MULTI-MODAL FEATURE FUSION FOR ACTION RECOGNITION IN RGB-D SEQUENCES

Amir Shahroudy[†,‡], *Gang Wang*[†]

[†]School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
{*amir3, wanggang*}*@ntu.edu.sg*

Tian-Tsong Ng[‡]

[‡]Institute for Infocomm Research
A*STAR, Singapore
ttng@i2r.a-star.edu.sg

ABSTRACT

Microsoft Kinect's output is a multi-modal signal which gives RGB videos, depth sequences and skeleton information simultaneously. Various action recognition techniques focused on different single modalities of the signals and built their classifiers over the features extracted from one of these channels. For better recognition performance, it's desirable to fuse these multi-modal information into an integrated set of discriminative features. Most of current fusion methods merged heterogeneous features in a holistic manner and ignored the complementary properties of these modalities in finer levels. In this paper, we proposed a new hierarchical bag-of-words feature fusion technique based on multi-view structured sparsity learning to fuse atomic features from RGB and skeletons for the task of action recognition.

Index Terms— Action Recognition, Kinect, Feature Fusion, Structured Sparsity

1. INTRODUCTION

Recent development of depth sensors had a noticeable impact on the research in machine vision field. After the introduction of Microsoft Kinect, a large volume of research has been done on depth signal analysis. Microsoft Kinect's output is a multi-modal signal which gives RGB videos, depth sequences and skeleton information [1] simultaneously; as a result, their features could be interdependent and complementary to each other on each frame. For better recognition performance, it's desirable to fuse these multi-modal information into an integrated and highly discriminative set of features.

The simplest fusion method is to concatenate heterogeneous features from different sources, like [2] which binded the skeletal and silhouette-based features, [3] who applied the same method for RGB+D "combing" or [4] in which proposed depth-based comparative coding descriptors were concatenated into BOW histograms of RGB data.

A more attentive way to fuse multi-modal features is to build separated kernels for different modalities of the data and apply a multi-kernel learning technique to aggregate their discriminative powers like [5, 6]. The main drawback of this method is its holistic approach to fuse the features. It ignores the frame-wise complementary properties of the features extracted from different modalities of the data. In this paper we are proposing a new method to fuse the heterogeneous features extracted from RGB+Skeleton signals of Kinect sensor. Unlike current methods, proposed fusion is not done in a holistic level. It utilizes the complementary properties of heterogeneous features in groups of neighboring frames and consequently leads into a better set of results.

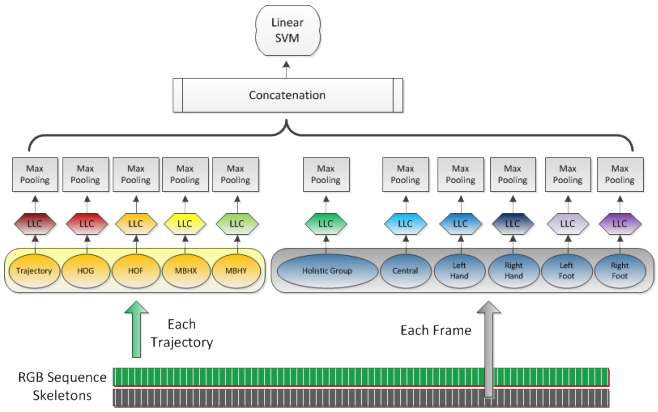

Fig. 1. Illustration of first baseline, "concatenation" method.

2. BASELINE FUSION METHODS

To extract features from the RGB channel, we use *dense trajectories* representation of the video sequences [7] with HOG [8], HOF, MBHX, and MBHY [9] descriptors for each single trajectory as proposed in [7]. We apply k-means clustering separately over all types of descriptors. Each descriptor could be regarded as a subchannel of RGB information. After building the feature-specific dictionaries, each trajectory could be represented as LLC codes [10] of its subchannel features.

On the other hand, 3 dimensional locations of skeleton joints are normalized against direction and position of the body. With the assumption of a perfect normalization, the movement of each limb would be independent of other body parts; therefore we can separate limbs from other joints and divide the skeleton into five subsets. Since we have the idea

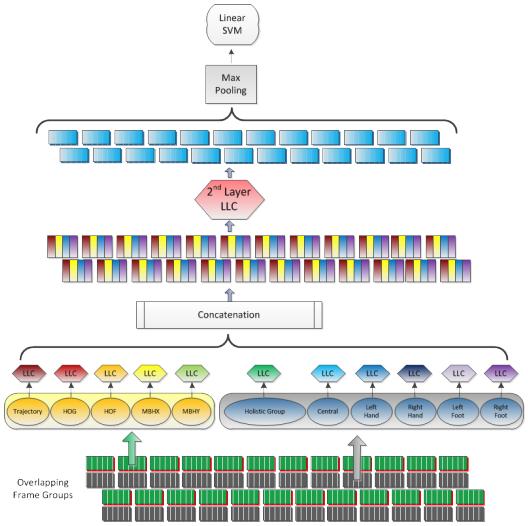

Fig. 2. Illustration of second baseline, "naïve hierarchical clustering" method.

of a proper fusion in next few steps, it makes sense to partition independent parts of the information into subchannels and prevent them from being concatenated. Like RGB subchannels, all groups of skeleton features go through k-means clustering and LLC encoding.

For the first baseline method, as shown in Fig. 1, we concatenate all the output codes of first layer's subchannels. We train an SVM on the result of max-pooling over all of the codes for each sample. Throughout this paper, we will call this method as *"concatenation"* and consider it as the first baseline in which the fusion is done in a holistic point of view.

To better represent video samples, we can divide them into overlapping M–frame segments. The feature representation of each segment would be the result of max-pooling for each subchannel. These segments are our target level for multi-modal feature fusion in contrast with holistic methods. Our second baseline approach to fuse the subchannel features is hierarchical clustering (Fig. 2). In the first layer, the concatenation is done over each segment and in the second layer we build a high level dictionary followed by an LLC encoding over the input codes from all segments. To represent each sample, a max-pooling over all if its segments is done and the final code is passed into a linear SVM for classification. We will call this method as *"naïve hierarchical clustering"*. As reported in [4], naïve hierarchical clustering method did not perform well in RGB+Depth data fusion and in this work we show its flaws and propose the proper way to fuse these heterogeneous features.

Notations. Throughout this paper, we use bold uppercase letters to represent matrices and bold lowercase letters for vectors. For a matrix $\mathbf{W}$, we denote its j-th row as $\mathbf{w}^j$ and its i-th column as $\mathbf{w}_i$. To represent the g-th group of features in i-th column we use bold superscripts like $\mathbf{w}_i^{\mathbf{g}}$.

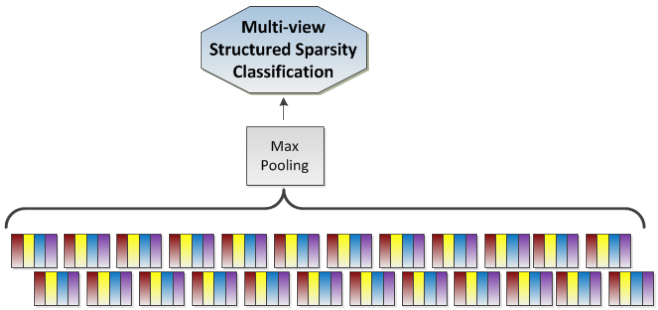

Fig. 3. Illustration of first proposed method: "Supervised Structured Sparsity Feature Fusion".

3. SUPERVISED STRUCTURED SPARSITY FEATURE FUSION

In naïve hierarchical clustering baseline, the output of first layer is still a concatenation of heterogeneous LLC codes and this means ignoring their difference and consider them as homogeneous features. The same misconception happened in the results of [4], where they tried to fuse the RGB+D features by cascading BOVW method and hierarchical clustering but with disappointing results.

An important observation here is: an individual subchannel or a suitably weighted combination of few number of subchannels could lead into a discriminative feature but when we concatenate them with all other groups, we diminish the discrimination and drop the performance of the entire framework. So, a proper way to handle this is to apply group sparsity over heterogeneous features of subchannels. It deactivates irrelevant groups and applies the desired weighting for informative set of subchannels. Multi-view structured sparsity [11] is a suitable method to apply group sparse coding over our multi-modal features. Here we adopt their integrated feature learning and classification scheme to improve our hierarchical framework for RGB+Skeleton feature fusion.

A key assumption in traditional MKL methods is that the features of the same group are equally important and then would be assigned the same weight in the final fusion. To overcome this limitation, [11] used a combination of group-l_1 norm and $l_{2,1}$ norm regularizers to emphasize on group-wise importance of features for each task and also take advantage of globally discriminative single features independently from their groups.

The output features of the first layer of naïve hierarchical clustering could be represented as: $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector of each sample. The d-dimensional feature representation $\mathbf{x}_i = [\mathbf{x}_i^1; ...; \mathbf{x}_i^k]$ consists of concatenated LLC codes of first level BOW from all k different subchannels of the data which are max-pooled over all the segments of each sample (Fig. 3). The labels of samples are given in a class assignment matrix $\mathbf{Y} \in \mathbb{R}^{n \times c}$, in which c is the number of action classes. Each row of $\mathbf{Y}$ contains $c-1$

zeros and a 1 representing the correct class label.

Feature learning is done through optimization of the cost function in Eq. 1.

$$\min_{\mathbf{W}} ||\mathbf{X}^T\mathbf{W} - \mathbf{Y}||_F^2 + \gamma_1||\mathbf{W}||_{G_1} + \gamma_2||\mathbf{W}||_{2,1} \quad (1)$$

in which $\mathbf{W} \in \mathbb{R}^{d\times c}$ is the weights matrix, and its columns can be identified with $\mathbf{w}_i = [\mathbf{w}_i^1; ...; \mathbf{w}_i^k]$, similar to $\mathbf{X}$.

Two regularization terms in Eq. 1 are defined as:

$$||\mathbf{W}||_{G_1} = \sum_{i=1}^{c}\sum_{g=1}^{k} ||\mathbf{w}_i^{\mathbf{g}}||_2 \quad (2)$$

$$||\mathbf{W}||_{2,1} = \sum_{j=1}^{d} ||\mathbf{w}^j||_2 \quad (3)$$

Note that in Eq. 2, l_2-norm of all subchannels of $\mathbf{W}$ are summed over both dimensions, but in Eq. 3 the summation is over the l_2-norms of all atomic rows of $\mathbf{W}$.

Group-l_1 norm regularizer (Eq. 2), forces the weights inside subchannels to be activated or deactivated together and applies sparsity over the number of active subchannels for each class. Second regularizer, $l_{2,1}$-norm (Eq. 3), gives chance to single features to be activated inside an inactive subchannel, but this applies to all classes at the same time.

To solve this optimization problem, [11] provided an efficient iterative algorithm which updates the columns of $\mathbf{W}$ one by one based on the gradient of the entire cost function regarding the columns of $\mathbf{W}$.

Upon convergence, each test sample $\mathbf{x}$ could be easily classified by finding the maximum value in the vector $\mathbf{x^T W}$.

4. UNSUPERVISED STRUCTURED SPARSITY FEATURE FUSION

A better modification over the naïve hierarchical clustering baseline, is to apply an unsupervised way to cluster the segments based on their first layer codes; therefore, in this section, columns of $\mathbf{X}$ represent LLC codes of all segments from entire samples. For unsupervised generalization of proposed method, [11] suggested to unfix the assignment matrix and update it on each iteration; as a result, columns of this matrix would represent cluster assignment values for training segments.

$$\min_{\mathbf{W},\mathbf{F}^T\mathbf{F}=\mathbf{I}} ||\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2 + \gamma_1||\mathbf{W}||_{G_1} + \gamma_2||\mathbf{W}||_{2,1} \quad (4)$$

Vector $\mathbf{b} \in \mathbb{R}^c$ is the intercept vector and makes the columns of $\mathbf{F} - \mathbf{1}_n\mathbf{b}^T$ centered.

To calculate the proper value for $\mathbf{F}$ matrix in each iteration, according to theorem 1 of [11], we have to apply singular value decomposition over $\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T$ and update $\mathbf{F}$ as:

$$\mathbf{U\Lambda V}^T = \mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T \quad \text{(svd)} \quad (5)$$

$$\mathbf{F} = \mathbf{U}\,[\mathbf{I}; \mathbf{0}]\,\mathbf{V} \quad (6)$$

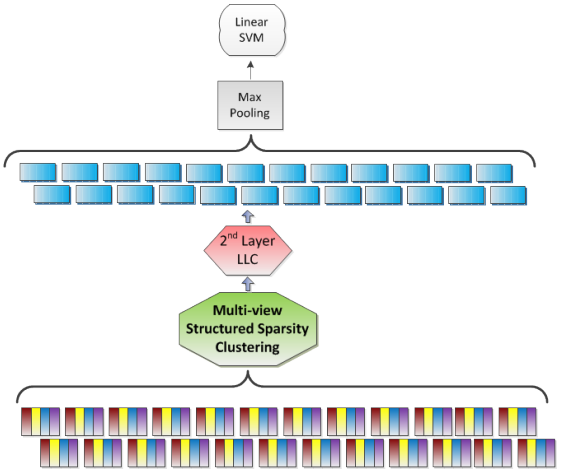

Fig. 4. Illustration of second proposed method: "Unsupervised Structured Sparsity Feature Fusion".

This guarantees the orthonormality constraint of $\mathbf{F}$ and fulfills the minimization task of $||\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2$. Since two regularization terms are independent of $\mathbf{F}$, they will not affect the optimization of the whole cost function; therefore, this method gives the optimal values for medium cluster assignment matrix on each iteration.

To initiate the $\mathbf{F}$ matrix, [11] suggested to apply k-means clustering over segments, but according to above discussion about the heterogeneous nature of the input features, we change this step and initiate it based on the class labels cues of segments. This modification leads into an apparent rise in the performance of the entire framework.

Upon convergence of the weights, we apply k-means over the final values of $\mathbf{F}$ and re-cluster them in their new space. After extracting the final cluster centers, we encode each segment using second layer LLC code and feed them into a linear SVM for classification. This framework is pictured in Fig. 4.

5. EXPERIMENTAL RESULTS

MSRDailyActivity3D dataset [12] is chosen as the benchmark in this paper. It includes 320 samples from 16 different action classes, and for each one, depth sequence, RGB video and skeleton information is provided.

First is to study the discriminative power of each subchannel in "concatenation" baseline. Fig. 5 shows the recognition rates for RGB and skeleton subchannels. "Entire skeleton" represents the LLC codes extracted from all the joints of the skeleton together, without any subchannel separation. On the other hand, "5 joint groups" shows the concatenation of LLC codes extracted from five separated parts of the skeleton in the first layer. We can compare these two with and without the presence of other subchannels. Although the concatenation of joint groups could not beat the entire skeleton subchannel per se, when fused with RGB subchannels, it outper-

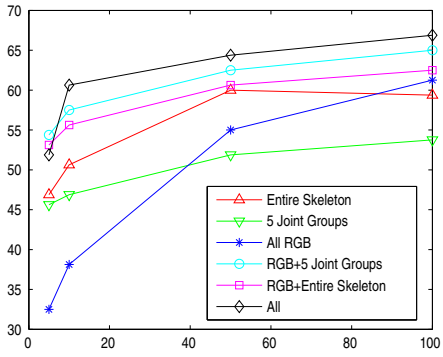

Fig. 5. Recognition rates for "concatenation" of different subchannels of features. Horizontal axis shows the first layer dictionary sizes and vertical axis shows the recognition rates in percents.

forms the "entire skeleton". This shows separating irrelevant subchannels of data (skeleton separation) could improve the performance of the final fusion. The best result is still from the concatenation of all subchannels including "entire skeleton" and all 5 groups of joints, so we keep all of them for the comparison of four different fusion methods.

The second experiment, compares the proposed fusion methods with baselines. Table 1 shows the results. As can be seen, naïve hierarchical clustering method could not even beat the simple concatenation, justifying the unsuccessful try in [4] to cascade hierarchical BOVW. The supervised version of the proposed method gives better results than baselines with a 7% margin on the last two dictionary sizes. Lastly, the unsupervised method outperforms all others with an outstanding margin for all dictionary sizes. This validates our hypothesis about the importance of atomic level mutual and complementary properties of multi-modal signals of Kinect and showed the deficiency of holistic fusion methods in the task of action recognition in RGB+Skeleton sequences.

6. CONCLUSION

We discussed the complementary properties of RGB and skeleton channels in Kinect sequences which is ignored in a lot of currently proposed recognition methods. We showed the drawbacks of holistic combination approaches and proposed two hierarchical supervised and unsupervised fusion methods to overcome these issues. The proposed techniques use structured sparsity to properly fuse groups of features from different subchannels of the data. We also showed to have a good fusion, the irrelevant binded sources of data should be discovered and separated in an appropriate way.

Our experimental results approved our hypothesis and shows our RGB+Skeleton fusion method has the potential to outperform current depth-only action recognition methods.

Table 1. Comparison of recognition rates for proposed fusion methods vs baselines on MSRDailyActivity3D dataset.

Dict. Size	Concat. Baseline	Naïve Hier. Clustering	Sup. SSFF	Unsup. SSFF
5	51.9%	57.5%	51.9%	**65%**
10	60.6%	59.4%	60.6%	**71.9%**
50	64.4%	64.4%	72.5%	**76.9%**
100	66.9%	64.4%	73.1%	**81.9%**

7. REFERENCES

[1] 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 *CVPR*, 2011.

[2] A. A. Chaaraoui, J. R. Padilla López, F. Flórez Revuelta, *et al.*, "Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices," in *ICCVW*, 2013.

[3] Y. Zhao, Z. Liu, L. Yang, and H. Cheng, "Combing rgb and depth map features for human activity recognition," in *APSIPA ASC*, 2012.

[4] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *ECCVW*, 2012.

[5] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *IROS*, 2011.

[6] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," *PAMI*, 2013.

[7] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *CVPR*, 2011.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[9] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *ECCV*, 2006.

[10] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.

[11] H. Wang, F. Nie, and H. Huang, "Multi-view clustering and feature learning via structured sparsity," in *ICML*, 2013.

[12] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012.