ &\leq E(p,q) + |d(p,q) - d(I_1,I_2)|, \end{aligned}$$

using the triangle inequality. We now show that $|d(p,q) - d(I_1,I_2)| \leq \|I_1 - p\|_2 + \|I_2 - q\|_2$. Let $\eta \in \mathcal{T}$. We have

$$\begin{aligned} \|U(\eta)I_1 - I_2\|_2 &= \|U(\eta)(p + I_1 - p) - (q + I_2 - q)\|_2 \\ &= \|U(\eta)p - q + U(\eta)(I_1 - p) - (I_2 - q)\|_2. \end{aligned}$$

Using the triangle inequality, we derive a lower and an upper bound as follows:

$$\begin{aligned} &\|U(\eta)p - q\|_2 - \|U(\eta)(I_1 - p)\|_2 - \|I_2 - q\|_2 \\ &\quad \leq \|U(\eta)I_1 - I_2\|_2 \leq \|U(\eta)p - q\|_2 + \|U(\eta)(I_1 - p)\|_2 + \|I_2 - q\|_2. \end{aligned}$$

As U is a unitary operator, we have $\|U(\eta)(I_1 - p)\|_2 = \|I_1 - p\|_2$. Hence, rewriting the previous equation, we get

(3.1)
$$\|U(\eta)p - q\|_2 - \|I_1 - p\|_2 - \|I_2 - q\|_2 \leq \|U(\eta)I_1 - I_2\|_2 \leq \|U(\eta)p - q\|_2 + \|I_1 - p\|_2 + \|I_2 - q\|_2.$$

Recall that $d(p,q) = \min_{\eta \in \mathcal{T}} \|U(\eta)p - q\|_2$ and $d(I_1,I_2) = \min_{\eta \in \mathcal{T}} \|U(\eta)I_1 - I_2\|_2$. Hence, by taking the minimum over all $\eta \in \mathcal{T}$, we obtain $|d(I_1,I_2) - d(p,q)| \leq \|I_1 - p\|_2 + \|I_2 - q\|_2$, which concludes the proof of the proposition. ∎

When most of the energy of I_1 and I_2 is captured by p and q (namely, when $\|I_1-p\|_2+\|I_2-q\|_2$ is small), the registration errors $E(p,q)$ and $E'(p,q,I_1,I_2)$ are equivalent. We suppose in the rest of this section that this condition is satisfied, and we measure the registration error with $E(p,q) = d_a(p,q) - d(p,q)$. Hence we focus exclusively in this analysis on the penalty induced by restricting the search space $\mathcal{T}$ to $\mathcal{T}_a^{p,q}$, that is, the penalty induced by relaxing the problem (P) into the problem ($\hat{\text{P}}$) in the above section.

Before studying the registration performance, we describe additional assumptions on the discretization of the transformation group $\mathcal{T}$. Recall that the transformation η_0 optimally aligns p and q in the L^2 sense in problem (P). We assume that it satisfies the following assumptions:

$$\eta_0 \circ \gamma_i \in \mathcal{T}_d \quad \text{for all } i \in \{1,\dots,K\}, \tag{3.2}$$

$$\eta_0^{-1} \circ \delta_i \in \mathcal{T}_d \quad \text{for all } i \in \{1,\dots,K\}, \tag{3.3}$$

where $\mathcal{T}_d$ is the discretization of $\mathcal{T}$ used to construct dictionary $\mathcal{D}$ as given in (2.1). These hypotheses state that the atoms of $U(\eta_0)p$ and $U(\eta_0^{-1})q$ belong to the dictionary, where $U(\eta_0)p$ is the optimal alignment of p with q and $U(\eta_0^{-1})q$ is the optimal alignment of q with p. As η_0 is obviously not known beforehand, it is difficult to verify this assumption in practice. However, we can assume that (3.2) and (3.3) hold when the parameter space used to design $\mathcal{D}$ is discretized finely.

Finally, the assumptions in our performance analysis can be summarized as follows:

$$\begin{aligned} &(\text{A}_1) \quad \|I_1 - p\|_2 + \|I_2 - q\|_2 \approx 0, \\ &(\text{A}_2) \quad \eta_0 \circ \gamma_i \in \mathcal{T}_d, \\ &\qquad\quad \eta_0^{-1} \circ \delta_i \in \mathcal{T}_d. \end{aligned}$$

3.2. Registration performance with exact pattern transformation. In our performance analysis, we first consider the special case where $d(p,q) = 0$. This means that there exists a transformation $\eta_0 \in \mathcal{T}$ for which $q = U(\eta_0)p$; i.e., the sparse image approximations can be aligned exactly. We show that in this case, our registration algorithm is able to recover the exact global transformation between p and q, as long as any subset of size $2K$ in $\mathcal{D}$ is linearly independent. We have the following proposition.

Proposition 3.2. *Suppose that any subset of size $2K$ in $\mathcal{D}$ is linearly independent. In this case, if $d(p,q) = 0$, then $E(p,q) = 0$.*

Proof. If $d(p,q) = 0$, then we have $\sum_{i=1}^K c_i \phi_{\eta_0\circ\gamma_i} - \sum_{i=1}^K d_i \phi_{\delta_i} = 0$. Thanks to the linear independence of any subset of size $2K$ in $\mathcal{D}$, for any γ_i there exists δ_j such that $\phi_{\eta_0\circ\gamma_i} = \phi_{\delta_j}$. Indeed, if this is not the case, we could write $\phi_{\eta_0\circ\gamma_i}$ as a linear combination of $2K-1$ atoms in $\mathcal{D}$ that are all different from $\phi_{\eta_0\circ\gamma_i}$ and that all belong to $\mathcal{D}$ thanks to assumption (A_2). This contradicts the assumption that any subset of $2K$ atoms in $\mathcal{D}$ is linearly independent. Then, since the mapping $\gamma \mapsto U(\gamma)\phi$ is a one-to-one function thanks to our dictionary design assumption, we have $\eta_0 \circ \gamma_i = \delta_j$. Thus, $\eta_0 = \delta_j \circ \gamma_i^{-1} \in \mathcal{T}_a^{p,q}$, and $d_a(p,q) = \min_{\eta\in\mathcal{T}_a^{p,q}} \|U(\eta)p - q\|_2 = d(p,q) = 0$. ∎

We can make the following remark about the design of the dictionary. The linear independence assumption guarantees that, when two K-sparse signals are equal, they have at least

one atom in common.[3] If this condition is violated, the patterns $U(\eta_0)p$ and q can have several decompositions in the dictionary with disjoint supports. In this case, all the features of the transformed patterns $U(\eta_0)p$ and q are distinct, which generally lead to $d_a(p,q) \neq d(p,q)$. Note that this assumption appears in many problems related to overcomplete dictionaries since it guarantees the uniqueness of K-sparse decompositions [6, 4, 32].

Finally, since Proposition 3.2 ensures that $E(p,q) = 0$ for an exactly transformed pattern, and we have $E'(p,q,I_1,I_2) \approx E(p,q)$ when the sparse approximation errors are not too large (assumption (A_1)), we can guarantee that the registration error $E'(p,q,I_1,I_2)$ is small in this case.

3.3. Registration performance in the general case.

3.3.1. Bound on the registration error. We now study the performance of our registration algorithm in the general case. The previous result applies only to an ideal scenario since the condition $d(p,q) = 0$ is rarely satisfied in practice. There is usually some slight innovation between the images (other than a transformation in $\mathcal{T}$), which results in a distance $d(p,q)$ that is nonzero. In addition, even when the original images are exactly related by a global transformation (i.e., $d(I_1,I_2) = 0$), there is no guarantee that the sparse approximations can be perfectly aligned (i.e., $d(p,q) = 0$), due to the discretization of the dictionary.

We study the general case where the sparse image approximations p and q have differences that cannot be explained by a global geometric transformation in $\mathcal{T}$. In more detail, when c and d denote, respectively, the coefficient vectors for patterns p and q following (2.2), we suppose that there exists a real number $\epsilon > 0$ such that $d(p,q) < \epsilon\sqrt{\|c\|_2^2 + \|d\|_2^2}$. The quantity ϵ therefore measures the normalized innovation between p and q.

We now turn to the main result of our paper, which is formulated in Theorem 3.3. This result relates the error of the registration algorithm in Algorithm 1 to the properties of the dictionary, namely *RLI* and the *transformation inconsistency*. It reads as follows.

Theorem 3.3. *If* $d(p,q) < \epsilon\sqrt{\|c\|_2^2 + \|d\|_2^2}$ *with* $\epsilon > 0$, *then*

$$E(p,q) \leq \alpha\rho \min\left(\|c\|_1, \|d\|_1\right),$$

when $\mathcal{D}$ *is* $(2K,\epsilon,\alpha)$*-robustly linearly independent (RLI) for some* $\alpha \in [0,\sqrt{2})$, *and* ρ *is the transformation inconsistency of* $\mathcal{D}$.

Theorem 3.3 shows that RLI with a small α and a small transformation inconsistency are key properties of the dictionary in order to guarantee the success of our algorithm. The RLI property can be thought of as an extension of the linear independence assumption to the case where $d(p,q) \neq 0$. Specifically, it guarantees the existence of two approximately similar features in $U(\eta_0)p$ and q when $d(p,q)$ is small. The transformation inconsistency captures the fact that geometrical transformations have a different effect on distinct atoms in the dictionary. We defer the proof of Theorem 3.3 to Appendix A, and we study in detail in the rest of this section the novel RLI and transformation inconsistency properties.

[3] The linear independence of any subset of size $2K$ in the dictionary actually guarantees a stronger result: it guarantees that any K-sparse signal has a unique decomposition in $\mathcal{D}$ [7]. In other words, it guarantees that when two K-sparse signals are equal, *all* the atoms are equal.

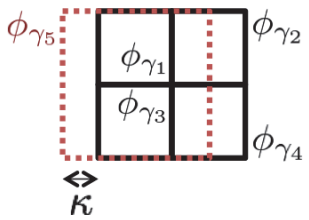


Figure 1. *Example of a linearly independent dictionary $\mathcal{D}$ that induces a large registration error $d_a(p,q) - d(p,q)$, when $p = 1/2(\phi_{\gamma_1} + \phi_{\gamma_2} + \phi_{\gamma_3} + \phi_{\gamma_4})$ and $q = \phi_{\gamma_5}$. We note that for $\epsilon > \sqrt{\kappa}$, this dictionary is not RLI unless $\alpha \geq 1$. Note that ϵ can be made very small since κ can be chosen to be any positive real number.*

3.3.2. Robust linear independence. We study now in more detail the novel dictionary properties. We first show that the linear independence assumption introduced in section 3.2 is no longer sufficient to bound the registration performance in the case where $d(p,q) \neq 0$ (but close to zero). To see this, we construct a linearly independent dictionary $\mathcal{D}$ and two sparse patterns p and q for which $d(p,q)$ can be made arbitrarily close to zero (i.e., $\epsilon \to 0$) yet the registration error is large. As illustrated in Figure 1, we consider a dictionary $\mathcal{D}$ containing four square atoms and an additional big square atom parametrized by its position κ with respect to ϕ_{γ_1}. Clearly, when $\kappa \neq 0$, the dictionary $\mathcal{D}$ is linearly independent since one cannot write an atom as a linear combination of the four other atoms. We consider the patterns $p = \frac{1}{2}\sum_{i=1}^{4} \phi_{\gamma_i}$ and $q = \phi_{\gamma_5}$. When κ is small, the transformation that best aligns p and q is the identity transformation.[4] All relative transformations between features in p and q are, however, combinations of dilations and translations, which result in an estimated transformation $\hat{\eta}$ in our algorithm that is significantly different from the identity. Hence we obtain a large registration error $d_a(p,q) - d(p,q)$ in this example. This example shows that the linear independence assumption defined in section 3.2 is fragile: it does not allow us to bound the registration error even when $d(p,q)$ is very small. One needs a more *robust* condition in order to guarantee a small registration error even in cases where the innovation between images is small (but nonzero).

Therefore, we propose to extend the notion of linear independence to a novel property called *robust linear independence (RLI)* to characterize sets of vectors. It is formally defined as follows.

Definition 3.4. *Let $(H, \|\cdot\|)$ be a normed space and $K \geq 1$. A family of vectors $(v_1, \ldots, v_K) \in H^K$ is* (ϵ, α)-robustly linearly independent *(RLI) if the following implication holds for any vector $a \in \mathbb{R}^K$:*

$$\left\| \sum_{i=1}^{K} a_i v_i \right\| < \epsilon \|a\|_2 \implies \exists i, j \text{ with } a_i, a_j \neq 0 \text{ such that } \left\| \frac{a_i v_i}{\|a_i v_i\|} + \frac{a_j v_j}{\|a_j v_j\|} \right\| \leq \alpha. \tag{3.4}$$

In other words, when ϵ and the parameter α are small, any linear combination of vectors that nearly vanishes in an RLI vector set contains at least two vectors that approximately

[4] If we look among all possible transformations, the optimal transformation η_0 is a translation that exactly aligns p and q. However, this transformation does not satisfy the assumptions in (3.2) and (3.3). To illustrate the main issue here, we consider only transformations that satisfy these assumptions. For small κ, the optimal transformation is therefore the identity.

cancel each other.

We now discuss the relation between RLI and linear independence. While linear independence prevents having collinear vectors, it is natural in our registration framework to allow collinear vectors in the dictionary since they represent essentially the same feature. Specifically, as the underlying transformation parameter of collinear atoms is the same, selecting one atom or the other is not important for the purpose of registration.[5] The notion of linear independence where collinear vectors are allowed can be written as follows. For any $a \in \mathbb{R}^K$ such that $a \neq 0$,

$$\sum_i a_i v_i = 0 \implies \exists i, j \text{ with } a_i, a_j \neq 0 \text{ such that } \frac{a_i v_i}{\|a_i v_i\|} + \frac{a_j v_j}{\|a_j v_j\|} = 0.$$

Note that this essentially corresponds to the notion of RLI in the case where $\alpha, \epsilon = 0$. Since we want to study the behavior of the algorithm for *nonzero* innovation between the images, we naturally extend the notion of linear independence (where collinear vectors are allowed) to Definition 3.4; if a linear combination of vectors has a small magnitude (where ϵ quantifies the magnitude), there exist two vectors that approximately cancel each other (where α quantifies this approximation). Note that, for a fixed α, the RLI gets harder to satisfy for a larger ϵ. In addition, for a fixed ϵ, the condition is harder to satisfy for a smaller α.

The following toy example illustrates the notion of RLI in $\mathbb{R}^3$.

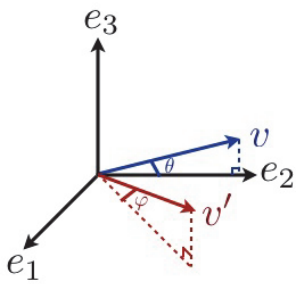


Figure 2. *Illustration of RLI property in* $\mathbb{R}^3$.

Example 1. Consider the setting of Figure 2 with $\theta = \varphi = \pi/20$. Then, for $\epsilon = 0.2$, we have the following:

1. (e_1, e_2, v) is RLI with $\alpha = 0.2$.
2. (e_1, e_2, v') is *not* RLI unless $\alpha \geq 0.78$.

The proof of Example 1 is straightforward from simple trigonometry. The set of vectors (e_1, e_2, v) has a better behavior in terms of RLI than (e_1, e_2, v'). The underlying reason is that v is very close to the vector e_2 (i.e., $\|e_2 - v\|_2$ is close to zero), while v' is close to a *linear combination* of e_1 and e_2 (but not to e_1 or e_2). While it is acceptable to have vectors that are close to each other, the RLI property prevents having a vector that is close to a linear combination of the other vectors. This can also be readily seen in the example of Figure 1.

[5]Note that this is in contrast to recovery problems (e.g., compressed sensing) where collinear vectors (in the measurement matrix) are not allowed, since it will not be possible then to recover the active component of the signals.

The definition of RLI can be extended to dictionaries as follows.

Definition 3.5. *A dictionary $\mathcal{D}$ is (K, ϵ, α)-RLI if any subset of size K in $\mathcal{D}$ is (ϵ, α)-RLI.*

The dictionary in the example of Figure 1 is *not* K-RLI for $K = 5$, with small ϵ (unless α is large). Indeed, by choosing a vector of coefficients $a = [0.5, 0.5, 0.5, 0.5, -1]^T$, we obtain $\|\sum_i a_i v_i\|_2 \approx 0$, yet $\alpha = 1$. Note that the RLI property on the dictionary has to be satisfied in order to obtain a good registration performance, as it ensures the existence of two approximately similar features (in the L^2 sense) in $U(\eta_0)p$ and q, when $d(p, q)$ is small.

We study now in more detail the RLI property on dictionaries. In particular, we examine the main difference between RLI and the well-known RIP [4]. The restricted isometry condition assumes that a collection of vectors behaves almost like an orthonormal system but only for sparse linear combinations. Specifically, the RIP with constant δ_K implies that any linear combination of K elements in the dictionary satisfies

$$\left\| \sum_{i=1}^{K} a_i v_i \right\|^2 \geq (1 - \delta_K) \|a\|_2^2 .$$

By imposing an RIP property on the dictionary $\mathcal{D}$ with $\delta_K \ll 1$, the norm of any sparse linear combination of atoms is guaranteed to be large (i.e., larger than $\sqrt{(1-\delta_K)\|a\|_2}$). In our case, contrary to the RIP, we are interested in linear combinations of atoms that nearly vanish. The RLI property imposes in this case the existence of two atoms that approximately cancel each other in the signal support. Consequently, RLI can be seen as a weak form of RIP, where we allow the norm of linear combinations to be close to zero, provided that two atoms approximately cancel each other in the sense of (3.4). In particular, any dictionary $\mathcal{D}$ that satisfies the RIP property with a parameter δ_K will be $(K, \sqrt{1-\delta_K}, 0)$-RLI. Indeed, since $\|\sum_{i=1}^{K} a_i v_i\| \geq \sqrt{1-\delta_K}\|a\|_2$ holds for any subset of K dictionary elements, the left-hand side of (3.4) cannot be satisfied when $\epsilon = \sqrt{1-\delta_K}$.

Let us consider a simple example to compare the new RLI property with the common ways of characterizing dictionaries, namely, the coherence [32] and the RIP [4].[6]

Example 2 (dictionary of translated box functions). Let $H = L^2(\mathbb{R})$, and define the box function

$$v(t) = \begin{cases} 1 & \text{if } t \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

We consider the infinite-size dictionary $\mathcal{D}_{\text{box}} = \{T_\tau v = v_\tau : \tau \in \mathbb{R}\}$, where T_τ is the translation operator by τ. The dictionary has the following properties:

- $\mathcal{D}_{\text{box}}$ is RIP with a constant $\delta_K(\mathcal{D}_{\text{box}})$ equal to 1 for any $K \geq 2$.
- The coherence of $\mathcal{D}_{\text{box}}$ is equal to 1.
- $\mathcal{D}_{\text{box}}$ is $(K, \epsilon, \epsilon\sqrt{\frac{2}{3}(4^K - 1)})$-RLI for $K \geq 1$ and $\epsilon \in (0, \sqrt{\frac{3}{4^K - 1}})$.

We give in the following a brief overview of the proof of the RLI of $\mathcal{D}$. The interested reader can find the complete proof in our technical report [10].

[6] Even though the definitions of RIP and coherence are originally for vectors in $\mathbb{R}^N$, we consider here a straightforward extension of the definitions of RIP and coherence to the case where vectors are in L^2.

Sketch of the proof. Without loss of generality, suppose that a is a vector in $\mathbb{R}^K$ of unit ℓ_2 norm such that $\|\sum_i a_i v_{\tau_i}\|_2 < \epsilon$ with $\epsilon < Y = \sqrt{\frac{3}{4^K-1}}$. Let i^* be the smallest integer that satisfies $|a_{i^*}| \geq 2^{i^*-1}Y$. Note that such an index always exists, as it would contradict $\|a\|_2 = 1$ otherwise. Let j^* be the smallest integer larger than i^* that satisfies $a_{i^*}a_{j^*} < 0$. Using the definition of i^*, one can show that, for t in $(\tau_{j^*-1}, \tau_{j^*})$, we have $|\sum_i a_i v_{\tau_i}(t)| \geq Y$. We can therefore show that the box functions $v_{\tau_{j^*-1}}$ and $v_{\tau_{j^*}}$ cancel each other: $\epsilon^2 > \|\sum_i a_i v_{\tau_i}\|_2^2 \geq Y^2(\tau_{j^*} - \tau_{j^*-1})$, which leads to $(\tau_{j^*} - \tau_{j^*-1}) \leq \epsilon^2/Y^2$. By relating the quantity $(\tau_{j^*} - \tau_{j^*-1})$ to $\|v_{\tau_{j^*}} - v_{\tau_{j^*-1}}\|_2$, we finally obtain the value of α for the specified value of ϵ. ■

The RIP constant $\delta_K(\mathcal{D}_{\text{box}})$ is 1 for any $K \geq 2$ as two box functions can be made arbitrarily close to each other. Similarly, the coherence of this dictionary is equal to 1. Nevertheless, the dictionary $\mathcal{D}_{\text{box}}$ is RLI with a coefficient α that is proportional to ϵ. One can understand the fact that $\mathcal{D}_{\text{box}}$ is RLI intuitively; if a linear combination of box functions has a small norm, there exist at least two box functions that nearly cancel each other. Note, however, that $\alpha = \epsilon\sqrt{\frac{2}{3}(4^K - 1)}$ grows very fast with K for a fixed ϵ.

Even if the dictionary $\mathcal{D}_{\text{box}}$ hardly satisfies the RIP and is highly coherent, it is still an interesting one in our framework. Indeed, it satisfies the key property that two sparse signals that are close in the L^2 sense have at least two approximately similar features. When applied to our registration problem, this guarantees the existence of two features that are related approximately by a transformation η_0 in the L^2 sense[7] when $d(p, q)$ remains small. This property is at the core of our registration algorithm since we infer the global transformation by looking at the relative transformations between the features.

We finally stress the differences between the proposed RLI property and other dictionary properties such as the RIP, coherence, or, more recently, the properties introduced in [3, 14, 26]. While the latter properties are specifically designed for the task of *signal recovery*, the proposed RLI property is introduced in the context of image registration. This explains in particular why a dictionary can be well behaved in terms of the RLI property despite having coherent atoms. In contrast, coherent columns are forbidden in the context of recovery problems (e.g., compressed sensing) as it is then difficult to distinguish between similar components in the signal reconstruction.

3.3.3. Transformation inconsistency. We now introduce a second dictionary property, *the transformation inconsistency*, that is key to studying the performance of our algorithm. The transformation inconsistency measures the difference in the effect of the same transformation on distinct atoms in the dictionary. It is formally defined as follows for parametric dictionaries given by (2.1).

Definition 3.6. *The transformation inconsistency ρ of a parametric dictionary $\mathcal{D}$ is equal to*

$$\rho = \sup_{\gamma,\gamma' \in \mathcal{T}_d} \sup_{\eta \in \mathcal{T} \setminus \{\mathbb{I}\}} \left\{ \frac{\|U(\eta)\phi_{\gamma'} - \phi_{\gamma'}\|_2}{\|U(\eta)\phi_\gamma - \phi_\gamma\|_2} \right\},$$

where $\mathbb{I}$ is the identity transformation.

The transformation inconsistency ρ is always greater than or equal to 1. Furthermore, when $\mathcal{T}$ is commutative, the transformation inconsistency takes its minimal value and is equal

[7] More precisely, this means that there exist a γ_i and a δ_j such that $\|U(\eta_0)\phi_{\gamma_i} - \phi_{\delta_j}\|_2$ is small.

to 1. Indeed, for any γ, γ' in $\mathcal{T}_d$ and $\eta \in \mathcal{T}$, we have

$$\frac{\|U(\eta)\phi_{\gamma'} - \phi_{\gamma'}\|_2}{\|U(\eta)\phi_{\gamma} - \phi_{\gamma}\|_2} = \frac{\|U(\gamma')(\phi_\eta - \phi)\|_2}{\|U(\gamma)(\phi_\eta - \phi)\|_2} = \frac{\|\phi_\eta - \phi\|_2}{\|\phi_\eta - \phi\|_2} = 1.$$

Hence, taking the supremum over all $\eta \in \mathcal{T}$ and atoms $\gamma, \gamma' \in \mathcal{T}_d$ results in having $\rho = 1$. This is expected since when $\mathcal{T}$ is commutative, a fixed transformation acts on all atoms similarly.

On the other hand, a large value of the transformation inconsistency ρ (i.e., $\rho \gg 1$) means that there exist two atoms in the dictionary that are affected in a very different way when they are subject to the same transformation. The transformation inconsistency plays a key role in our registration algorithm. Indeed, as the global transformation between two sparse patterns is estimated from one of the relative transformations between features, it is preferable that transformations act in a similar way on all the features of the sparse patterns for more consistent registration. That means that dictionaries with small transformation inconsistency provide better registration performance.

In order to outline the importance of this novel property in our registration framework, we give a few illustrative examples of dictionaries with different transformation inconsistency parameters.

Example 3 (dictionary with quasi-isotropic mother function, $\mathcal{T} = SE(2)$). We consider $\mathcal{T}$ to be the special Euclidean group ($\mathcal{T} = SE(2)$). That is, $\mathcal{T}$ accounts for translations, rotations, and combinations of those. We consider an ellipse-shaped mother function ϕ as shown in Figure 3(a) with anisotropy $r = \frac{l}{L}$. Then, we suppose for the sake of simplicity that $\mathcal{T}_d = \mathcal{T}$ (i.e., the dictionary is built by applying all transformations $\gamma \in \mathcal{T}$ to the generating function ϕ).

We illustrate in Figure 3(b) the effect of transformation η, which is a simple rotation, on two different atoms with parameters γ and γ' positioned at different points in the two-dimensional (2D) plane. While the rotation of the atom parametrized by γ induces a very slight change on it (when $r \approx 1$), the same rotation applied on the atom $\phi_{\gamma'}$ completely changes its position. This is due to the fact that translations and rotations do not commute. Hence, the transformation η has a very different impact on atoms ϕ_γ and $\phi_{\gamma'}$, and we get $\rho \to \infty$ from Definition 3.6. Therefore, when the generating function ϕ approaches isotropy, the transformation inconsistency grows to infinity.

In this example, our registration algorithm is not guaranteed to have a small error. To illustrate it, let us consider the patterns p and q illustrated in Figure 3(c), which are each composed of two atoms whose coefficients are all equal. The distance $d(p, q)$ between the patterns can be made arbitrarily small with a generating function that is close to isotropic (i.e., $r \to 1$), while the minimal distance $d_a(p, q)$ in our algorithm remains large. Indeed, since our algorithm considers only relative transformations between pairs of atoms, the estimated global transformation between the patterns can only be equal to a combination of a translation and a rotation of $\frac{\pi}{2}$. However, when $r \approx 1$, the optimal transformation is clearly the identity, which cannot be selected with our algorithm: this results in a large registration error $d_a(p, q) - d(p, q)$. Note that the error here is entirely related to the fact that the transformation inconsistency ρ is large, and not to the RLI property since the dictionary under consideration here is RLI for small values of the sparsity K.

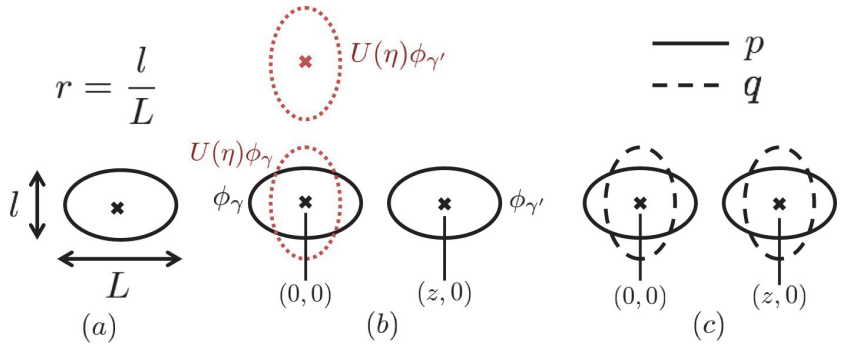


Figure 3. *Example of a dictionary where the transformation inconsistency ρ is large.* (a) *Mother function of the dictionary with anisotropy $r = \frac{l}{L}$.* (b) *Atoms ϕ_γ, $\phi_{\gamma'}$ and a transformation η that leads to a large transformation inconsistency ρ.* (c) *Examples of patterns p (atoms represented with solid line) and q (atoms represented with dashed line), where our algorithm has a large registration error $d_a(p,q) - d(p,q)$.*

To be complete, we should note that the one-to-one mapping assumption defined in section 2.2 for the function $\gamma \mapsto U(\gamma)\phi$ is not satisfied in Example 3, since ϕ has a rotational symmetry of π. In this case, a slightly more complicated definition of the transformation inconsistency ρ has to be made to avoid having $\rho = \infty$ (with the definition of ρ given in Definition 3.6, we obtain $\rho = \infty$ by setting η to be a rotation of π and γ to be the identity and choosing any γ' different from γ). The main intuitions of the transformation inconsistency ρ, as defined in Definition 3.6, however, hold when ϕ has a finite number of symmetries. We study in detail the generalization of the transformation inconsistency ρ to the case where ϕ has symmetries in $\mathcal{T}$ in Appendix B.

Example 4 (dictionary built with translation and isotropic dilations, $\mathcal{T} = \mathcal{T}_d = \mathbb{R}^2 \times \mathbb{R}_*^+$). In this example, we let $\mathcal{T}$ be the group of translations and isotropic dilations. The generating function of the dictionary can have any form, as long as its support is much smaller than the dimension of the image. For example, we can choose a circle-shaped mother function, as depicted in Figure 4(a). Then, we consider the scenario where the two atoms ϕ_γ and $\phi_{\gamma'}$ are separated by z (where z is considered to be very large), as illustrated in Figure 4(b). A transformation η that consists of a small isotropic dilation has a very different effect on both atoms since translations and dilations do not commute. In particular, the transformation η applied to $\phi_{\gamma'}$ results in an atom that has no intersection with $\phi_{\gamma'}$, while the same transformation has almost no effect on ϕ_γ; i.e., $U(\eta)\phi_\gamma \approx \phi_\gamma$. Thus, the transformation inconsistency is very high and $\rho \approx \infty$ according to Definition 3.6. In Figure 4(c), we illustrate why this may cause a problem in our registration algorithm: we consider the two sparse patterns p and q composed of two features each, where the coefficients of all the atoms are equal. It is not hard to see that the optimal global transformation between both patterns is the identity. At the same time, our algorithm can only estimate a global transformation that is a dilation (combined possibly with a translation) since all transformations between pairs of atoms in p and q consist of combinations of dilation and translation.

We refer the reader to our technical report [10] for more examples. Overall, the above examples suggest that, whenever the transformation inconsistency of the dictionary is large, one may construct an example where our registration algorithm poorly approximates the

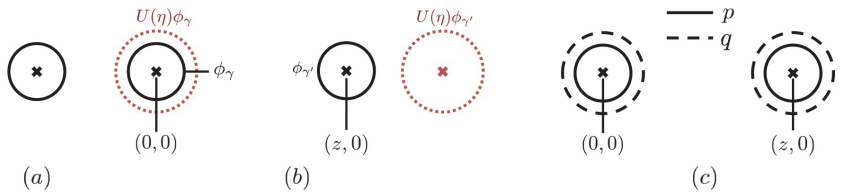


Figure 4. *Example of a dictionary where the transformation inconsistency ρ is large.* (a) *Mother function of the dictionary.* (b) *Atoms ϕ_γ, $\phi_{\gamma'}$ and transformation η that causes ρ to be large.* (c) *Examples of patterns p (atoms represented with solid line) and q (atoms represented with dashed line), where our algorithm has a large registration error $d_a(p,q) - d(p,q)$.*

transformation-invariant distance. It is worth mentioning that even though the previous examples consider localized atoms with finite support, our approach is not constrained to such atoms. In the general setting where $\mathcal{T}$ is any transformation group (and $\mathcal{T}_d = \mathcal{T}$ for the sake of simplicity), an example of failure could be constructed as follows. The basic idea is to build two patterns p and q of the form $p = \phi_\gamma + \phi_{\gamma'}$ and $q = U(\eta_1)\phi_\gamma + U(\eta_2)\phi_{\gamma'}$ for which (i) $p \approx q$, and (ii) $\|U(\eta_1)p - q\|_2$ and $\|U(\eta_2)p - q\|_2$ are large (with respect to $\|p - q\|_2$). The optimal transformation between p and q is then simply the identity, whereas the transformations considered in our algorithm (namely η_1 and η_2, along with $\eta_1 \circ \gamma \circ \gamma'^{-1}$ and $\eta_2 \circ \gamma' \circ \gamma^{-1}$) result in a poor registration performance as they all differ from the identity transformation.

In more detail, when $\rho \gg 1$ we know that there exist two atoms ϕ_γ and $\phi_{\gamma'}$ with $\gamma \in \mathcal{T}$ and $\gamma' \in \mathcal{T}$, along with a transformation η_1 for which $\|U(\eta_1)\phi_\gamma - \phi_\gamma\|_2 \approx 0$, while $\|U(\eta_1)\phi_{\gamma'} - \phi_{\gamma'}\|_2$ is large. By setting $\eta_2 = (\gamma' \circ \gamma^{-1}) \circ \eta_1 \circ (\gamma \circ (\gamma')^{-1})$, we get that $\|U(\eta_2)\phi_{\gamma'} - \phi_{\gamma'}\|_2 = \|U(\eta_1)\phi_\gamma - \phi_\gamma\|_2 \approx 0$. Hence, the norm $\|p - q\|_2$ is necessarily small since $\|p - q\|_2 = \|\phi_\gamma + \phi_{\gamma'} - U(\eta_1)\phi_\gamma - U(\eta_2)\phi_{\gamma'}\|_2 \leq \|\phi_\gamma - U(\eta_1)\phi_\gamma\|_2 + \|\phi_{\gamma'} - U(\eta_2)\phi_{\gamma'}\|_2$. Besides, we know by construction that $\|U(\eta_1)\phi_{\gamma'} - \phi_{\gamma'}\|_2$ is large and $\|U(\eta_2)\phi_\gamma - \phi_\gamma\|_2$ is also generally large since the group $\mathcal{T}$ is noncommutative. This gives us, in general, large values of $\|U(\eta_1)p - q\|_2$ and $\|U(\eta_2)p - q\|_2$. This construction shows that, when the dictionary has a large inconsistency parameter, one can find patterns for which the registration algorithm fails to recover the right global transformation.

In general, the above examples show that it is better to choose a dictionary with a small transformation inconsistency (i.e., ρ small) to have good registration performance irrespective of the patterns to be aligned.

The performance of the registration algorithm depends on the transformation inconsistency as well as on the RLI of the dictionary, as shown in Theorem 3.3. The success of our registration algorithm for *all* sparse signals in the dictionary is guaranteed when the RLI and transformation inconsistency conditions are satisfied. Note that the conditions on the dictionary properties are essentially tight, as one can construct an example where our algorithm fails whenever one of the parameters is large enough. The performance bound should be interpreted more in a qualitative way than in a quantitative way. It provides two rather intuitive conditions for our algorithm to provide low registration error. In order to use this bound quantitatively, however, one has to be able to compute explicitly the newly defined properties on generic dictionaries. We outline here the fact that such a bound could not have been established with traditional measures for characterizing dictionaries, namely coherence

or the RIP constant. Finally, we remark that the result in Theorem 3.3 can be used to bound the registration error $E'(p, q, I_1, I_2)$ thanks to Proposition 3.1. The price to pay in this case is the approximation error $\|I_1 - p\|_2 + \|I_2 - q\|_2$.

4. Image registration experiments. In this section, we evaluate the performance of our algorithm in image registration experiments. First we describe the implementation choices in our registration algorithm. Then we study its performance for different dictionaries and put the results in perspective with the theoretical guarantees in section 3. Finally, we present illustrative image registration and classification experiments with simple test images and handwritten digits.

4.1. Algorithm implementation. In all the experiments of section 4.3, we focus on achieving invariance to translation, rotation, and scaling. Invariance to these transformations is indeed considered to be a minimal requirement in invariant pattern recognition. These three operations generate the *group of similarities* that we denote by $\mathcal{T} = SIM(2)$. Any element in $\mathcal{T}$ is therefore indexed by four parameters: a translation vector $b = (b_x, b_y)$, dilation a, and rotation θ. We now describe the sparse approximation algorithm and the dictionary design used in our experiments.

4.1.1. Sparse approximation algorithm. There are many methods for constructing sparse approximations of images. In our experiments, we use a modified implementation of the matching pursuit (MP) [21] algorithm, as MP is a pretty simple algorithm that works relatively well in practice. It is an iterative algorithm that successively identifies the atoms in $\mathcal{D}$ that best match the image to be approximated. More precisely, MP iteratively computes the correlation between the atoms in $\mathcal{D}$ and the signal residual, which is obtained by subtracting the contributions of the previously chosen atoms from the original image. At each iteration, the atom with the highest correlation is selected, and the residual signal is updated. While the standard MP algorithm solves the sparse approximation problem without positivity constraint on the coefficients, we propose a slightly modified algorithm (that we call nonnegative matching pursuit (NMP)) in order to select atoms that have the highest positive correlation with the residual signal. This choice is driven by the objective of having a part-based signal expansion, where each feature participates in constructing the signal representation. The NMP algorithm is formally defined in Algorithm 2.

One way to choose the sparsity K consists in controlling the approximation error of I_1 and I_2. Specifically, we can impose a stopping criterion in the NMP algorithm of the form $\|r_K\|_2 \leq e$, where r_K is the residual at iteration K and e is a fixed threshold controlling the approximation error. When e is chosen to be small enough, this guarantees a relatively small sparse approximation error.

Note that the complexity of NMP is governed by the selection step; hence $O(K|\mathcal{D}|)$ operations need to be performed. In addition, the complexity of solving $(\hat{\mathrm{P}})$ using Algorithm 1 is $O(K^2N)$ with $N = \max(N_1, N_2)$ with N_1 and N_2, respectively, the dimensions of the discretized images corresponding to p and q. Therefore, if the sparse approximation step is necessary for registration, the complexity of the overall registration algorithm is $O(K|\mathcal{D}| + K^2N)$. Depending on the factor $\frac{|\mathcal{D}|}{KN}$, the complexity of the global registration algorithm might be governed by either the sparse approximation step or the registration step. Overall, the choice of K results from a trade-off between approximation error (hence registration performance)

Algorithm 2. Nonnegative matching pursuit (NMP) for feature extraction.

Input: image I, sparsity K, dictionary $\mathcal{D}$.

Ensure: coefficients c, support Γ.

1. Initialization of the residual: $r_0 \leftarrow I$ and support: $\Gamma \leftarrow \emptyset$.

2. While $1 \leq i \leq K$, do:

2.1. Selection step:

$$\gamma_i \leftarrow \underset{\gamma \in \mathcal{T}_d}{\operatorname{argmax}} \langle r_{i-1}, \phi_\gamma \rangle ,$$
$$\Gamma \leftarrow \Gamma \cup \{\gamma_i\}.$$

2.2. If $\langle r_{i-1}, \phi_{\gamma_i} \rangle \leq 0$, go to **3**.

2.3. Update step:

$$c_i \leftarrow \langle r_{i-1}, \phi_{\gamma_i} \rangle ,$$
$$r_i \leftarrow r_{i-1} - \langle r_{i-1}, \phi_{\gamma_i} \rangle \phi_{\gamma_i}.$$

3. Return c, Γ.

and computational complexity. Finally, note that in applications involving the registration of a *test* image with possibly many *training* images, the sparse approximations of the training images are computed offline. Hence, only the sparse approximation of the test image needs to be computed during the test phase.

4.1.2. Choice of the dictionary. We now discuss the choice of the dictionary $\mathcal{D}$ that is used in our experiments. As pointed out in (2.1), the dictionary $\mathcal{D}$ is simply constructed by applying geometric transformations $\gamma \in \mathcal{T}_d$ to a mother function ϕ. We thus need to choose appropriately the mother function ϕ as well as the discretization for constructing the subset $\mathcal{T}_d$ of $\mathcal{T}$. In light of the derived analytical results, ideally we would like to design a dictionary that satisfies the following constraints:

- Images should have a good sparse approximation in the dictionary (assumption (A_1) of the analysis).
- The dictionary should be RLI (Theorem 3.3).
- The transformation inconsistency parameter of the dictionary should not be too large (Theorem 3.3).

We propose using an anisotropic Gaussian generating function, as it has been shown to provide good approximation results in natural images [12]. It is defined as follows:

$$\phi(x, y) = \frac{1}{\xi} \exp\left(- \left(\frac{x}{\nu} \right)^2 - y^2 \right),$$

where $\nu > 1$ controls the anisotropy and the normalization factor ξ is chosen to have $\|\phi\|_2 = 1$.[8] The choice of $\nu \approx 1$ results in an isotropic mother function that causes the transformation

[8]Formally, the Gaussian mother function does not satisfy the one-to-one mapping assumption of $\gamma \mapsto U(\gamma)\phi$. We circumvent this by slightly modifying the definition of $\mathcal{T}_a^{p,q}$. We define the stabilizer of ϕ to be the set that keeps the mother function unchanged: $\mathcal{S}_\phi = \{\gamma : U(\gamma)\phi = \phi\}$. Then we define $\mathcal{T}_a^{p,q} = \{\delta_i \circ \pi \circ (\gamma_j)^{-1} : 1 \leq i, j \leq K, \pi \in \mathcal{S}_\phi\}$. For more details, we refer the reader to Appendix B.

inconsistency ρ to be very large (see Example 3). The transformation inconsistency is also large when the value of ν is chosen to be large [10]. In our experiments, we have generally chosen an intermediate value $\nu = 4$ as a compromise between the two extreme values.

The dictionary $\mathcal{D}$ is built by transforming the generating function ϕ with all transformations in $\mathcal{T}_d$. In our experiments, we consider the following discretization:

- The translation parameters can take any positive integer value smaller than the image dimension.
- The rotation angles are uniformly discretized in $[0, \pi)$ with a step size of $\frac{\pi}{8}$. We have seen experimentally that this step size results in a good directional accuracy. A denser discretization comes at the expense of higher computational cost.
- The scaling parameters are sampled uniformly on a logarithmic scale with a step size of half an octave. This step size results in a compromise between the sparse approximation error and an oversampling of the scale space that might lead to wrong registration (and a too high computational complexity). We set the minimum scale to 1, and the maximum scale is designed to have 99% of the energy of a centered atom inside the image domain.

Figure 5 illustrates several examples of part-based representations obtained with NMP and a dictionary of Gaussian atoms, as described above. We observe that the part-based decomposition manages to approximate well the main geometric characteristics of the image. Furthermore, the same features are used in the different approximations, up to some geometrical transformation that corresponds to the relative transformation between the different versions of the original image. This is exactly the property that is at the core of our registration algorithm.

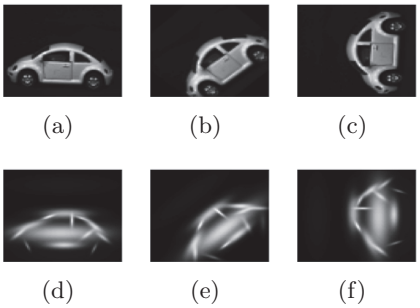


Figure 5. *Sparse approximations of transformed versions of the Car image, computed with NMP and a dictionary constructed from a Gaussian generating function, with $\nu = 4$. The first row shows the original images, size 75×75 pixels. The second row shows the corresponding sparse approximations with a sparsity of $K = 15$ atoms.*

4.1.3. Registration refinement. Our registration algorithm estimates a transformation in the set $\mathcal{T}_a^{p,q} \subset \{\gamma \circ \delta^{-1} : \gamma, \delta \in \mathcal{T}_d\}$, where $\mathcal{T}_d$ is the chosen discretization of the parameter space $\mathcal{T}$. In order to reduce the registration error that is due to the discretization of the dictionary, we have chosen in the experiments to extend our registration algorithm with a gradient descent technique that refines the estimated transformation. Hence, even if the optimal transformation η_0 is not located on the lattice formed by the discretization of the

Figure 6. *Gaussian mother functions with different values of the anisotropy ν.*

transformation parameter space, the additional local optimization step allows us to converge to the optimal transformation if it lies close to the estimation computed by our registration algorithm.

Specifically, the problem consists in minimizing the objective function $J(\eta) = \|U(\eta)p - q\|_2^2$, where the unknown transformation η is constrained to be in $\mathcal{T}$. Following the same approach as the authors in [15], we consider the gradient descent induction given by

$$\tau_{i+1} = \tau_i - w\nabla J(\tau_i) \quad \text{for } i \geq 0,$$

where the gradient is defined by

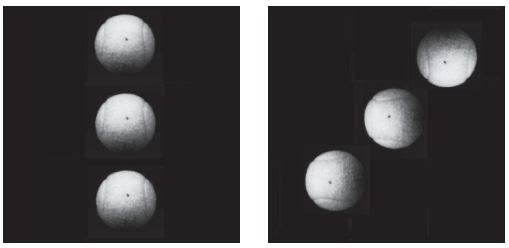
$$\nabla J(\tau_i) = G_{\tau_i}^{-1} \begin{bmatrix} \partial_1 J(\tau_i) \\ \vdots \\ \partial_P J(\tau_i) \end{bmatrix},$$

with $G_\gamma = (\langle \partial_i \phi_\gamma, \partial_j \phi_\gamma \rangle)_{1 \leq i,j \leq P}$ (for any $\gamma \in \mathcal{T}$), and w defines the step size. For more details on the derivation of this gradient descent scheme, we refer the reader to [10].

4.2. Influence of the dictionary on the registration performance. In a first set of experiments, we examine the influence of the dictionary choice on the registration performance.[9] Here we fix here the transformation group $\mathcal{T}$ to be the special Euclidean group $SE(2)$ (containing translations and rotations). We consider that the dictionary mother function is a 2D anisotropic Gaussian function. We vary the anisotropy parameter ν of the mother function to generate a class of different dictionaries. Several generating functions obtained by varying the anisotropy parameter ν are illustrated in Figure 6. Note that the discretization of the parameter space $\mathcal{T}_d$ is kept fixed for all dictionaries. We now study the registration performance of each of these dictionaries.

In the first experiment, we test our registration algorithm with the images I_1 and I_2 illustrated in Figure 7 for our class of dictionaries. We first represent in Figure 8 the mean sparse approximation error $\frac{1}{2}(\|I_1 - p\|_2 + \|I_2 - q\|_2)$ for decompositions with $K = 3$ atoms, when the anisotropy parameter ν in the dictionary mother function varies. For the same class of dictionaries, we also measure the registration performance $|\|U(\eta_0)I_1 - I_2\|_2 - \|U(\hat{\eta})I_1 - I_2\|_2|$, where η_0 and $\hat{\eta}$ are, respectively, the optimal transformation (namely a rotation of $\pi/4$) and the estimated transformation. Note that we used this notion of error instead of $E'(p, q, I_1, I_2)$ in order to focus exclusively on the error due to a wrong estimate of the transformation. The registration performance is illustrated in Figure 8. One can see clearly that the sparse approximation error is increasing with the anisotropy of the mother function. Indeed, when

[9] In this set of experiments, we apply our registration algorithm without the gradient descent refinement. We do so in order to focus exclusively on the performance of Algorithm 1 in terms of the considered dictionary.



(a) I_1 (b) I_2

Figure 7. *Original images used in the first experiment.*

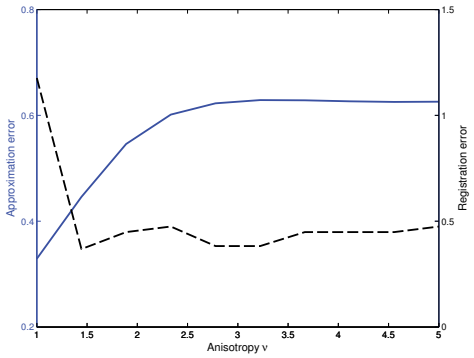


Figure 8. *Approximation error (solid) and registration error (dashed) for images in Figure 7 as a function of the anisotropy of the dictionary generating function. The sparsity K is fixed to 3. The sparse approximation error is given by $\frac{1}{2}(\|I_1 - p\|_2 + \|I_2 - q\|)$, and the registration error by $|\|U(\eta_0)I_1 - I_2\|_2 - \|U(\hat{\eta})I_1 - I_2\|_2|$.*

the mother function approaches isotropy, the dictionary approximates well the tennis balls in images I_1 and I_2. The registration performance, however, has an opposite behavior: the error decreases with increasing values of the anisotropy. This suggests that the sparse approximation error is not the only quantity controlling the performance of the registration algorithm, as predicted by our theoretical performance analysis. Indeed, using the same arguments as in Example 3, we know that the transformation inconsistency parameter goes to infinity when the mother function is isotropic: this explains the poor registration performance for generating functions that are close to isotropic.

As the transformation inconsistency parameter looks crucial in the registration performance, we estimate its value for the same class of dictionaries. This estimation is performed by applying the definition of the transformation inconsistency,[10] where the infinite set $\mathcal{T}$ is finely discretized. In a final step of the estimation, the transformation $\eta \in \mathcal{T}$ that maximizes the transformation inconsistency is refined with a local gradient descent search. Figure 9 shows the estimated value of the transformation inconsistency parameter with respect to the anisotropy of the generating function. One can see that the evolution of the transformation inconsistency parameter is consistent with the theoretical analysis in section 3. For near-isotropic atoms, the parameter ρ is large (Example 3). Similarly, when ν is large, the

[10]Note that we applied the definition in (B.1) since these atoms have a rotational symmetry of π.

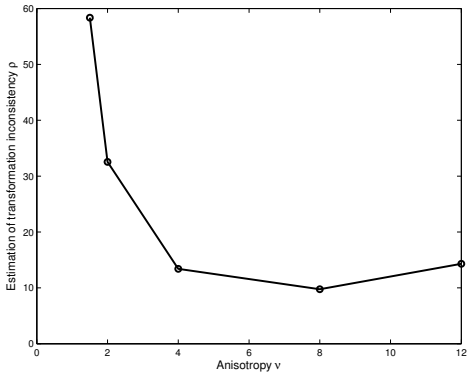


Figure 9. *Estimation of the transformation inconsistency parameter for dictionaries built on Gaussian mother functions with different anisotropy ν.*

(a) Original image I_1 (b) Transformed image I_2

Figure 10. *Original images used in the second experiment.*

transformation inconsistency increases as shown in [10]. Even though our estimation of the transformation inconsistency may not be perfectly accurate (due to the discretization of $\mathcal{T}$), it confirms the tendencies described earlier in the theoretical analysis. It further contributes to explaining the trade-off between approximation and registration error that has been illustrated in Figure 8.

We now study a second experiment where we consider that the transformation group is $\mathcal{T} = \mathbb{R}^2 \times \mathbb{R}_*^+$. That is, $\mathcal{T}$ contains transformations that can be written as combinations of translation and isotropic dilation. We construct another class of dictionaries by fixing the generating function to be an isotropic Gaussian (as shown in Figure 6, $\nu = 1$), but we vary the step size that is used for the discretization of the dilation parameter. More precisely, the set of transformations $\mathcal{T}_d$ that is used to build the dictionary is constructed from $\mathcal{T}$ by imposing a fixed uniform discretization of the translation parameter and a uniform discretization of the dilation parameter whose step size Δ_s can take different values. Note that the minimum and maximum scales are kept fixed in all dictionaries and only the space Δ_s between two consecutive scale parameters is varied. We finally measure the sparse approximation performance with $K = 3$, as well as the registration accuracy that can be obtained with this second class of dictionaries for the images I_1 and I_2 shown in Figure 10. Both sparse approximation and registration errors are computed similarly to the previous experiment. They are illustrated in Figure 11 as a function of the different values of the scale step size Δ_s.

We observe in Figure 11 that the sparse approximation and registration errors have opposite behaviors with respect to the scale space discretization. This is in line with our observations on the first experiment above. Indeed, a fine discretization leads to a small approximation error. At the same time, the registration is less accurate when the discretization is fine. Conversely, coarser discretization of the scale parameter results in a less compact dictionary,

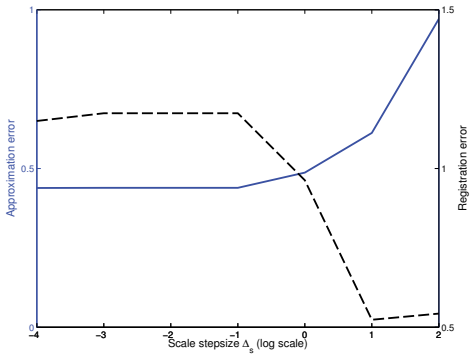


Figure 11. *Approximation error (solid line) and registration error (dashed line) for images in Figure 10 as a function of the scale step size used for constructing the dictionary. The sparsity K is fixed to 3.*

(a) Duck (b) Car (c) Bear

Figure 12. *Test images [13]. All images are resized to be of dimension 75×75 pixels.*

hence in larger approximation errors but better registration performance. These tendencies can be explained using the arguments developed in Example 4.

In summary, these two experiments show that constructing a dictionary that guarantees a small approximation error of the images is not enough to have a low registration error. As we have seen earlier in section 3, crucial parameters such as RLI and transformation inconsistency have to be taken into account in the design of the dictionary in order to reach good registration performance.

4.3. Illustrative examples. In this section we propose some illustrative experiments that study the performance of our registration algorithm for determining the transformation between pairs of images, or for image classification. We further compare the properties of our registration algorithm to other baseline solutions for computing transformation-invariant distances.

In our first set of experiments, we consider the test images shown in Figure 12, which have been collected from the Amsterdam Library of Object Images (ALOI) dataset [13]. We generate 100 random transformations and apply them to the test images. Each of the transformations belongs to $\mathcal{T}$ and consists of a combination of translation, rotation, and isotropic scaling. Both components of the translation vector are smaller than half the image size, and the isotropic scaling parameter is constrained to be in $[0.5, 1.5]$. These restrictions guarantee that most of the image energy lies in the image space, possibly with some occlusions. We put no specific restrictions on the rotation angle. Figure 13 illustrates some examples of transformed images.

We first examine the accuracy of our algorithm in estimating the correct global transformation between pairs of images. We register each of the transformed test images with

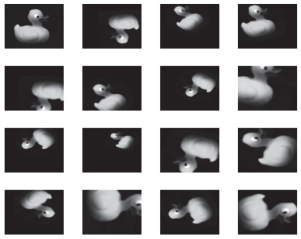
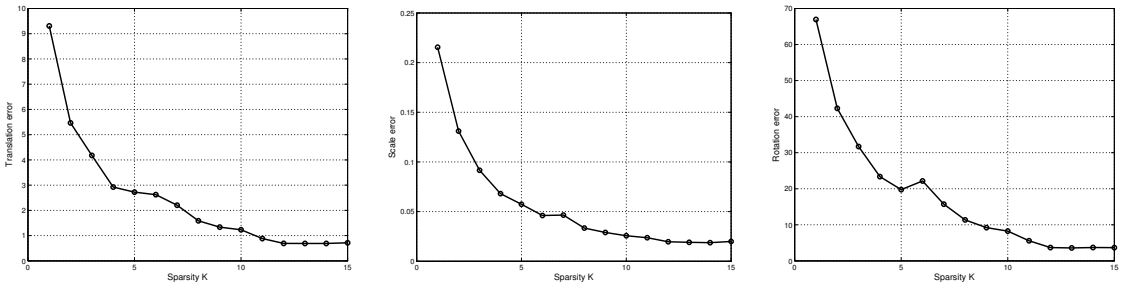


Figure 13. *Sample set of test images built by applying random geometric transformations to the Duck image.*



(a) Translation error: $\|\hat{b} - b'_0\|_2$ (b) Scale error: $|\hat{a} - a'_0|$ (c) Rotation error: $\min(|\hat{\theta} - \theta'_0|, 180 - |\hat{\theta} - \theta'_0|)$

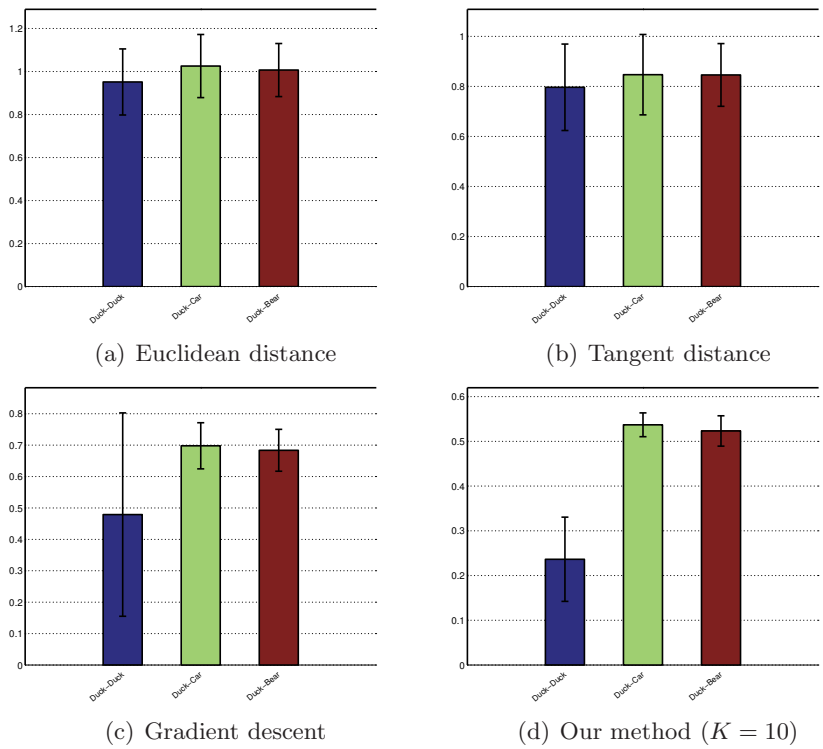
Figure 14. *Errors in translation, scaling, and rotation (in degrees) versus the sparsity K in the approximation of Duck images. The parameter of the optimal transformation obtained by solving* (P′) *is denoted with $\eta'_0 = (b'_0, a'_0, \theta'_0)$, and the estimated transformation with $\hat{\eta} = (\hat{b}, \hat{a}, \hat{\theta})$. The results are averaged over* 100 *tests.*

the original image and compute the average registration accuracy over 100 such operations. Figure 14 shows the average error in the translation, scaling, and rotation parameters when registering pairs of Duck images for different numbers of features K in the sparse image approximations. We see that for $K \geq 10$ our algorithm determines a very good approximation $\hat{\eta} = (\hat{b}, \hat{a}, \hat{\theta})$ of the optimal transformation $\eta'_0 = (b'_0, a'_0, \theta'_0)$. That is, we have on average a translation error of approximately 1 pixel, a scaling error of 0.02, and an angle error of 10 degrees.

We now compare our method with several baseline algorithms for computing distances that are invariant to transformations. The first of these methods is based on the tangent distance [29] that approximates the transformation-invariant distance between the images with the distance between two linear subspaces that can be easily computed. Specifically, the authors in [29] approximate the distance $d(I_1, I_2)$ with

$$d_{TD}(I_1, I_2) = \min_{I'_1 \in T(I_1), I'_2 \in T(I_2)} \|I'_1 - I'_2\|_2,$$

where $T(I_1)$ and $T(I_2)$ are the tangent planes to the manifold of transformed images of I_1 and I_2, respectively, evaluated at I_1 and I_2. The equations of $T(I_1)$ and $T(I_2)$ can be explicitly computed, and the original problem of computing the transformation-invariant distance reduces to solving a least squares problem [29]. We also compare our method with an approach that solves the original problem (P′) using a simple gradient descent technique starting from



(a) Euclidean distance (b) Tangent distance

(c) Gradient descent (d) Our method ($K = 10$)

Figure 15. *Average and standard deviation of intra- and interclass distances for different methods. The blue color denotes intraclass distance, while the green and red colors refer to the distance between the Duck and Car images and the Duck and Bear images, respectively. The distance has been computed between one reference image Duck and* 100 *randomly generated transformed images in each class. The intraclass distance should ideally be at zero. For our approach, the transformation-invariant distance* $d_a(p, q)$ *is computed based on the sparse approximations, while the original images are used in the other methods.*

the identity transformation. Finally, the last comparative scheme is simply based on the computation of the regular Euclidean distance between the images I_1 and I_2. Note that in all three competing solutions, the distances are computed directly on the original images, whereas, in our approach we use only *the sparse image approximations* to compute the distance. We choose to do so since our aim here is to show that our method can be used without explicitly using the complete images in the transformation estimation: a good sparse approximation is indeed sufficient to obtain accurate registration results.

We extend the previous experiments to classification of images. In particular, we compare the transformation-invariant distance for images of the same class to the same distance computed between images of different classes. Ideally, the first one (the intraclass distance) should be smaller than the latter (the interclass distance) in order to obtain good classification performance. We start with a simple scenario where the reference image is chosen to be the Duck image in Figure 12. We then compute the transformation-invariant distance between the reference image and the transformed versions of images in the same class (Duck) and in the other classes (Car and Bear). Figure 15 shows the average of the transformation-invariant distances computed with the different methods. One can see that the Euclidean

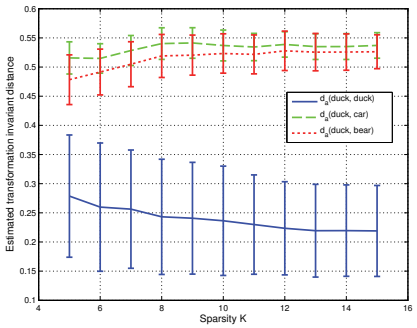


Figure 16. *Evolution of the intraclass and interclass transformation-invariant distance $d_a(p, q)$ in the proposed algorithm as a function of the number of features in the sparse images.*

distance between images of the same class is not significantly different from the distance between images of different classes. The tangent distance does not improve the performance since this method provides only local invariance to transformations. Similarly, the gradient descent approach converges to the correct transformation only when it is close enough to the initial transformation. As this happens rarely, this approach does not provide results that are significantly different for intra- and interclass comparisons. In our method, however, one can see that the intraclass distance is significantly smaller than the interclass distance. Figure 16 further shows the evolution of the transformation-invariant distance with respect to the sparsity of the images. We see that the intraclass distance is always smaller than any of the interclass distances in our algorithm, even for very small values of the sparsity K. This provides a confirmation that salient geometric features in sparse images are crucial for proper registration. Hence, without having a very accurate sparse representation of the patterns, our registration algorithm succeeds in having an approximation of the distances that allows us to at least classify the simple patterns under test. Note that this observation does not contradict the worst-case theoretical analysis in which we assume that the sparse approximation error is small. We observe in practice that, even when this assumption does not hold, one can still obtain a good registration accuracy that is sufficient for the classification of simple signals. Finally, we note that the results are essentially the same if we repeat the same experiments with a different reference image in our dataset. Overall, our illustrative experiments so far show that, with a coarse approximation of the original images in the dictionary, our approach succeeds in obtaining an accurate estimation of the transformation, and the computation of the distances show that the intra- and interclass images are well distinguished. This is an interesting property toward the development of registration algorithms in applications where access to the original (high quality) images is not possible.

We extend the simple classification experiments proposed above and now study the performance of our registration method in a more challenging task of transformation-invariant handwritten digit classification. We use the digits 0 to 5 from the standard MNIST database of handwritten digits [18]. We construct the training data by randomly choosing 100 images for each digit, which results in 600 training images. The test data is constructed similarly: 100 images are taken in each class in order to generate 600 test images. Note that the test data does not contain any of the training images. Finally, we apply to each test image a random

transformation built on translation, rotation, and isotropic scaling. Our classifier then works as follows: each test data is assigned the label of the digit in the training set that best aligns with it or, equivalently, that minimizes the transformation-invariant distance to the test image. In other words, the label of a test image is chosen to be the label of its nearest neighbor in the training dataset, up to a geometrical transformation. We compare the classification results when the transformation-invariant distance is computed with the different methods proposed above. Moreover, for completeness, we also compare our method to an approach that first extracts maximally stable extremal regions (MSER) [22], followed by a similarity normalization [5] that transforms the image into a common system of coordinates. The transformation-invariant distance between two digits is then defined as the distance between the normalized images.

The classification results are shown in Table 2.

Table 2

Handwritten digit classification accuracy for different approaches in computing transformation-invariant distances.

	Classification accuracy
Euclidean distance	14%
Tangent distance	33%
Gradient descent	62%
MSER + similarity normalization	75%
Proposed registration algorithm ($K = 10$)	86%

One can see that using the Euclidean distance on the transformed test images results in a very poor classifier, whose performance is actually close to that of a random classifier. Using the tangent distance results in some improvement, but it is still far from the desired performance. This is due to the fact that the tangent distance is appropriate only for local transformations, while the transformations that we consider are generally of large magnitude. Similarly, the gradient descent approach does not perform well, since it is only guaranteed to reach a local minimum. The MSER-based approach outperforms these local methods and achieves a classification performance of 75%. Using our registration method, however, we achieve a relatively high classification rate, which is by far the best performance among the compared methods. It is worth noting that the performance of our algorithm (86% of classification accuracy) is only slightly worse than the performance of a Euclidean nearest neighbor classifier with aligned images (i.e., no transformations are applied on the test data), which reaches a classification accuracy of 94%. The latter classifier provides an upper bound on the performance that we could achieve in our settings where test images are transformed.

Finally, note that existing methods in the literature achieve close to zero error rate when applied to the MNIST database [18]. However, unlike the proposed approach, these methods generally do not support invariance to large transformations. Furthermore, our method is general in the sense that it is not specific to handwritten digit classification and can be used in any application involving image alignment.

5. Conclusions. We have proposed in this paper a simple registration algorithm based on the sparse representation of the input images in a parametric dictionary of geometric functions. Our method is general in the sense that we can achieve invariance to *any* transformation

group, provided that the geometric dictionary is properly constructed. We define novel properties of dictionaries, namely RLI and transformation inconsistency, in order to characterize the registration performance, which cannot be done with usual properties such as the coherence or the RIP. We show that our algorithm has low registration error when RLI and transformation inconsistency take small values. We also show that the proposed registration algorithm compares favorably with other baseline registration methods from the literature in illustrative alignment and classification experiments on simple visual objects and handwritten digits. To the best of our knowledge, this paper constitutes the first theoretically motivated work for image registration through sparse approximations in parametric dictionaries. We plan to extend our study to also account for the information conveyed by the coefficients of the sparse approximation, to further guide the registration process. Moreover, one future research direction consists in extending the algorithm to a scenario where different parts of the image can undergo different transformations. Finally, it is interesting to use the theoretical findings of this paper to study the design of proper dictionaries that behave well with respect to the newly introduced properties.

Appendix A. Proof of Theorem 3.3.

We recall that η_0 denotes the optimal transformation between p and q and that p and q are given by

$$p = \sum_{i=1}^{K} c_i \phi_{\gamma_i},$$
$$q = \sum_{i=1}^{K} d_i \phi_{\delta_i}.$$

We can write

$$d_a(p,q) - d(p,q) = \min_{\eta \in \mathcal{T}_a^{p,q}} \|U(\eta)p - q\|_2 - \|U(\eta_0)p - q\|_2 \tag{A.1}$$
$$= \min_{\eta \in \mathcal{T}_a^{p,q}} \|U(\eta)p - U(\eta_0)p + U(\eta_0)p - q\|_2 - \|U(\eta_0)p - q\|_2 \tag{A.2}$$
$$\leq \min_{\eta \in \mathcal{T}_a^{p,q}} \|U(\eta)p - U(\eta_0)p\|_2 \tag{A.3}$$
$$= \min_{\eta \in \mathcal{T}_a^{p,q}} \left\| \sum_{i=1}^{K} c_i \phi_{\eta \circ \gamma_i} - \sum_{i=1}^{K} c_i \phi_{\eta_0 \circ \gamma_i} \right\|_2 . \tag{A.4}$$

Let (i^*, j^*) be the indices of the most correlated atoms when the two decompositions are optimally aligned, i.e.,

$$(i^*, j^*) = \operatorname*{argmin}_{1 \leq i,j \leq K} \left\| \phi_{\eta_0 \circ \gamma_i} - \phi_{\delta_j} \right\|_2 , \tag{A.5}$$

and let $\tilde{\eta}$ be the transformation between the corresponding features, i.e.,

$$\tilde{\eta} = \delta_{j^*} \circ \gamma_{i^*}^{-1}.$$

By definition, $\tilde{\eta}$ belongs to the set of feature-to-feature transformations $\mathcal{T}_a^{p,q}$.

If $\|\phi_{\eta_0\circ\gamma_{i^*}} - \phi_{\delta_{j^*}}\|_2 = 0$, then we have $\phi_{\eta_0\circ\gamma_{i^*}} = \phi_{\delta_{j^*}}$. Since we suppose that $\gamma \mapsto U(\gamma)\phi$ is a bijective mapping, we have $\eta_0 \circ \gamma_{i^*} = \delta_{j^*}$ and we finally get $\eta_0 = \tilde{\eta}$. Hence, we have in this case a registration error $d_a(p,q) - d(p,q) = 0$.

We now focus on the case $\|\phi_{\eta_0\circ\gamma_{i^*}} - \phi_{\delta_{j^*}}\|_2 > 0$. Thanks to (A.4), we have

$$d_a(p,q) - d(p,q) \leq \left\| \sum_{i=1}^{K} c_i \phi_{\tilde{\eta}\circ\gamma_i} - \sum_{i=1}^{K} c_i \phi_{\eta_0\circ\gamma_i} \right\|_2 \tag{A.6}$$

$$= \left\| \sum_{i=1}^{K} c_i \left(\phi_{\tilde{\eta}\circ\gamma_i} - \phi_{\eta_0\circ\gamma_i}\right) \right\|_2 \tag{A.7}$$

$$\leq \sum_{i=1}^{K} |c_i| \|\phi_{\tilde{\eta}\circ\gamma_i} - \phi_{\eta_0\circ\gamma_i}\|_2 \tag{A.8}$$

by using the triangle inequality. Since $\|\phi_{\delta_{j^*}} - \phi_{\eta_0\circ\gamma_{i^*}}\|_2 > 0$, we factorize the previous expression as follows:

$$d_a(p,q) - d(p,q) \leq \|\phi_{\delta_{j^*}} - \phi_{\eta_0\circ\gamma_{i^*}}\|_2 \sum_{i=1}^{K} |c_i| \frac{\|\phi_{\tilde{\eta}\circ\gamma_i} - \phi_{\eta_0\circ\gamma_i}\|_2}{\|\phi_{\tilde{\eta}\circ\gamma_{i^*}} - \phi_{\eta_0\circ\gamma_{i^*}}\|_2} \tag{A.9}$$

$$\overset{(*)}{=} \|\phi_{\delta_{j^*}} - \phi_{\eta_0\circ\gamma_{i^*}}\|_2 \sum_{i=1}^{K} |c_i| \frac{\left\|U(\eta_0^{-1}\circ\tilde{\eta})\phi_{\gamma_i} - \phi_{\gamma_i}\right\|_2}{\left\|U(\eta_0^{-1}\circ\tilde{\eta})\phi_{\gamma_{i^*}} - \phi_{\gamma_{i^*}}\right\|_2} \tag{A.10}$$

$$\leq \|\phi_{\delta_{j^*}} - \phi_{\eta_0\circ\gamma_{i^*}}\|_2 \sum_{i=1}^{K} \rho |c_i| \tag{A.11}$$

$$= \rho \|\phi_{\delta_{j^*}} - \phi_{\eta_0\circ\gamma_{i^*}}\|_2 \|c\|_1, \tag{A.12}$$

where we have used in $(*)$ the fact that U is unitary. ρ is the transformation inconsistency parameter introduced in Definition 3.6.

We now focus on bounding $\|\phi_{\delta_{j^*}} - \phi_{\eta_0\circ\gamma_{i^*}}\|_2$. In order to do so, we note that $\phi_{\delta_{j^*}}$ and $\phi_{\eta_0\circ\gamma_{i^*}}$ are, respectively, features in q and $U(\eta_0)p$. Since we assume that $d(p,q) = \|U(\eta_0)p - q\| < \epsilon\sqrt{\|c\|_2^2 + \|d\|_2^2}$, by appropriately using the RLI property (Definition 3.4), we readily obtain an upper bound on $\|\phi_{\delta_{j^*}} - \phi_{\eta_0\circ\gamma_{i^*}}\|_2$. Formally, let e be the vector of length $2K$ constructed from the concatenation of the coefficient vectors c and $-d$, and define $\{\chi_j\}_{j=1}^{2K}$ as follows:

$$\chi_i = \eta_0 \circ \gamma_i,$$
$$\chi_{K+i} = \delta_i.$$

Using this definition, we have $U(\eta_0)p - q = \sum_{i=1}^{K} c_i \phi_{\eta_0\circ\gamma_i} - \sum_{i=1}^{K} d_i \phi_{\delta_i} = \sum_{i=1}^{2K} e_i \phi_{\chi_i}$ and $\|e\|_2 = \sqrt{\|c\|_2^2 + \|d\|_2^2}$. Since $d(p,q) < \epsilon\|e\|_2$ by hypothesis, and $\mathcal{D}$ is $(2K, \epsilon, \alpha)$-RLI with $\alpha < \sqrt{2}$, there exist i, j for which

$$\left\| \frac{e_i \phi_{\chi_i}}{|e_i|} + \frac{e_j \phi_{\chi_j}}{|e_j|} \right\|_2 \leq \alpha, \tag{A.13}$$

as the atoms in the dictionary are normalized. If both i and j are not larger than K, the above inequality implies that

$$\left\| \frac{c_i}{|c_i|}\phi_{\eta_0\circ\gamma_i} + \frac{c_j}{|c_j|}\phi_{\eta_0\circ\gamma_j} \right\|_2 \overset{(a)}{=} \left\|\phi_{\eta_0\circ\gamma_i} + \phi_{\eta_0\circ\gamma_j}\right\|_2 \overset{(b)}{\geq} \sqrt{2}, \tag{A.14}$$

where (a) is obtained thanks to the positivity of c and (b) is a consequence of the positivity of the atoms. Since we assume that $\alpha < \sqrt{2}$, (A.13) and (A.14) cannot hold together. Hence, we exclude the case where $i \leq K$ and $j \leq K$. For the exact same reasons, it is easy to see that we cannot have $i \geq K+1$ and $j \geq K+1$. Therefore, the only possibility is $i \leq K$ and $j \geq K+1$ (or $j \leq K$ and $i \geq K+1$, which is identical up to the relabeling of i and j). Thus, by rewriting (A.13) we get

$$\left\|\phi_{\eta_0\circ\gamma_i} - \phi_{\delta_{j-K}}\right\|_2 \leq \alpha,$$

thanks to the positivity of c and d. Since i^* and j^* are by definition chosen to minimize the error between two features in $U(\eta_0)p$ and q (A.5) we have $\|\phi_{\eta_0\circ\gamma_{i^*}} - \phi_{\delta_{j^*}}\|_2 \leq \|\phi_{\eta_0\circ\gamma_i} - \phi_{\delta_{j-K}}\|_2 \leq \alpha$. Plugging this inequality into (A.12), we get

$$d_a(p,q) - d(p,q) \leq \alpha\rho\|c\|_1. \tag{A.15}$$

It is not hard to see that $d_a(p,q) = d_a(q,p)$ and $d(p,q) = d(q,p)$. Hence, we get

$$d_a(p,q) - d(p,q) \leq \alpha\rho\|d\|_1. \tag{A.16}$$

By combining (A.15) and (A.16), we conclude that

$$d_a(p,q) - d(p,q) \leq \alpha\rho \min\left(\|c\|_1, \|d\|_1\right).$$

Appendix B. Detailed study of the case where $\gamma \mapsto U(\gamma)\phi$ is not bijective. We study in this appendix the case where $\gamma \mapsto U(\gamma)\phi$ is not a one-to-one mapping. In other words, we assume here that the generating function ϕ has symmetries in $\mathcal{T}$. More precisely, let $\mathcal{S}_\phi$ be defined by

$$\mathcal{S}_\phi = \{\gamma \in \mathcal{T} : U(\gamma)\phi = \phi\}.$$

In group theory, $\mathcal{S}_\phi$ is known as the *stabilizer* of ϕ in $\mathcal{T}$. Note that $\mathcal{S}_\phi$ is a subgroup of $\mathcal{T}$. Moreover, it is easy to see that the stabilizer of any atom ϕ_δ can be obtained from S_ϕ with $\mathcal{S}_{\phi_\delta} = \delta \circ \mathcal{S}_\phi \circ \delta^{-1} = \{\delta \circ \pi \circ \delta^{-1} : \pi \in \mathcal{S}_\phi\}$. Hence, given any $\delta \in \mathcal{T}$, the set of elements γ in $\mathcal{T}$ that satisfy $\phi_\delta = \phi_\gamma$ is equal to $\delta \circ \mathcal{S}_\phi$.

When $\gamma \mapsto U(\gamma)\phi$ is a bijective mapping, $\mathcal{S}_\phi$ is equal to the trivial group. When $\mathcal{T} = SE(2)$ and ϕ is an ellipse-shaped generating function (Figure 3), the stabilizer contains two elements, namely the identity transformation and the rotation of angle π. Note that when ϕ is exactly circular, ϕ is symmetric with respect to all rotations; we get $\mathcal{S}_\phi = SO(2)$.

In general, we avoid choosing a generating function whose stabilizer in $\mathcal{T}$ is an infinite subgroup, since the mother function should be discriminative enough for different transformations if we hope to recover the underlying transformation in $\mathcal{T}$. Our goal in this section is to show the modifications we need to perform in order to extend the assumption $|\mathcal{S}_\phi| = 1$ to $|\mathcal{S}_\phi| < \infty$; that is, we need to assume that a limited number of symmetries exist in atom transformations.

B.1. Modified algorithm. The main challenge of having $|\mathcal{S}_\phi| > 1$ is that several features can have the exact same appearance although they correspond to different transformations of the mother function. Clearly, arbitrarily choosing the transformation generally results in a wrong registration. The only way of solving this problem exactly is to examine all transformations that potentially generate a feature and test accordingly all feature-to-feature transformations. Formally, let ϕ_γ and ϕ_δ be, respectively, arbitrary features in p and q. As we mentioned earlier, the set of parameters that generate features having the same appearance as ϕ_γ is $\gamma \circ \mathcal{S}_\phi$. The same result holds for ϕ_δ. Hence, the set of transformations that map features of appearance ϕ_γ to features of appearance ϕ_δ is given by

$$\{\delta \circ \pi \circ (\pi')^{-1} \circ \gamma^{-1} : \pi, \pi' \in \mathcal{S}_\phi\} = \{\delta \circ \pi \circ \gamma^{-1} : \pi \in \mathcal{S}_\phi\}.$$

We thus extend the set of feature-to-feature transformations $\mathcal{T}_a^{p,q}$ to

$$\mathcal{T}_a^{p,q} = \{\delta_i \circ \pi \circ \gamma_j^{-1} : 1 \leq i, j \leq K, \pi \in \mathcal{S}_\phi\}.$$

Note that the only difference with respect to the set $\mathcal{T}_a^{p,q}$ defined in section 2 is that we compose in the middle of the expression with all transformations in the stabilizer group of ϕ. Hence, the cardinality of $\mathcal{T}_a^{p,q}$ is equal to $|\mathcal{S}_\phi|K^2$. The rest of the algorithm (Algorithm 1) remains unchanged.

B.2. Modified analysis. We now turn to the analysis of the modified algorithm. First, it can be shown that in the case where images can be perfectly aligned, Proposition 3.2 holds for the modified algorithm when $|\mathcal{S}_\phi| < \infty$, when there is a finite number of symmetries.

We then extend the analysis of the modified algorithm to the case where images cannot be perfectly aligned, but where the innovation is limited by $d(p,q) < \epsilon\sqrt{\|c\|_2^2 + \|d\|_2^2}$. The main difficulty of the analysis lies in the fact that we have the transformation inconsistency ρ (as defined in Definition 3.6) equal to infinity when the mother function is symmetric (we can see this, for example, by considering the same setting as in Example 3, which is illustrated in Figure 3 with η a rotation of π). We take into account the symmetries of the generating function in the following new definition of ρ:

$$\rho = \sup_{\eta \in \mathcal{T}} \sup_{\substack{\eta' \in \mathcal{T}_d \\ \eta' \notin \eta \circ \mathcal{S}_\phi}} \inf_{\pi \in \mathcal{S}_\phi} \sup_{\gamma \in \mathcal{T}_d} \frac{\left\|U(\eta \circ \pi \circ (\eta')^{-1})\phi_\gamma - \phi_\gamma\right\|_2}{\|U(\eta)\phi - U(\eta')\phi\|_2}, \tag{B.1}$$

where $\eta \circ \mathcal{S}_\phi$ denotes the set $\{\eta \circ \gamma, \gamma \in \mathcal{S}_\phi\}$. Note that by constraining η' to be outside the set $\eta \circ \mathcal{S}_\phi$, the denominator of the above equation is never equal to zero. Therefore, this new definition of the transformation inconsistency solves the problem that we have observed in Example 3 for the particular case of generating functions having a symmetry of π in $\mathcal{T} = SE(2)$.

Note also that when $\mathcal{S}_\phi$ is the trivial group, the above definition of ρ reduces to Definition 3.6, since it is easy to check that

$$\frac{\left\|U(\eta \circ \pi \circ (\eta')^{-1})\phi_\gamma - \phi_\gamma\right\|_2}{\|U(\eta)\phi - U(\eta')\phi\|_2} = \frac{\left\|U(\eta \circ (\eta')^{-1})\phi_\gamma - \phi_\gamma\right\|_2}{\left\|U(\eta \circ (\eta')^{-1})\phi_{\eta'} - \phi_{\eta'}\right\|_2} \leq \sup_{\gamma, \gamma' \in \mathcal{T}_d} \sup_{\eta \in \mathcal{T} \setminus \{\mathbb{I}\}} \left\{ \frac{\|U(\eta)\phi_{\gamma'} - \phi_{\gamma'}\|_2}{\|U(\eta)\phi_\gamma - \phi_\gamma\|_2} \right\},$$

and the reverse inequality also holds. Hence, this definition can be seen as an extension to the case where the generating function has intrinsic symmetries in $\mathcal{T}$. Intuitively, the transformation inconsistency ρ is small whenever two transformations η and η', applied on the generating function and yielding atoms similar in appearance, are such that $\eta \circ \pi \circ (\eta')^{-1}$ does not induce a large change in the appearance of *any* atom in the dictionary $\mathcal{D}$ for some $\pi \in \mathcal{S}_\phi$.

Using this new definition of ρ, we obtain the same bound of Theorem 3.3 for the modified algorithm. In the following, we give the main differences in the proof of this statement with respect to the proof of Theorem 3.3 given in Appendix A.

Proof. Let (i^*, j^*) be the indices defined in (A.5), and let $\tilde{\eta} = \delta_{j^*} \circ \pi \circ \gamma_{i^*}^{-1}$ for any $\pi \in \mathcal{S}_\phi$. Clearly, we have $\tilde{\eta} \in \mathcal{T}_a^{p,q}$.

In the case where $\|\phi_{\eta_0 \circ \gamma_{i^*}} - \phi_{\delta_{j^*}}\|_2 = 0$, there exists $\pi \in \mathcal{S}_\phi$ such that $\eta_0 \circ \gamma_{i^*} = \delta_{j^*} \circ \pi$; thus $\eta_0 = \delta_{j^*} \circ \pi \circ \gamma_{i^*}^{-1} \in \mathcal{T}_a^{p,q}$. Hence, in this case $d_a(p,q) = d(p,q)$.

We now consider the case where $\|\phi_{\eta_0 \circ \gamma_{i^*}} - \phi_{\delta_{j^*}}\|_2 > 0$. By using the same series of inequalities as in (A.1)–(A.8), we know that

$$\begin{aligned} d_a(p,q) - d(p,q) &\leq \|\phi_{\delta_{j^*}} - \phi_{\eta_0 \circ \gamma_{i^*}}\|_2 \sum_{i=1}^{K} |c_i| \frac{\|U(\eta_0^{-1} \circ \tilde{\eta})\phi_{\gamma_i} - \phi_{\gamma_i}\|_2}{\|\phi_{\delta_{j^*}} - \phi_{\eta_0 \circ \gamma_{i^*}}\|_2} \\ &\leq \|\phi_{\delta_{j^*}} - \phi_{\eta_0 \circ \gamma_{i^*}}\|_2 \sup_{\gamma \in \mathcal{T}_d} \left\{ \frac{\|U(\eta_0^{-1} \circ \tilde{\eta})\phi_\gamma - \phi_\gamma\|_2}{\|\phi_{\delta_{j^*}} - \phi_{\eta_0 \circ \gamma_{i^*}}\|_2} \right\} \sum_{i=1}^{K} |c_i|. \end{aligned}$$

Since this inequality is valid for any $\tilde{\eta}$ of the form $\delta_{j^*} \circ \pi \circ \gamma_{i^*}^{-1}$, where $\pi \in \mathcal{S}_\phi$, we deduce from the previous inequality that

$$\begin{aligned} d_a(p,q) - d(p,q) &\leq \|\phi_{\delta_{j^*}} - \phi_{\eta_0 \circ \gamma_{i^*}}\|_2 \inf_{\pi \in \mathcal{S}_\phi} \sup_{\gamma \in \mathcal{T}_d} \left\{ \frac{\|U(\eta_0^{-1} \circ \delta_{j^*} \circ \pi \circ \gamma_{i^*}^{-1})\phi_\gamma - \phi_\gamma\|_2}{\|\phi_{\gamma_{i^*}} - \phi_{\eta_0^{-1} \circ \delta_{j^*}}\|_2} \right\} \|c\|_1 \\ &\leq \|\phi_{\delta_{j^*}} - \phi_{\eta_0 \circ \gamma_{i^*}}\|_2 \sup_{\eta \in \mathcal{T}} \sup_{\substack{\eta' \in \mathcal{T}_d \\ \eta' \notin \eta \circ \mathcal{S}_\phi}} \inf_{\pi \in \mathcal{S}_\phi} \sup_{\gamma \in \mathcal{T}_d} \left\{ \frac{\|U(\eta \circ \pi \circ (\eta')^{-1})\phi_\gamma - \phi_\gamma\|_2}{\|U(\eta)\phi - U(\eta')\phi\|_2} \right\} \|c\|_1 \\ &= \|\phi_{\delta_{j^*}} - \phi_{\eta_0 \circ \gamma_{i^*}}\|_2 \rho \|c\|_1. \end{aligned}$$

By using the same upper bound on $\|\phi_{\delta_{j^*}} - \phi_{\eta_0 \circ \gamma_{i^*}}\|_2$ in the exact same way as in Appendix A (thanks to the RLI property), we obtain the desired result. ∎

Acknowledgments. We thank the anonymous reviewers for their valuable comments that helped improve the quality of the paper. We also thank Prof. Laurent Jacques and Elif Vural for fruitful discussions.

REFERENCES

[1] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani, *Hierarchical model-based motion estimation*, in Proceedings of the European Conference on Computer Vision (ECCV), Springer, Berlin, Heidelberg, 1992, pp. 237–252.

[2] A. Bronstein and M. Bronstein, *Spatially-sensitive affine-invariant image descriptors*, in Proceedings of the European Conference on Computer Vision (ECCV), Springer, Berlin, Heidelberg, 2010, pp. 197–208.

[3] E. Candès, Y.C. Eldar, D. Needell, and P. Randall, *Compressed sensing with coherent and redundant dictionaries*, Appl. Comput. Harmon. Anal., 31 (2011), pp. 59–73.

[4] E.J. Candes and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203–4215.

[5] F. Cao, J.-L. Lisani, J.-M. Morel, P. Musé, and F. Sur, *A Theory of Shape Identification*, Lecture Notes in Math. 1948, Springer-Verlag, Berlin, 2008.

[6] M.A. Davenport, M.F. Duarte, Y.C. Eldar, and G. Kutyniok, *Introduction to Compressed Sensing*, in Compressed Sensing. Theory and Applications, Cambridge University Press, Cambridge, UK, 2012, pp. 1–64.

[7] D.L. Donoho and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization*, Proc. Natl. Acad. Sci., USA, 100 (2003), pp. 2197–2202.

[8] M. Elad and M. Aharon, *Image denoising via sparse and redundant representations over learned dictionaries*, IEEE Trans. Image Process., 15 (2006), pp. 3736–3745.

[9] A. Fawzi and P. Frossard, *A geometric framework for registration of sparse images*, in Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 1976–1980.

[10] A. Fawzi and P. Frossard, *Image Registration with Sparse Approximations in Geometric Dictionaries*, preprint, arXiv:1301.6646 [Cs.CV], 2013.

[11] M. Ferraro and T.M. Caelli, *Relationship between integral transform invariances and Lie group theory*, J. Opt. Soc. Amer. A, 5 (1988), pp. 738–742.

[12] R.M. Figueras, I. Ventura, P. Vandergheynst, and P. Frossard, *Low-rate and flexible image coding with redundant representations*, IEEE Trans. Image Process., 15 (2006), pp. 726–739.

[13] J.M. Geusebroek, G.J. Burghouts, and A.W.M. Smeulders, *The Amsterdam Library of Object Images*, Int. J. Comput. Vis., 61 (2005), pp. 103–112.

[14] R. Giryes, S. Nam, M. Elad, R. Gribonval, and M.E. Davies, *Greedy-Like Algorithms for the Cosparse Analysis Model*, preprint, arXiv:1207.2456 [math.FA], 2012.

[15] L. Jacques and C. De Vleeschouwer, *A geometrical study of matching pursuit parametrization*, IEEE Trans. Signal Process., 56 (2008), pp. 2835–2848.

[16] R. Kimmel, C. Zhang, A. Bronstein, and M. Bronstein, *Are MSER features really interesting?*, IEEE Trans. Pattern Anal. Mach. Intell., 33 (2011), pp. 2316–2320.

[17] E. Kokiopoulou and P. Frossard, *Minimum distance between pattern transformation manifolds: Algorithm and applications*, IEEE Trans. Pattern Anal. Mach. Intell., 31 (2009), pp. 1225–1238.

[18] Y. LeCun, C. Cortes, and C.J.C. Burges, *The MNIST Database of Handwritten Digits*, http://yann.lecun.com/exdb/mnist/.

[19] D.G. Lowe, *Distinctive image features from scale-invariant keypoints*, Int. J. Comput. Vis., 60 (2004), pp. 91–110.

[20] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, *Supervised dictionary learning*, in Advances in Neural Information Processing Systems (NIPS), 21, MIT Press, Cambridge, MA, 2009, pp. 1033–1040.

[21] S.G. Mallat and Z. Zhang, *Matching pursuits with time-frequency dictionaries*, IEEE Trans. Signal Process., 41 (1993), pp. 3397–3415.

[22] J. Matas, O. Chum, M. Urban, and T. Pajdla, *Robust wide baseline stereo from maximally stable extremal regions*, in Proceedings of the British Machine Vision Conference, Vol. 1, 2002, pp. 384–393.

[23] K. Mikolajczyk and C. Schmid, *A performance evaluation of local descriptors*, IEEE Trans. Pattern Anal. Mach. Intell., 27 (2005), pp. 1615–1630.

[24] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, *A comparison of affine region detectors*, Int. J. Comput. Vis., 65 (2005), pp. 43–72.

[25] J.-M. Morel and G. Yu, *ASIFT: A new framework for fully affine invariant image comparison*, SIAM J. Imaging Sci., 2 (2009), pp. 438–469.

[26] T. Peleg and M. Elad, *Performance Guarantees of the Thresholding Algorithm for the Co-Sparse Analysis Model*, preprint, arXiv:1203.2769 [cs.IT], 2012.

[27] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, *RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 763–770.

[28] J. Segman, J. Rubinstein, and Y.Y. Zeevi, *The canonical coordinates method for pattern deformation: Theoretical and computational considerations*, IEEE Trans. Pattern Anal. Mach. Intell., 14 (1992), pp. 1171–1183.

[29] P.Y. Simard, Y.A. Le Cun, J.S. Denker, and B. Victorri, *Transformation invariance in pattern recognition: Tangent distance and propagation*, Int. J. Imaging Syst. Technol., 11 (2000), pp. 181–197.

[30] R. Szeliski, *Image alignment and stitching: A tutorial*, Found. Trends Comput. Graph. Vis., 2 (2006), pp. 1–104.

[31] R. Szeliski, *Computer Vision: Algorithms and Applications*, Springer-Verlag, New York, 2010.

[32] J.A. Tropp, *Greed is good: Algorithmic results for sparse approximation*, IEEE Trans. Inform. Theory, 50 (2004), pp. 2231–2242.

[33] N. Vasconcelos and A. Lippman, *A multiresolution manifold distance for invariant image similarity*, IEEE Trans. Multimedia, 7 (2005), pp. 127–142.

[34] E. Vural and P. Frossard, *Analysis of Descent-Based Image Registration*, preprint, arXiv:1302.3785 [cs.CV], 2013.

[35] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, *Robust face recognition via sparse representation*, IEEE Trans. Pattern Anal. Mach. Intell., 31 (2009), pp. 210–227.

[36] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma, *TILT: Transform invariant low-rank textures*, Int. J. Comput. Vis., 99 (2012), pp. 1–24.

[37] B. Zitova and J. Flusser, *Image registration methods: A survey*, Image Vision Comput., 21 (2003), pp. 977–1000.