Hypergraph Regularized Autoencoder for 3D Human Pose Recovery

Chaoqun Hong, Jun Yu(✉), You Jane, and Xuhui Chen

School of Computer and Information Engineering,
Xiamen University of Technology, 361024 Xiamen, China
yujun@hdu.edu.cn

Abstract. Image-based human pose recovery is usually conducted by retrieving relevant poses with image features. However, semantic gap exists for current feature extractors, which limits recovery performance. In this paper, we propose a novel feature extractor with deep learning. It is based on denoising autoencoder and improves traditional methods by adopting locality preserved restriction. To impose this restriction, we introduce manifold regularization with hypergraph Laplacian. Hypergraph Laplacian matrix is constructed with patch alignment framework. In this way, an automatic feature extractor for silhouettes is achieved. Experimental results on two datasets show that the recovery error has been reduced by 10% to 20%, which demonstrates the effectiveness of the proposed method.

Keywords: Human pose recovery · Deep learning · Manifold regularization · Hypergraph · Patch alignment framework

1 Introduction

3D human pose recovery tries to generate a visually pleasing and semantically correct human skeleton with sensor data. Traditionally, it is usually achieved by motion capture system. These systems are expensive and require attached markers. Markerless solutions currently draw plenty of attention. The most successful of them is Microsoft Kinect. It makes use of RGBD data. However, RGBD cameras are still not commonly used. In this way, researchers devote themselves into image-based 3D human pose recovery over recent years as the demand for high-quality and accurate poses in vision systems has increased [9][7].

Typical routine of image-based pose recovery rely on the same three-step procedure: 1) extracting the visual features from the 2D images (usually silhouettes); 2) mapping the 2D visual features to the 3D poses using a specified learning algorithm; 3) reconstructing the 3D poses based on the mapping function obtained. Well-designed feature should be discriminative with respect to 2D images and 3D poses. Until now, quite a lot of features have been proposed for human pose analysis, such as shape context[2], histograms of oriented gradients[4], Hierarchical centroid[12] and so on. However, feature descriptors are still ambiguous due

H. Zha et al. (Eds.): CCCV 2015, Part I, CCIS 546, pp. 66–75, 2015.
DOI: 10.1007/978-3-662-48558-3_7

to the so-called semantic gap between images and features, since they cannot completely represent the semantic content and information of images.

Deep learning architectures [11] have been useful for exploring hidden representations in natural images and have proven success in a variety of vision tasks. In the current big data era, the extensive availability of training images enables deep models to be generic and flexible. Inspired by the learning capability and capacity of the deep learning model, we hypothesized that deep architectures would be perfectly suited to seeking the proper representations for 2D images and 3D poses and modelling their relationship. Current solutions generally learn multilevel representations by deep learning [5]. For instance, autoencoder [13] is an unsupervised feature-learning scheme in which the internal layer acts as a generic extractor of inner image representations. A double-layer structure, which can efficiently map the input data onto appropriate outputs, is obtained by using a multilayer perceptron. However, in these methods, the locality of features is lost. This makes similar pose silhouettes being described by totally different hidden vectors, and unstable performance in pose reconstruction. In order to solve this problem, one possible solution is to add an additional locality-preserving term to the formulation of deep learning [14].

In this paper, a novel approach is proposed to recover 3D human poses from silhouettes with hypergraph regularized autoencoder (HRA). It is based on marginalized denoising autoencoders (MDA) [3]. Different from previous works, it makes use of locality information of samples. The main contribution of this work is two-fold:

- The state-of-the-art work in pose recovery with autoencoders is improved by imposing locality preserved restriction. To impose this restriction, an Laplacian matrix is constructed to describe the internal relationship of samples.
- The construction of Laplacian matrix is further improved by using hypergraph. This process is based on a real-valued form of combinatorial optimization problem. The weights of hyperedges for the whole alignment are computed by statistics of distances between neighboring pairs.

The remainder of this paper is organized as follows. The proposed hypergraph regularized autoencoder is presented in Section 2. Then, experimental results on human pose recovery and comparisons with other state-of-the-art methods are presented in Section 3. Finally, we conclude the paper in Section 4.

2 Hypergraph Regularized Autoencoder

2.1 Marginalized Denoising Autoencoders

In denoising autoencoders, inputs $x_1, ..., x_n$ are corrupted by random feature removal. $\hat{x}_i$ is denoted as the corrupted version of x_i and $W : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is denoted as the mapping of reconstructing the corrupted inputs. In this way, we can defined the squared reconstruction loss as:

$$\frac{1}{2n} \sum_{i=1}^{n} \parallel x_i - W\hat{x}_i \parallel^2 . \tag{1}$$

The solution to (1) depends on which features of each input are randomly corrupted. To lower the variance, MDA [3] perform multiple passes over the training set, each time with different corruption. In this way, the overall squared loss is defined as:

$$\frac{1}{2mn}\sum_{j=1}^{m}\sum_{i=1}^{n} \parallel x_i - W\hat{x}_{i,j} \parallel^2, \tag{2}$$

where $\hat{x}_{i,j}$ represents the jth corrupted version of the original input x_i and m is the number of layers.

For the matrix form, $X = [x_1, ..., x_n] \in \mathbb{R}^{d\times n}$ is denoted as the data matrix, $\overline{X} = [X, .., , X]$ is denoted as the m-times repeated version and $\hat{X}$ is defined as the corrupted version of $\overline{X}$. Then, the loss in (2) is reduced to:

$$\frac{1}{2mn} tr[(\overline{X} - W\hat{X})^T(\overline{X} - W\hat{X})]. \tag{3}$$

The minimization to (3) can be expressed as the well-known closed-form solution for ordinary least squares:

$$W = PQ^{-1} \text{ with } Q = \hat{X}\hat{X}^T \text{ and } P = \overline{X}\hat{X}. \tag{4}$$

In our implementation, we further reduce the stacked form of MDA to a overlapped form. Instead of concatenating the output of each layer, we simply use the output of each layer as the input of the next layer. In this way, (2) can be rewritten as:

$$\frac{1}{2n}\sum_{i=1}^{n} \parallel x_i - W\hat{x}_{i,m} \parallel^2 . \tag{5}$$

(3) can be also rewritten as:

$$\frac{1}{2n} tr[(\overline{X}_m - W\hat{X})^T(\overline{X}_m - W\hat{X})]. \tag{6}$$

2.2 Manifold Regularization

As mentioned before, due to the loss of locality information, similar features can be encoded as totally different hidden representation, which may bring about the loss of the locality information of the features to be encoded. To preserve such locality information, we introduce manifold regularization to (5). Then, the reconstruction loss can be defined by:

$$\frac{1}{2n}(\sum_{i=1}^{n} \parallel x_i - W\hat{x}_{i,m} \parallel^2 + \alpha \sum_{i,k}^{n} \parallel x_i - x_k \parallel^2 \omega_{i,k}), \tag{7}$$

where α indicates the weights of locality reservation term and $\omega_{i,k}$ represents the similarity between sample i and sample k. With the introduction of locality reservation term, the matrix form can be defined as:

$$\frac{1}{2n}tr[(\overline{X}_m - W\hat{X})^T(\overline{X}_m - W\hat{X}) + \alpha\overline{X}_m^T L\overline{X}_m], \tag{8}$$

where L is known as Laplacian matrix.

2.3 Hypergraph Optimization

The key to solve (8) is to construct Laplacian matrix L. Traditional methods assumed that the relationships among images are pairwise. However, this assumption is over-simplified and a lot of information is lost. To avoid this problem, hypergraph representation is proposed [15]. Different from traditional graph-based representation, one edge is able to connect more than two vertices in hypergraph representation. In other words, vertices connected by an edge are thought as a subset of vertices in the graph. Therefore, hypergraph representation is much more descriptive and powerful. The definitions are shown in Table 1.

Table 1. Definition of symbols in the hypergraph.

Symbol	Definition
u, v	Vertices in the hypergraph
e	Edges in the hypergraph
$\omega(e)$	The weight of an edge e
$\delta(e)$	The degree of an edge, e. It illustrates how many vertices are connected by e. In traditional graph representation, $\delta(e) = 2$.
$d(v)$	The degree of a vertex, v. It is calculated by summing the weighting values of edges connected to this vertex.
D_v	The diagonal matrix containing the vertex degrees
D_e	The diagonal matrix containing the edge degrees
H	In this matrix, $H(v, e) = 1$ if $v \in e$
Ω	The diagonal matrix containing the weights of hyperedges
V	The set of vertices
E	The set of edges

In our method, we construct hypergraph Laplacian matrix inspired by patch alignment framework [6], which consists of two steps.

1. Part Optimization: We define one patch to be the vertices connected by one hyperedge. Thus, the patch in the proposed learning process is defined by:

$$\arg\min_{f \in R^{|V|}} \sum_{m,n \subset e} \frac{w(e)}{\delta(e)} (\frac{y_m}{\sqrt{d_m}} - \frac{y_n}{\sqrt{d_n}})^2 \tag{9}$$

For one patch, we should compute:

$$\sum_{m,n \subset e} \frac{w(e)}{\delta(e)} (\frac{y_m}{\sqrt{d_m}} - \frac{y_n}{\sqrt{d_n}})^2, \tag{10}$$

which means that we randomly choose two vertices in the subset of vertices contained by a hyperedge, e, and sum the value of

$$\frac{w(e)}{\delta(e)}(\frac{y_m}{\sqrt{d_m}} - \frac{y_n}{\sqrt{d_n}})^2. \tag{11}$$

Expanding (9) and combining items, we can get the patch optimization for each hyperedge:

$$\frac{1}{2}\sum_{v\subset e}\frac{F}{DV_v^{\frac{1}{2}}}EH'_e\frac{\Omega}{DE}H_eE'\frac{F}{DV_v^{\frac{1}{2}}}. \tag{12}$$

Matrix E is

$$\begin{bmatrix} -\boldsymbol{e}^T \\ I \end{bmatrix} \tag{13}$$

where $\boldsymbol{e} = [1, ..., 1]^T$, I is an $n \times n$ identity matrix.

2. Whole alignment: In the hypergraph, the weight of a hyperedge is computed by summing the similarity scores of all the pairs of vertices contained in this hyperedge. The similarity score of any pair of vertices is defined as the distance of image features:

$$S(u, v) = exp(-\frac{1}{\sigma}dist(feat(u), feat(v))), \tag{14}$$

where $feat(u)$ represents the image feature vector of vertex u, $dist(x, y)$ is usually set to be the L2 distance and σ is the standard deviation of all distances. With the hyper edge weighting matrix, the multi-view hypergraph Laplacian can be computed by summing the patch optimization defined in (11) of all the hyperedges:

$$\frac{1}{2}\sum_{e\in E}\sum_{v\in e}\frac{F}{DV_v^{\frac{1}{2}}}EH'_e\frac{\Omega}{DE}H_eE'\frac{F}{DV_v^{\frac{1}{2}}}. \tag{15}$$

One hyperedge is defined to contain one sample and its k nearest neighbors. In this way, the computational complexity of hypergraph-based manifold regularization can be divided into two parts. The first part is finding nearest neighbors with Euclidean distances, which is $O(k \times n \times d^2)$. The second part is computing Laplacian matrix, which is n^2. The introduction of manifold regularization may reduce the speed of extracting features.

3 Experimental Evaluation

3.1 Datasets and Settings

In our experiments, we use two datasets to evaluate the performance and emphasize the advantage of the proposed HRA.

The first dataset is that used in [1]. In this dataset, a person is walking in a spiral pattern, and we name this dataset Walking. The training data consists of all the pose vectors taken from sequences 01-07. All sequences are concatenated to give 1691 training pose vectors. Sequence 08 is used for testing, and contains 418 testing poses. Mocap data are retrieved for a 54 degrees of freedom body model, with three angles for each of 18 joints, including body orientation with respect to the camera. For evaluation, the mean RMS absolute difference errors between the true and estimated joint angle vectors are reported in degrees:

$$d_{degree}(y, y^r) = \frac{1}{M} \sum_{i=1}^{M} |(y - y_i^r) \bmod \pm 180^\circ)|, \tag{16}$$

where y is the ground truth, y^r is the recovered degree, $M = 54$ is the number of degrees and $(\bullet) \bmod \pm 180^\circ)(\bullet + 180^\circ) \bmod 360^\circ - 180^\circ$ reduces angles to the interval $[-180^\circ, +180^\circ]$. The training silhouettes are created by using POSER to render the Mocap poses.

The second dataset is the HumanEva-I dataset, which is widely used in evaluating the performance of pose recovery [10]. This dataset contains five motion types performed by four subjects. A 3D pose is encoded as a collection of joint coordinates in 3D space and there are 14 joints in the HumanEva data set, therefore each 3D action data is represented by a $14 \times 3 = 42$-dimensional feature vector. For evaluation, Trial 1 of Subjects 1 and 2 is used. Since there are many invalid motions in Mocap data, we collect all the valid frames. The frames in Trial 1 of Subject 1 are used as the training set. The number of frames is 701. The frames in Trial 1 of Subject 1 are used as the testing set. The number of frames is 604. For evaluation, the retrieval error is computed. The distance between two poses is then calculated as the average Euclidean distance between corresponding joint markers:

$$d_{pose}(y, y^r) = \frac{1}{M} \sum_{i=1}^{M} \parallel m_i(y) - m_i(y^r) \parallel, \tag{17}$$

where $\parallel \bullet \parallel$ computes the 3D distance between two markers which are represented by 3D coordinates:

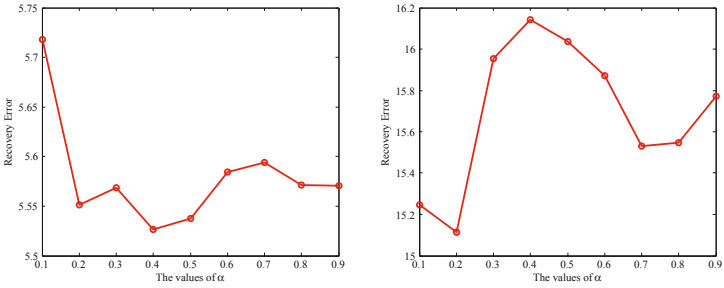
$$\parallel m_i(y) - m_i(y^r) \parallel = \sum_{j=1}^{3} \parallel m_i(y_j) - m_i(y_j^r) \parallel . \tag{18}$$

All the images are resized to be $128 \times 128 = 16384$ for fairness of comparison. With silhouettes features, we get recovered poses by relevance vector machine [1].

3.2 Optimization of Autoencoders

When we adopt manifold regularization in (7), parameter α is introduced to balance reconstruction loss and locality loss. We show the performance with different settings of α to look into its influence. The curve is shown in Fig. 1.

We can figure out that the proposed method performs the best when $\alpha = 0.4$ for Walking dataset while it performs the best when $\alpha = 0.2$ for HumanEva-I dataset.



(a) Results of Walking (b) Results of HumanEva

Fig. 1. The influence of α

3.3 Comparison with State-of-the-Arts

In this subsection, we compare the proposed HRA with the following existing features.

- Hierarchical centroid (HC) [12]: The number of dimensions is $n \times \sum_{i=1}^{n}(2 \times 2^{2\times(i-1)})$ in which n is the level of the computed centroids. In the experiments, $n = 3$.
- Histograms of oriented gradients (HOG) [4]. The number of HOG windows per bound box is 3×3 and the number of histogram bins is 9. Each image is thus represented by a feature vector with 81 dimensions.
- Shape context (SC) [2]: 200 points are sampled for each silhouette image. In each image, the histogram of shape contexts is constructed from 12 angular bins and 5 radial bins. To match the shape context features efficiently, the shape signature histograms known as shapemes, proposed by Mori et al. [8] are applied. Theoretically speaking, a shapeme is the bag-of-feature form of shape contexts. A codebook containing 200 codes is trained, giving the shapeme a dimensionality of 200.
- Marginalized Denoising Autoencoder (MDA) [3]: MDA approximates the expected loss function of traditional denoising autoencoders with its Taylor expansion. The dimensionality of hidden representation is the same as the size of original images.

To compare the performance of these methods, we show the average recovery error of each pose component. The Walking poses consist of a body model with 54 degrees of freedom, so the average recovery error for each degree is shown.

Table 2. The performance on each marker of Walking.

index	HC	HOG	SC	MDA	HRA	index	HC	HOG	SC	MDA	HRA
1	4.20	3.74	4.05	3.91	3.66	28	4.69	3.57	4.52	4.37	3.93
2	3.78	2.84	2.79	1.89	1.49	29	12.20	12.76	10.29	10.90	10.07
3	31.38	23.37	22.51	11.83	12.02	30	7.87	8.15	7.74	7.24	6.46
4	3.48	2.83	3.19	3.60	3.09	31	9.06	10.10	9.77	8.91	8.26
5	2.72	2.86	2.75	2.90	2.71	32	4.97	4.08	4.71	4.83	4.38
6	3.31	3.69	3.67	3.24	2.84	33	13.18	12.16	11.06	10.94	10.04
7	3.42	3.17	3.53	3.76	3.45	34	4.61	4.23	4.03	4.52	3.83
8	2.24	1.92	2.68	2.67	2.23	35	3.00	3.44	3.01	3.11	2.77
9	4.72	3.81	5.06	5.03	4.58	36	0.84	1.14	1.23	1.12	0.73
10	9.20	8.80	9.44	9.30	8.99	37	13.62	13.56	10.74	11.08	10.68
11	4.46	4.03	4.23	3.92	3.35	38	4.44	4.14	4.10	4.35	3.85
12	4.19	3.38	4.30	4.67	4.40	39	4.69	4.03	3.80	4.11	3.89
13	2.01	1.61	2.06	2.36	1.95	40	16.39	14.95	13.49	13.35	12.80
14	0.03	0.33	0.59	0.45	0.05	41	5.92	5.65	5.28	5.41	4.88
15	0.53	0.68	0.91	0.85	0.44	42	8.07	8.14	8.04	7.71	7.67
16	3.93	4.04	4.37	3.79	3.38	43	10.44	10.53	8.93	9.60	9.54
17	11.45	11.07	10.14	9.53	9.72	44	12.42	12.63	12.40	12.51	11.44
18	9.43	8.23	8.09	6.75	6.64	45	10.06	9.80	9.50	10.10	9.05
19	8.91	8.41	7.85	8.12	7.51	46	11.22	10.88	9.74	8.94	8.34
20	1.76	1.64	1.96	1.85	1.40	47	4.71	4.54	4.19	3.84	3.64
21	11.96	10.57	10.77	12.19	11.34	48	7.29	6.46	6.44	6.88	6.65
22	3.08	2.71	2.92	3.14	2.69	49	16.34	16.68	15.77	15.44	14.80
23	2.06	2.29	2.31	2.29	1.90	50	3.87	3.77	3.92	3.65	3.23
24	0.26	0.54	0.72	0.64	0.25	51	18.83	17.07	17.31	17.40	15.51
25	1.06	1.21	1.44	1.43	0.98	52	10.56	9.94	9.24	10.67	10.05
26	2.32	2.14	2.32	2.08	1.57	53	1.57	1.76	2.03	1.95	1.52
27	0.34	0.58	0.77	0.66	0.24	54	8.81	8.97	8.85	8.04	7.50
AVG	6.78	6.36	6.21	6.00	5.53						

The HumanEva-I poses consist of 14 3D joint coordinates, so the average recovery error for each joint position is shown. The results are shown in Table 2 and Table 3. Average recovery errors for all the items are also shown at the end of tables. In each row, the smallest error is highlighted. Of the 52 items in Table 2, HRA performs the best in 37 items (68.52%). Of the 14 items in Table 3, HRA performs the best in 12 items (85.71%). This illustrates that HRA outperforms the other

methods in most cases. Its average performance is also the best. Thanks to the descriptive power of autoencoders, HRA works well for joints that are easily occluded such as hands. Further more, HRA usually outperforms MDA due to the introduction of manifold regularization.

Table 3. The performance on each marker of HumanEva-I.

index	HC	HOG	SC	MDA	HRA	index	HC	HOG	SC	MDA	HRA
1	18.68	24.65	19.32	17.76	17.62	8	26.43	24.35	19.08	18.97	17.73
2	14.20	23.75	18.60	22.18	20.93	9	18.22	25.78	20.22	12.40	9.57
3	14.89	25.23	19.79	9.73	9.21	10	25.87	23.03	18.02	18.52	15.99
4	25.33	24.17	18.93	18.09	16.55	11	26.43	24.38	19.11	18.96	18.77
5	26.03	23.33	18.26	22.64	21.05	12	18.18	21.06	16.45	12.36	7.09
6	27.12	17.55	13.64	19.51	5.79	13	18.62	23.03	18.02	16.71	15.99
7	25.71	24.19	18.95	18.39	16.51	14	26.29	24.38	19.11	18.85	18.78
AVG	22.29	23.49	18.39	17.51	15.11						

Some recovery results are shown in Fig 2. Due to the limitation of paper space, we only show the results of Walking. We can see that the proposed HRA gives recovered poses more close to the original images.

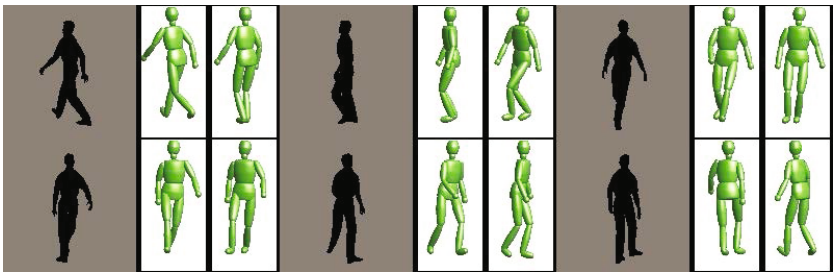


Fig. 2. Recovery results of Walking. For each set of images, the first image is the original image, the second image is the result of HRA and the third image is the result of MDA

4 Conclusion

In this paper, a novel approach of 3D pose recovery with 2D silhouettes is proposed. It improves the previous approach of image feature extractors with denoising autoencoders by introducing locality sensitive constriction. Locality reservation is able to keep the mutual dependency in the encoding procedure and makes

similar silhouettes from the same pose grouped together. Hypergraph regularization with patch alignment framework is adopted to impose locality reservation, which improves the descriptive power of autoencoders and reduces ambiguity of extracted features. Experimental results on both Walking and HumanEva-I datasets show that the proposed method outperforms previous method on recovery performance.

References

1. Agarwal, A., Triggs, B.: Recovering 3D human pose from monocular images. IEEE Trans. Pattern Anal. Mach. Intell. **28**(1), 44–58 (2006)
2. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **24**(4), 509–522 (2002)
3. Chen, M., Weinberger, K.Q., Sha, F., Bengio, Y.: Marginalized denoising autoencoders for nonlinear representations. In: IEEE International Conference on Machine Learning, pp. 1476–1484. IEEE (2014)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 886–893. IEEE Press (2005)
5. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computing **18**(7), 1527–1554 (2006)
6. Hong, C., Yu, J., Li, J., Chen, X.: Multi-view hypergraph learning by patch alignment framework. Neurocomputing **118**(22), 79–86 (2013)
7. Hong, C., Yu, J., Tao, D., Wang, M.: Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval. IEEE Transactions on Industrial Electronics **62**(6), 3742–3751 (2015)
8. Mori, G., Belongie, S., Malik, J.: Efficient shape matching using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **27**(11), 1832–1837 (2005)
9. Shen, J., Liu, G., Chen, J., Fang, Y., Xie, J., Yu, Y., Yan, S.: Unified structured learning for simultaneous human pose estimation and garment attribute classification. IEEE Transactions on Image Processing **23**(11), 4786–4798 (2014)
10. Sigal, L., Balan, A.O., Black, M.J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision **87**(1–2), 4–27 (2010)
11. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1958 (2014)
12. Yang, M., Qiu, G., Huang, J., Elliman, D.: Near-duplicate image recognition and content-based image retrieval using adaptive hierarchical geometric centroids. In: Proceedings of the IEEE International Conference on Pattern Recognition, pp. 958–961. IEEE Press (2006)
13. Yoshua, B.: Learning deep architectures for AI. Foundations and Trends in Machine Learning **2**(1), 1–127 (2009)
14. Yuan, Y., Mou, L., Lu, X.: Scene recognition by manifold regularized deep learning architecture. IEEE Transactions on Neural Networks and Learning Systems (2015)
15. Zhou, D., Huang, J., Scholkopf, B.: Learning with hypergraphs: clustering, classification, and embedding. In: Advances in Neural Information Processing Systems. vol. 19, pp. 1601–1608. MIT Press (September 2007)