Supervised Descriptor Learning for Multi-Output Regression

Xiantong Zhen
University of Western Ontario, London, ON, Canada
xzhen7@uwo.ca

Zhijie Wang
GE Healthcare, London, ON, Canada
zhijie@ualberta.ca

Mengyang Yu
Northumbria University, Newcastle, United Kingdom
mengyang.yu@northumbria.ac.uk

Shuo Li
GE Healthcare, London, ON, Canada
shuo.li@ge.com

Abstract

Descriptor learning has recently drawn increasing attention in computer vision, Existing algorithms are mainly developed for classification rather than for regression which however has recently emerged as a powerful tool to solve a broad range of problems, e.g., head pose estimation. In this paper, we propose a novel supervised descriptor learning (SDL) algorithm to establish a discriminative and compact feature representation for multi-output regression. By formulating as generalized low-rank approximations of matrices with a supervised manifold regularization (SMR), the SDL removes irrelevant and redundant information from raw features by transforming into a low-dimensional space under the supervision of multivariate targets. The obtained discriminative while compact descriptor largely reduces the variability and ambiguity in multi-output regression, and therefore enables more accurate and efficient multivariate estimation. We demonstrate the effectiveness of the proposed SDL algorithm on a representative multi-output regression task: head pose estimation using the benchmark Pointing'04 dataset. Experimental results show that the SDL can achieve high pose estimation accuracy and significantly outperforms state-of-the-art algorithms by an error reduction up to 27.5% . The proposed SDL algorithm provides a general descriptor learning framework in a supervised way for multi-output regression which can largely boost the performance of existing multi-output regression tasks.

1. Introduction

Descriptor learning, which is also known as feature representation learning, has recently become attractive in visual recognition [33, 25, 37]. Most of existing algorithms are specifically developed for classification tasks while not directly applicable to regression due to the continuous multivariate targets rather than discrete class labels [4]. However, multi-output regression has recently emerged and extensively studied for many computer vision tasks, *e.g.*, head pose estimation [15], human body pose estimation [29] and viewpoint estimation [28]. Moreover, many researchers have found their applications, *e.g.*, camera relocalization [24, 13] and cardiac volume estimation [1, 38], can be elaborately solved by transferring the original problem into a multi-output regression task, which not only substantially outperforms conventional approaches but also offers a more compact and exquisite mathematical formulation to circumvent the difficulty in conventional approaches, *e.g.*, the inverse problems [13].

Great challenges in multi-output regression arise from the complex relationship between the high-dimensional input feature descriptors and multivariate targets. Images with the same targets often exhibit great variability due to illumination changes, geometrical complexity and inter-subject variations. Meanwhile, images with different targets can also share very similar appearance which causes large ambiguity. The variability and ambiguity pose great challenges in multi-output regression tasks and designing discriminative feature descriptors to reduce the variability and ambiguity becomes the bottleneck for accurate and efficient multivariate estimation [9]. Handcrafted descriptors, e.g., HOG, are mainly used and can obtain good results in many multi-output regression tasks [14, 15]. However, the guidance of observed regression targets that represent high-level concepts in the input data [15] is completely ignored, which leads to indiscriminate and lengthy representations. Incorporating the supervision of targets into descriptor learning to achieve discriminative feature representations is imperative and highly desired for more accurate and efficient multivariate estimation, which has not been addressed.

In this paper, we propose a novel supervised descriptor learning (SDL) algorithm for multi-output regression to achieve more accurate and efficient multivariate estima-

tion. The proposed SDL is formulated as generalized low-rank approximations of matrices with a supervised manifold regularization (SMR). The SDL seeks low-dimensional feature representations under the supervision of regression targets achieving discriminative and compact descriptors. By systematically assembling the generalized low-rank approximation and the newly proposed SMR to leverage their strengths in dimension reduction and supervised manifold learning, the SDL provides a novel general framework to effectively learn compact and discriminative feature representations for multi-output regression. The SDL can be built on image intensity, handcrafted features and deeply learned representations, and can also work seamlessly in conjunction with existing regressors. The obtained compact and discriminative feature representation not only substantially boosts multivariate estimation for better performance but also enables to conduct in a more efficient way.

In order to take advantage of prior knowledge to capture edges and gradient structures [7], we propose building our SDL on gradient orientation matrices (GOM) rather than pixel intensity. This is partially inspired by previous work showing that combining gradient orientations with supervised learning can boost classification performance [33, 2] and replacing pixel intensities with gradient orientations offers reliable subspace estimation [30]. The GOM of an image can be constructed by stacking histogram of oriented gradients (HOG) in a matrix with rows of orientations and columns of spatial cells. The SDL provides an effective strategy to differentially explore spatial layout and orientation information in the GOM to achieve data-driven representations for specific tasks.

To demonstrate the ability of learned compact and discriminative representations for multivariate estimation, we evaluate the SDL on a representative multi-output regression tasks: head pose estimation, a challenging problem in computer vision. Experimental results show that the SDL is able to generate compact while discriminative feature representations for head poses, and achieves high estimation accuracy which significantly outperforms the state-of-the-art algorithms including both descriptors and dimensionality reduction techniques.

2. Related Work

We will briefly review related work on discriminative descriptor learning and multi-output regression both of which have recently drawn considerable attention in computer vision for diverse applications.

Discriminative descriptor learning can either be built on well-established feature descriptors, *e.g.*, the SIFT [20, 6], LBP [22, 12], HOG [2] descriptors or be reformulated to learn those descriptors in a discriminative way [33, 18, 25]. The performance of handcrafted descriptors can be improved by applying discriminative learning techniques [6, 12, 2]. Cai et al. [6] proposed to learn linear discriminant projections for dimensionality reduction of local SIFT descriptors by minimizing the distance between matched pairs of descriptors while maximizing those between unmatched pairs. Guo et al. [12] developed a three-layered model to extract discriminative and robust features based on the LBP descriptors. Ahmend et al. [2] proposed a learning architecture to select relative discriminative filters from a pool of candidate HOG filters based on their incremental contributions to the performance of object detection.

By reformulating and parameterizing the handcrafted descriptors, e.g., SIFT, HOG and LBP descriptors, [33, 25, 18], to learn new descriptors discriminatively has gained great popularity recently. These methods can take advantages of prior knowledge and discriminative learning techniques to improve classification performance. To increase the discriminate ability of kernel descriptors [5], Wang et al. [33] presented a supervised framework to embed the image label information into the design of patch level kernel descriptors. The obtained supervised kernel descriptors (SKDES) achieve impressive performance on image classification benchmarks. Simonyan et al. [25] reformulated the learning of the configuration of SIFT-like spatial pooling regions as the problem of selecting a few regions among a large set of candidate ones. They solved a convex optimization objective function with a sparse and low-rank regularization to learn a new SIFT-like descriptor. Lei at al. [18] developed a discriminant face descriptor (DFD) for face recognition. They follow an LBP-like feature extraction mechanism by extending from a pixel in LBP to a local image patch. Both image filters and neighborhood sampling weights are parameterized and simultaneously learned in a data-driven way.

Multi-output regression has recently generated increasing interest and been used to solve conventional problems, *e.g.*, camera pose estimation in [13, 24], which substantially outperforms previous approaches based on an inverse problem. To avoid tedious segmentation, cardiac bi-ventricular volume estimation has also been formulated as a multi-output regression problem, which achieved much better results [38] than conventional segmentation based methods. Moreover, it is also shown in recent work that exploring the multivariate output space can largely improve regression performance. By incorporating the geometric structure of the output manifold into the regression process, Liu et al. [19] provided a novel mechanism named local linear transform (LLT) to redefine the loss functions, which largely improves the performance of support vector regression (SVR) [26]. Sohn et al. [27] proposed a new approach for multi-output regression that can jointly learn both the output structure and regression coefficients via using inverse-covariance regularization. Rai et al. [23] presented a multiple-output regression model that simultaneously uses the covariance

structure of the latent model parameters and the conditional covariance structure of the observed outputs. Both of these methods have shown improved performance for their applications.

Most of the existing multi-output regression tasks use handcrafted descriptors [24, 15] which are lack of sufficient discrimination for accurate multivariate estimation. Regression targets are explored for specific regressors [19]. Moreover, supervised descriptor learning algorithms are mainly developed for classification tasks, which are not directly applicable to regression due to the continuous multivariate targets rather than the discrete class labels [4]. In this work, we focus on supervised descriptor learning for multi-output regression by exploring the multivariate targets to achieve discriminative and compact feature representations.

3. Supervised Descriptor Learning

Our supervised descriptor learning (SDL) algorithm is to learn low-rank approximations of matrices from oriented gradients for image representations. The gradient orientation matrices (GOM) are first constructed from pyramid histogram of gradients (PHOG) for images, which takes advantages of prior knowledge to capture the key image characteristics: spatial layout and edges of local shapes. The GOMs of training samples associated with targets are fed into the proposed SDL which is formulated as generalized low-rank approximations of matrices with a supervised manifold regularization (SMR). By integrating the proposed SMR to explore the manifold of the target space, the SDL incorporates the supervision of targets to achieve discriminative feature representations. As a consequence, the SDL is able to effectively find low-rank approximations of matrices [34] to obtain compact and discriminative feature representations for efficient and accurate multivariate estimation.

3.1. Preliminaries

Given a set of annotated training data $\{X_1, \ldots, X_L\}$ and the corresponding multivariate targets $\{Y_1, \ldots, Y_L\}$, where L is the number of training samples and $Y_i \in \mathbb{R}^d$, our task is to learn discriminative and compact representations of matrices. Instead of using a vectorized input space, we consider matrix representations, *i.e.*, $X_i \in \mathbb{R}^{M \times N}$, which could be any matrix representations of images, *e.g.*, raw pixel intensities. We propose using gradient orientations matrices (GOM) because of its physically-meaningful representation of images by capturing spatial layout and orientation structures. We will find a discriminative low-rank representation of each X_i by differentially exploring the spatial and orientation information in the GOM. The obtained low-rank representations of matrices are then vectorized as our final descriptors.

3.2. Generalized Low-Rank Approximation

We propose using the generalized low-rank approximation of matrices due to its efficient computation of dimension reduction of matrices [34]. This is to find two-side 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 the following optimization problem of minimizing the reconstruction errors:

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

where $\|\cdot\|_F$ is the Frobenius norm of a matrix, I_m is an identity matrix of size $m \times m$ and the constraints $W^{\mathrm{T}}W = I_m$ and $V^{\mathrm{T}}V = I_n$ ensure that W and V have orthogonal columns to avoid redundancy in the approximations.

According to (1), we know that D_i is the low-rank approximation of X_i in terms of the transformations of W and V, and it is worth to mention that the matrices $D_1, \ldots, D_L$ are not required to be diagonal. D_i is the compact representation of X_i which will reduce regression complexity for efficient multivariate estimation. The objective function in (1) only minimize the approximation errors in the low-rank space leading to indiscriminate representations $\{D_i\}_{i=1}^L$. To be more discriminative, we consider incorporating regression targets for supervised learning to achieve discriminative low-rank representations.

3.3. Supervised Manifold Regularization

In order to to achieve discriminative representations, we propose a supervised manifold regularization (SMR) to explore the manifold of the multivariate target space for supervised descriptor learning.

We impose discrimination on the low-rank representation $\{D_i\}_{i=1}^L$ by integrating the proposed SMR into (1). To this end, we first construct a weighted graph $G = (V, E)$ using the ϵ-neighborhood method [16], namely, nodes Y_i and Y_j are connected if $||Y_i - Y_j||^2 < \epsilon$, where $\epsilon \in R$. V and E respectively represent L vertices and edges between vertices. The graph is built on the multivariate targets $(Y_1, \ldots, Y_L)$ rather than on inputs in conventional manifold regularization [3, 36], which naturally induces the supervision of regression targets.

We denote $S \in \mathbb{R}^{L \times L}$ as the symmetric similarity matrix with non-negative elements corresponding to the edge weights of the graph G, where each element S_{ij} is computed by a heat kernel with a parameter: σ:

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

We set the diagonal elements of S to be zeros, *i.e.*, $S_{ii} =$

0. In the low-rank space, we would like to minimize the following term

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

Since the similarity matrix S characterizes the manifold structure of the multivariate target space, low-rank approximations $\{D_i\}_{i=1}^L$ preserve the intrinsic local geometrical structure of the target space and are therefore automatically aligned to their regression targets. The discrimination is then naturally injected into the low-rank representations $\{D_i\}_{i=1}^L$. An intuitive consequence of minimizing the regularization term is that in the low-dimensional space, data points with similar targets are forced to be close while these with different targets tend to be far apart, which therefore increases the discriminative ability of new representations.

3.4. Descriptor Learning with SMR

Combining (3) and (1), we obtain the compact objective function of GLRAM with the proposed SMR in terms of W, V and $\{D_i\}_{i=1}^L$ as follows:

$$\begin{aligned}\underset{\substack{W,V,D_1,\ldots,D_L \\ W^{\mathrm{T}}W=I_m, V^{\mathrm{T}}V=I_n}}{\arg\min} & \underbrace{\frac{1}{L}\sum_{i=1}^{L} \|X_i - WD_iV^{\mathrm{T}}\|_F^2}_{\text{Low-rank approximation errors}} \\ & + \underbrace{\beta \sum_{i,j} \|D_i - D_j\|_F^2 S_{ij}}_{\text{Supervised manifold regularization}}\end{aligned} \tag{4}$$

where $\beta \in (0, \infty)$ is a tuning parameter to balance the tradeoff between approximation errors and discrimination of the low-rank approximations, which also serves to keep the flexibility of the model.

In the objective function of (4), the first term guarantees the reconstruction fidelity in the low-rank approximation while the second SMR term introduces the discrimination to learned new representations. The SDL takes advantages of the strengths of the GLRAM in dimension reduction of matrices and the SMR in supervised manifold learning, and provides an effective and compact formulation to efficiently learn low-dimensional but highly discriminative feature representations.

3.5. Alternate Optimization

The objective function in (4) can not be solved straightforwardly using existing methods since we have to solve for the projections: W, V and the low-rank approximations: $\{D_i\}_{i=1}^L$ simultaneously. We seek an alternative objective function which can be efficiently solved by an iterative algorithm via an alternate optimization. To this end, we can rewrite the low-rank approximation error term in the objective function (4) in term of traces of matrices as

$$\begin{aligned}&\frac{1}{L}\sum_{i=1}^{L} \|X_i - WD_iV^{\mathrm{T}}\|_F^2 \\ =&\frac{1}{L}\sum_{i=1}^{L} \mathrm{Tr}((X_i - WD_iV^{\mathrm{T}})^{\mathrm{T}}(X_i - WD_iV^{\mathrm{T}})) \\ =&\frac{1}{L}\sum_{i=1}^{L} (\mathrm{Tr}(X_i^{\mathrm{T}}X_i) - \mathrm{Tr}(VD_i^{\mathrm{T}}W^{\mathrm{T}}X_i) \\ &- \mathrm{Tr}(X_i^{\mathrm{T}}WD_iV^{\mathrm{T}}) + \mathrm{Tr}(D_i^{\mathrm{T}}D_i)).\end{aligned} \tag{5}$$

By the fact that $\mathrm{Tr}(Z) = \mathrm{Tr}(Z^{\mathrm{T}})$, we know

$$\mathrm{Tr}(X_i^{\mathrm{T}}WD_iV^{\mathrm{T}}) = \mathrm{Tr}(VD_i^{\mathrm{T}}W^{\mathrm{T}}X_i). \tag{6}$$

The first term: $\sum_{i=1}^L \|X_i\|_F^2$ in (5) is a constant given the data $\{X_i\}_{i=i}^L$, and therefore the minimization of (5) is equivalent to minimizing

$$\sum_{i=1}^{L} (\mathrm{Tr}(D_i^{\mathrm{T}}D_i) - 2\,\mathrm{Tr}(VD_i^{\mathrm{T}}W^{\mathrm{T}}X_i)) \tag{7}$$

Setting the derivatives of (7) w.r.t. D_i to be 0, we have

$$D_i = W^{\mathrm{T}}X_iV, \quad i = 1, \ldots, L. \tag{8}$$

Thus, given the W and V, for any i, D_i is uniquely determined by $D_i = W^{\mathrm{T}}X_iV$ which is the compact representation of X_i in the low-rank space.

By substituting (8) into (5) and dropping the constant $\sum_{i=1}^L \|X_i\|_F^2$, the minimization of (5) is equivalent to the following maximization problem:

$$\underset{\substack{W,V \\ W^{\mathrm{T}}W=I_m, V^{\mathrm{T}}V=I_n}}{\arg\max} \frac{1}{L}\sum_{i=1}^{L} \|W^{\mathrm{T}}X_iV\|_F^2 \tag{9}$$

Therefore, by changing the sign of the second term in original objective function (4), we have an alternative objective function as follows:

$$\begin{aligned}\underset{\substack{W,V \\ W^{\mathrm{T}}W=I_m, V^{\mathrm{T}}V=I_n}}{\arg\max} & \frac{1}{L}\sum_{i=1}^{L} \|W^{\mathrm{T}}X_iV\|_F^2 \\ & - \beta \sum_{i,j} \|W^{\mathrm{T}}(X_i - X_j)V\|_F^2 S_{ij},\end{aligned} \tag{10}$$

which is our final optimization problem to solve, and has no closed-form solution. To seek the optimal solutions of W and V iteratively, we employ an alternate optimization procedure to compute one variable by fixing the other. In

other words, we optimize W by fixing V, and we optimize V by fixing W.

We rewrite the objective function in (10) in terms of traces of matrices as

$$\begin{aligned}\underset{\substack{W,V\\ W^{\mathrm{T}}W=I_m, V^{\mathrm{T}}V=I_n}}{\arg\max} \quad & \frac{1}{L}\operatorname{Tr}(\sum_{i=1}^{L} W^{\mathrm{T}}X_i V V^{\mathrm{T}} X_i^{\mathrm{T}} W) \\ & -\beta \operatorname{Tr}(\sum_{i,j} W^{\mathrm{T}}(X_i - X_j) V S_{ij} V^{\mathrm{T}} (X_i - X_j)^{\mathrm{T}} W).\end{aligned} \tag{11}$$

The above formulation naturally avoids the rank deficit issue in a trace ratio form in which the estimation would be unstable and overly sensitive to the sample in hand with less training samples [17].

Given a V, to find an optimal W is to solve

$$\underset{W}{\arg\max} \operatorname{Tr}(W^{\mathrm{T}}AW), \text{ s.t. } W^{\mathrm{T}}W = I_m. \tag{12}$$

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

$$\begin{aligned}A = & \frac{1}{L}\sum_{i=1}^{L} X_i V V^{\mathrm{T}} X_i^{\mathrm{T}} \\ & - \beta \sum_{i,j} (X_i - X_j) V S_{ij} V^{\mathrm{T}} (X_i - X_j)^{\mathrm{T}}.\end{aligned} \tag{13}$$

Similarly, given a W, the solution of $V \in R^{N\times n}$ can be found by solving

$$\underset{V}{\arg\max} \operatorname{Tr}(V^{\mathrm{T}}BV), \text{ s.t. } V^{\mathrm{T}}V = I_n, \tag{14}$$

and consists of the n eigenvectors of B with the n largest eigenvalues, where B is computed by

$$\begin{aligned}B = & \frac{1}{L}\sum_{i=1}^{L} X_i^{\mathrm{T}} W W^{\mathrm{T}} X_i \\ & - \beta \sum_{i,j} (X_i - X_j)^{\mathrm{T}} W S_{ij} W^{\mathrm{T}} (X_i - X_j).\end{aligned} \tag{15}$$

The optimal solutions of W and V are obtained by iteratively solving the optimization problems in (12) and (14), respectively. Both are the standard eigen decomposition problem and can be efficiently solved by the singular value decomposition (SVD) which is used in our algorithm due to the fact that the truncated SVD achieves the best approximation with respect to the Frobenius norm for given matrices [34, 36].

4. Experiments and Results

We demonstrate the effectiveness of the proposed SDL on a typical multi-output regression task: head pose estimation. Automatic head pose estimation from images is challenging due to illumination, facial expression and inter-subject variations, etc.

4.1. The Pointing'04 Dataset

The Pointing'04 dataset [11] is a widely used benchmark for head pose estimation which is a typical multi-output regression problem. The dataset is challenging and contains 2790 images of 15 subjects, and each subject has two series of 93 images with different head poses represented by *yaw* and *pitch*, namely, each image with a two-dimensional target. Bounding boxes associated with images indicating the head regions are provided with the dataset. We crop the images with the bounding boxes and resize them into 64×64 pixels. To benchmark with existing methods [15, 9, 14] on the this dataset, we employ two validation protocols, *i.e.*, even training/test split and five-fold cross validations.

4.2. Experimental Settings

The gradient orientation matrix (GOM) is constructed from oriented gradients with different spatial and orientation divisions. To fully capture sufficient spatial information, we use a three-level pyramid HOG (PHOG) of cell sizes: 8×8, 16×16 and 32×32 in 9 signed and unsigned orientations obtaining a matrix of size 84×31 from an image of 64×64 pixels [8].

To show the advantage of our SDL algorithm, we have also compared with widely-used descriptors, *e.g.*, GIST [21] and histogram of LBP [32] both of which are implemented with a similar spatial pyramid to the PHOG descriptor, and dimensionality reduction methods, *e.g.*, generalized principal component analysis (GPCA) [35] and principal component analysis (PCA). Note that our final descriptor (800d) learned by the SDL is of much lower dimensionality than the compared descriptors (LBP: 4872d and GIST: 4096d), which will dramatically reduce the computational complexity in regression.

To leverage the strength of random forests for regression tasks [38], we use the adaptive K-cluster regression forests (AKRF) recently proposed in [15] as for multivariate estimation. The AKRF has shown large advantages over other regressors, *e.g.*, support vector regression (SVR), traditional random forests and kernel partial least squares [15]. We use the same experimental settings as in [15] to establish fair comparisons. The free parameter β that keeps the tradeoff between reconstruction fidelity and discrimination in the approximated low-rank space can be obtained by cross validation in the training stage. The performance of head pose estimation is measured by the commonly-used mean absolute error (MAE).

Table 1. The comparison results for head pose estimation on the Pointing'04 dataset using even training/test split. The SDL achieves an improvement of 27.5%.

Methods	yaw	pitch	average
SDL	**4.58**	**3.03**	**3.81**
PHOG (Baseline)	5.35	4.23	4.79
GPCA [35]	5.28	3.60	4.44
PCA	5.46	4.43	4.94
LBP [22]	5.36	4.49	4.92
GIST [21]	6.21	5.30	5.76
AKRF [15]	5.49	4.18	4.83

Table 2. The comparison results for head pose estimation on the Pointing'04 dataset using five-fold cross validation. The SDL achieves an improvement of 22.3%.

Methods	yaw	pitch	average
SDL	**4.12**	**2.09**	**3.11**
PHOG (Baseline)	5.30	3.34	4.32
GPCA[35]	5.11	3.13	4.12
PCA	5.25	3.53	4.39
LBP [22]	5.53	3.37	4.45
GIST [21]	6.06	5.03	5.55
AKRF [15]	5.50	3.41	4.46
Geng et al. [10]	4.24	2.69	3.47
Fenzi et al. [9]	5.94	6.73	6.34
Haji [14]	6.56	6.61	6.59

4.3. Head Pose Estimation

The proposed SDL produces high estimation accuracy for both *yaw* and *pitch* and significantly outperforms the state-of-the-art algorithm in [15] with a large reduction of the MAE up to 27.5% (*pitch*) as shown in Table 1 using even training/test split validation. Furthermore, the SDL also produces consistently better results than the state-of-the-art methods using the same five-fold cross validation with a reduction of MAE up to 22.3% (*pitch*) as shown in Table 2. The overall results using five-fold cross validation are better than those using even training/test split validation in Table 1 due to that more samples are available for training.

The strength of the proposed SDL is further shown by comparing with the handcrafted PHOG, LBP and GIST descriptors and the unsupervised dimensionality reduction techniques: GPCA (800d) and PCA (800d). The performance improvement over the PHOG descriptor indicates the advantage of the induced learning in the proposed SDL algorithm. The effectiveness of the supervised learning, *i.e.*, the proposed supervised manifold regularization (SMR), is shown by superb performance over GPCA without supervision. Moreover, the significant improvement over PCA applied to the PHOG indicates the advantages of our gradient orientation matrix (GOM) representation based on which descriptors are learned.

We look into the results of the SDL by the visualization in a low-dimensional space as illustrated in Fig. 1. Even with only two dimensions (a: $m = 2$ and $n = 1$ and b: $m = 1$ and $n = 2$), the learned descriptor demonstrates highly discriminative ability. As a result, data points of head poses with similar orientations are clustered while those with very different orientations tend to be scattered faraway, which is due to the supervision of the regression targets incorporated by the supervised manifold regularization (SMR). The SDL can effectively extract the most discriminative features closely related to regression targets, which makes the data points discriminatively aligned according to their targets even in a very low-dimensional space. Interestingly, Fig. 1 (a) and (b) exhibit different clustering patterns, which demonstrates the different roles of orientation and spatial features in the obtained image representations. The different effects of spatial cells and orientation bins on the clustering patterns validate the use of the GOM rather than the HOG vector in which spatial and orientation information is equally treated for feature representations. The SDL allows to effectively investigate the different physical meanings of orientation and spatial layout for more useful image representations.

Moreover, we have also investigated the performance of the proposed SDL with varied dimensionality to show the effectiveness in low dimensions. Since the dimensionality reduction is induced by both W and V, we test one by keeping the other fixed. Fig. 2 (a) and (b) are the results of head pose estimation using the even training/split validation. The SDL reaches the best results for both head pose with very low dimensions, which shows the effectiveness of the SDL in learning compact but discriminative descriptors. We use $m = 40$ and $n = 20$ in all our experiments for our final results.

In addition, Fig. 2 also shows the advantages of differentially exploring spatial and orientation information of images by using the GOM rather than the vectorized HOG descriptor. The performance demonstrates different variation patterns head pose with the changes of W and V. The performance is more affected by orientation bins than the spatial bins, which would be due to that the orientation information is more characteristic in representing head poses. This finding again validates individual investigation of oriental and spatial information for image representations.

5. Conclusion

In this paper, we have presented a novel, general supervised descriptor learning (SDL) algorithm for multi-output regression. The proposed SDL algorithm can obtain a com-

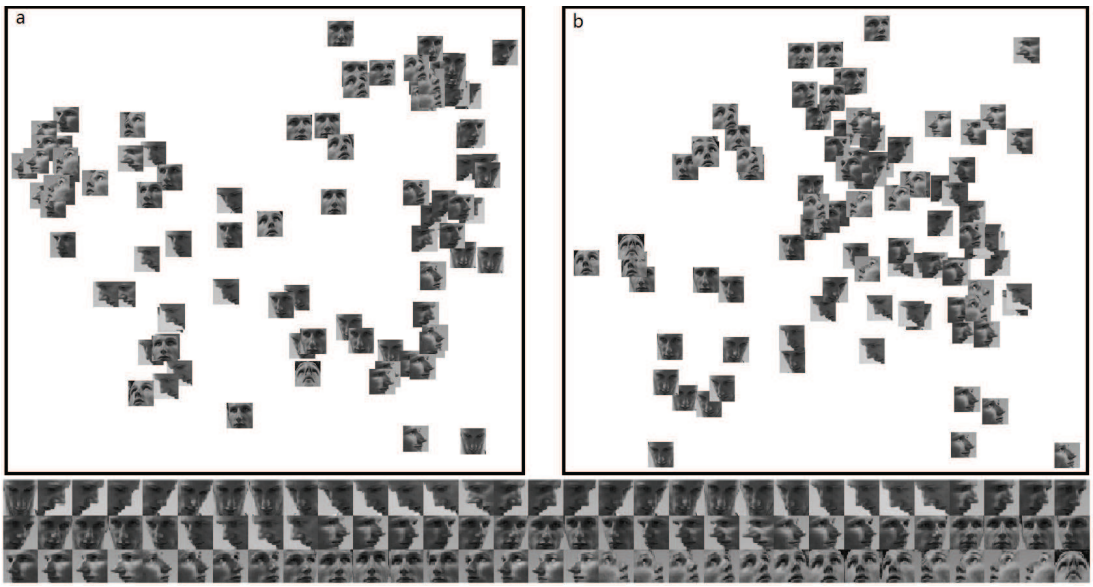

Figure 1. Illustration of head pose images in a two-dimensional space (a: $m = 2$ and $n = 1$ and b: $m = 1$ and $n = 2$). Head poses with similar orientations tend to be clustered while these with distinctive orientations are scattered away. The visualization is implemented using the method provided in [31]. The bottom are the images with all 93 different pose orientations.

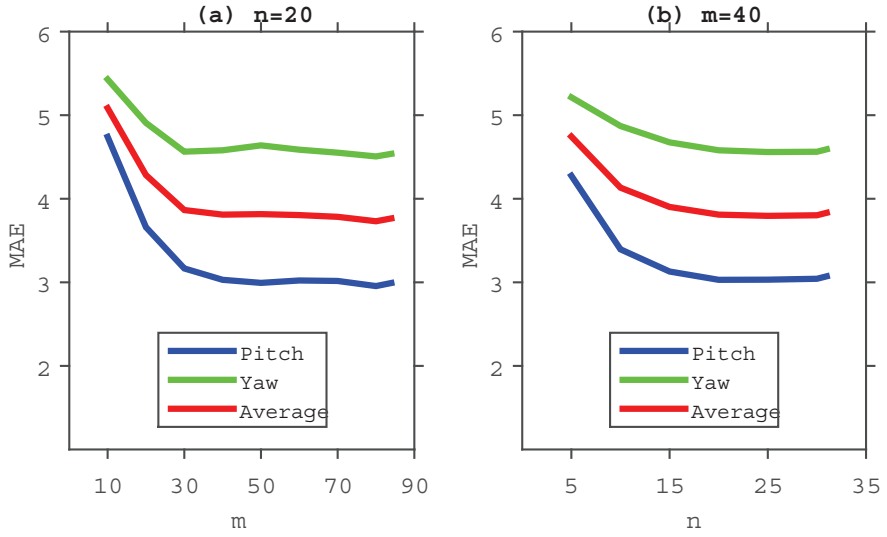

Figure 2. The estimation errors with different dimensions for head pose estimation in (a) and (b) using the even train/test split validation. m is the reduced dimensionality by W and n is the reduced dimensionality by V.

pact and highly discriminative feature representation, which enables more accurate and efficient multivariate estimation. Based on a matrix representation of images, the SDL is formulated as a generalized low-rank approximation of matrices with a supervised manifold regularization, which achieves a compact while effective dimensionality reduction algorithm. The SDL offers a general supervised descriptor learning framework which can be widely built on image intensity, handcrafted features and representations by deep learning algorithms. The SDL can not only boost the performance of existing multi-output regression tasks, but also allows to conduct more efficiently. Experimental results on a representative multivariate estimation task: head pose estimation using the Pointing'04 dataset demonstrate the effectiveness of the SDL for image representations in multi-output regression.

References

[1] M. Afshin, I. Ben Ayed, K. Punithakumar, M. Law, A. Islam, A. Goela, T. Peters, and S. Li. Regional assessment of cardiac left ventricular myocardial function via MRI statistical features. *IEEE TMI*.

[2] E. Ahmed, G. Shakhnarovich, and S. Maji. Knowing a good hog filter when you see it: Efficient selection of filters for detection. In *ECCV*. 2014.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006.

[4] C. BenAbdelkader. Robust head pose estimation using supervised manifold learning. In *ECCV*, pages 518–531, 2010.

[5] L. Bo, X. Ren, and D. Fox. Kernel descriptors for visual recognition. In *NIPS*, pages 244–252, 2010.

[6] H. Cai, K. Mikolajczyk, and J. Matas. Learning linear discriminant projections for dimensionality reduction of image descriptors. *IEEE TPAMI*, 33(2):338–352, 2011.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, volume 1, pages 886–893, 2005.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010.

[9] M. Fenzi, L. Leal-Taixé, B. Rosenhahn, and J. Ostermann. Class generative models based on feature regression for pose estimation of object categories. In *IEEE CVPR*, pages 755–762, 2013.

[10] X. Geng and Y. Xia. Head pose estimation based on multivariate label distribution. In *IEEE CVPR*, pages 1837–1842, 2014.

[11] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial structures. In *ICPR, International Workshop on Visual Observation of Deictic Gestures*, pages 1–9, 2004.

[12] Y. Guo, G. Zhao, and M. Pietikäinen. Discriminative features for texture description. *Pattern Recognition*, 45(10):3834–3843, 2012.

[13] A. Guzman-Rivera, P. Kohli, B. Glocker, J. Shotton, T. Sharp, A. Fitzgibbon, and S. Izadi. Multi-output learning for camera relocalization. In *IEEE CVPR*, pages 1114–1121, 2014.

[14] M. A. Haj, J. Gonzalez, and L. S. Davis. On partial least squares in head pose estimation: How to simultaneously deal with misalignment. In *IEEE CVPR*, pages 2602–2609, 2012.

[15] B. Hara and R. Chellappa. Growing regression forests by classification: Applications to object pose estimation. In *ECCV*. 2014.

[16] X. He and P. Niyogi. Locality preserving projections. In *NIPS*, volume 16, page 153, 2004.

[17] N. Karampatziakis and P. Mineiro. Discriminative features via generalized eigenvectors. *ACM ICML*, 2014.

[18] Z. Lei, M. Pietikäinen, and S. Z. Li. Learning discriminant face descriptor. *IEEE TPAMI*, 36(2):289–302, 2014.

[19] G. Liu, Z. Lin, and Y. Yu. Multi-output regression on the output manifold. *PR*, 42(11):2737–2743, 2009.

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

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

[22] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen. *Computer vision using local binary patterns*, volume 40. Springer, 2011.

[23] P. Rai, A. Kumar, and H. Daume. Simultaneously leveraging output and task structures for multiple-output regression. In *NIPS*, pages 3185–3193, 2012.

[24] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *IEEE CVPR*, pages 2930–2937, 2013.

[25] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning local feature descriptors using convex optimisation. *IEEE TPAMI*, 2(4), 2014.

[26] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[27] K.-A. Sohn and S. Kim. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *AISTATS*, pages 1081–1089, 2012.

[28] M. Torki and A. Elgammal. Regression from local features for viewpoint and pose estimation. In *IEEE ICCV*, pages 2603–2610, 2011.

[29] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. *IEEE CVPR*, 2014.

[30] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Subspace learning from image gradient orientations. *IEEE TPAMI*, 34(12):2454–2466, 2012.

[31] L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *JMLR*, 9(2579-2605):85, 2008.

[32] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

[33] P. Wang, J. Wang, G. Zeng, W. Xu, H. Zha, and S. Li. Supervised kernel descriptors for visual recognition. In *IEEE CVPR*, pages 2858–2865, 2013.

[34] J. Ye. Generalized low rank approximations of matrices. *Machine Learning*, 61(1-3):167–191, 2005.

[35] J. Ye, R. Janardan, and Q. Li. GPCA: an efficient dimension reduction scheme for image compression and retrieval. In *ACM SIGKDD*, pages 354–363, 2004.

[36] Z. Zhang and K. Zhao. Low-rank matrix approximation with manifold regularization. *IEEE TPAMI*, 35(7):1717–1729, 2013.

[37] X. Zhen, L. Shao, and F. Zheng. Discriminative embedding via image-to-class distances. In *BMVC*, 2014.

[38] X. Zhen, Z. Wang, A. Islam, I. Chan, and S. Li. Direct estimation of cardiac bi-ventricular volumes with regression forests. In *MICCAI*. 2014.