Descriptor Learning via Supervised Manifold Regularization for Multioutput Regression

Xiantong Zhen, Mengyang Yu, *Student Member, IEEE*, Ali Islam, Mousumi Bhaduri, Ian Chan, and Shuo Li*

***Abstract*— Multioutput regression has recently shown great ability to solve challenging problems in both computer vision and medical image analysis. However, due to the huge image variability and ambiguity, it is fundamentally challenging to handle the highly complex input-target relationship of multioutput regression, especially with indiscriminate high-dimensional representations. In this paper, we propose a novel supervised descriptor learning (SDL) algorithm for multioutput regression, which can establish discriminative and compact feature representations to improve the multivariate estimation performance. The SDL is formulated as generalized low-rank approximations of matrices with a supervised manifold regularization. The SDL is able to simultaneously extract discriminative features closely related to multivariate targets and remove irrelevant and redundant information by transforming raw features into a new low-dimensional space aligned to targets. The achieved discriminative while compact descriptor largely reduces the variability and ambiguity for multioutput regression, which enables more accurate and efficient multivariate estimation. We conduct extensive evaluation of the proposed SDL on both synthetic data and real-world multioutput regression tasks for both computer vision and medical image analysis. Experimental results have shown that the proposed SDL can achieve high multivariate estimation accuracy on all tasks and largely outperforms the algorithms in the state of the arts. Our method establishes a novel SDL framework for multioutput regression, which can be widely used to boost the performance in different applications.**

***Index Terms*— Descriptor learning, generalized low-rank approximation, manifold learning, multioutput regression.**

I. Introduction

MULTIOUTPUT regression [1] has shown great effectiveness in many computer vision tasks, e.g., object detection [2], pose estimation [3], and viewpoint estimation [4]. Traditionally, challenging tasks, e.g., camera relocalization [5], can be elaborately solved by formulating as a multioutput regression problem, which can avoid the difficulty, e.g., the inverse problem [5], in the conventional approaches.

Moreover, multioutput regression has started to generate increasing interest in the medical image analysis, e.g., multiple organ localization and direct volume estimation [6]–[8]. Multioutput regression outperforms the previous approaches and more importantly offers a more compact and exquisite mathematical formulation to tackle the obstacles in the traditional methods [5]. Multioutput regression allows to deploy advanced machine learning techniques to facilitate medical image analysis, which provides an effective tool to automate the analysis of medical imagining data and, therefore, enables accurate and efficient diagnosis in clinical practice [9].

However, it is challenging to discover the underlying nonlinear relationship between the high-dimensional input descriptor space and the multivariate target space in multioutput regression. Inputs, e.g., images, associated with similar targets often exhibit varied appearances due to illumination changes, geometrical complexity, and intersubject variations, which causes high variability; while those associated with distinctive targets could share similar appearance, which induces huge ambiguity. It becomes the bottleneck to design discriminative descriptors that can reduce the variability and ambiguity for accurate and efficient multivariate estimation [10]. Handcrafted features widely used in existing multioutput regression tasks neglect the guidance of regression targets [3], which results in indiscriminate and lengthy representations, and therefore less accurate multivariate estimation. It is, therefore, imperative and highly desirable to explore the targets to achieve discriminative descriptors, which remains unaddressed for multioutput regression tasks.

To fill this gap, we propose a novel supervised descriptor learning (SDL) algorithm for multioutput regression. The SDL seeks low-dimensional feature representations aligned to regression targets to obtain discriminative and compact descriptors that enable more accurate and efficient estimation of multiple targets. We formulate the SDL in a framework of generalized low-rank approximations of matrices (GLRAM) [11]. By incorporating the supervision of multiple regression targets, we propose a supervised manifold regularization (SMR) to achieve discriminative learning.

By integrating the SMR into the framework of GLRAM, the SDL possesses multiple attractive merits.

Manuscript received January 7, 2016; revised April 20, 2016; accepted May 19, 2016. This work was supported in part by the Southern Ontario Smart Computing Innovation Platform Consortium through the Ontario Government and in part by the Federal Economic Development Agency for Southern Ontario. The work of X. Zhen was supported by the National Science Foundation of China under Grant 61571147. *(Asterisk indicates corresponding author)*

X. Zhen is with the Department of Medical Biophysics, University of Western Ontario, London, ON N6A 3K7, Canada (e-mail: zhenxt@gmail.com).

M. Yu is with Northumbria University, Newcastle upon Tyne NE1 8ST, U.K. (e-mail: ymymath@gmail.com).

A. Islam is with St. Joseph's Health Care, London, ON N6A 4V2, Canada (e-mail: ali.islam@sjhc.london.on.ca).

M. Bhaduri and I. Chan are with the London Healthcare Sciences Centre, London, ON N6A 5W9, Canada (e-mail: mousumi.bhaduri@lhsc.on.ca; ian.chan@lhsc.on.ca).

*S. Li is with the Departments of Medical Imaging and Medical Biophysics, University of Western Ontario, London, ON N6A 3K7, Canada (e-mail: slishuo@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2016.2573260

1) The generalized low-rank approximation operates on matrix representations of images, which reduces time and space costs for more accurate and efficient computation of low-rank matrices than on vecterized representations [11], finds low-dimensional feature representations that can substantially reduce complexity of multioutput regression, and allows to explore distinctive physical meanings, e.g., spatial layout and orientation structure, in matrices for optimal representations.
2) The SMR encodes the intrinsic local geometrical structure of the multivariate target space and naturally incorporates the supervision to realize discriminative descriptor learning. The learned descriptors are aligned to regression targets achieving discriminative representations. In contrast to the conventional manifold learning [12], [13], the SMR makes full use of regression targets for SDL, which is first studied for multioutput regression.
3) By seamlessly integrating the newly proposed SMR into the GLRAM, the proposed SDL leverages their strengths in supervised manifold learning and subspace learning and, therefore, offers a novel general framework to effectively generate discriminative but compact low-dimensional descriptors for multioutput regression.

The obtained objective function of SDL is novel and can be efficiently solved by our newly proposed iterative algorithm via an alternate optimization. Furthermore, the proposed SDL algorithm is flexible to work in conjunction with diverse matrix inputs. We develop our SDL on gradient orientation matrices (GOMs) rather than directly on raw image intensity in order to capture edges and gradient structures [14]. It is partially inspired by the previous work to build on image gradients, which has shown that replacing pixel intensities with gradient orientations offers reliable subspace estimation [15] and combining gradient orientations with supervised learning can improve the performance for classification [16], [17]. The GOM of an image is obtained by storing the gradients of a pyramid histogram of oriented gradients (PHOGs) in a matrix, which contains the orientations in rows and spatial cells in columns. Therefore, the SDL can explore spatial layout and orientation information discriminately in the GOM, which enables optimal data-driven representations to cater different applications. A preliminary test conducted in [18] and [19], based on which we make new contributions on both methodology and experiment validation in this paper.

Contributions: We propose a novel SDL algorithm for multioutput regression, which can work in conjunction with and improve the performance of existing multioutput regressors. The SDL achieves compact and discriminative feature representation, which cannot only substantially boost performance of multivariate estimation but also enables more efficient multioutput regression. The SDL is solved by a newly proposed alternating optimization algorithm, which is proved to converge fast to a guaranteed global optimum. The proposed SDL has successfully solved challenging multioutput regression tasks, i.e., head pose estimation and face alignment, for computer vision and cardiac chamber area estimation for the medical image analysis. It is worth to mention that our work is the first to directly estimate four-chamber areas, which establishes a useful clinical tool for more accurate medical data prediction.

II. Related Work

Multioutput regression has recently shown great success in solving many conventional problems, which mainly use handcrafted descriptors. Discriminative descriptor learning algorithms have been developed primarily in image classification/retrieval and face recognition tasks while not yet for multioutput regression.

A. Multioutput Regression

Multioutput regression has recently been used for both computer vision and medical image analysis tasks. Camera pose estimation in [5] is formulated as multioutput regression problem, which significantly outperforms the conventional methods based on inverse problems. Cardiac biventricular volume estimation has also been formulated as a multioutput regression problem to avoid tedious segmentation, which achieved remarkable results [7].

It has been shown in recent work [20] that regression performance can be largely improved by exploring the multivariate target space. The geometric structure of the output manifold is explored and incorporated into the regression process in [20], which refines the loss functions by local linear transform (LLT). By working in conjunction with support vector regression (SVR) [21], the LLT can improve the regression performance of SVR. By jointly learning both the output structure and regression coefficients via inverse-covariance regularization, Sohn and Kim [22] proposed the simultaneous estimation of structured sparsity and output structure for multioutput regression. Recently, multioutput regression with output and task structures (MROTSs) [1] was proposed to jointly explore the covariance structure of latent model parameters and the conditional covariance structure of multiple targets. The MROTS has achieved improved performance in different multioutput regression tasks.

B. Descriptor Learning

A recent trend is to design descriptor learning algorithms by building on well-established basic handcrafted descriptors, e.g., the SIFT [23]–[25], LBP [26], and HOG [16], [17], [27], [28], which, however, were mainly developed for computer vision tasks, such as image classification, retrieval, and face recognition rather than for multioutput regression.

It has been shown that the performance of handcrafted descriptors can be improved further by discriminant learning [17], [24], [26]. A local discriminant projection algorithm was proposed in [24] for dimensionality reduction of local descriptors, i.e., SIFT. The projections are learned to minimize the distances between matched pairs of descriptors while maximizing those between unmatched pairs. Similarly, a three-layered model based on the LBP descriptors was developed in [26] to extract discriminative and robust features. Also based on handcrafted feature descriptors, a learning architecture was developed in [17] to select discriminative filters

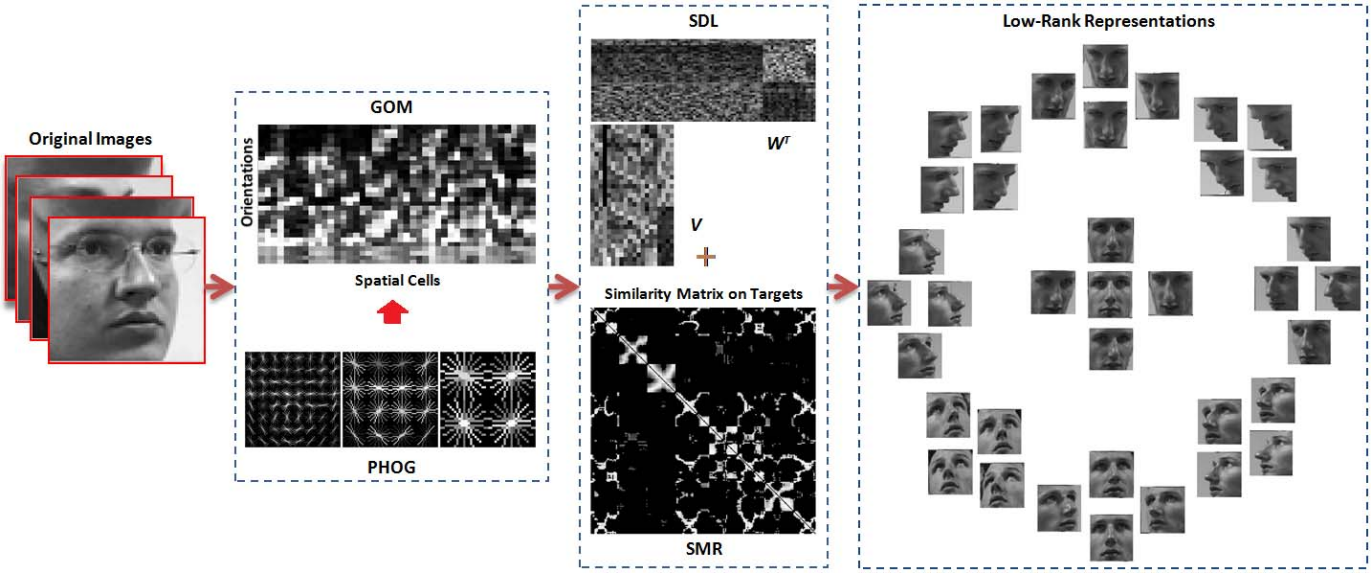


Fig. 1. Flowchart of descriptor learning using the proposed SDL algorithm. The transformations W and V are learned from training samples and used to compute low-rank approximations. In a low-rank space, samples tend to be aligned to their targets and similar head poses tend to be close to each other.

from a set of candidate HOG filters based on their incremental contributions to the performance of object detection.

Recent work on descriptor learning is focused on the reformulation of the handcrafted feature descriptors [16], [27], [28] to use the prior knowledge by capturing edge and gradient information. Discriminative learning algorithms are employed to learn new descriptors, which mainly targets to improve classification tasks. Based on the idea of kernel descriptors [29], the image label information has been taken into consideration and embedded into the design of patch level kernel descriptors, obtaining supervised kernel descriptors. Recently, the configuration of SIFT-like spatial pooling regions is reformulated in [28] as the problem of selecting a few regions from a set of candidate ones. A convex optimization objective function with a sparse and low-rank regularization is solved to learn new descriptors for image retrieval. Similarly, a discriminant face descriptor (DFD) was designed in [27] for the face recognition task. In the DFD, the feature extraction in LBP is parameterized by extending from a pixel to a local image patch. The configurations, including image filters and neighborhood sampling weights, are learned simultaneously from data in a discriminant way.

Most of the existing multioutput regression tasks use handcrafted descriptors, e.g., HOG [3], among which multivariate targets are only explored for particular regressors [20] to improve estimation performance. Moreover, those existing discriminative descriptor learning algorithms are mainly designed for image classification, retrieval, and face recognition tasks, while it is nontrivial to apply them to multioutput regression tasks due to the multivariate continuous outputs rather than discrete labels in classification or recognition. In this paper, we propose a novel SDL algorithm to generate discriminative and compact descriptors for more accurate and efficient multioutput regression.

III. Supervised Descriptor Learning

The schematic flowchart of descriptor learning by the proposed SDL algorithm is shown in Fig. 1. Based on the GOMs obtained from PHOGs, the proposed SDL, which is formulated as GLRAM with an SMR, is applied to generate discriminative and compact descriptors for accurate and efficient multioutput regression.

A. Problem Statement

We are given a set of training samples $\{X_1, \ldots, X_L\}$ and the corresponding multivariate targets $\{Y_1, \ldots, Y_L\}$, where L denotes the number of training samples and $Y_i \in \mathbb{R}^d$ is a continuous multivariate variable. Our aim is to generate highly discriminative but compact feature descriptors for multioutput regression.

We start with more natural matrix representations rather than vectors as inputs, i.e., $X_i \in \mathbb{R}^{M\times N}$, which can be any forms of matrix representations, e.g., raw image pixel intensities and gradient orientation-based features. Inspired by the recent work [15], we propose using the GOMs, which takes the advantage of prior knowledge to capture spatial layout and orientation structures in input images. To achieve more accurate and efficient multioutput regression, we would like to find a new discriminative but compact low-dimensional representation of a GOM X_i by distinctively exploring its spatial and orientation information in rows and columns of X_i. The final descriptor is obtained by vectorizing the low-rank approximation of matrices X_i and fed into multioutput regressors.

B. Generalized Low-Rank Approximation of Matrices

To achieve low-dimensional and therefore compact representations, we propose building the SDL on the GLRAM to leverage its great strength and computational efficiency

in dimension reduction of matrices [11]. In contrast to the conventional low-rank approximations of matrices [13], [30], [31], the GLRAM is to learn two transformations: $W \in \mathbb{R}^{M\times m}$ and $V \in \mathbb{R}^{N\times n}$ with $m \ll M$ and $n \ll N$, and L matrices $D_i \in \mathbb{R}^{m\times n}$, such that WD_iV^T is an appropriate approximation of each X_i, $i = 1, \ldots, L$. We solve for W, V, and D_i by the following minimization problem:

$$\underset{\substack{W,V,D_1,\ldots,D_L \\ W^TW=I_m, V^TV=I_n}}{\arg\min} \frac{1}{L}\sum_{i=1}^{L} \|X_i - WD_iV^T\|_F^2 \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, I_m is an identity matrix of size $m \times m$, and the constraints $W^TW = I_m$ and $V^TV = I_n$ guarantee that W and V have orthogonal columns, which avoids redundancy in the low-rank approximations.

Since the objective function in (1) only minimizes the reconstruction error, the obtained in the low-rank representations $\{D_i\}_{i=1}^L$ tend to be lack of discriminative ability. In order to increase the discrimination of each D_i, we will consider to explore the multivariate regression target for supervised learning. In addition, we would like to highlight that unlike traditional low-rank approximation [13], the matrices $D_1, \ldots, D_L$ are not required to be diagonal.

C. Supervised Manifold Learning

To achieve the discriminative representations of $\{D_i\}_{i=1}^L$, we propose incorporating an SMR for SDL by exploring the multivariate target space. In particular, we impose discrimination on $\{D_i\}_{i=1}^L$ in (1) by adding the SMR term.

We construct a weighted graph $G = (V, E)$ based on Y_i using the ϵ-neighborhood method [32]. To be more precise, nodes Y_i and Y_j are connected if $||Y_i - Y_j||^2 < \epsilon$, where $\epsilon \in R$. V and E denote L vertices and edges, respectively. Note that the graph G is constructed on the multivariate targets $(Y_1, \ldots, Y_L)$ rather than on inputs used in conventional manifold learning [12], [13].

We use $S \in \mathbb{R}^{L\times L}$ to denote the symmetric similarity matrix, in which nonnegative elements correspond to the edge weights of the graph G. The element S_{ij} of S is computed by a heat kernel in (2) with the parameter σ

$$S_{ij} = \exp\left(\frac{-\|Y_i - Y_j\|^2}{2\sigma^2}\right), \quad i, j = 1, \ldots, L. \tag{2}$$

The diagonal elements of S are set to be zeros, i.e., $S_{ii} = 0$. With the SMR term, we minimize

$$\sum_{i,j} \|D_i - D_j\|_F^2 S_{ij}. \tag{3}$$

The similarity matrix S reflects the proximity of data points (images) with respect to their targets and characterizes the manifold structure of the multivariate target space. Therefore, minimizing (3) will achieve low-rank approximations $\{D_i\}_{i=1}^L$ automatically aligned to their regression targets by preserving the intrinsic local geometrical structure of the target space. The discrimination is naturally injected into the low-rank representations $\{D_i\}_{i=1}^L$. In the low-dimensional space, data points with similar targets tend to fall close to each other, while those with dissimilar targets are forced to be apart. Therefore, the discriminative ability of learned low-dimensional representations is largely increased.

D. Learning With SMR

Combining (3) with (1), we obtain the objective function of the GLRAM with the SMR

$$\underset{\substack{W,V,D_1,\ldots,D_L \\ W^TW=I_m, V^TV=I_n}}{\arg\min} \frac{1}{L}\sum_{i=1}^{L} \|X_i - WD_iV^T\|_F^2 + \beta \sum_{i,j} \|D_i - D_j\|_F^2 S_{ij} \tag{4}$$

where the first term finds a low-rank space for input matrices $\{X_i\}_i^L$, the second term injects the supervision of targets to make the low-rank approximations $\{D_i\}_i^L$ to be discriminative, and $\beta \in (0, \infty)$ is a regularization parameter which serves to balance the tradeoff between the reconstruction errors and the discrimination of the low-rank approximations. β can be obtained by cross validation to cater different applications.

Rather than directly solving the above objective function in (4) due to the challenge of the joint optimization with W, V, and D, we will find an alternative objective function based on which we build our final objective function for SDL.

Since the columns of both W and V are orthogonal, the first term in (4) can be rewritten as

$$\begin{aligned}
&\frac{1}{L}\sum_{i=1}^{L} \|X_i - WD_iV^T\|_F^2 \\
&= \frac{1}{L}\sum_{i=1}^{L} \mathrm{Tr}((X_i - WD_iV^T)^T(X_i - WD_iV^T)) \\
&= \frac{1}{L}\sum_{i=1}^{L} \big(\mathrm{Tr}\left(X_i^TX_i\right) - \mathrm{Tr}\left(VD_i^TW^TX_i\right) \\
&\qquad - \mathrm{Tr}\left(X_i^TWD_iV^T\right) + \mathrm{Tr}\left(D_i^TD_i\right)\big).
\end{aligned} \tag{5}$$

Based on the fact that $\mathrm{Tr}(Z) = \mathrm{Tr}(Z^T)$, we have

$$\mathrm{Tr}\left(X_i^TWD_iV^T\right) = \mathrm{Tr}\left(VD_i^TW^TX_i\right).$$

Since given the data $\{X_i\}_{i=i}^L$, $\sum_{i=1}^L \|X_i\|_F^2$ is a constant, to minimize (5) is equivalent to minimizing

$$\sum_{i=1}^{L} \big(\mathrm{Tr}\left(D_i^TD_i\right) - 2\,\mathrm{Tr}\left(VD_i^TW^TX_i\right)\big). \tag{6}$$

Setting the derivatives of (6) with respect to D_i to be 0 gives

$$D_i = W^TX_iV. \tag{7}$$

We further notice that for any i, given the W and V, D_i can be uniquely determined by (7), which is the compact representation of X_i. With (7), we can deal solely with W and V shared by $\{X\}_{i=L}^L$ rather than jointly with

$\{D\}_{i=L}^{L}$ in our objective function, which can be solved efficiently.

1) Substituting (7) into (5), we obtain an alternative minimization term for the first term in (4) as

$$\underset{\substack{W,V\\ W^TW=I_m, V^TV=I_n}}{\arg\min} \quad \frac{1}{L}\left(\sum_{i=1}^{L}\|X_i\|_F^2 - \sum_{i=1}^{L}\|W^TX_iV\|_F^2\right). \tag{8}$$

By dropping the constant $\sum_{i=1}^{L}\|X_i\|_F^2$ and changing the sign of the above optimization problem in (8), we, therefore, have an equivalent maximization problem as follows:

$$\underset{\substack{W,V\\ W^TW=I_m, V^TV=I_n}}{\arg\max} \quad \frac{1}{L}\sum_{i=1}^{L}\|W^TX_iV\|_F^2. \tag{9}$$

By solving the above function, we can obtain W and V, which are used to compute the low-rank approximation D_i of X_i for $i = 1, \ldots, L$.

2) Substituting (7) into the second term in (4), we obtain the SMR in terms of $\{X\}_{i=L}^{L}$, W, and V as follows:

$$\sum_{i,j}\|W^T(X_i - X_j)V\|_F^2 S_{ij} \tag{10}$$

which regularizes the learning of W and V to be supervised by the targets encoded in the similarity matrix S, obtaining discriminative low-rank approximations of $\{X\}_{i=L}^{L}$.

By combining the SMR term in (9) and (10), we solve the final objective function that takes the following forms:

$$\underset{\substack{W,V\\ W^TW=I_m, V^TV=I_n}}{\arg\max} \quad \frac{1}{L}\sum_{i=1}^{L}\|W^TX_iV\|_F^2 - \beta\sum_{i,j}\|W^T(X_i - X_j)V\|_F^2 S_{ij}. \tag{11}$$

In (11), the first term of minimizing construction errors guarantees the reconstruction fidelity in the low-rank approximation, while the second SMR term introduces supervision to enhance the discrimination of the learned representations. By fully leveraging the strengths of the GLRAM and the SMR in dimension reduction of matrices and in supervised manifold learning, the SDL offers an effective and compact formulation to learn highly discriminative, low-dimensional descriptors for multioutput regression.

E. Iterative Solutions via Alternate Optimization

It is not straightforward to solve the objective function in (11) using the existing methods. We seek an iterative algorithm via an alternate optimization, which can efficiently solve the objective function. To facilitate the derivation, the objective function (11) can be further rewritten in terms of traces of matrices

$$\underset{\substack{W,V\\ W^TW=I_m, V^TV=I_n}}{\arg\max} \quad \frac{1}{L}\operatorname{Tr}\left(\sum_{i=1}^{L}W^TX_iVV^TX_i^TW\right) - \beta\operatorname{Tr}\left(\sum_{i,j}W^T(X_i - X_j)VS_{ij}V^T(X_i - X_j)^TW\right). \tag{12}$$

The obtained objective function in (12) avoids the rank-deficit problem in the trace ratio form [33], which would lead to unstable estimation and over sensitivity to limited training samples in hand [34].

The solutions of the objective functions (12) do not have closed forms. We propose to find the optimal solutions of W and V in an iterative way. In particular, we propose an alternate optimization algorithm to solve one by fixing the other, that is, we optimize W by fixing V and, optimize V by fixing W.

1) Fixing V, we find the optimal W by solving

$$\arg\max_{W}\operatorname{Tr}(W^TAW), \quad \text{s.t. } W^TW = I_m. \tag{13}$$

The solution of $W \in \mathbb{R}^{M\times m}$ consists of the m eigenvectors of matrix A corresponding to the m largest eigenvalues, where A is obtained by

$$A = \frac{1}{L}\sum_{i=1}^{L}X_iVV^TX_i^T - \beta\sum_{i,j}(X_i - X_j)VS_{ij}V^T(X_i - X_j)^T. \tag{14}$$

2) Fixing W, we find the optimal $V \in R^{N\times n}$ by solving

$$\arg\max_{V}\operatorname{Tr}(V^TBV), \quad \text{s.t. } V^TV = I_n. \tag{15}$$

The solution of W consists of the n eigenvectors of B associated with the n largest eigenvalues, where B is obtained by

$$B = \frac{1}{L}\sum_{i=1}^{L}X_i^TWW^TX_i - \beta\sum_{i,j}(X_i - X_j)^TWS_{ij}W^T(X_i - X_j). \tag{16}$$

We obtain the optimal solutions of W and V by solving the alternate optimization problems of (13) and (15) iteratively. We adopt the singular value decomposition (SVD) to solve the standard eigendecomposition problems, since it has been shown that the best approximation of given matrices with respect to the Frobenius norm can be obtained by the truncated SVD [11], [13]. The pseudocode for our iterative algorithm is illustrated in Algorithm 1.

F. Convergence Analysis

Although alternating optimization has been used in the previous work, it does not necessarily converge. We provide a rigorous mathematical proof on the convergence of the proposed alternating optimization algorithm. The procedure in

Algorithm 1 Supervised Descriptor Learning

Input: Data matrices $X_1, \ldots, X_L$ and their corresponding targets $Y_1, \ldots, Y_L$.

Output: The projection matrices W and V.

1: Calculate the similarity matrix S using the output targets $Y_1, \ldots, Y_L$;
2: Initialize $V^{(0)} = (I_n, 0)^{\mathrm{T}}$ and set $i \leftarrow 1$;
3: **repeat**
4: Calculate the matrix A by using Eq. (14) and $V^{(i-1)}$;
5: Compute the m eigenvectors $\{\phi_j^W\}_{j=1}^m$ of A corresponding to the m largest eigenvalues;
6: $W^{(i)} = [\phi_1^W, \phi_2^W, \ldots, \phi_m^W]$;
7: Calculate the matrix B by using Eq. (16) and $W^{(i)}$;
8: Compute the n eigenvectors $\{\phi_j^V\}_{j=1}^m$ of B corresponding to the n largest eigenvalues;
9: $V^{(i)} = [\phi_1^V, \phi_2^V, \ldots, \phi_m^V]$;
10: $i \leftarrow i + 1$;
11: **until** Convergence.
12: **return** $W \leftarrow W^{(i-1)}$ and $V \leftarrow V^{(i-1)}$.

Algorithm 1 is effective if and only if it converges, which is guaranteed in Theorem 1.

Theorem 1: Let $L(W, V)$ denote the objective function in (12). Then, $L(W, V)$ is bounded and monotonically increases after every optimization step for W and V, and hence, it converges.

Proof: Since $L(W, V)$ is a continuous function and district $\{(W, V)|W^T W = I_m, V^T V = I_n\}$ is closed, then $L(W, V)$ is bounded. For the tth step in the iteration, the solutions are computed by $W^{(t)} = \arg\max_W \ L(W, V^{(t-1)})$ and $V^{(t-1)} = \arg\max_V \ L(W^{(t)}, V)$. This procedure ensures that we have the following inequality:

$$\begin{aligned} \cdots \leq L(W^{(t-1)}, V^{(t-1)}) &\leq L(W^{(t)}, V^{(t-1)}) \\ &\leq L(W^{(t)}, V^{(t)}) \leq \cdots . \end{aligned}$$

Consequently, $L(W^{(t)}, V^{(t)})$ is monotonically increasing as $t \to \infty$. Therefore, it converges according to the monotone convergence theorem. □

The convergence with iterations for both head pose and area estimation is shown in Fig. 2, in which we can see the objective functions converge very fast within few steps, showing the efficiency of the iterative solution via alternate optimization. In practice, we can check the convergence of the algorithm by computing the value of the objective function $L(W, V)$ in (15) after each iteration. In particular, we check whether $L(W^{(t)}, V^{(t)}) - L(W^{(t-1)}, V^{(t-1)}) < \eta$, for some threshold $\eta > 0$. We set $\eta = 0.001$ in our experiments.

IV. Experiments and Results

We evaluate the proposed SDL on both synthetic data and real-world data sets for multioutput regression tasks: head pose estimation, face alignment for computer vision, and four-chamber cardiac ventricular area estimation for the medical image analysis. Head pose estimation and face alignment are extremely challenging due to illumination, facial expression, and intersubject variations. Cardiac ventricular area estimation is one of the most important procedures daily used in clinical heart disease diagnosis [7]. Simultaneously estimating the areas of four chambers is clinically very useful but challenging due to huge appearance and geometrical structure variations across subjects. Moreover, the complex interaction between four chambers and temporal variations across a cardiac cycle makes it even more challenging. This problem is still unsolved in the medical image analysis.

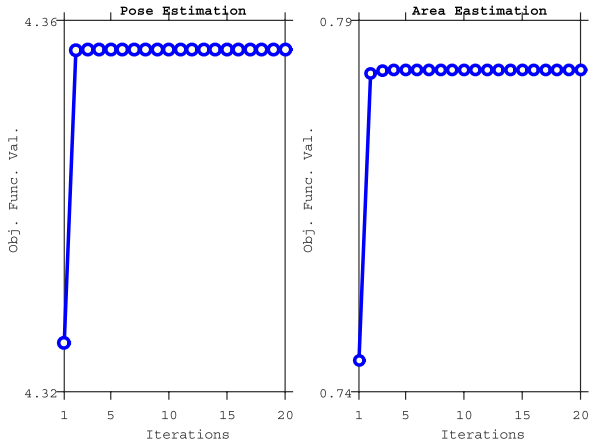


Fig. 2. Convergence of the objective function with iterations. The iterative solution via alternate optimization converges fast within few iterations on both tasks.

A. Experimental Settings

The SDL can work independently with different regressors. We adopt the adaptive K-cluster regression forests (AKRFs) [3] for multivariate estimation. We follow the experimental settings of AKRF in the original work [3] and keep the same for all methods to establish fair comparison. In particualr, we train 20 trees to build the regression forests, set the splitting number to be 2, and use the minimum leaf size of 5, which means that a node is not split any more if the number of training instances associated with the node is less than or equal to 5; the splitting is conducted by a linear support vector machine in which the parameters are set by default as in [3]; the sample rate is set to be 1 for training each tree. We set the tradeoff parameter $\beta = 1$ to keep both the reconstruction fidelity and discriminative ability.

To show the effectiveness of the proposed SDL for dimensionality reduction, we have conducted comprehensive experiments with other representative dimensionality reduction techniques, including principal component analysis (PCA), generalized PCA (GPCA) [11], locality preserving projection (LPP), 2-D LPP [35], and canonical correlation analysis (CCA) [40]. Note that the input image representation to the 1-D algorithms: PCA, LPP, and kernel CCA (KCCA) is a vector, e.g., the vectorized PHOG descriptor, while the input to the 2-D algorithms: GPCA, 2-D-LPP, GLRAM + MR (manifold regularization), and SDL is a matrix representation, e.g., the GOM. We summarized their properties and applicability in Table I to highlight the novelty of the proposed SDL. In contrast to the existing algorithms,

TABLE I
SUMMARY OF DIFFERENT DIMENSIONALITY REDUCTION METHODS

	1D Vector	2D Matrix	Unsupervised	Supervised	Input Manifold	Output Manifold
SDL (Our method)		√		√		√
LPP [32]	√		√		√	
2D-LPP [35]		√	√		√	
GLRAM+MR*		√	√		√	
PCA [36]	√		√			
GPCA [11] (2D-PCA) [37]		√	√			
2D-LDA [38]		√		√		
2D-CCA [39]		√	√			
CCA/KCCA [40]	√		√	√		

*The GLRAM with manifold regularization (MR) is implemented by constructing the Lapalcian matrix using input features rather than regression targets (label information).

the proposed SDL is the first supervised dimensionality reduction algorithm with 2-D inputs for multioutput regression. GPCA shares the similar idea with the 2-D PCA [37] by working on 2-D image representations [41]. However, in the 2-D PCA, a linear transformation is applied only to the right-hand side of image matrices so the image data are projected in one mode only, resulting in poor dimensionality reduction compared with GPCA, which applies to both the left- and right-hand sides [11]. The 2-D CCA [39], which requires pairs of 2-D inputs, is not applicable to our multioutput regression task due to fact that the targets are continuous multivariate values rather than discrete class labels. We implement the KCCA, which outperforms CCA [40]. The 2-D linear discriminant analysis (2-D LDA) [38] is also a well-known supervised dimensionality reduction method for 2-D matrix inputs, which, however, is not applicable to multioutput regression, because 2-D LDA needs discrete class labels as supervision to divide samples into the same classes and different classes.

B. Experiments on Synthetic Data

We conduct simulated experiments on the synthetic data to show the ability of the proposed SDL to learn discriminative descriptors for multioutput regression. The experimental results demonstrate that the proposed SDL significantly improves the baseline of raw inputs for different multioutput tasks and consistently outperforms the state-of-the-art algorithms.

1) Synthetic Data: We adopt the simulation methods in [4], [5], [30], and [32], which were used for testing multioutput regressors, to generate the synthetic data for testing the 2-D reduction algorithms for multioutput regression tasks. In particular, we generate data by $\hat{Y}_i = W\hat{X}_i + \boldsymbol{\epsilon}$, where $\hat{Y}_i \in R^{Q\times d}$ and $\hat{X}_i \in R^{d\times d}$ whose rows are sampled independent identically distributed from multivariate normal distributions, the weight matrix $W \in R^{Q\times d}$ is generated randomly and fixed for both training and test samples, and $\boldsymbol{\epsilon}$ is the added Gaussian noise. Differently from [4], [5], [30], and [32], our input $\hat{X}_i$ is 2-D matrix and we take $Y_i = (1/d)\sum \hat{Y}_i^j \in R^Q$ as the multiple outputs associated with $\hat{X}_i$, where $\hat{Y}_i^j$ is the jth column of $\hat{Y}_i$. As a consequence, rows and columns of the input matrix correlated distinctively to multiple outputs, which fulfills the practical applications. To demonstrate the ability of the proposed SDL in learning discriminative but compact descriptors from inputs with redundant information, we add the random values by concatenating random values to each $\hat{X}_i$ along both rows and columns to obtain X_i of the size $D \times D$, where $D > d$, and we set $d = 20$ and $D = 40$. To further show the effectiveness of the SDL for diverse multioutput regression tasks, we experiment with different numbers of outputs: $Q = 5, 10$, and 50. We generate 3000 and 1000 samples (Y_i, X_i) for training and test, respectively.

TABLE II
COMPARISON OF THE PROPOSED SDL WITH OTHER DIMENSIONALITY REDUCTION TECHNIQUES ON THE SYNTHETIC DATA (AVERAGE CORRELATION COEFFICIENT BETWEEN PREDICTED AND GROUND TRUTH VALUES OF Q OUTPUTS)

Method \ Number of Outputs	Q = 5	Q = 10	Q = 50
SDL	**0.952**	**0.949**	**0.946**
Baseline	0.841	0.839	0.834
GPCA (2D-PCA)	0.887	0.886	0.873
PCA	0.876	0.871	0.864
2D-LPP	0.877	0.876	0.863
LPP	0.856	0.851	0.845
GLRAM+MR	0.887	0.876	0.871
KCCA	0.875	0.869	0.864

2) Results: Table II shows the experimental results on the synthetic data in terms of the average correlation coefficient of multiple outputs between the prediction and ground truth. The leftmost column corresponds the representative feature learning methods, and the three columns on the right correspond to multioutput regression tasks of different numbers Q of outputs. The baseline is the raw inputs without using feature learning. The simulated multioutput regression task is challenging due to the highly redundant information of 75% in the inputs. The proposed SDL yields high estimation performance on all the multioutput regression tasks and substantially outperforms the baseline of raw inputs without the feature learning and state-of-the-art dimensionality reduction algorithms.

The proposed SDL significantly improves the baseline of raw inputs without feature learning by large margins up to 13.4% in terms of correlation coefficients, as shown in Table II. By supervised learning, the proposed SDL is able to extract informative features that are related to multivariate targets while removing redundant information, obtaining highly

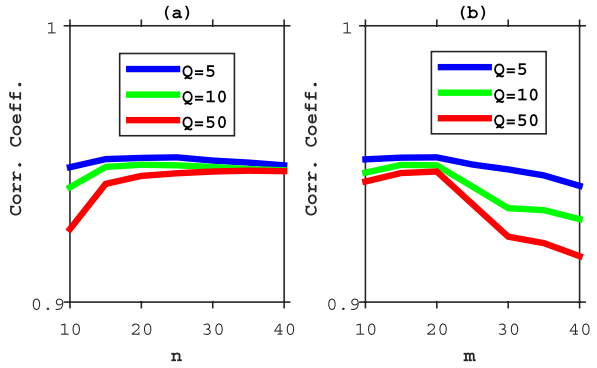


Fig. 3. Performance with different dimensionalities. Q is the number of outputs. m is the reduced dimensionality by W and n is the reduced dimensionality by V. (a) $m = 20$. (b) $n = 25$.

discriminative but compact descriptors. The consistently high improvement of the SDL over the baseline for all tasks of different numbers of outputs demonstrates the great strength of the proposed SDL to learn discriminative but compact descriptors for multioutput regression tasks.

The effectiveness of the proposed SDL is further demonstrated by the much better performance than widely used representative feature learning algorithms. The SDL consistently outperforms other dimensionality reduction algorithms by large margins by up to 12%. The advantage of our supervised feature learning has been validated by the great improvement over the unsupervised learning algorithms including PCA, 2-D PCA, LPP, 2-D LPP, and GLRAM; the large improvement over the supervised learning algorithm KCCA which also uses output information demonstrates the effectiveness of the proposed SMR in learning discriminative descriptors for multioutput regression.

The proposed SDL can generate highly discriminative but compact descriptors by largely removing redundant information from raw inputs. We show the effect of dimensionality learned by the proposed SDL on the performance in Fig. 3. The SDL can achieve the best performance with the dimensions of 500 ($n = 25$ and $m = 20$), which is much lower compared with the raw input of 1600 dimensions, which demonstrates its great capability of learning discriminative compact descriptor for multioutput regression. More importantly, as shown in Fig. 3, the different performance variation patterns with the dimensions of rows and columns show the great advantages of the proposed SDL by learning two separate projection matrices for descriptor learning. By treating rows and columns of input matrices separately, the SDL can effectively explore the distinctive meanings residing in rows and columns to achieve optimal low-dimensional descriptors.

C. Experiments on Real Data

We apply the SDL to head pose estimation and face alignment for the typical computer vision task and to area estimation for medical image analysis.

1) Data Sets and Settings: The Pointing'04 data set [42] is a widely used benchmark for head pose estimation in computer vision. The dataset contains 2790 images from 15 subjects.

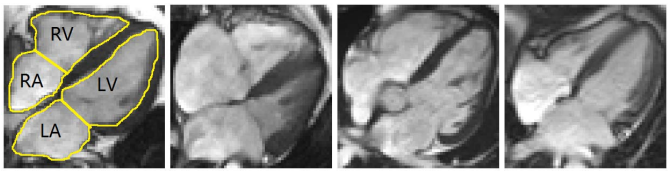


Fig. 4. Illustration of MR images with four chambers including the left ventricle/atrium (LV/RV) and the right ventricle/atrium (RV/RA).

Each subject has two series of 93 images with different head poses represented by yaw and pitch, i.e., each image has a 2-D target. Bounding boxes associated with images that indicate the head regions are provided with the data set. We follow the original work to crop the images into 64×64 pixels based on the bounding boxes. We adopt the commonly used mean absolute error (MAE) to measure the performance for head pose estimation. compared with the existing methods [3], [10], [43] on the this data set, we employ a fivefold cross validation.

The Cardiac MR data set is our newly collected multioutput regression benchmark for direct area estimation of four-chamber cardiac ventricular, which is still unaddressed in the medical image analysis due to the high complexity of four chambers. It contains 3125 MR images of cardiac four chambers from 125 subjects each of which has 25 images across a temporal cardiac cycle. Example images are shown in Fig. 4. The ground truth is obtained from manual segmentation by human experts. Note that we use the normalized areas as the targets, i.e., the number of pixels in a chamber divided by the total number of pixels of the images. The leave-one-subject-out cross validation is used for evaluation. To evaluate the area estimation performance, we measure the performance by the correlation coefficient between the ground truth and the estimation.

The 300-W data set [44] has been widely used for the validation of face alignment methods. The 300-W data set is created from several reannotated data sets, including LFPW [45], AFW [46], Helen [47], and IBUG [44]. The number of landmarks is 68. We follow the common settings in the previous work [48], [49] and build the training set using AFW, the training set of LFPW, and the training set of Helen, with 3148 images in total. The testing set consists of IBUG, the testing set of LFPW, and the testing set of Helen, with 689 images in total. We also follow common evaluation settings on: 1) fullset: all images of the testing set; 2) common subset: testing sets of Helen and LFPW; and 3) challenging subset: the IBUG data set. Following these work, we evaluate the performance using the average landmark error normalized by interpupil distance:

$$\text{error} = \frac{1}{N}\sum_{i=1}^{N}\frac{\frac{1}{P}\sum_{p=1}^{P}\sqrt{\left(x_p^{(i)}-\hat{x}_p^{(i)}\right)^2+\left(y_p^{(i)}-\hat{y}_p^{(i)}\right)^2}}{d^i_{\text{pupils}}}. \tag{17}$$

We use a PHOG descriptor, which is widely used in computer vision tasks, e.g., head pose estimation [3], as the baseline method for performance comparison. The PHOG (2604d)

TABLE III
COMPARISON RESULTS FOR HEAD POSE ESTIMATION

Methods	yaw	pitch	average
SDL	**4.12±0.17**	**2.09±0.12**	**3.11**
PHOG (Baseline)	5.30±0.29	3.34±0.21	4.32
Intensity*	5.21±0.24	3.17±0.14	4.19
GLRAM [11]	5.09±0.25	3.03±0.16	4.06
LBP [52]	5.53±0.29	3.37±0.23	4.45
GIST [51]	6.06±0.44	5.03±0.35	5.55
AKRF [3]	5.50	3.41	4.46
Geng et al. [53]	4.24	2.69	3.47
Fenzi et al. [10]	5.94	6.73	6.34
Haji [43]	6.56	6.61	6.59

*This results are obtained by the proposed SDL learned from image intensity.

is obtained from a three-level pyramid [50]. The parameter ϵ is set to be the largest distance of the 1% shortest pairwise distances of all training samples. This ensures only very similar samples are connected in the graph. Note that 1% is found by cross validation from training sample and works generally well on all data sets.

To further demonstrate the advantage of the SDL in learning discriminative descriptors, we have compared our SDL with widely used descriptors, e.g., GIST [51] and histogram of LBP [50]. We implement them with a similar spatial pyramid to PHOG to establish fair comparison. Our final descriptor learned by the SDL is of 800 dimensions, which is much lower than the LBP (4872d) and GIST (4096d) descriptors. The compact low-dimensional descriptors learned by the SDL can dramatically reduce the computational cost to achieve more efficient multioutput regression. The contributions of the used GOM representation and the SMR are investigated, respectively, by applying the SDL to intensity and by implementing GLRAM without the SMR.

2) Head Pose Estimation: The proposed SDL produces high performance for head pose estimation with high accuracy for both yaw and pitch, which outperforms the algorithms in the state of the arts [3] by a large reduction of the MAE up to 22.3% (pitch). The comparison results using the validation of even training and test split are summarized in Table III.

The advantages of the proposed SDL is further shown by the much better performance than the handcrafted descriptors, including PHOG, LBP and GIST, and previous methods applied to this task. The significant improvement over the baseline using the PHOG descriptor indicates the advantage of the induced supervised manifold learning by the SDL. By the SMR, the SDL can effectively extract the most discriminative features that are closely related to regression targets while removing irrelevant information from raw features. Moreover, the regression can be conducted more efficient than that based on original vectorized PHOG descriptor due the large dimension reduction. Note that the better performance of the baseline PHOG than the existing methods is largely due to the use the advanced AKRF regression method recently proposed in [3], which outperforms other regression methods, e.g., SVR and conventional random forests.

The better performance of the baseline PHOG with AKRF than [10] and [43] is largely due to the use of the AKRF, because both [10] and [43] use the PHOG features but different regressors. Our SDL is much better than all the algorithms with the same AKRF regressor, which shows the highly discriminative ability of the descriptor learned by our SDL.

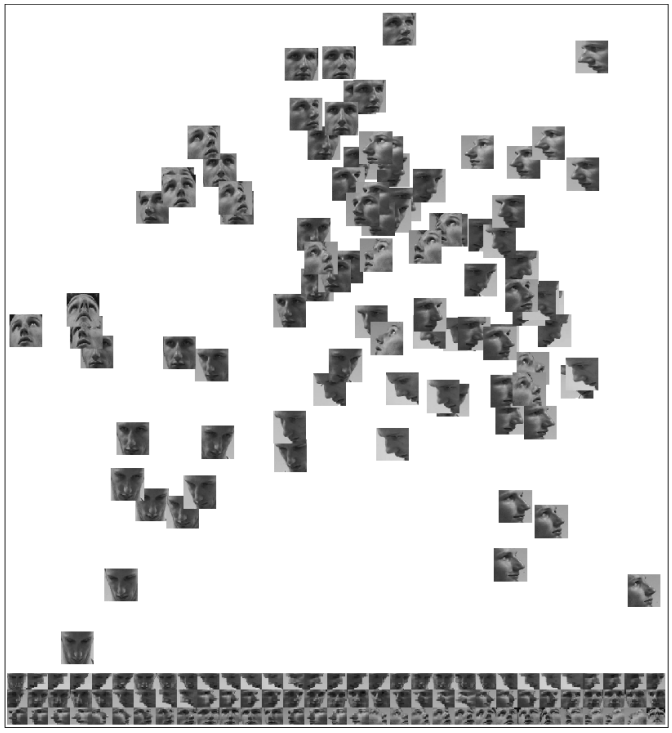


Fig. 5. 2-D illustration ($m = 1$ and $n = 2$) of head pose images using the method in [54].

The improvement of SDL over that directly on image intensity by SDL shows the benefit of the used GOM representation. The GOM not only leverages the prior knowledge to capture edges and gradient information but also allows to explore the distinctive physical meanings of residing in orientation information and spatial layout. The increase over GLRAM validates the effectiveness of the proposed SMR in exploring the supervision of regression targets for feature learning. Even on the raw image intensity, the SDL can produce the impressive results that are comparable with and even better than other carefully handcrafted descriptors, e.g., PHOG, LBP, and GIST, showing the great power of the SDL in descriptor learning.

The effectiveness of the SDL is further demonstrated by the low-dimensional visualization, as shown in Fig. 5. The learned descriptors by the SDL demonstrate highly discriminative ability with only two dimensions. The illustration is obtained by setting $m = 1$ and $n = 2$. We can see in Fig. 5 that images with similar head poses are clustered, while those with very different head pose orientations are separated and tend to be scattered away. The high discrimination stems from the supervision of the regression targets incorporated by the proposed SMR. The SDL is able to effectively extract the most discriminative features closely related to head pose orientations. As a consequence, in a very low-dimensional space, images are discriminatively aligned to their head poses, which guarantees more accurate head pose estimation.

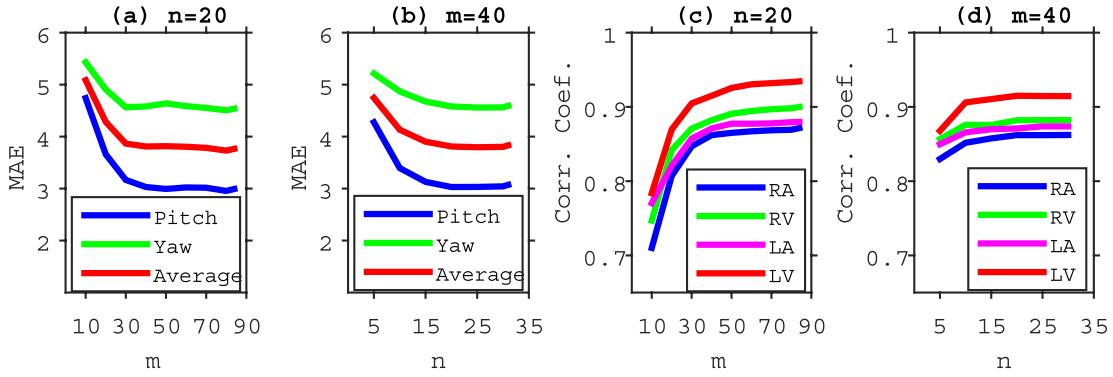


Fig. 6. Estimation performance under different dimensions for (a) and (b) head pose estimation and (c) and (d) area estimation. m and n are the reduced dimensions by W and V, respectively.

TABLE IV
COMPARISON RESULTS FOR FOUR-CHAMBER CARDIAC VENTRICULAR AREA ESTIMATION

Methods	LV	LA	RV	RA
SDL	**0.915**	**0.871**	**0.882**	**0.862**
PHOG (Baseline)	0.869	0.819	0.832	0.811
Intensity*	0.879	0.824	0.848	0.818
GLRAM [11]	0.894	0.841	0.856	0.835
LBP [52]	0.868	0.799	0.827	0.794
GIST [51]	0.864	0.828	0.843	0.815

*This results are obtained by the proposed SDL learned from image intensity.

TABLE V
COMPARISON OF THE PROPOSED SDL WITH OTHER DIMENSIONALITY REDUCTION TECHNIQUES FOR HEAD POSE ESTIMATION

Methods	yaw	pitch	average
SDL	**4.12±0.17**	**2.09±0.12**	**3.11**
GPCA (2D-PCA)	5.11±0.24	3.13±0.11	4.12
PCA	5.25±0.27	3.53±0.18	4.39
2D-LPP	5.19±0.27	3.95±0.22	4.57
LPP	5.31±0.31	4.12±0.26	4.72
GLRAM+MR	4.88±0.15	2.95±0.10	3.92
KCCA	8.71±1.33	7.25±1.21	7.98

TABLE VI
COMPARISON OF THE PROPOSED SDL WITH OTHER DIMENSIONALITY REDUCTION TECHNIQUES FOR AREA ESTIMATION

Methods	LV	LA	RV	RA
SDL	**0.915**	**0.871**	**0.882**	**0.862**
GPCA (2D-PCA)	0.885	0.838	0.843	0.822
PCA	0.871	0.812	0.825	0.807
2D-LPP	0.892	0.833	0.861	0.825
LPP	0.887	0.820	0.830	0.809
GLRAM+MR	0.902	0.843	0.864	0.845
KCCA	0.815	0.760	0.807	0.744

3) Area Estimation: The areas of the four chambers, i.e., the left/right ventricles (LV/RV) and left/right atriums (LA/RA) in a single image are estimated simultaneously. Despite of the great challenge in the estimation of continuous area values, the SDL produces high accuracy for all the four chambers with large advantages over other methods. Especially for the LV, the SDL can achieve a correlation coefficient of 91.5%, which is clinically significant indicating its potential use in practical diagnosis of heart deceases [7].

The SDL substantially outperforms other methods, including both the state-of-the-art descriptors, e.g., LBP and GIST, and other techniques applied to this task by up to 7.2%, which demonstrates the effectiveness of the SDL for continuous multivariate estimation. The comprehensive comparison with other methods is reported in Table IV. The SDL can achieve very close area estimation for all the four chambers. Consistently to head pose estimation, both GOM and SMR contribute the good performance of the proposed SDL, which is shown by comparison results with intensity and GLRAM in Table IV.

The proposed SDL outperforms representative dimensionality reduction algorithms, including PCA, GPCA, LPP, 2-D-LPP, and kernel CCA (KCCA). The comparison results for both head pose and cardiac chamber area estimation are reported in Tables V and VI, respectively. The advantages of the proposed SDL stem from the systematical integration of the generalized low-rank approximation of matrices (GLRAM) and SMR, which can find a highly discriminative but compact low-dimensional space.

The favorable benefit of the incorporated SMR in the SDL is shown by the better performance than unsupervised algorithms: PCA, GPCA, LPP, and 2-D LPP. The SDL outperforms compared dimensionality reduction algorithms with large margins, which demonstrates the effectiveness of the SDL in combing the ideas of supervised manifold learning. Moreover, the much better performance of SDL than GLRAM + MR, e.g., an average error reduction of 22.1% for head pose estimation, shows the advantages of constructing the Laplacian matrix using regression targets instead of using input features, which also demonstrates the effectiveness of combining GLRAM with the SMR in a single objective function for feature learning. In addition, KCCA produces the worst results due to huge information loss, because the dimensionality of learned descriptors is restricted by output dimensions [40], [55].

The effectiveness of the proposed SDL in learning compact descriptors is shown by the impressive results with very low dimensionality, as shown in Fig. 6. We conduct experiments on

Fig. 7. Blue dots and yellow circles represent the ground truth and estimated landmarks, respectively.

the effects of dimension reduction by W and V by testing one with the other fixed. Fig. 6(a) and (b) shows the performance with different values of W and V, respectively, for head pose estimation. Fig. 6(c) and (d) shows the performance with different values of W and V, respectively, for area estimation. The peak performance by the SDL for both head pose and area estimation happens with very low dimensions of approximately $m = 40$ and $n = 20$.

Moreover, the distinctive performance variation patterns in Fig. 6 demonstrate the advantages using the GOM, which allows to differentially explore the orientation and spatial layout information, rather than the vectorized PHOG descriptor. The performance also shows different variation patterns between head pose and area estimation. The estimation results are more effected by spatial cells than orientation bins for area estimation, which would be due to that spatial layouts of the four chambers carries the discriminative appearance information. However, the orientation bins exhibit more effects on head pose estimation than area estimation, which could be explained by the fact that the orientation information is more characteristic for head poses. The different performance effected by spatial cells and orientation bins between area and head pose estimation validates the use of the GOM rather than the PHOG vector in which spatial and orientation information is treated equally in feature representations.

4) Face Alignment: The experimental results on the 300 W data set are reported in Table VII and Fig. 7. As shown in Table VII, compared with the methods published recently, our SDL achieves competitive results although our method does not rely on any initialization and Cascade regression models. The large improvement of the SDL over baseline PHOG shows the strength of our SDL for feature learning in face alignment. We have also shown the face alignment results for randomly-selected samples from the data set in Fig. 7 for intuitive illustration. Our method can accurately predict the landmarks very close to ground truth.

TABLE VII
COMPARISON ON THE 300 W DATA SET (68 LANDMARK POINTS)

Method	Year	Common Subset	Challenging Subset	Fullset
[46]	2012	8.22	18.33	10.20
[56]	2013	5.57	15.40	7.50
[48]	2014	4.95	11.98	6.30
[49]	2015	4.79/4.73	10.92/9.98	5.99/5.76
PHOG		8.34	15.68	9.78
SDL		4.74	10.89	5.95

The dramatic improvement of the SDL over PHOG is due to the great ability of the SDL to extract discriminative features by supervised learning. The great challenges of face alignment stem from the high intersubject variations, illumination, occlusion, and facial expression variability. Handcrafted features, e.g., PHOG, demonstrate severe limitation on this task, while learning-based algorithms especially supervised learning, e.g., the proposed SDL, can largely handle those challenges. Our SDL exhibits its great capability of learning discriminative descriptors by taking into account the supervision of regression outputs, i.e., the landmarks on the face. The SDL can extract discriminative and informative features that are directly related to the landmarks while removing redundant noisy information, which therefore significantly improves the performance of face alignment.

In general, the competitive performance of our SDL algorithm on challenging face alignment demonstrates its

effectiveness in supervised feature learning for more extensive applications. Together with experiments conducted in this paper, we have successfully validated our SDL algorithm for feature learning in multioutput regression.

V. Conclusion

In this paper, we have presented a novel SDL algorithm to obtain a compact and highly discriminative descriptors for multioutput regression. We formulate the SDL as GLRAM with an SMR, which leverages their strengths in dimensionality reduction and supervised manifold learning. Being able to work on diverse inputs, e.g., image intensity and handcrafted features, the SDL provides a general framework of SDL for multioutput regression, which cannot only improve the performance of multivariate estimation, but also enables more efficient multioutput regression. Extensive experiments have been conducted on challenging multioutput regression tasks: head pose estimation, face alignment, and four-chamber cardiac ventricular area estimation. The achieved high performance demonstrates the strength of the SDL for diverse multivariate estimation applications in both computer vision and medical image analysis.

Acknowledgment

The authors would like to thank Dr. J. Qin for the assistance with the computing environment. Computations were performed using the data analytics Cloud at SHARCNET (www.sharcnet.ca) provided through the Southern Ontario Smart Computing Innovation Platform.

References

[1] P. Rai, A. Kumar, and H. Daume, "Simultaneously leveraging output and task structures for multiple-output regression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 3185–3193.

[2] S. Schulter, C. Leistner, P. Wohlhart, P. M. Roth, and H. Bischof, "Alternating regression forests for object detection and pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 417–424.

[3] B. Hara and R. Chellappa, "Growing regression forests by classification: Applications to object pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 552–567.

[4] M. Torki and A. Elgammal, "Regression from local features for viewpoint and pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2603–2610.

[5] A. Guzman-Rivera *et al.*, "Multi-output learning for camera relocalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1114–1121.

[6] R. Gauriau, R. Cuingnet, D. Lesage, and I. Bloch, "Multi-organ localization combining global-to-local regression and confidence maps," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2014, pp. 337–344.

[7] X. Zhen, Z. Wang, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Direct estimation of cardiac bi-ventricular volumes with regression forests," in *Proc. 17th Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.*, 2014, pp. 586–593.

[8] X. Zhen, Z. Wang, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation," *Med. Image Anal.*, vol. 30, pp. 120–129, May 2015.

[9] S. Wang and R. M. Summers, "Machine learning and radiology," *Med. Image Anal.*, vol. 16, no. 5, pp. 933–951, Jul. 2012.

[10] M. Fenzi, L. Leal-Taixé, B. Rosenhahn, and J. Ostermann, "Class generative models based on feature regression for pose estimation of object categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 755–762.

[11] J. Ye, "Generalized low rank approximations of matrices," *Mach. Learn.*, vol. 61, nos. 1–3, pp. 167–191, 2005.

[12] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

[13] Z. Zhang and K. Zhao, "Low-rank matrix approximation with manifold regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1717–1729, Jul. 2013.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2005, pp. 886–893.

[15] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Subspace learning from image gradient orientations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2454–2466, Dec. 2012.

[16] P. Wang, J. Wang, G. Zeng, W. Xu, H. Zha, and S. Li, "Supervised kernel descriptors for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2858–2865.

[17] E. Ahmed, G. Shakhnarovich, and S. Maji, "Knowing a good HOG filter when you see it: Efficient selection of filters for detection," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 80–94.

[18] X. Zhen, Z. Wang, M. Yu, and S. Li, "Supervised descriptor learning for multi-output regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1211–1218.

[19] X. Zhen, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Direct and simultaneous four-chamber volume estimation by multi-output regression," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2015, pp. 669–676.

[20] G. Liu, Z. Lin, and Y. Yu, "Multi-output regression on the output manifold," *Pattern Recognit.*, vol. 42, no. 11, pp. 2737–2743, Nov. 2009.

[21] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

[22] K.-A. Sohn and S. Kim, "Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2012, pp. 1081–1089.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[24] H. Cai, K. Mikolajczyk, and J. Matas, "Learning linear discriminant projections for dimensionality reduction of image descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 338–352, Feb. 2011.

[25] M. Yu, L. Shao, X. Zhen, and X. He, "Local feature discriminant projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, doi: 10.1109/TPAMI.2015.2497686.

[26] Y. Guo, G. Zhao, and M. Pietikäinen, "Discriminative features for texture description," *Pattern Recognit.*, vol. 45, no. 10, pp. 3834–3843, Oct. 2012.

[27] Z. Lei, M. Pietikäinen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 289–302, Feb. 2014.

[28] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1573–1585, Aug. 2014.

[29] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 244–252.

[30] Y. Deng, Q. Dai, R. Liu, Z. Zhang, and S. Hu, "Low-rank structure learning via nonconvex heuristic recovery," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 3, pp. 383–396, Mar. 2013.

[31] K. Tang, R. Liu, Z. Su, and J. Zhang, "Structure-constrained low-rank representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2167–2179, Dec. 2014.

[32] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Neural Inf. Process. Syst.*, vol. 16. 2004, p. 153.

[33] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[34] N. Karampatziakis and P. Mineiro, "Discriminative features via generalized eigenvectors," in *Proc. ACM Int. Conf. Mach. Learn.*, 2014, pp. 1–8.

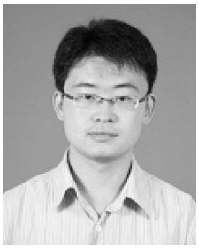
[35] S. Chen, H. Zhao, M. Kong, and B. Luo, "2D-LPP: A two-dimensional extension of locality preserving projections," *Neurocomputing*, vol. 70, nos. 4–6, pp. 912–921, 2007.

[36] I. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Wiley, 2002.

[37] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.

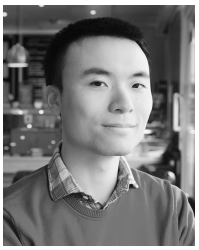
[38] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 1569–1576.

[39] S. H. Lee and S. Choi, "Two-dimensional canonical correlation analysis," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 735–738, Oct. 2007.
[40] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[41] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Uncorrelated multilinear principal component analysis for unsupervised multilinear subspace learning," *IEEE Trans. Neural Netw.*, vol. 20, no. 11, pp. 1820–1836, Nov. 2009.
[42] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *Proc. Int. Workshop Vis. Observat. Deictic Gestures ICPR*, 2004, pp. 1–9.
[43] M. A. Haj, J. Gonzàlez, and L. S. Davis, "On partial least squares in head pose estimation: How to simultaneously deal with misalignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2602–2609.
[44] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Dec. 2013, pp. 397–403.
[45] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2930–2940, Dec. 2013.
[46] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.
[47] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 679–692.
[48] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1685–1692.
[49] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4998–5006.
[50] A. Vedaldi and B. Fulkerson. (2008). *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. [Online]. Available: http://www.vlfeat.org/
[51] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[52] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen, *Computer Vision Using Local Binary Patterns*, vol. 40. Springer Science & Business Media, 2011.
[53] X. Geng and Y. Xia, "Head pose estimation based on multivariate label distribution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1837–1842.
[54] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, nos. 2579–2605, p. 85, 2008.
[55] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Dept. Statist., Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. 688, 2005.
[56] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.



Xiantong Zhen received the B.S. and M.E. degrees from Lanzhou University, Lanzhou, China, in 2007 and 2010, respectively, and the Ph.D. degree from the Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, U.K., in 2013.

He is currently a Post-Doctoral Fellow with the University of Western Ontario, London, ON, Canada. His current research interests include machine learning, computer vision, and medical image analysis.



Mengyang Yu (S'14) received the B.S. and M.S. degrees from the School of Mathematical Sciences, Peking University, Beijing, China, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K.

His current research interests include computer vision, machine learning, and data mining.

Ali Islam, photograph and biography not available at the time of publication.

Mousumi Bhaduri, photograph and biography not available at the time of publication.

Ian Chan, photograph and biography not available at the time of publication.

Shuo Li received the Ph.D. degree in computer science from Concordia University, Montréal, QC, Canada, in 2006.

He was a Research Scientist and a Project Manager of General Electric (GE) Healthcare, London, ON, Canada, for nine years. He is currently an Associate Professor with the Department of Medical Imaging and Medical Biophysics, University of Western Ontario, London, and a Scientist with the Lawson Health Research Institute, London. He is the Founder and has been the Director of the Digital Imaging Group of London, London, since 2006, which is a highly dynamic and interdisciplinary group. He has authored or co-authored over 100 publications and edited five Springer books. His current research interests include the development of intelligent analytic tools to facilitate physicians and hospital administrative to handle the big medical data, centered with medical images.

Dr. Li was a recipient of several GE internal awards. His Ph.D. thesis received the Doctoral Prize giving to the most deserving graduating student in the Faculty of Engineering and Computer Science. He serves as a Guest Editor and an Associate Editor in several prestigious journals. He severed as a Program Committee Member in top conferences.