論文 / 著書情報
Article / Book Information

Title	Graph Regularized Implicit Pose for 3D Human Action Recognition
Authors	Tommi Kerola, Nakamasa Inoue, Koichi Shinoda
Citation	Proc. APSIPA, , , pp. 155-159
Pub. date	2016, 12
DOI	https://doi.org/10.1109/APSIPA.2016.7820717
URL	http://www.ieee.org/index.html
Copyright	(c)2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.
Note	このファイルは著者（最終）版です。 This file is author (final) version.

Graph Regularized Implicit Pose for 3D Human Action Recognition

Tommi Kerola*†, Nakamasa Inoue* and Koichi Shinoda*
* Tokyo Institute of Technology, Tokyo, Japan
E-mail: {kerola,inoue}@ks.cs.titech.ac.jp, shinoda@cs.titech.ac.jp

***Abstract*—We present a novel feature descriptor for 3D human action recognition using graph signal processing techniques. A linear subspace is learned using graph total variation and graph Tikhonov regularizers, transforming 3D time derivative information into a representation that is robust against noisy skeleton measurements. The graph total variation regularizer learns an action representation that encourages piece-wise constantness, which helps discriminating between different action classes. Graph Tikhonov regularization ensures the searched low-rank subspace is similar to the original feature. Experiments show that our approach learns a good representation of an action due to the explicit graph structure, and achieves a statistically significant improvement over the baseline moving pose method, resulting in a 93.5% accuracy on the challenging MSRAction3D dataset.**

I. Introduction

Thanks to the recent developments in machine learning, we are now able to demand even higher performance for difficult tasks in computer vision than ever before. One such task is human action recognition, which has several applications in robotics, health care, and the security industry.

Action recognition based on RGB images is challenging due to problems such as background clutter and varying lighting. These problems are however easily solved using cheap, affordable depth cameras, such as the Microsoft Kinect [4], which also includes tracking of skeletons [11]. Although noisy, the estimated skeletons are much more discriminative than the raw depth data.

Previous work in 3D human action recognition can usually be divided into three types [3]. Approaches that use tracked skeletons [15], [6], approaches that use the raw depth data [7], [9], and finally methods than use both [13], [14]. While raw depth data is useful for recognition of objects, leveraging skeletons usually results in better recognition performance due to the explicit semantic knowledge of skeleton joint locations. However, the tracked skeletons are noisy, and most existing approaches do not handle this explicitly. Even the approaches that do, tend to use only elementary Gaussian smoothing [15] or low-frequency Fourier components [13] for reducing the noise in the input. This brings us to the motivation of this paper, which is to use graphs to regularize features gotten from the tracked skeletons in order to created an enhanced action descriptor.

The first author acknowledges the Japanese Government (Monbukagakusho:MEXT) scholarship support for carrying out this research.

†T.K. is now at Preferred Networks Inc.

Using graphs for regularizing optimization problems have been previously explored to some extent. Zhang *et al.* [16] proposed to learn a dictionary that is derived from the structure of a graph, while observed data is used for learning the parameters. Their approach does not, however, result in an efficient implementation. Consequently, recent work by Thanou *et al.* [12] presented a dictionary learning algorithm that both incorporates the graph topology and is computationally efficient due to expressibility through Chebyshev polynomials. Adding graph structure to dimensionality reduction using principal component analysis (PCA) has been done by using the graph Laplacian matrix for doing graph Tikhonov regularization [5]. Tikhonov regularization ensures smoothness following the graph; computational efficiency is granted through a closed-form solution. Following this work, Shahid *et al.* [10] created PCA-GTV, with added graph total variation (GTV) regularization, which demonstrated that the regularizer helps learning a subspace that is very robust against noise, and also more discriminative, as it has an automatic grouping effect [10].

In this paper, we illustrate the advantage of graph regularizers for learning a linear subspace embedding suitable for KNN-based action recognition. We improve the established moving pose (MP) descriptor [15] framework, and create an action representation that is more discriminative than the baseline MP descriptor, while the fast running time of the baseline method is retained at test time. The graph regularization approach is shown to perform better than classic PCA.

Our motivations for using graph regularization for action recognition are the following:

- KNN methods are sensitive to the local structure of the data. Our graphs encourage locality, by learning a subspace with a local community structure.
- KNN methods are sensitive to the high dimensionality of the data [1]. We find a low-rank representation that allows data representation using only a few components.
- GTV is quite robust against noise [10], which helps as the Kinect skeletons are often erroneous due to tracking failures.
- Our enhanced method is just a matrix multiplication at test time, which keeps the running time of the action classification method low.

The rest of this paper is organized as follows. Section II introduces our proposed graph-regularized framework for action recognition. Section III discusses how to construct the graphs

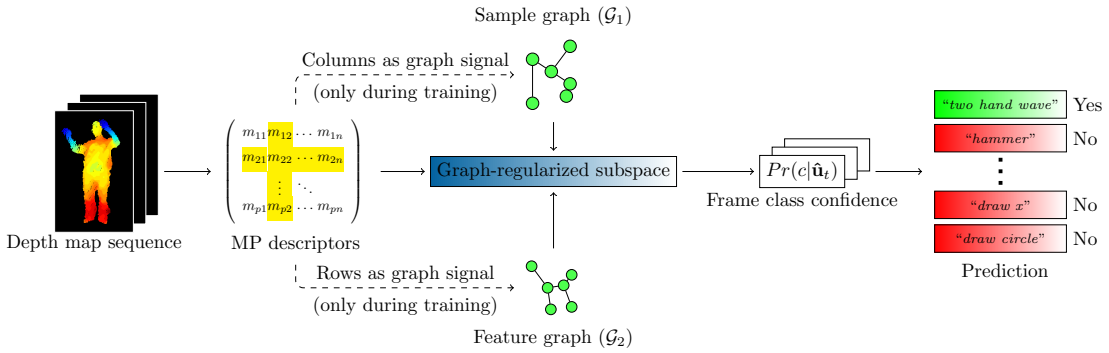

Figure 1. Overview of the proposed action recognition system. The sample and feature graphs regularize the subspace embedding of the moving pose (MP) descriptors, which leads to improved performance. The graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are only used for learning the subspace and are not used during test time.

used in our method. Finally, experimental results are shown in Sec. IV, and Sec. V concludes the paper.

II. Graph Regularized Implicit Pose (GRIP)

In this section, we present GRIP, our proposed method for human action recognition. We propose a framework that learns a linear subspace suitable for fast KNN-classifier recognition. Our approach uses the moving pose descriptor [15], and then performs dimensionality reduction with graph regularizers [10] for learning our subspace. An overview of the proposed system is shown in Fig. 1. In the following, we describe the steps of our method.

A. Moving Pose Descriptor

Given a depth video of a human action containing T frames, 20 tracked skeleton joints [11] are extracted from each frame t. The limb lengths are normalized as in [15], while keeping joint angles intact. Let $\mathbf{p}_{j,t} \in \mathbb{R}^3$ denote the 3D position of the j-th joint in frame t. First, the position of the center hip joint is subtracted from all the other joints, and then the position of each joint is convolved with a 5×1 Gaussian kernel. Next, first- and second-order derivatives $\partial \mathbf{p}_{j,t} \approx \mathbf{p}_{j,t+1} - \mathbf{p}_{j,t-1}$ and $\partial^2 \mathbf{p}_{j,t} \approx \mathbf{p}_{j,t+2} + \mathbf{p}_{j,t-2} - 2\mathbf{p}_{j,t}$ are calculated. In order to gain invariance against speed-variations, the derivatives are ℓ_2-normalized as $\partial \tilde{\mathbf{p}}_{j,t} = \partial \mathbf{p}_{j,t} / \sqrt{\sum_{q=1}^{20} \|\partial \mathbf{p}_{q,t}\|^2}$, and similarly for the second-order derivative. The moving pose (MP) descriptor [15] for frame t is then the vector $\mathbf{m}_t = [\mathbf{p}_{1,t}; \ldots; \mathbf{p}_{20,t}; \alpha\partial\tilde{\mathbf{p}}_{1,t}; \ldots; \alpha\partial\tilde{\mathbf{p}}_{20,t}; \beta\partial^2\tilde{\mathbf{p}}_{1,t}; \ldots; \beta\partial^2\tilde{\mathbf{p}}_{20,t}]$, where $\alpha = 0.75$ and $\beta = 0.6$ are parameters controlling the relative importance of the derivatives.

B. Subspace Transformation

Classification of human actions is challenging due to intra-class and inter-subject variations, along with skeleton noise. Therefore, to ease classification, we propose to transform each MP descriptor $\mathbf{m}_t$ from the original feature space $\mathbb{R}^p$ into a low-rank representation $\mathbf{u}_t$ with smaller variability that implicitly represents the original pose.

Assume we have a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, with vertex set $\mathcal{V}$, edge set $\mathcal{E}$ and weight matrix $\mathbf{W}$, where $W(i,j)$ stores the weights of edge $(v_i, v_j) \in \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. The degree vector $\mathbf{d} = \mathbf{W}\mathbf{1}$ stores the degrees of the vertices. We define a graph signal on $\mathcal{G}$ as a function $f : \mathcal{V} \to \mathbb{R}^Z$ that maps a vector to each vertex. Such a signal can be represented by the matrix $\mathbf{F} = [\mathbf{f}_1^\mathsf{T}; \ldots; \mathbf{f}_{|\mathcal{V}|}^\mathsf{T}] \in \mathbb{R}^{|\mathcal{V}| \times Z}$. The graph gradient of $\mathbf{F}$ with respect to $\mathcal{G}$ is defined as

$$(\nabla_{\mathcal{G}} \mathbf{F})(e) = \sqrt{\mathrm{W}(i,j)} \left(\frac{\mathbf{f}_i}{\sqrt{d(i)}} - \frac{\mathbf{f}_j}{\sqrt{d(j)}} \right) , \qquad (1)$$

which captures the weighted difference of the signal at an edge $e = (v_i, v_j)$. It can be represented as a linear operator using the matrix $\nabla_{\mathcal{G}} \in \mathbb{R}^{|\mathcal{E}| \times |\mathcal{V}|}$, which sparsely factorizes the well-known graph Laplacian matrix $\boldsymbol{\mathcal{L}} = \nabla_{\mathcal{G}}^\mathsf{T} \nabla_{\mathcal{G}}$.

Next, we define a graph $\mathcal{G}_1$ where each vertex is an MP descriptor $\mathbf{m}_t$, and connect vertices that should have the same representation, once variability is removed. This means that for $\mathcal{G}_1$, the graph signal is $\mathbf{F}_{\mathcal{G}_1} = \mathbf{U}^\mathsf{T}$, where $\mathbf{U}$ contains the hypothesis for $\mathbf{u}_t$ on each column, $\forall t$.

Further, if two features, *e.g.* $\mathbf{p}_{1,t}$ and $\mathbf{p}_{2,t}$ are similar in the original MP descriptor space, then we wish this to hold for $\mathbf{U}$ as well. In order to preserve such feature similarities, we define a graph $\mathcal{G}_2$ that connects individual features *i.e.* $\mathrm{m}_t(i)$ to each other. The graph signal for $\mathcal{G}_2$ then becomes $\mathbf{F}_{\mathcal{G}_2} = \mathbf{U}$.

For learning our representation space, we turn to the PCA-GTV framework [10], which is a dimensionality reduction method using graph regularizers. We create a data matrix $\mathbf{M} \in \mathbb{R}^{p \times n}$ by stacking all p-dimensional MP descriptors along the columns. Therefore, n is equal to the total number of frames in the training set. The prior assumptions modeled by $\mathcal{G}_1$ and $\mathcal{G}_2$ are encouraged by solving

$$\arg\min_{\mathbf{U}} \ \|\mathbf{M} - \mathbf{U}\|_1 + \underbrace{\gamma_1 \|\nabla_{\mathcal{G}_1} \mathbf{U}^\mathsf{T}\|_1}_{\text{GTV}} + \underbrace{\gamma_2 \|\nabla_{\mathcal{G}_2} \mathbf{U}\|_F^2}_{\text{GTIK}} . \qquad (2)$$

Here, the graph total variation (GTV) regularizer encourages the signal to become piece-wise constant, *i.e.*, connected MP descriptors become equal, thus minimizing variation. The graph Tikhonov (GTIK) regularizer, on the other hand, encourages our assumption of feature similarity, preserving the sense of closeness in the learned subspace. The construction process of $\mathcal{G}_1$ and $\mathcal{G}_2$ is deferred to Sec. III. The optimization problem (2) is convex, and can be efficiently solved using the forward-backward primal dual algorithm [2].

As (2) only shrinks the singular values of $\mathbf{U}$, the linear subspace $\hat{\mathbf{V}}$ is gotten by the singular value decomposition $\mathbf{U} = \mathbf{V}\mathbf{\Sigma}\mathbf{Q}^{\mathsf{T}}$, where we only keep the columns of $\mathbf{V}$ corresponding to dimensions d with singular value $\Sigma(d,d) > \tau\|\mathbf{\Sigma}\|_\infty$ to get $\hat{\mathbf{V}}$, where $\tau = 0.1$ is a parameter. We then get a subspace embedding of the MP descriptors $\mathbf{M}$ by $\hat{\mathbf{U}} = \hat{\mathbf{V}}^{\mathsf{T}}\mathbf{M}$.

C. Frame Class Confidence and Classification

As we have argued, the transformed features $\hat{\mathbf{u}}$ in the matrix $\hat{\mathbf{U}}$ should be well-suited for a KNN classifier. Similar to previous research [15], we use a frame-based KNN approach for action classification. The membership confidence of class c for each frame descriptor $\hat{\mathbf{u}}$ in a depth video Ξ_r, is modeled by

$$Pr(c|\hat{\mathbf{u}}) \approx \frac{n_c}{n} \ , \tag{3}$$

where n_c is the number of nearest neighbors of $\hat{\mathbf{u}}$ having class c and n is the total number of neighbors. The approximation holds well in practice if the number of nearest neighbors is large [15]. We ascertain this by creating for each Ξ_r the set of nearest neighbors $\{\hat{\mathbf{u}}_i | \hat{\mathbf{u}}_i \in \mathcal{N}_{\hat{\mathbf{u}}}^{(2,q)}, \forall q \neq r, \forall \hat{\mathbf{u}} \in \Xi_r\}$, where $\mathcal{N}_{\hat{\mathbf{u}}}^{(2,q)}$ is the set of the two nearest neighbors of $\hat{\mathbf{u}}$ in the q-th training sequence.

Action classification of a test sequence $\{\hat{\mathbf{u}}_{1,\text{test}}, \ldots, \hat{\mathbf{u}}_{T,\text{test}}\}$ with T frames is then done by the frame-based KNN approach

$$c^\star = \arg\max_c \sum_{t=1}^{T} \sum_{\hat{\mathbf{u}}_i \in \mathcal{N}_{\hat{\mathbf{u}}_{t,\text{test}}}^{(K,\text{all})}} Pr(c|\hat{\mathbf{u}}_i) \ , \tag{4}$$

where $c^\star$ is the predicted action class.

III. Graph Construction

A. KNN Graphs

KNN graphs are created by connecting each vertex to its $K = 10$ nearest neighbors. The graph aims to capture the similarity structure between the descriptors. Weights are set by the Euclidean distance kernel $\mathrm{W}(i,j) = \exp(-\mathrm{D}(i,j)/\sigma^2)$ if $(i,j) \in \mathcal{E}$, where $\mathrm{D}(i,j) = \|\mathbf{m}_i - \mathbf{m}_j\|_2^2$, and $\sigma = 1$ controls the severeness of the regularization penalty induced by the edges of the graph, and is set to unity due to the input skeletons being normalized.

B. Confidence Graphs

Confidence graphs make use of the class membership probability (3), along with the ground-truth class label y_i for each descriptor $\mathbf{m}_i$ in each frame. The graph aims to create a sparse community structure as given by the training set labels, in which the similarities inside the communities are expressed by the edge weights. Descriptors from sequences of the same class are connected and weights are set by

$$\mathrm{W}(i,j) = \begin{cases} c_i c_j \exp\left(-\frac{\mathrm{D}(i,j)}{\sigma_{y_i}^2}\right) & \text{if } y_i = y_j \wedge \mathrm{D}(i,j) \leq \frac{1}{2}\sigma_{y_i}^2 \\ 0 & \text{otherwise} \end{cases} \ , \tag{5}$$

where we use the shorthand $c_i = Pr(c|\mathbf{m}_i)$, and σ_{y_i} is set to the mean class distance, *i.e.* the mean of $\{\mathrm{D}(i,j)\}_{\forall i<j \text{ s.t. } y_i=y_j}$. The distance thresholding is done to increase graph sparseness and promote community structure.

IV. Experiments

In this section, we present our experimental evaluation of the proposed method. We use the established MSRAction3D benchmark dataset [7], which consists of 20 different actions performed by 10 different subjects, performed up to three times each. Some actions are quite spatially similar, *e.g.* "*draw x*" and "*draw circle*". The cross-subject setting proposed by Li *et al.* [7] is used, and half of the subjects are used for training, and the rest for testing. We carefully implemented the MP descriptor and PCA-GTV in Python. For $\mathcal{G}_1$, both KNN and confidence graphs are tested, in order to create a discriminative low-rank representation. Only KNN graphs are used for $\mathcal{G}_2$, as the feature closeness prior does not need knowledge of class memberships.

PCA-GTV reduced the dimension of the MP descriptor to 26. Results are shown in Table I. The best result was gotten using KNN graphs for $\mathcal{G}_1$, using $\gamma_1 = 2^2$ and $\gamma_2 = 2^{-3}$, for which the improvement over the baseline MP descriptor is statistically significant (p-value < 0.05 using McNemar's test). The confusion matrix is shown in Fig. 3. While confidence graphs also work, it does not perform as well as using KNN graphs. This is probably because low-confidence descriptors will become isolated vertices in the graph, which effectively makes them unregularized. The proposed method improves over the baseline and shows that graph regularization helps improving recognition performance. Further, we can see that while adding standard PCA (reduced to 10 dimensions) helps due to the KNN classifier's sensitivity to high dimensionality, PCA-GTV is better due to the robustness against noise, which was previously demonstrated by Shahid *et al.* [10].

Actions involving human-object interaction, such as "*hammer*" and "*hand catch*" are difficult to capture using pure skeletons, and are sometimes mistaken for spatially similar actions such as "*high throw*" and "*draw x*". We can see that our graph regularized subspace representation gives a small but significant improvement.

We note here that while more heavily computable methods obtaining slightly higher accuracies (96.7%) on this dataset do exist [8], getting the best result on this dataset is not the purpose of this paper. Rather, we wish to illustrate that graph regularization can help improve existing methods for human action recognition. Learning our subspace took 359.8 seconds, but test time execution speed became faster compared to the baseline method, due to the reduced dimensionality (see Table II).

In order to show the effect of the piece-wise constantness encouragement provided by the GTV regularizer, we show the accuracy as a function of γ_1 in Fig. 4. As can be seen, a higher γ_1 places more importance on the regularizer and is shown to lead to higher accuracy, although diminishing returns are shown earlier for the confidence graph.

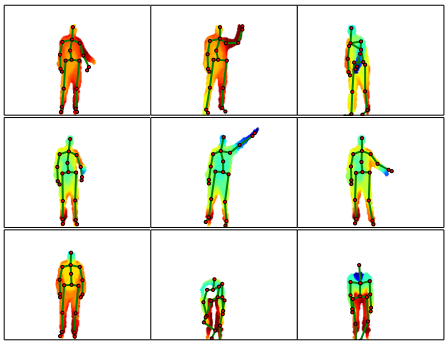

Figure 2. Example depth frames and skeletons of the actions "*hammer*" (top), "*hand catch*" (middle) and "*bend*" (bottom) in the MSRAction3D dataset.

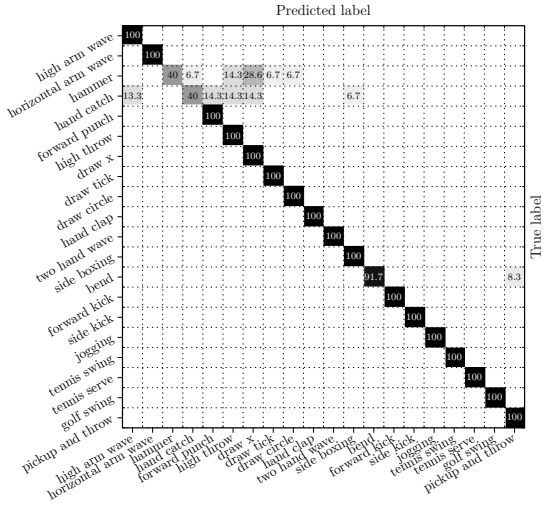

Figure 3. Confusion matrix for using our method on the MSRAction3D dataset. Each cell shows classification accuracy (%) from white (0) to black (100) in the cross-subject setting. The average accuracy is 93.5%.

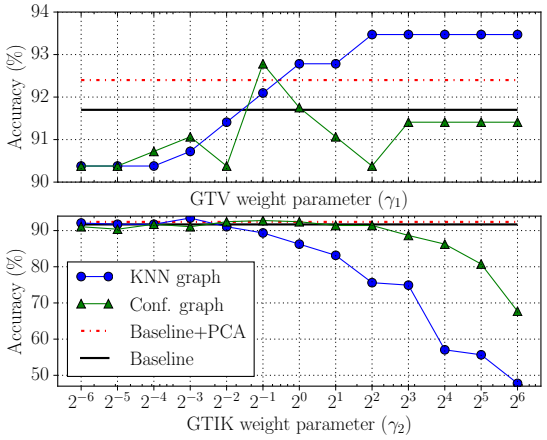

Figure 4. Accuracy per the weight parameters γ_1, γ_2 on the MSRAction3D dataset.

We also found that using a too large value for γ_2 (> 1) resulted in a vast deterioration of recognition performance (see Fig. 4). This is probably due to the GTIK regularizer causing the learned subspace to follow the original feature representation too closely, thus overfitting to noise while neglecting the ℓ_1 reconstruction term in (2).

Table I
RECOGNITION PERFORMANCE ON THE MSRACTION3D DATASET.

Method	Accuracy (%)
MP + KNN (baseline) [15]	91.7
MP + PCA + KNN	92.4
MP + PCA-GTV(Conf. Graph) + KNN	**92.8**
MP + PCA-GTV(KNN Graph) + KNN	**93.5**

Table II
TEST-TIME EXECUTION SPEED ON THE MSRACTION3D DATASET.

Method	Total time for test set (s)
MP + KNN (baseline) [15]	20.71
MP + PCA + KNN	0.55
MP + PCA-GTV(Conf. Graph) + KNN	**2.07**
MP + PCA-GTV(KNN Graph) + KNN	**1.90**

V. CONCLUSION

We have proposed GRIP, an enhanced feature descriptor using graph regularization for action recognition with tracked 3D skeletons. Experimental evaluation demonstrated that the method improves over the baseline MP approach, and resulted in a 93.5% accuracy on the MSRAction3D benchmark action recognition dataset using KNN graphs. This shows that graph regularization helps improving performance despite noisy skeleton estimates. In the future, we will investigate other types of graph representations and regularizers.

REFERENCES

[1] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Database theory—ICDT'99*. 1999.

[2] Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *JOTA*, 158(2), 2013.

[3] Fei Han, Brian Reily, William Hoff, and Hao Zhang. Space-time representation of people based on 3D skeletal data: a review. *arXiv preprint arXiv:1601.01006v2*, 2016.

[4] Jungong Han, Ling Shao, Dong Xu, and J. Shotton. Enhanced computer vision with Microsoft Kinect sensor: A review. *IEEE TC*, 43(5), 2013.

[5] Bo Jiang, Chris Ding, Bio Luo, and Jin Tang. Graph-laplacian pca: Closed-form solution and robustness. In *CVPR*, 2013.

[6] Tommi Kerola, Nakamasa Inoue, and Koichi Shinoda. Spectral graph skeletons for 3D action recognition. In *ACCV*, 2014.

[7] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *CVPR Workshops*, 2010.

[8] Jiajia Luo, Wei Wang, and Hairong Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *ICCV*, 2013.

[9] Omar Oreifej, Zicheng Liu, and WA Redmond. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, 2013.

[10] Nauman Shahid, Nathanaël Perraudin, Vassilis Kalofolias, Benjamin Ricaud, and Pierre Vandergheynst. Pca using graph total variation. In *ICASSP*. IEEE, 2016.

[11] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1), 2013.

[12] Dorina Thanou, David I Shuman, and Pascal Frossard. Parametric dictionary learning for graph signals. In *IEEE GlobalSIP*, 2013.

[13] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[14] Jiang Wang and Ying Wu. Learning maximum margin temporal warping for action recognition. In *ICCV*, 2013.

[15] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *ICCV*, 2013.

[16] Xuan Zhang, Xiaowen Dong, and Pascal Frossard. Learning of structured graph dictionaries. In *ICASSP*. IEEE, 2012.