Jointly Learning Heterogeneous Features for RGB-D Activity Recognition

Jian-Fang Hu[†], Wei-Shi Zheng[‡⋆], Jianhuang Lai[‡], and Jianguo Zhang[◇]

[†]School of Mathematics and Computational Science, Sun Yat-sen University, China; [‡]School of Information Science and Technology, Sun Yat-sen University, China; [⋆]Guangdong Province Key Laboratory of Computational Science, Guangzhou, China; [◇]School of Computing, University of Dundee, United Kingdom.

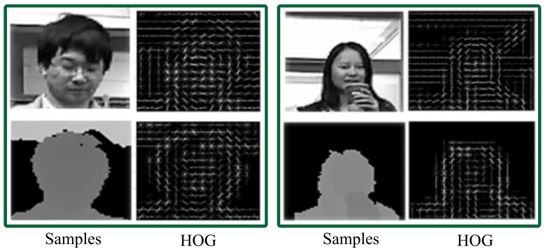

Figure 1: Visualization of HOG features from two activity snapshots in RGB (gray) and depth channel respectively. As shown, the HOG features from both channels of the same activity unveils similar "gist" structure of that activity.

The emergence of low-cost depth sensors (e.g., the Microsoft Kinect) opens a new dimension to address the challenge of human activity recognition. Compared to the conventional use of RGB videos, the information from depth channel is insensitive to illumination variations, invariant to color and texture changes, and more importantly reliable for estimating body silhouette and skeleton (human posture) [3]. Bearing on these merits, developing discriminative depth descriptors [2, 5] and fusing RGB channel and Depth channel together [6] become two emerging branches of activity recognition.

However, the majority of these methods neither seek to jointly learn the features extracted from RGB and depth channels simultaneously nor model their underlying connections. As can be observed from Figure 1, features from different channels may share visual structures. The benefits of exploring both shared and specific structures for classification have been demonstrated by multi-task learning, since it can significantly reduce the effective complexity of the task and transfer knowledge between related tasks [1, 4]. However, these methods assume that the features employed by different tasks are homogeneous, thus not applicable for mining shared and feature-specific structures among heterogeneous features.

In this paper, we propose a heterogeneous feature learning model for RGB-D activity recognition. Our model is built on mining a set of subspaces (one subspace for each heterogeneous feature) such that features with different dimensionality can be compared directly, and their shared and specific components can be easily encoded. We introduce a feature-specific projection matrix as a linear transformation for each feature, and then formulate our subspaces mining and shared features learning in the framework of multi-task learning. Therefore, the optimal solutions for projection matrices and shared-specific structures can be jointly derived, which is illustrated in Fig. 2. Specifically, our model is formulated as

$$\begin{aligned}\min_{W_0,\{W_i\},\{\Theta_i\}} \sum_{i=1}^{S} & (\|(W_0+W_i)^T\Theta_i^T X_i - Y\|_F^2 + \beta\|W_i\|_F^2 - \gamma\|X_i - \Theta_i\Theta_i^T X_i\|_F^2) \\ & + \alpha\|W_0\|_F^2 \\ & s.t. \Theta_i^T\Theta_i = I, i = 1,2,\ldots S\end{aligned}$$

where $\|\bullet\|_F$ denotes the Frobenius matrix norm. The regularization terms $\|W_i\|_F^2$ and $\|W_0\|_F^2$ are defined in the way such that a reliable generalization and effective closed-form solution can be obtained for our joint learning model. α and β are two parameters to control the trade-off between the shared and specific components.

Our model is casted as a least-square problem with both *prediction* (first term) and *reconstruction* loss (third term). It intends to jointly learn the common subspaces, shared and feature-specific components in a unified framework. The prediction loss item minimizes the empirical risk of each feature and thus guides our shared-specific structures learning for the purpose of better recognition. The reconstruction loss term is employed to ensure that a good reconstruction (controlled by the parameter γ) can be derived in the learned subspace using the projection matrix during optimization, which leads to a meaningful solution of the model. Here, an orthogonal constraint $\Theta_i^T\Theta_i = I$ is imposed on the projection matrix Θ_i in order to reduce the redundancy to certain extent while preserving data information. We call the proposed model the **j**oint heter**o**geneous feat**u**res **le**arning (JOULE) model.

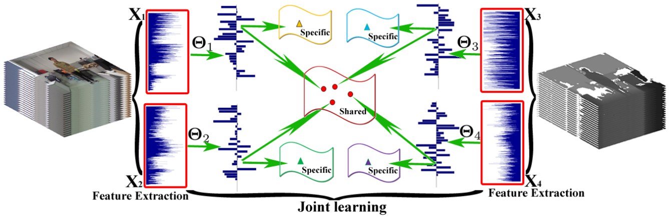

Figure 2: A graphic illustration of our joint learning framework. In this framework, all the projection matrices $\{\Theta_i\}_{i=1,2,3,4}$, shared structures and specific structures are jointly learned for the purpose of recognition. For simplicity, we only show the process of joint learning RGB and Depth channels in this example. However, the skeleton channel is also utilized in our experiments. Please see our full paper for more details.

DataSet	MSR Daily	CAD 60	SYSU 3D HOI (s-1)	SYSU 3D HOI (s-2)
HON4D [2]	80	72.7	73.39	79.22
Actionlets [5]	85.75	74.7	–	–
Our method	95	84.1	79.63	84.89

Table 1: Part of comparison results in the main paper.

In addition to the proposed joint heterogeneous features learning framework, we also collected a new RGB-D activity dataset (SYSU 3D HOI set) focusing on complex activities involved human-object interactions. For constructing this set, 40 subjects were asked to perform 12 different activities. For each activity, each participants manipulate one of the six different objects: phone, chair, bag, wallet, mop and besom. Compared to existing datasets, our set presents new challenges: 1) the involved motions and the manipulated objects' appearance are highly similar between some activities; 2) the number of participants is at least four times larger than that of existing ones; 3) two different protocols (s-1 and s-2) are employed to benchmark different methods. To the best of our knowledge, this is the most complete set to date for studying 3D activities in terms of the number of subjects. Our extensive experiments on two public available sets and the newly collected set have demonstrated the efficacy of the proposed method (ref. Table 1).

[1] Yonatan Amit, Michael Fink, Nathan Srebro, and Shimon Ullman. Uncovering shared structures in multiclass classification. In *ICML*, 2007.

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

[3] 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. In *CVPR*, 2011.

[4] A. Torralba, K.P. Murphy, and W.T. Freeman. Sharing visual features for multiclass and multiview object detection. *TPAMI*, 29(5):854–869, 2007.

[5] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Learning actionlet ensemble for 3d human action recognition. *TPAMI*, 36(5):914–927, 2014.

[6] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. Modeling 4d human-object interactions for event and object recognition. In *ICCV*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.