2014 IEEE International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)
October 18-19, 2014, Wuhan, Hubei, China

Integrating Joint and Surface for Human Action Recognition in Indoor Environments

Qingyang Li, Yu Zhou, Anlong Ming
School of Computer Science
Beijing University of Posts and Telecommunications, Beijing, 100876
Email:liqingyang9512@gmail.com, yuzhou@bupt.edu.cn, mal@bupt.edu.cn

***Abstract*—Action recognition has a long research history, despite several contributed approaches have been introduced, it remains a challenging task in computer vision. In this paper, we present a uniform fusion framework for action recognition, which integrates not only the local depth cues but also the global depth cues. Firstly, the action recognition task is formulated as the maximize the posterior probability, and then the observation for the original action is decomposed into the sub-observations for each individual feature representation strategy of the original action. For the local depth cues, the joints inside the human skeleton is employed to model the local variation of the human motion. In addition, the normal of the depth surface is utilized as the global cue to capture the holistic structure of the human motion. Rather than using the original feature directly, the support vector machine model learning both the discriminative local cue (i.e., the joint) and the discriminative global cue (i.e., the depth surface), respectively. The presented approach is validated on the famous MSR Daily Activity 3D Dataset. And the experimental results demonstrate that our fusion approach can outperform the baseline approaches.**

I. Introduction

Recognizing the human activity in the cluttered indoor environments is a critical issue in computer vision. It has several practical daily life applications, e.g. the human-robot interfaces [1], video surveillance, *et al*. Moreover, other applications of computer vision can also benefit from the accuracy of the human action recognition, e.g., video object tracking [2], [3] event detection and recognition. Although several contributed research has been introduced in the literatures nowadays, it remains a challenging task. The critical issues include the intra-class variations, e.g., human pose variation, deformation, self-occlusion, *et al* and the extra-class noises, e.g., different action may have similar appearance in practice.

The early action recognition approaches mainly work on the color videos, e.g., [4], *et al*. In these approaches, the invariant key points are frequently employed as the local feature to capture the action of the target. However, the information supplied by the color video is frequently insufficient to recognize the action of the human accurately in practice.

The research reported in this paper was supported in part by the National Key Technology Research and Development Program of the Ministry of Science and Technology of China under Grant No. 2014BAK14B03; the Fundamental Research Funds for the Central Universities under Grant No. 2013PT13, 2013XZ1.2

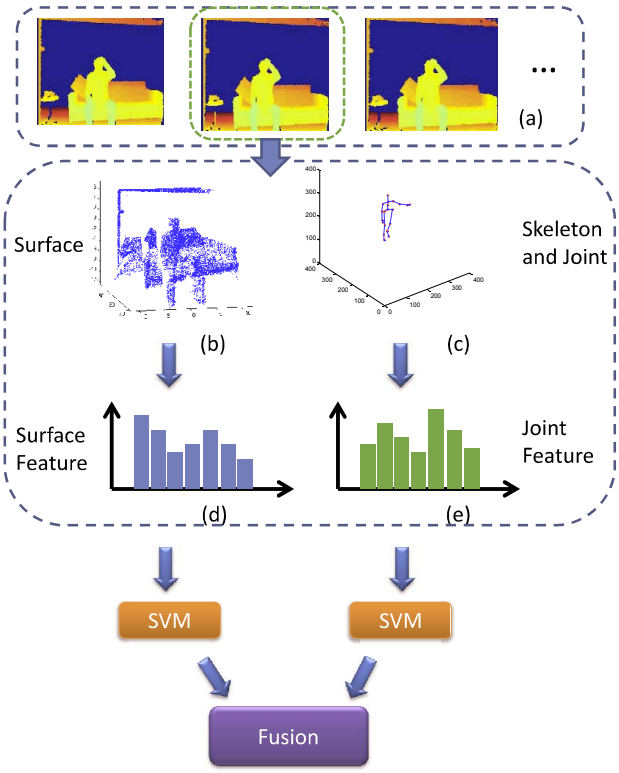

Fig. 1. Flow chart of our approach. (a) The input depth video sequence; (b) The surface of a video frame; (c) The skeleton of a target; (d) Feature representation for the surface; (e) Feature representation for the joints.

In recent years, the cost-effective depth cameras, e.g., the Kinect RGB-D sensor [5], have attracted much attention.Such camera can provide the 3D depth cues of the scene. And hence the activity recognition can naturedly benefit from such depth information.

In the depth based action recognition, since the powerful 3D joint position of the human skeleton can be easily obtained [6], the features based on the 3D joints of the skeleton are frequently employed to capture the invariant characteristics of the human, e.g., [7] utilized the spatial distance between

978-1-4799-5353-0/14/$31.00 ©2014 IEEE

pairwise joints and the local occupancy pattern feature as the feature representations. Since the skeleton based feature can exploit rough structural information of the target, it can naturally address the non-rigid deformation caused by the free human motion.

In addition, the dense 3D point cloud captured by the Kinect sensor can supply the accurate depth information of the human. Hence the geometrical surface is also informative for recognition. The surface can be interpreted as the 3D contour of the human, which supplies sufficient shape information.

However, as we observed, these two kinds of features which we mentioned above are frequently discussed separately in the recent literatures. As mentioned in [8], the skeleton cue and the contour cue are complementary to each other for shape classification. Consequently, in this paper, we propose a novel fusion framework to integrate the joint feature of the skeleton and the surface feature. The flow chart of our approach is shown in Fig. 1, (a) is the input depth sequence captured by the Kinect sensor, (b) is the surface of the scene, and (c) is the skeleton (blue line) of the human captured by [6], the red dots are the joints. In (d) and (e), the quantized features are shown. We collect the joint features and the surface features for the whole sequences, and train the SVM model respectively. And then we fuse them into a uniform framework based on a novel posteriori decomposition. The SVM model can select the most discriminative features, i.e., the support vector, which can be interpreted as the salient action part of the human. Hence our model are informative to reflect the specificality of each human motion. Moreover, since that we fuse both the local and the global cues of the human, our model can naturally address the commonly challenging situations in human activity recognition.

Since we focus on the indoor environment, the presented approach is validated on the famous MSR Daily Activity 3D dataset, and the experimental results demonstrate that our novel fused action recognition approach can obtain promising recognition accuracy when compared with the existing approaches. The rest of the paper is organized as follows: Section 2 reviews the related work. Section 3 describes the problem formulation. In Section 4, we describe the methods to get the features. Section 5 shows the experimental results.Finally,conclusion is given in section 6 of the paper.

II. Related Work

Action recognition has a long research history, a detailed survey can be found in [9]. Since the skeleton and surface feature are utilized in this paper, which are closely related to the 2D shape representation, the detailed survey is referred to in [10].

The skeleton based representations have been well studied in the 2D shape, e.g., [11], [12], [13], [14]. Since the reliable 3D structure (3D skeleton) of the human can be easily obtained by the cost-effective depth sensor [15], [16], it attaches much attention to the human action recognition, e.g., [17] utilized the Hidden Markov model to reflect the transition probability of the 3D joints. [18] employed the conditional random field to model the 3D joint positions. [7] presented a novel joint based image feature to represent the action. In this approach, the pairwise distance between pairs of the joints in 3D space are utilized. And the local occupancy patterns are also employed to reflect the critical properties of the human motion, which counts the number of the points that fall into the cells around the corresponding joint.

The contour based representation are also well discussed in the literatures, e.g., [19], [20], [21], [22], [23], [24], [25], [26]. And in the early research literatures, reliable key point based strategies are frequently utilized, e.g., [27]. And then the motion trajectory based approaches are introduced, e.g., [28], [29]. Furthermore,the holistic approaches are becoming popular recently, e.g., [30] summarizes the whole depth sequences into a motion map, and then the HOG [31] features are extracted from the motion map to reflect the whole sequence. [32] divides the whole depth sequences into several spatiotemporal grids, and the global occupancy patterns are employed to represent the whole sequences.

In contrast to those approaches, our approach mainly focus on presenting a uniform framework to fuse different action cues, e.g., the local cue (joint) and the global cue (surface) in this paper. Based on the posterior decomposition, we can easily obtain the fused confidence for each action class. In addition, our approach does not depend on any specific action representation. The aforementioned approaches can be naturally fused into our framework to further improve the recognition accuracy.

III. Problem Formulation

Given the observation video set $\mathbf{V} = \{V_i | i = 1, 2, ..., N\}$, the goal of action recognition is to infer the class label $l_i \in \{1, 2, ..., L\}$ for each video $V_i \in \mathbf{V}$. In some special cases, a video may contain several human activities. For simplification, we assume that each video contains only one human activity in this paper. Hence, the object of the human action recognition is to compute the class label l^* which maximizes the posteriori probability.

$$l^* = \arg\max p(l|V_i) \tag{1}$$

In practice, several discriminative features can be extracted from the video, e.g.,the skeleton based feature and the surface based feature that are utilized in this paper. Since the features lie in different feature spaces, we assumed that they are conditionally independent on l. Hence the posteriori probability in Eq. (1) can be further factorized into each feature representation. We thus have,

$$p(l|V_i) = \prod_{j=1}^{N} p(l_j|f_{i,j}) \tag{2}$$

where $p(l_j|f_{i,j})$ is the label posterior based on the feature type $f_{i,j}$, N is the total number of the feature types, e.g., $N = 2$ in this paper. Hence, our goal lies in computing the $p(l_j|f_{i,j})$ for each feature type. In the following section, we first introduce the features utilized in this paper, and then present the strategy to obtain the label posterior for each feature type.

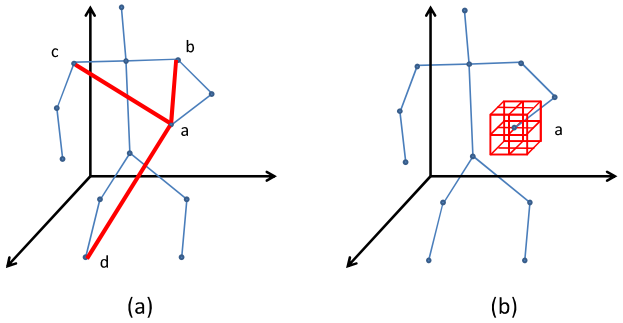

Fig. 2. The local feature utilized in this paper. (a) The feature to measure the pairwise distance, a b, c and d are all the joints and the red solid lines reflect the pairwise distances in spatial space; (b) The local that utilized in this paper, for a joint a, its surrounding cube region is shown in red color.

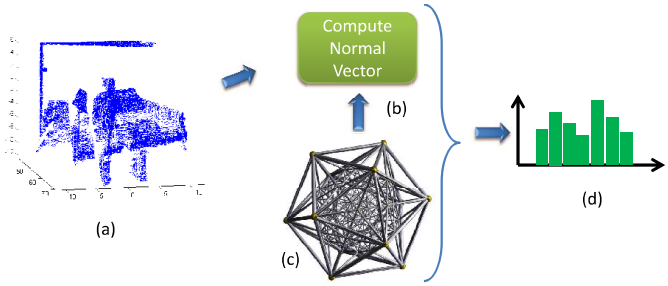

Fig. 3. The holistic feature that is utilized in this paper. (a) The input surface data; (b) The process to compute the normal feature; (c) The visualization of the polychoron; (d) The final holistic action descriptor.

IV. Features

This section gives the detailed description of the two types of features that are utilized to represent the actions: the local feature and the holistic feature. The local feature represents the position information of each joint and the regional information around each joint in the depth map while the holistic feature represents the geometrical surface information of the target.

1) The local feature: For each depth sequence V_i with T frames, in order to address the local variances, ,we obtain the 3D coordinates and the screen coordinates plus depth of the joint i in the frame $t, t \in \{1, ..., T\}$ using the skeleton tracker [6].And the coordinates are normalized.With such normalization, the 3D coordinates remain invariant to the variation of the absolute body position, e.g., the initial body orientation or the body size.

Assume the number of the joints detected by the skeleton tracker is Q (In general, the number of the joints is 20). Each joint $q \in \{1, ..., Q\}$ has three coordinates, i.e., $q = (x_{q,t}, y_{q,t}, z_{q,t})$ We compute the pairwise distance between pairs of the joints, i.e.,

$$f_j(q, g) = |x_{q,t} - x_{g,t}| + |y_{q,t} - y_{g,t}| + |z_{q,t} - z_{g,t}| \quad (3)$$

Hence $f_j(q, g)$ represents l_1 distance between the joint pairs (q, g), where $q \in \{1, ..., Q\}, g \in \{1, ..., Q\}$ and $q \neq g$. Consequently, we can obtain the distance between joint q and the other $Q-1$ joints, and put them into a $1 \times (n-1)$ dimension vector. An example is shown in Fig.2(a), the blue dots are the detected joints on the skeleton, and the red solid lines reflect the l_1 distance between pairs of the joints.

In addition, the 3D point cloud around each joint is employed to accurately distinguish actions with interaction between joints and the objects. For example, in the eating action, interaction exists among the food, the hands and the face. In order to gain such discriminative information, we generate the 3D point cloud from the frame t of the depth sequences, e.g., Fig.1(b), After locating the position of the joint q in the 3D point cloud, we get a cube region with $N_w \times N_h \times N_d$ pixels which originates from the joint q. The cube region is then divided into $w \times h \times d$ cells. Each cell c_e includes $x \times y \times z$ pixels ($x = N_w/w, y = N_h/h, z = N_d/d$). Hence, we count the number N_e of points lie in each cell c_e, i.e.:

$$N_e = \sum I_k \quad (4)$$

For the c_e, if the 3D point cloud has a point in the pixel $k(k \in c_e)$,$I_k = 1$ and $I_k = 0$ otherwise. In the frame t, every corresponding N_e of all the cells of joint$i$ is put into a $w \times h \times d$ dimensions vector $O(t, i)$,which represents regional information of the joint i in the frame t.Then,we let $L(t, i) = (P(t, i), O(t, i))$ denote the local feature of the joint i in the frame t. Finally, we put local features of all the joints in all the frames of an action into the vector L in sequence. $L = \{L(t, i)|i \in [1, n], t \in [1, m], i \in Z, t \in Z\}$ is the local feature for an action sample.

2) The holistic feature: In order to obtain the global information of the action, the distribution of 4D surface normal orientation is employed as the holistic feature [33]. In order to get the holistic feature, firstly, we generate the 4D space from the depth sequences V_i, and then divide the 4D space into $w \times h \times t$ spatiotemporal cells. In each cell of the spatiotenporal space, we compute a set of unit normal $N = \{\hat{\boldsymbol{n}}_j$ as well as quantize the cell using 120 vertices of the polychoron, i.e., $P = \{\boldsymbol{p_i}\}$, which extends from a 2D polygon. We refer to each vertice as a "projector", and denote $\boldsymbol{p_i}$ the 4D coordinate of each "projector".Then,we compute the inner product between each $\hat{\boldsymbol{n}}_j$ and each $\boldsymbol{p_i}$: $c(\hat{\boldsymbol{n}}_j, \boldsymbol{p_i}) = max(0, \hat{\boldsymbol{n}}_j^T \boldsymbol{p_i})$. Consequently, the original uniform distribution of the 4D normal orientation can be obtained as

$$\Pr(\boldsymbol{p_i}|\boldsymbol{N}) = \frac{\sum_{j \in \boldsymbol{N}} \boldsymbol{c}(\hat{\boldsymbol{n}}_j, \boldsymbol{p_i})}{\sum_{\boldsymbol{p_v} \in \boldsymbol{p}} \sum_{j \in \boldsymbol{N}} \boldsymbol{c}(\hat{\boldsymbol{n}}_j, \boldsymbol{p_v})} \quad (5)$$

The original holistic feature is $Ho = (\Pr(\boldsymbol{p_i}|N) \mid \boldsymbol{p_i} \in P)$. However, in order to gain the final holistic feature for training and testing, we have to refine the original holistic feature according to the method provided in [33] because the uniform quantization is not optional. In Fig.3, an example of the holistic feature is shown. (a) is the obtained surface map, we compute the norm vector (b) based on the surface map. (c) is the visualization of the polychoron, and (d) is the final holistic action descriptor.

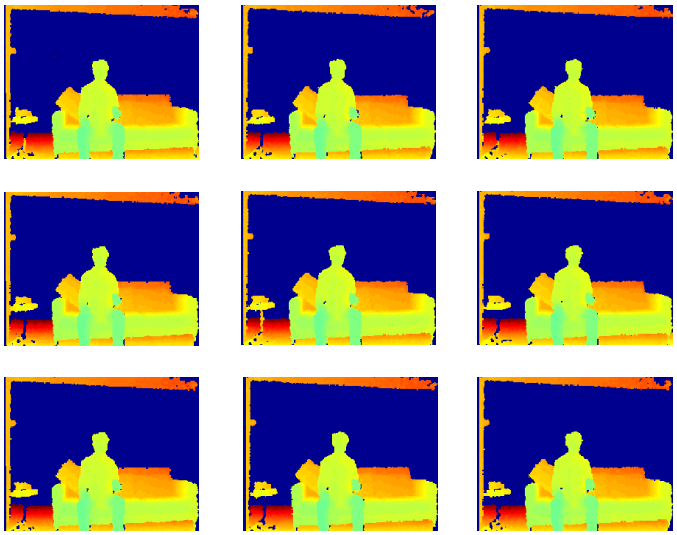

Fig. 4. Example frames from the MSR Daily Activity 3D dataset

A. Classifiers and Learning Algorithm

Let $f_{i,j}$ denote the feature vector described in the previous section, where the subscript i indicates that $f_{i,j}$ represents the video V_i and the subscript j indicates that $f_{i,j}$ is obtained based on the feature type j. Then the posterior introduced in section III can be trained by the support vector machine (SVM) model, i.e.,

$$p(l = c|f_{i,j}) \propto \exp(s(f_{i,j})) \tag{6}$$

where the $s(f_{i,j})$ is obtained by:

$$s(f_{i,j}) = \sum_{i=1}^{l} (\alpha_i K(w_{i,j}, f_{i,j}) + b). \tag{7}$$

x_k is the testing vector, $w_s$and b can be gained through the training process. The kernel $K(w_s, x_k) = w_s' \times x_k$ is linear. $X = x_k$, $l \in X$.

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, ..., l \end{aligned} \tag{8}$$

Eq.(8) can be solved by iteratively optimizing ξ and $\mathbf{w}, b$ the LIBSVM [34] is utilized to obtain the final solution. In all our implementation, all the classifiers are learned in one vs. all way. For instance, when the action class *eating* is regarded as the positive class, then all the other action classes in the training set are treated as the negative class.

V. Experimental Results

In this section, we validate the presented approach on the famous MSR Daily Activity 3D Dataset [1]. It is a set of depth sequences captured by the Kinect sensor, and it contains sixteen action classes, i.e.: *drink*, *eat*, *read book*, *call cellphone*, *write on a paper*, *use laptop*, *use vacuum cleaner*, *cheer up*, *sit still*, *toss paper*, *play game*, *lie down on sofa*, *walk*, *play guitar*, *stand up*, *sit down*. For each action,there are ten subjects to complete it twice. The first accomplishment of the action is in a sitting position, and the second accomplishment of the action is in a standing position. A sofa is set in the scene. The data used in the proposed experiment in this paper are the depth maps, There are 320 action videos all together in the dataset. the resolution of all the depth sequences are 320×240. Example depth video sequences are given in Fig.4.

TABLE I. Recognition Accuracy Comparison for MSR-DailyActivity3D Dataset

Method	Accuracy
Only LOP feature	0.43
Dynamic Temporal Warping	0.54
Random Occupancy Pattern	0.64
Only Joint Position Features	0.68
Actionlet Ensemble on LOP Features	0.61
Actionlet Ensemble on Joint Features	0.74
Our	**0.76**

For the experimental parameters setting, the spatiotemporal cells is set to be $5 \times 4 \times 3$ in width, height, and the number of frames, respectively. In each cell of the depth sequences, we utilize 120 vertices of the 600-cell polychoron to initialize the projectors. The surrounding region of each joint is divided into $12 \times 12 \times 4$ cells. and each cell has $6 \times 6 \times 80$ pixels.

Followed [33], half of the data are utilized for training, while the others are employed as the testing data. The cross-subject setting is employed to quantitative comparison, as shown in Tab. I. By utilizing the fused framework, our approach achieve the recognition accuracy of 76%, and if only the LOP feature is utilized to represent the action, the recognition rate is only 42.5%. If we utilize the joint feature individually, the recognition rate can raise to 68%. When the Dynamic Temporal Warping is employed, the recognition accuracy is 54%. Followed [7], when we combine Actionlet ensemble on the LOP feature and the Joint feature, the recognition accuracy is 61% and 74%, respectively. In contrast, the presented approach can achieve the recognition accuracy by 76%, which demonstrates that the proposed method is very effective.

VI. Conclusion

In this paper, we present a novel framework to integrate not only the local joint feature but also the global depth surface feature. The action recognition task is formulated to maximize the posterior probability, which is further decomposed into the sub-observation for each individual feature representation strategy for the action. Instead of directly utilizing the original feature vector, we train the support vector machine for each action class, which selects the most discriminative feature for each class. The presented approach is validated on the famous MSR daily activity 3D dataset, and the experimental results demonstrate that the presented approach can outperform the baseline approaches.

[1] http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/

REFERENCES

[1] Y. Zhou, Y. Yang, M. Yi, X. Bai, W. Liu, and L. J. Latecki, "Online multiple person detection and tracking from mobile robot in cluttered indoor environments with depth camera," *International Journal of Pattern Recognition and Artificial Intelligence(IJPRAI)*, vol. 28, no. 1, 2014.

[2] Y. Zhou, X. Bai, W. Liu, and L. J. Latecki, "Fusion with diffusion for robust visual tracking," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 2978–2986.

[3] Y. Zhou, C. Rao, Q. Lu, X. Bai, and W. Liu, "Multiple feature fusion for object tracking," in *Proceedings of the Intelligent Science and Intelligent Data Engineering*, 2012, pp. 145–152.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie., "Behavior recognition via sparse spatio-temporal features," in *IEEE International Conference on Computer Vision (ICCV)*, 2005, pp. 1–6.

[5] http://www.xbox.com/en US/Kinect.

[6] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1, 2, 8.

[7] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012, pp. 1, 2, 3, 5, 6,7, 8.

[8] X. Bai, W. Liu, and Z. Tu, "Integrating contour and skeleton for shape classification," in *In ICCV Workshops on Non-Rigid Shape Analysis and Deformable Image Alignment (NORDIA)*, 2009.

[9] W. Li, Z. Zhang, and Z. Liu, "Expandable data-driven graphical modeling of human actions based on salient postures," *IEEE Transactions on Circuits and Systems for Video Technology*, p. 2, 2008.

[10] Y. Zhou, J. Liu, and X. Bai, "Research and perspective on shape matching," *Acta Automatic Sinica*, vol. 38, no. 6, pp. 889–910, 2012.

[11] X. Bai, L. J. Latecki, and W. Liu, "Skeleton pruning by contour partitioning with discrete curve evolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI))*, vol. 29, no. 3, pp. 449–462, 2007.

[12] X. Bai and L. J. Latecki, "Path similarity skeleton graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI))*, vol. 30, no. 7, pp. 1282–1292, 2008.

[13] W. Shen, X. Bai, R. Hu, H. Wang, and L. J. Latecki, "Skeleton growing and pruning with bending potential ratio," *Pattern Recognition (PR)*, vol. 44, no. 2, pp. 196–209, 2011.

[14] X. Bai, X. Wang, L. J. Latecki, and W. L. Z. Tu, "Active skeleton for non-rigid object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[15] W. Shen, K. Deng, X. Bai, T. Leyvand, B. Guo, and Z. Tu, "Exemplar-based human action pose correction," *IEEE Trans. Cybernetics*, vol. 44, no. 7, pp. 1053–1066, 2014.

[16] W. Shen, R. Lei, D. Zeng, and Z. Zhang, "Regularity guaranteed human pose correction," in *ACCV*, 2014.

[17] F. Lv and R. Nevatia, "Recognition and segmentation of 3-d human action using hmm and multi-class adaboost," in *European Conference on Computer Vision (ECCV)*, 2006, p. 2.

[18] L. Han, X. Wu, W. Liang, G. Hou, and Y. Jia, "Discriminative human action recognition in the learned hierarchical manifold space," *Image and Vision Computing*, p. 2, 2010.

[19] Y. Zhou, J. Wang, Q. Zhou, X. Bai, and W. Liu, "Shape matching using co-occurrence pattern," in *Proceedings of the IEEE International Conference on Image and Graphics (ICIG)*, 2011, pp. 344–349.

[20] J. Ma, J. Zhao, Y. Zhou, and J. Tian, "Mismatch removal via coherent spatial mapping," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2012, pp. 1–4.

[21] X. Bai, C. Rao, and X. Wang, "Shape vocabulary: A robust and efficient shape representation for shape matching," *IEEE Transactions on Image Processing (TIP)*, vol. 23, no. 9, pp. 3935–3949, 2014.

[22] J. Ma, J. Zhao, J. Tian, X. Bai, and Z. Tu, "Regularized vector field learning with sparse approximation for mismatch removal," *Pattern Recognition*, vol. 46, no. 12, pp. 3519–3532, 2013.

[23] J. Wang, Y. Zhou, X. Bai, and W. Liu, "Shape matching and recognition using group-wised points," in *Advances in Image and Video Technology*, 2012, pp. 393–404.

[24] J. Ma, J. Zhao, J. Tian, Z. Tu, and A. L. Yuille, "Robust estimation of nonrigid transformation for point set registration," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2147–2154.

[25] J. Ma, J. Zhao, and J. Tian, "Nonrigid image deformation using moving regularized least squares," *IEEE Signal Processing Letters*, vol. 20, no. 10, pp. 988–991, 2013.

[26] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, "Regularized vector field learning with sparse approximation for mismatch removal," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1706–1721, 2014.

[27] I. Laptev, "On space-time interst points," *IJCV*, pp. 1, 2, 6, 2005.

[28] H. Wang, A. Klser, C. Schmid, and C. Liu, "ction recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR)*, 2011, p. 2.

[29] S. Wu, O. Oreifej, and M. Shah, "Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories," in *ICCV*, 2011.

[30] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *ACM Multimedia*, 2012, pp. 2, 3, 6, 7, 8.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 2, 4.

[32] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. M.Campos., "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," in *In 17th Iberoamerican Congress on Pattern Recognition (CIARP)*, 2012, pp. 2, 3, 6, 8.

[33] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *CVPR*, 2013.

[34] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.