APPROXIMATION OF PATTERN TRANSFORMATION MANIFOLDS WITH PARAMETRIC DICTIONARIES

Elif Vural and Pascal Frossard

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Laboratory - LTS4
CH-1015 Lausanne, Switzerland

ABSTRACT

The construction of low-dimensional models explaining high-dimensional signal observations provides concise and efficient data representations. In this paper, we focus on pattern transformation manifold models generated by in-plane geometric transformations of 2D visual patterns. We propose a method for computing a manifold by building a representative pattern such that its transformation manifold accurately fits a set of given observations. We present a solution for the progressive construction of the representative pattern with the aid of a parametric dictionary, which in turn provides an analytical representation of the data and the manifold. Experimental results show that the patterns learned with the proposed algorithm can efficiently capture the main characteristics of the input data with high approximation accuracy, where the invariance to the geometric transformations of the data is accomplished due to the transformation manifold model.

Index Terms— Pattern transformation manifolds, manifold learning, dimensionality reduction, matching pursuit, sparse representations

1. INTRODUCTION

A set of signals that can be locally mapped to a lower dimensional Euclidean space constitutes a manifold in the original high-dimensional signal space. Specifically, a family of signals that are globally definable by a small set of parameters is a signal manifold. The recovery of the manifold structure underlying a set of data has been a popular research topic of the recent years. In manifold learning an efficient representation of the data with reduced dimensionality is searched, which also helps understanding the structure of the signal. It is possible to retrieve the manifold structure of data in various ways, such as by describing a global parameterization based on geodesic distances as in ISOMAP [1], or via locally linear representations as in LLE [2].

In this work, we propose a method for building a pattern transformation manifold (PTM) that fits a given set of observed data. The transformation manifold of a visual pattern consists of images that are geometrically transformed versions of the pattern. For instance, the images obtained by the rotation and translation of a pattern form a pattern transformation manifold. Given a set of observations of a signal exposed to geometric transformations, we study the problem of learning a representative pattern whose transformation manifold constitutes an accurate approximation of the observations. We

---

build the pattern as a linear combination of some atoms in a parametric dictionary. The utilization of parametric atoms in the pattern construction not only brings desirable properties such as differentiability, but also provides compact data representations, which is one of the main concerns in dimensionality reduction. Our solution is based on constructing the representative pattern progressively by optimizing the atom selection at each step, where we use a locally linear approximation of the manifold. Experimental evaluation of the proposed algorithm shows that the data approximation error is minimized gradually with the refinement of the representative pattern, and the constructed pattern transformation manifold fits the data well.

Since we select atoms in the construction of the representative pattern in a greedy manner, our method bears a resemblance to sparse approximation algorithms such as Matching Pursuit (MP) [3] or Simultaneous Orthogonal Matching Pursuit (SOMP) [4]. Although our problem is also related to sparse signal representations, it is essentially different from these approaches in the following sense. In SOMP, the simultaneous sparse approximation of a group of signals is achieved by expressing signals as linear combinations of selected dictionary atoms, where atom coefficients can be different for each signal. However, in our problem we would like to approximate signals as transformed versions of the same representative pattern, where the atom coefficients in the construction of the pattern are fixed and it is the transformation parameters that varies among different signals. Hence, the solution of our problem cannot be simply achieved by employing a simultaneous matching pursuit algorithm with an input set enriched with transformations. Instead, we propose a geometric solution to the problem through the transformation manifold model, where we minimize the total distance of the data to the constructed manifold in atom and coefficient optimization.

2. PROBLEM FORMULATION

We formulate here the transformation manifold learning problem. Let $p \in \mathbb{R}^n$ be a visual pattern. A point on the transformation manifold $\mathcal{M}_p$ of p is an image $U_\lambda(p)$ that is obtained by applying the geometric transformation U_λ to p, i.e., $\mathcal{M}_p = \{U_\lambda(p), \lambda \in \Lambda\} \subset \mathbb{R}^n$, where λ is a parameter vector specifying the applied transformation, and Λ is the domain of parameter vectors on which the manifold is defined. For instance, if the transformation consists of a two dimensional translation of p in the image plane, the parameter vectors are in the form $\lambda = (t_x, t_y)$, the parameter domain can be taken as $\Lambda = \mathbb{R}^2$, and each manifold point is a translated version of p.

Let $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^n$ be a set of observations of a geometrically transformed visual signal. We would like to describe the

observations $\{x_i\}$ by the transformations of a representative pattern as accurately as possible. Hence, the problem consists in determining a representative pattern p such that each point x_i is approximated by a transformation of the pattern p as $x_i = U_{\lambda_i}(p) + e_i$. Here, the term e_i denotes the deviation of x_i from the transformation manifold $\mathcal{M}_p$ of p. We assume that the parameter domain Λ is known, and given a parameter vector λ, the transformation function U_λ corresponding to λ is available. However, the representative pattern p is to be learned, as well as the parameter vectors $\{\lambda_i\}$ corresponding to the observations $\{x_i\}$.

We construct p as a linear combination of atoms of a parametric dictionary $D = \{a_j\}_{j=1}^K$, where each atom is available in an analytical form $a_j = \phi_\gamma$ as some geometrically transformed version of a generating mother function ϕ. The representative pattern is expressed as $p = \sum_{j=1}^K c_j a_j$, where the vector c of coefficients is a sparse vector, i.e., it has a small number $S \ll K$ of nonzero coefficients. In the selection of the atoms that build p, the objective is to approximate the observations X accurately. We represent the approximation accuracy in terms of the distance of data points to the transformation manifold $\mathcal{M}_p$. Therefore, we formulate the problem as follows. Given N observations $\{x_i\}_{i=1}^N$ and a dictionary $D = \{a_j\}_{j=1}^K$, determine a coefficient vector $c \in \mathbb{R}^K$ and transformation parameters $\{\lambda_i\}_{i=1}^N$ such that the total data approximation error

$$E = \sum_{i=1}^N \|x_i - U_{\lambda_i}\big(\sum_{j=1}^K c_j a_j\big)\|^2 \qquad (1)$$

is minimized. The notation $\|.\|$ stands for the l_2-norm throughout the text.

3. PTM APPROXIMATION ALGORITHM

In this section we describe an algorithm for building the representative pattern that approximates the given observation set in a transformation-invariant way. We construct the pattern progressively by selecting atoms from the dictionary in a greedy manner, where in each iteration we add a new atom and optimize the coefficient of the selected atom such that the data approximation error E is minimized. In the minimization of the error, we use a locally linear approximation of the manifold near each data point. For this reason, we initialize the algorithm by assigning a tentative set of parameter vectors $\{\lambda_i\}$ to the data points $\{x_i\}$. One possible way of achieving the parameterization is to select a reference atom $a \in D$ with a suitable coefficient α and then project the observations on the transformation manifold of αa, where each data point x_i is assigned the parameter vector $\lambda_i = \arg\min_{\lambda \in \Lambda} \|U_\lambda(\alpha a) - x_i\|$. Following this, we initialize the pattern p. Note that it is possible to initialize p in several ways, such as a multiple of the reference atom as $p = \alpha a$ or simply as $p = 0$. In the former case, the reference atom a can be selected in a manner similar to an iteration of MP, for instance, it can be chosen as the atom that has the highest correlation with an average image of the data set. Since the learning procedure is based on successive refinements of p, the initialization of the pattern is not expected to affect its final structure much. However, it may have an influence on the speed of convergence.

Let us now discuss the selection of a new atom in the construction of p. Let p_l denote the pattern that consists of l atoms after the lth iteration. In the lth iteration, we would like to choose an atom $a_j \in D$ and a coefficient $c_j \neq 0$ such that the approximation error

$$E = \sum_{i=1}^N d^2(x_i, \mathcal{M}_{p_l}) \qquad (2)$$

is minimized, where $p_l = p_{l-1} + c_j a_j$, and the notation $d(x, \mathcal{M})$ denotes the distance between a point x and a manifold $\mathcal{M}$. Notice that the current assignment of the parameter vectors $\{\lambda_i\}$ to $\{x_i\}$ does not necessarily correspond to the points of projection on the transformation manifold of the new pattern p_l after the update, i.e., $\|x_i - U_{\lambda_i}(p_l)\| \neq d(x_i, \mathcal{M}_{p_l})$ in general. In the minimization of the error E in (2), it is not easy to formulate analytically and compute the exact distance between an observation x_i and the manifold $\mathcal{M}_{p_l}$. Therefore, in order to minimize E we make use of a linear approximation of the manifold near each observation, which is based on the transformation parameters of the previous iteration. Let d be the dimension of the manifold, i.e., the number of transformation parameters that define the PTM. The first order approximation of the manifold $\mathcal{M}_{p_l}$ around the manifold point $U_{\lambda_i}(p_l)$ is given by

$$\mathcal{M}_{p_l} \approx \mathcal{S}_{p_l} = \{U_{\lambda_i}(p_l) + T_{ij}\,\beta : \beta \in \mathbb{R}^d\}, \qquad (3)$$

where T_{ij} is an $n \times d$ matrix consisting of tangent vectors. The kth column of T_{ij} is the tangent vector $\partial/\partial_k\, U_{\lambda_i}(p_l)$, which is the derivative of the manifold point $U_{\lambda_i}(p_l)$ with respect to the kth transformation parameter. We approximate the exact manifold distance $d(x_i, \mathcal{M}_{p_l})$ by the distance between x_i and $\mathcal{S}_{p_l}$. The orthogonal projection of x_i on $\mathcal{S}_{p_l}$ is given by $\hat{x}_i = U_{\lambda_i}(p_l) + T_{ij}\beta^*$, where the coefficient vector β^* of the projection is $\beta^* = (T_{ij}^{\mathrm{T}} T_{ij})^{-1} T_{ij}^{\mathrm{T}}(x_i - U_{\lambda_i}(p_l))$. Hence, the difference vector $\hat{e}_i$ between x_i and $\hat{x}_i$ is

$$\begin{aligned} \hat{e}_i = x_i - \hat{x}_i &= x_i - U_{\lambda_i}(p_l) \\ &\quad - T_{ij}(T_{ij}^{\mathrm{T}} T_{ij})^{-1} T_{ij}^{\mathrm{T}}(x_i - U_{\lambda_i}(p_l)), \end{aligned} \qquad (4)$$

which gives an approximation of the error term in (2) as $E \approx \hat{E} = \sum_{i=1}^N \|\hat{e}_i\|^2$. In order to minimize the error, we propose the following. For each atom $a_j \in D$, we compute the minimum value of the error $\hat{E}$ optimized over the coefficient c_j of the atom. Then, we pick the atom that provides the highest reduction in the error. Now, for an atom a_j, let us discuss the determination of the optimal coefficient c_j. In order to analyze the dependency of $\hat{E}$ on c_j, first notice that $U_{\lambda_i}(p_l) = U_{\lambda_i}(p_{l-1} + c_j a_j) = U_{\lambda_i}(p_{l-1}) + c_j U_{\lambda_i}(a_j)$, due to the linearity of the 2D pattern transformations. The kth column of the matrix T_{ij} is thus

$$\frac{\partial}{\partial_k} U_{\lambda_i}(p_l) = \frac{\partial}{\partial_k} U_{\lambda_i}(p_{l-1}) + c_j \frac{\partial}{\partial_k} U_{\lambda_i}(a_j).$$

This gives a decomposition of the matrix T_{ij} as $T_{ij} = P_i + c_j A_{ij}$, where P_i and A_{ij} are respectively the matrices consisting of the tangents to the manifolds $\mathcal{M}_{p_{l-1}}$ and $\mathcal{M}_{a_j}$ at the points $U_{\lambda_i}(p_{l-1})$ and $U_{\lambda_i}(a_j)$. Now, defining $y_i = x_i - U_{\lambda_i}(p_{l-1})$, we arrange $\hat{E}$ as

$$\hat{E} = \sum_{i=1}^N \|y_i - c_j U_{\lambda_i}(a_j) - T_{ij}(T_{ij}^{\mathrm{T}} T_{ij})^{-1} T_{ij}^{\mathrm{T}}(y_i - c_j U_{\lambda_i}(a_j))\|^2.$$

$$(5)$$

In equation (5), $(T_{ij}^{\mathrm{T}} T_{ij})^{-1}$ is a matrix whose entries are rational functions of c_j, and the other terms are polynomial in c_j. Therefore, the total error $\hat{E}$ is a rational function of c_j of the form $\hat{E} = \sum_{i=1}^N f_i(c_j)/g_i(c_j)$, where $f_i(c_j)$ and $g_i(c_j)$ are polynomial functions of c_j, and the denominator $g_i(c_j)$ is equal to the determinant of the matrix $T_{ij}^{\mathrm{T}} T_{ij}$, therefore it is positive. Since $\hat{E}$ is a sum of rational functions with positive denominators, it is a DC function of c_j, i.e., it can be expressed as the difference of two convex functions [5]. The special form of DC functions allows their minimization via DC solver algorithms. Although DC solvers are guaranteed to find the

global minimum of the function, in the experiments we have used a descent-based optimization procedure, since it is simpler than a DC solver and gives close results for this specific problem.

Once the new atom and the corresponding optimal coefficient is determined, the pattern p is updated. However, due to the approximation of the original manifold by a linear model, an update that decreases $\hat{E}$ may not decrease the original error E. In this case, we do not accept the update on p, and try the remaining atoms in the order given by the amount of decrease that they yield in $\hat{E}$. Whenever a new atom is added to the pattern p, we recompute the projections of the points $\{x_i\}$ on the refined manifold $\mathcal{M}_p$ and update the parameter vectors $\{\lambda_i\}$ such that they correspond to the new manifold points of projection. We repeat this until convergence. Due to the resemblance of the atom selection step to that of the Matching Pursuit algorithm, we call this method Transformation-Invariant Matching Pursuit (TIMP) and summarize it in Algorithm 1. Since the error computation is repeated for all atoms and observations, the complexity of the algorithm is $O(K \cdot N \cdot P(d))$, where $P(d)$ is a polynomial term in the dimension d of the manifold.

Algorithm 1 Transformation-Invariant Matching Pursuit

1: **Input:**
   $X = \{x_i\}_{i=1}^{N}$: Set of observations
   $D = \{a_j\}_{j=1}^{K}$: Dictionary
   Λ: Feasible domain of parameter vectors
   U: Mapping from parameter domain Λ to the manifold
2: **Initialization:**
3: Determine a set of parameter vectors $\{\lambda_i\}$ by projecting $\{x_i\}$ on the transformation manifold of a reference atom $a \in D$.
4: $l = 0$.
5: Initialize the representative pattern p_0.
   Repeat 6-10 until convergence
6: $l = l + 1$.
7: Determine the atom $a_j \in D$ and the optimal coefficient c_j that minimizes the approximate total error $\hat{E}$ in (5).
8: Update $p_l = p_{l-1} + c_j a_j$.
9: Update parameter vectors $\{\lambda_i\}$ by projecting $\{x_i\}$ to the transformation manifold $\mathcal{M}_{p_l}$ of p_l.
10: Check if the update on p reduces the exact total error E. If not, reject the update, remove the chosen atom a_j from the current list, and go back to 7.
11: **Output**:
   $p = p_l$: A representative pattern whose transformation manifold $\mathcal{M}_p$ fits the data X

---

4. EXPERIMENTAL RESULTS

We now present the experimental results obtained with the proposed algorithm. In all experiments we use a dictionary

$$D = \{\phi_{\gamma_k}\}_{k=1}^{K/2} \cup \{\varphi_{\gamma_k}\}_{k=K/2+1}^{K}, \qquad (6)$$

where the atoms $\{\phi_{\gamma_k}\}$ and $\{\varphi_{\gamma_k}\}$ are derived respectively from the 2D Gaussian function $\phi(x,y) = \sqrt{2/\pi}\exp(-x^2 - y^2)$ and the AnR (anisotropic refinement) function $\varphi(x,y) = \sqrt{2/(3\pi)}(4x^2 - 2)\exp(-x^2-y^2)$. Gaussian atoms represent smooth components of images, whereas AnR atoms capture edge-like features. Dictionary atoms are generated by applying random geometric transformations $\{\gamma_k\}$ to the mother functions ϕ and φ. These transformations are composed of rotation, translation in horizontal and vertical directions, and anisotropic scale change in horizontal and vertical directions.

In all experiments we measure the approximation accuracy of the learned representative pattern by the data approximation error, which is the average squared distance of input data points to the transformation manifold of the output pattern. The data approximation error in the plots is normalized with respect to the average norm of input images.

First, we conduct experiments on transformation manifolds defined by the rotation and translation of patterns. Given a pattern p, we define its transformation manifold by

$$\mathcal{M}_p = \{U_\lambda(p) : \lambda = (\theta, t_x, t_y) \in \Lambda\}, \qquad (7)$$

where θ denotes the rotation and t_x and t_y are the translations in x and y directions. As an initial experiment, we examine the approximation accuracy of the learned pattern with respect to the noise level of the input data set. We form a synthetical pattern r that is composed of randomly selected 5 Gaussian and 3 AnR type atoms. Then we generate a data set X of 50 different patterns by applying r random geometric transformations of the form (7). From X we obtain several data sets of different noise levels by corrupting its images with additive Gaussian noise. Then for each data set we run the TIMP algorithm and try to recover the representative pattern r, where the algorithm is terminated after 40 iterations in all runs. We use a dictionary of $K = 1000$ atoms. In Fig. 1 the data approximation error is plotted with respect to the noise variance of the data set. As the deviation of the input data from the transformation manifold of r is controlled by the noise level, the total error E in (2) is linearly proportional to the noise variance. The plot in Fig. 1 confirms that the data approximation error has a linear dependence on the noise variance.

Then, we conduct another experiment for the same type of pattern transformation manifolds on a data set consisting of 200 top-view images of two different objects captured under different orientation and positions. The images are converted to greyscale, down-sampled, and the background pixels are set to the intensity value of 0. Some example images from the data set are shown in Fig. 2(a). A representative pattern is obtained with the TIMP algorithm that consists of 20 atoms, where the dictionary consists of 2000 atoms. The constructed pattern is shown in Fig. 2(b). In order to evaluate the performance of the proposed algorithm, we compare it with two other reference approaches. In the first one, we approximate the data set images by the transformation manifold of an average pattern in the data set, where the average pattern is chosen as the pattern that has the smallest distance to the centroid of all images. In the second approach, we construct a progressive approximation of this average pattern with the MP algorithm, and compute the data approximation error yielded by the transformation manifold of each intermediate pattern in the approximation. In Fig. 2(c), the data approximation error is plotted with respect to the number of atoms in the construction of the representative pattern with the TIMP algorithm and the MP construction of the average pattern, where the error obtained with the transformation manifold of the original average pattern is also indicated. The figures show that the TIMP algorithm efficiently builds a representative pattern whose transformation manifold provides a model for the data, yielding higher representation accuracy than the transformation manifold of the average pattern and its approximations. Interestingly, the approximations of the average pattern with MP gives less error than the average pattern itself, which suggests that the regularized approximations of a nonsmooth manifold may perform better than itself in terms of data representation.
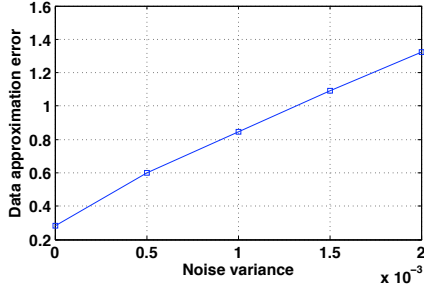
Fig. 1. Dependence of the approximation accuracy on the data noise

(a) Images from objects data set   (b)   Learned pattern
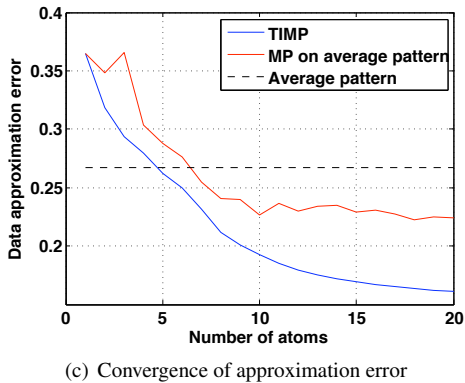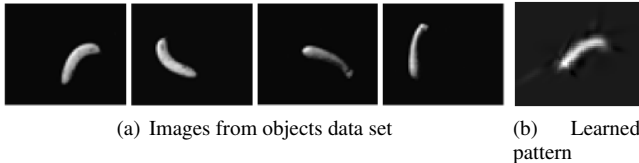

(c) Convergence of approximation error

Fig. 2. Results obtained on the object images data set

Finally, we test the proposed algorithm on a 4-dimensional pattern transformation manifold, where we include also the scale change in the transformations. Therefore, this time the transformation manifold of the pattern p is given by

$$\mathcal{M}_p = \{U_\lambda(p) : \lambda = (\theta, t_x, t_y, s) \in \Lambda\}, \qquad (8)$$

where s denotes the scale change in addition to the parameters of the model in (7). For this experiment, we create a data set of 500 images by applying random transformations of the form (8) to 500 randomly selected images of the digit '5' from the MNIST handwritten digits database[1]. Some examples from the data set are shown in Fig. 3(a). We use a dictionary consisting of 1000 atoms. The representative pattern learned after 20 iterations is shown in Fig. 3(b). In Fig. 3(c) the data approximation error is plotted with respect to the number of atoms used in the construction of the pattern with the TIMP algorithm and the construction of the average pattern of the data set with the MP algorithm, together with the error given by the transformation manifold of the average pattern. The results are in agreement with the results of the previous experiment. In spite of the challenge caused by the diversity of the handwritten data, the main features characterizing the digit '5' are well captured in the built pattern, and

---

[1] http://yann.lecun.com/exdb/mnist/

(a) Images from digits data set   (b)   Learned pattern
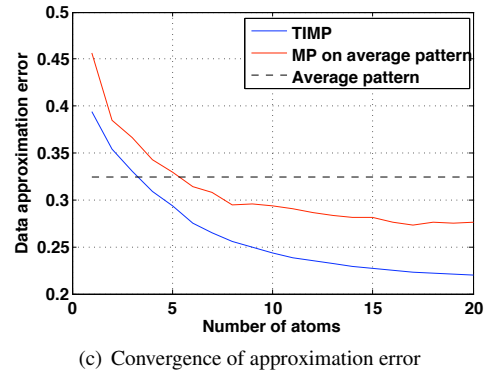

(c) Convergence of approximation error

Fig. 3. Results obtained on the handwritten digits data set

the proposed method outperforms the other baseline approaches in terms of representation accuracy.

5. CONCLUSIONS

We have proposed a method for generating a visual pattern such that its transformation manifold forms an accurate model for a given set of observed images. We build the representative pattern gradually in terms of the atoms of a parametric dictionary, where we present a geometric solution for the atom selection. Our method is general and applicable to various kinds of 2D geometric pattern transformations. We demonstrate the performance of the proposed algorithm on data sets conforming to transformation manifold models of different dimensions. The experimental results show that the approximation error is reduced efficiently during the progressive construction of the representative pattern, and the learned pattern captures the input data characteristics accurately.

6. REFERENCES

[1] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction.," *Science*, vol. 290, no. 5500, pp. 2319–2323, December 2000.

[2] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[3] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec 1993.

[4] Joel A. Tropp, Anna C. Gilbert, and Martin J. Strauss, "Algorithms for simultaneous sparse approximation part I: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572–588, 2006.

[5] E. Kokiopoulou and P. Frossard, "Minimum distance between pattern transformation manifolds: Algorithm and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1225–1238, 2009.