Deep Multimodal Feature Analysis for Action Recognition in RGB+D Videos

Amir Shahroudy, *Student Member, IEEE*, Tian-Tsong Ng, *Member, IEEE*, Yihong Gong, *Senior Member, IEEE*, and Gang Wang, *Member, IEEE*

Abstract—Single modality action recognition on RGB or depth sequences has been extensively explored recently. It is generally accepted that each of these two modalities has different strengths and limitations for the task of action recognition. Therefore, analysis of the RGB+D videos can help us to better study the complementary properties of these two types of modalities and achieve higher levels of performance. In this paper, we propose a new deep autoencoder based shared-specific feature factorization network to separate input multimodal signals into a hierarchy of components. Further, based on the structure of the features, a structured sparsity learning machine is proposed which utilizes mixed norms to apply regularization within components and group selection between them for better classification performance. Our experimental results show the effectiveness of our cross-modality feature analysis framework by achieving state-of-the-art accuracy for action classification on five challenging benchmark datasets.

Index Terms—Multimodal analysis, RGB+D, action recognition, structured sparsity

1 Introduction

Recent development of range sensors had an indisputable impact on research and applications of machine vision. Range sensors provide depth information of the scene and objects, which helps in solving problems that are considered hard for RGB inputs [1].

Human activity recognition is one of the active fields in computer vision and has been explored extensively. Recent advances in hand-crafted [2], [3] and convnet-based [4] feature extraction and analysis of RGB action videos achieved impressive performance. They generally recognize action classes based on appearance and motion patterns in videos. The major limitation of RGB sequences is the absence of 3D structure from the scene. Although some works are done towards this direction [5], recovering depth from RGB in general is an underdetermined problem. As a result, depth sequences provide an exclusive modality of information about the 3D structure of the scene, which suits the problem of activity analysis [6], [7], [8], [9], [10], [11], [12]. This complements the textural and appearance information from RGB. Our goal in this work is to analyze the multimodal RGB+D signals for identifying the strengths of respective modalities through teasing out their shared and modality-specific components and to utilize them for improving the classification of human actions.

Having multiple sources of information, one can find a new space of common components which can be more robust than any of the input features. Through linear projections, canonical correlation analysis (CCA) [13], [14] gives us the correlated form of input modalities which in essence is a robust representation of multimodal signals. However, the downside of CCA is the linearity limitation. Kernel canonical correlation analysis (KCCA) [15] extended this idea into nonlinear kernel-based projections, which is still limited to the representation capacity of the kernel's space and is not able to disentangle the high-level nonlinear complexities between the input modalities. Further, the traditional solutions of CCA and KCCA are to solve the maximization of correlation between the projected vectors analytically, which does not scale well with the size of the data.

To overcome these limitations, a new deep autoencoder-based nonlinear common component analysis network is proposed to discover the shared and informative components of input RGB+D signals.

Besides the shared components, each input modality has specific features which carry discriminative information for the recognition task. In this respect, we can enhance the representation by incorporating the modality-specific components of respective modalities [16], [17]. Based on this intuition, at each layer our deep network factorizes the multimodal input features into their shared and modality-specific components. By stacking such layers, we further decode the complex and highly nonlinear representations of the input modalities in a nonlinear fashion.

- *A. Shahroudy is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, and the Institute for Infocomm Research, 1 Fusionopolis Way, Singapore 138632. E-mail: amir3@ntu.edu.sg.*
- *G. Wang is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798. E-mail: wanggang@ntu.edu.sg.*
- *T.-T. Ng is with the Institute for Infocomm Research, 1 Fusionopolis Way, Singapore 138632. E-mail: ttng@i2r.a-star.edu.sg.*
- *Y. Gong is with the Institute of Artificial Intelligence and Robotics, Xián Jiaotong University, Shaanxi 710048, China. E-mail: ygong@mail.xjtu.edu.cn.*

Manuscript received 15 Mar. 2016; revised 24 Dec. 2016; accepted 5 Mar. 2017. Date of publication 4 Apr. 2017; date of current version 10 Apr. 2018. Recommended for acceptance by T. Darell, C. Lampert, N. Sebe, Y. Wu, and Y. Yan.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2017.2691321

Across the layers, our deep multimodality analysis extracts a set of structured features which consist of hierarchically factorized multimodal components. The common components are robust against noise and missing information between the modalities, and the modality-specific components carry the remaining informative features which are irrelevant to the other modality. To effectively perform recognition tasks on our structured features, we design a structured sparsity-based learning framework. With different mixed norms, features of each component can be grouped together and group selection can be applied to learn a better classifier. We also show that the advantage of our learning framework is more significant as network gets deeper.

The contributions of this work are two-fold: first we introduce a new deep learning network for hierarchical shared-specific factorization of RGB+D features. Second, a structured sparsity learning machine is proposed to explore the structure of hierarchical factorized representations for effective action classification.

The rest of this paper is organized as follows. Section 2 explores the related work. Section 3 introduces the proposed deep component factorization network. Section 4 describes our classification framework for factorized components. Section 6 provides our experimental results, and Section 7 concludes the paper.

2 Related Work

There are other works which applied deep networks to multimodal learning. The work in [18], [19] used DBM for finding a common space representation for two input modalities, and predict one modality from the other. Andrew et al. [20] proposed a deep canonical correlation analysis network with two stacks of deep embedding followed by a CCA on top layer. Our method is different from these works in two major aspects. First, the previous work performed the multimodal analysis in just one layer of the deep network, but our proposed method performs the common component analysis in every single layer. Second, we incorporate modality-specific components in each layer to maintain all the informative features, at each layer.

Jia et al. [21] factorized the input features to shared and private components by applying structured sparsity, for the task of multi-view learning on human pose estimation, with linearity assumption. Cai et al. [16] proposed a nonlinear factorization of the features into common and individual components, towards a better representation of features for action recognition. They utilized mixture models to add nonlinearity to linear probabilistic CCA [22]. Our proposed technique stacks layers of nonlinear shared component analysis to progressively disentangle highly nonlinear correlations between the input features.

While learning frameworks in [23], [24], [25], [26] applied structured sparsity for other similar tasks, our structured sparsity learning machine extends the sparse selection into two levels of concurrent component and layer selection, which is more suited to the hierarchical outputs from our deep factorization network.

Recent single modality action recognition methods on depth signals can be divided into two major groups: depth map analysis methods [6], [11], [27], [28] and skeleton based methods [8], [9], [10], [29], [30].

The first group extract the action descriptors directly from depth map sequences. The idea of spatio-temporal interest points [31] was applied in depth videos by [11]. They also proposed depth cuboid similarity features to represent local patches. HON4D [6] represents depth sequences as histograms of 4D oriented normals of local patches, quantized on the vertices of a regular polychoron. Rahmani et al. [7], [28] achieved higher levels of robustness against viewpoint variations by using histograms of oriented principle components. Lu et al. [27] proposed binary range-sample descriptors based on τ tests on depth patches. The work of [32] applied convolutional networks for learning action classes on depth maps. Rahmani and Mian [33], [34] introduced a nonlinear knowledge transfer model to transform different views of human actions to a canonical view.

The second group of methods represent actions based on the 3D positions of major body joints, which are available for most of depth based action datasets. Luo et al. [29] proposed a novel skeleton-based discriminative dictionary learning method, utilizing group sparsity and geometry constraints. Vemulapalli et al. [8] represented skeletons as points and actions as curves in a Lie group using the 3D relative geometry between body parts. Evangelidis et al. [30] proposed a compact and view-invariant representation of body poses calculated from joint positions. Wang et al. [35] introduced a mining technique to find part-based mid-level patterns (frequent local parts) and aggregated the local representations as bag-of-FLPs to be classified by a SVM. Veeriah et al. [36] extended the structure of the long short-term memory (LSTM) units [37] to learn differential patterns in skeleton locations. The work of [38] introduced a hierarchy of recurrent networks to learn part-based motion patterns and combine them for action classification. Zhu et al. [39] proposed a new regularization term for learning co-occurrences of motion patterns among different joint groups. The work of [40] introduced a new part-aware LSTM structure to discover the long-term motion patterns of skeleton-based body parts separately and learn the action classes based on these representations. In [9], motion and local depth based appearance of each body joint was encoded using Fourier transform over a temporal pyramid. They also proposed a mining method to find the set of most representative body joints for each action class. Shahroudy et al. [41] formulated the discriminative joint selection by introducing a hierarchical mixed norm. The work of [42] combined different spatio-temporal depth and skeleton based gradient features and applied a random decision forest for action classification. Meng et al. [43] proposed a real-time action recognition method by applying random forest classifier on a set of distance values between the body joints and interacted objects. Wang and Wu [10] applied max-margin time warping to match the descriptors of skeletons over the temporal axis and learn phantom templates for each action class. An extensive review of different approaches and techniques on 3D skeletal data is done by [44]. The fusion of various depth based features is also studied by [45].

Multimodality analysis of RGB+D action videos is studied by [46], [47], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56]. Ni et al. [46] introduced a RGB+D fusion method by concatenating depth descriptors to RGB based representations of STIP points. Liu and Shao [47] introduced a genetic

programming framework to improve the RGB and depth descriptors and their fusion simultaneously through an iterative evolution. The work of [48] solved the problem of RGB+D action recognition by utilizing RGB information for better tracking of interest point trajectories and describe them by depth-base local surface patches. Hu et al. [49] proposed a heterogeneous multitask feature learning framework to mine shared and modality-specific RGB+D features. The work of [50] applied projection matrices to the common and independent spaces between RGB and depth modalities. They learned their model by minimizing the rank of their proposed low-rank bilinear classifier.

The work of [51] also extracted STIPs from RGB and combined their HOG and HOF descriptors from RGB channel with local depth patterns (LDP) features from depth channel to fuse the two modalities. Depth-induced multiple channel STIPs [52], also added depth distances into GMM-based STIP representations. In [53] the STIP detection is done separately on RGB and depth and the HOGHOF descriptors are fused by combining the BOW representations of LLC codes of local features. Tsai et al. [54] used depth channel to segment the human body into known parts. STIPs with descriptors on RGB and depth channels are aggregated for each part by BOF representation over temporal pyramids. They assigned higher weights to non-occluded body parts to achieve a more robust global representations for action recognition. Multistream fused hidden Markov model was utilized to fuse pixel change history feature from RGB with MHI feature from depth channel by [55]. The work of [56] proposed a structured sparsity based fusion for RGB+D local descriptors. An evolutionary programming RGB+D fusion method was proposed by [47]. The proposed RGB+D analysis frameworks, are different from these methods, since our focus is on studying the correlation between the two modalities in the local level features and factorizing them to their correlated and independent components.

Recent advances of visual recognition in digital images using deep convolutional networks [57], [58], [59], [60] also inspired the research in video analysis. Ng et al. [61] studied two techniques of feeding videos to convnets for video classification. They proposed temporal pooling of the convnet-based features of frames to aggregate video descriptors from frame features. They also studied the advantages of utilizing a long short-term memory [37] network stacked over a convnet for video classification. Simonyan and Zisserman [4] fed a fixed length video sequence and its optical flow to a two-streamed convnet and fused the scores of the two streams in the end to classify the action labels. Wang et al. [62] combined the advantages of hand-crafted trajectory-based features and deep convnet learning based methods by applying [4]'s network along the motion trajectories of input videos. A novel deep convoultional framework for video event detection and evidence recounting was proposed by [63]. They introduced a back pass technique to localize the key evidences of the interested events in spatial-temporal domain. Tran et al. [64] studied the fully three dimensional convnet based video analysis and evaluated their proposed framework on various video analysis tasks including action recognition.

The applications of recurrent neural networks for 3D human action recognition were explored very recently. Du et al. [38] applied a hierarchical RNN to discover common 3D action patterns in a data-driven learning method. They divided the input 3D human skeletal data to five groups of joints and fed them into a separated bidirectional RNN. The output hidden representation of the first layer RNNs were concatenated to form upper-body and lower-body mid-level representation and these were fed to the next layer of bidirectional RNNs. The holistic representation for the entire body was obtained by concatenating the output hidden representations of these second layer RNNs and it was fed to the last RNN layer. The output hidden representation of the final RNN was fed to the softmax classifier for action classification. Differential RNN [36] added a new gating mechanism to the traditional LSTM to extract the derivatives of internal state (DoS). The derived DoS was fed to the LSTM gates to learn salient dynamic patterns in 3D skeleton data. The work of [39] introduced an internal dropout mechanism applied to LSTM gates for stronger regularization in the RNN-based 3D action learning network. To further regularize the learning, a co-occurrence inducing norm was added to the network's cost function which enforced the learning to discover the groups of co-occurring and discriminative joints for better action recognition. Liu et al. [65] extended the recurrent network based sequence analysis towards sequences of body joints. They added a new dimension dimension to the structure of a deep LSTM-based framework to learn the features over time and over the sequences of joints concurrently. To apply ConvNet-based learning to this domain, [66] used synthetically generated data and fitted them to real mocap data. Their learning method was able to recognize actions from novel poses and viewpoints.

Different from other methods, the proposed framework analyzes the components between the two modalities in a deep network, and factorizes the input RGB+D features into their shared and specific components in a hierarchy of nonlinear layers. Our solution is general and can be applied on any type of multimodal features to analyze their cross-modality components.

3 Deep Shared-Specific Component Analysis

We have two sets of features extracted from different modalities of data (RGB and depth signals) as our input for the task of action classification. State-of-the-art RGB based features [2], [67] include 2D motion patterns and appearance information of objects and scenes. On the other hand, various depth-based features [6], [7], [9], [11] encode 3D shape and motion information, without appearance and texture details. Consequently, it is beneficial to fuse the complementary RGB and depth-based features for better performance in action analysis.

There are different techniques for feature fusion. The choice of fusion strategy should rely on dependency of features. When features have very high dependency, descriptor-level fusion gives the best outcome, and when multiple groups of features have very low interdependency, kernel-level fusion performs better [17]. Since RGB and depth based features encode an entangled combination of common and modality-specific information of the observed action, they are neither independent nor fully correlated. Therefore, it is reasonable to embed the input data into a space of factorized

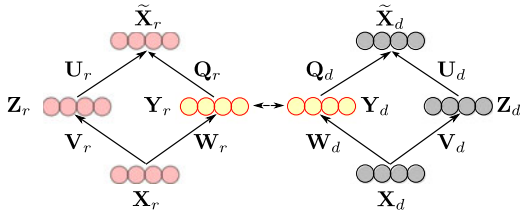


Fig. 1. Illustration of the proposed single layer shared-specific component analysis. $\mathbf{X}_r$ and $\mathbf{X}_d$ are input RGB and depth based features. We factorize each input feature into shared ($\mathbf{Y}$) and specific ($\mathbf{Z}$) components by forcing the $\mathbf{Y}$ vectors to be close, and the input features to be reconstructible from derived components.

common and modality-specific components. The combination of the shared and specific components in input features can be very complex and highly nonlinear. To disentangle them, we stack layers of nonlinear autoencoder-based component factorization to form a deep shared-specific analysis network.

In this section, we first introduce our basic framework of shared-specific analysis for multimodal signal factorization, then describe the deep network of stacked layers, where each layer performs factorization analysis and collectively produce a hierarchical set of shared and modality-specific components.

3.1 Single Layer Shared-Specific Analysis

Let us notate input RGB features by $\mathbf{X}_r$ and depth features by $\mathbf{X}_d$. We propose to factorize each input feature pattern into two spaces: first, common component space which corresponds to the highest correlation with the other modality $(\mathbf{Y}_r, \mathbf{Y}_d)$, and second, its modality-specific feature component space $(\mathbf{Z}_r, \mathbf{Z}_d)$

$$\begin{bmatrix} \mathbf{Y}_r \\ \mathbf{Y}_d \\ \mathbf{Z}_r \\ \mathbf{Z}_d \end{bmatrix} = g(\mathbf{X}_r, \mathbf{X}_d; \Omega), \tag{1}$$

where Ω is the set of model parameters that will be learned from the training data. We propose a sparse autoencoder-based network as the $g(\cdot)$ function, as illustrated in Fig. 1.

Feature vectors of each modality are factorized into $\mathbf{Y}$ and $\mathbf{Z}$ which represent shared and individual components of each modality respectively. Each component is derived from a linear projection of the input features followed by a nonlinear activation. Mathematically

$$\mathbf{Y}_r = f(\mathbf{W}_r\mathbf{X}_r + \mathbf{b}_{Y_r}\mathbf{1}^n) \tag{2}$$

$$\mathbf{Z}_r = f(\mathbf{V}_r\mathbf{X}_r + \mathbf{b}_{Z_r}\mathbf{1}^n), \tag{3}$$

in which $f(\cdot)$ is a nonlinear activation function. We use sigmoid scaling in our implementation. $\mathbf{b}_{Y_r}$ and $\mathbf{b}_{Z_r}$ are bias terms. Similarly, for the depth based input, we have

$$\mathbf{Y}_d = f(\mathbf{W}_d\mathbf{X}_d + \mathbf{b}_{Y_d}\mathbf{1}^n) \tag{4}$$

$$\mathbf{Z}_d = f(\mathbf{V}_d\mathbf{X}_d + \mathbf{b}_{Z_d}\mathbf{1}^n). \tag{5}$$

To prevent output degeneration, we expect the original features to be reconstructible from their factorized components [68]

$$\begin{aligned} \widetilde{\mathbf{X}}_r &= f([\,\mathbf{Q}_r\mathbf{U}_r\,]\begin{bmatrix} \mathbf{Y}_r \\ \mathbf{Z}_r \end{bmatrix} + \mathbf{b}_{\widetilde{X}_r}\mathbf{1}^n) \\ &= f(\mathbf{Q}_r\mathbf{Y}_r + \mathbf{U}_r\mathbf{Z}_r + \mathbf{b}_{\widetilde{X}_r}\mathbf{1}^n) \end{aligned} \tag{6}$$

$$\begin{aligned} \widetilde{\mathbf{X}}_d &= f([\,\mathbf{Q}_d\mathbf{U}_d\,]\begin{bmatrix} \mathbf{Y}_d \\ \mathbf{Z}_d \end{bmatrix} + \mathbf{b}_{\widetilde{X}_d}\mathbf{1}^n) \\ &= f(\mathbf{Q}_d\mathbf{Y}_d + \mathbf{U}_d\mathbf{Z}_d + \mathbf{b}_{\widetilde{X}_d}\mathbf{1}^n). \end{aligned} \tag{7}$$

Now we can formulate the desired component factorization into an optimization problem with the cost function

$$\begin{aligned} \Omega^* = \underset{\Omega}{argmin}\ & \Delta(\mathbf{Y}_r, \mathbf{Y}_d) + \lambda\|\Omega\|_2 \\ & + \zeta_r\Delta(\mathbf{X}_r, \widetilde{\mathbf{X}}_r) + \zeta_d\Delta(\mathbf{X}_d, \widetilde{\mathbf{X}}_d) \\ & + \alpha_r\Psi(\mathbf{Y}_r; \rho_Y) + \alpha_d\Psi(\mathbf{Y}_d; \rho_Y) \\ & + \beta_r\Psi(\mathbf{Z}_r; \rho_Z) + \beta_d\Psi(\mathbf{Z}_d; \rho_Z), \end{aligned} \tag{8}$$

where $\Omega = \{\mathbf{W}_\cdot, \mathbf{V}_\cdot, \mathbf{Q}_\cdot, \mathbf{b}_\cdot\}$ is the set of all parameters, and $[\lambda, \zeta_\cdot, \alpha_\cdot, \beta_\cdot]$ are hyper-parameters of trade-off between terms.

The first term in (8) forces the shared components of the two modalities ($\mathbf{Y}_r$ and $\mathbf{Y}_d$) to be as close as possible. We formulate this term as the Frobenius norm of the difference between two matrices

$$\Delta(\mathbf{Y}_r, \mathbf{Y}_d) = \|\mathbf{Y}_r - \mathbf{Y}_d\|_F. \tag{9}$$

The second term is the general weight regularization term, applied on the projection weights to prevent networks from overfitting training data.

The reconstruction costs are represented as $\Delta(\mathbf{X}_r, \widetilde{\mathbf{X}}_r)$ and $\Delta(\mathbf{X}_d, \widetilde{\mathbf{X}}_d)$ to prevent the model from degeneration. Here, we use Frobenius norm (the same as (9)) of the reconstruction error for the reconstruction cost term.

Last four terms of (8) are sparsity penalty terms over $\mathbf{Y}$ and $\mathbf{Z}$ outputs. It has been shown in [69], [70] that applying sparsity on the features of $\mathbf{Y}$ and $\mathbf{Z}$ will help to improve the learning capability, especially when components are overcomplete. As our sparsity penalty, we use KL-divergence term, applied between $\mathbf{Y}$ components and the sparsity parameters ρ_Y, as well as $\mathbf{Z}$ and ρ_Z.

It is worth pointing out, since the proposed framework is built on a sparse autoencoder-like scheme and has sigmoid scaling nonlinearity, it is necessary to apply PCA whitening on the input matrices $\mathbf{X}_r$ and $\mathbf{X}_d$ and scale their elements into the range of $[0, 1]$.

In this formulation, the disparity between $\mathbf{Z}_d$ and $\mathbf{Z}_r$ components is applied implicitly. The similarity inducing norm pushes the common components of the two modalities to move inside $\mathbf{Y}$ components. Therefore, we expect the remaining features in each of $\mathbf{Z}$ components to be highly different across the modalities.

3.2 Deep Shared-Specific Component Analysis

State-of-the-art RGB and depth based features for action recognition, are extracted by multiple linear and nonlinear layers of projection, embedding, spatial and temporal pooling, or statistical distribution encodings, e.g., BOvW [71] and FV [72] or Fourier temporal pyramids in [9]. Hence the common components between modalities can lie on highly complex and nonlinear subspaces of input data, and one

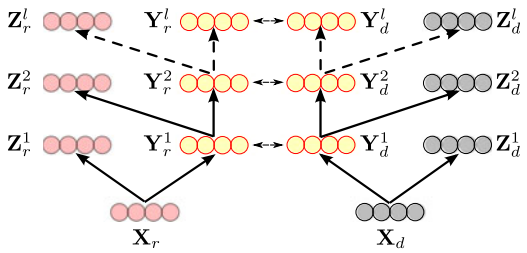


Fig. 2. Cascading factorization layers to a deep shared-specific network. To disentangle the highly nonlinear combination of shared-specific components, factorization layers are stacked by feeding the $\mathbf{Y}$ components of each layer as inputs of the next layer.

layer of the proposed shared-specific analysis cannot decode these complexities between the components.

By cascading multiple shared-specific analysis layers, we build a deep network to further factorize input features based on their higher orders of common and private information between modalities. To do so, we feed $\mathbf{Y}$ components of the previous layer as multimodal inputs of the current layer and apply the same method with new learning parameters in order to further factorize the features. As illustrated in Fig. 2, each layer extracts modality-specific components of the modalities and passes the shared ones for further factorization in the next layer

$$\begin{bmatrix} \mathbf{Y}_r^{(i)} \\ \mathbf{Y}_d^{(i)} \\ \mathbf{Z}_r^{(i)} \\ \mathbf{Z}_d^{(i)} \end{bmatrix} = \begin{cases} g(\mathbf{X}_r, \mathbf{X}_d; \mathbf{\Omega}^{(i)}) & \text{if } i = 1 \\ g(\mathbf{Y}_r^{(i-1)}, \mathbf{Y}_d^{(i-1)}; \mathbf{\Omega}^{(i)}) & \text{if } i > 1. \end{cases} \tag{10}$$

By applying this hierarchy on nonlinear layers, we expect the network to factorize more complex and higher order components of the inputs as it moves forward through the layers. Our deep network is trained greedily and layer-wise [73], [74], [75]. In other words, the optimization of each layer is started upon the convergence of the previous layer's training.

Upon training of the deep network, each input sample will be factorized into a pair of specific components $(\mathbf{Z}_r^{(i)}, \mathbf{Z}_d^{(i)})$ for each layer $i \in [1, .., l]$, plus the concatenation of last layer's shared components $(\mathbf{Y}_r^{(l)}, \mathbf{Y}_d^{(l)})$.

3.3 Convolutional Shared-Specific Component Analysis

The input features $(\mathbf{X}_r, \mathbf{X}_d)$ of the proposed deep shared-specific component analysis network are assumed to describe different representations of a multimodal entity. In our application, this entity is a RGB+D human action video and the inputs are RGB-based and depth-based features of it.

Since every input video can be regarded as a three dimensional cube (in x, y, t), it can be split to sub-cubes along all of its three dimensions, and the proposed multimodal analysis can be done on each of these sub-cubes separately. By limiting our analysis into holistic RGB+D features, we may lose discriminative local information in both modalities, because local features also have dependencies across modalities and their deep shared-specific component analysis (DSSCA) is beneficial. Therefore, as depicted in Fig. 3, we first train the local DSSCA network (DSSCA^L) on RGB+D features of the sub-cubes of training video samples.

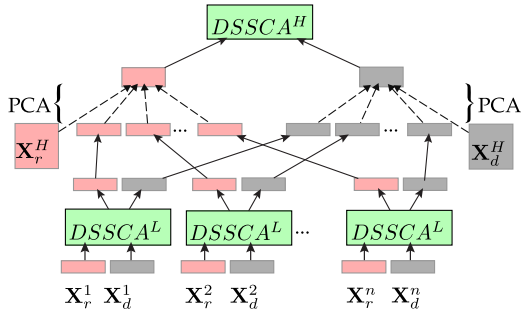


Fig. 3. Schema of our convolutional and holistic networks of deep shared-specific component analysis ($DSSCA$). We divide each video into n local cubes. Local features $\mathbf{X}_r^i$ and $\mathbf{X}_d^i$ are extracted from the ith cube. Convolutional network (denoted as $DSSCA^L$) is trained and then applied to decompose local features. The factorized components are then combined with holistic features $\mathbf{X}_r^H$ and $\mathbf{X}_d^H$. This combination undergoes PCA and is fed into the holistic network (denoted as $DSSCA^H$) as its multimodal input.

One can think of this stage as applying the same DSSCA^L network on all the sub-cubes of every input RGB+D video. At each step, we have a fixed-sized window over the current sub-cube in both RGB and depth channels of the input video and feed their corresponding sub-video representations to the DSSCA^L network as a single training sample. By convolving this window over all of the possible sub-cubes of every input video sample, we train the DSSCA^L network.

The learned DSSCA^L is then utilized to decompose the multimodal features of all the convolved sub-cubes. For every input video sample, we concatenate all of the factorized components of its sub-cubes. The resulting representation is then put together with the holistic multimodal features of video sample, to build the input for the holistic DSSCA network (DSSCA^H), similar to [76]. The inputs of DSSCA^H are PCA whitened and scaled into the range of $[0, 1]$.

Overall, we have $L = l_1 + l_2$ layers of factorization where l_1 and l_2 are the number of layers in DSSCA^L and DSSCA^H networks respectively. By applying the trained local-holistic networks into the features of each video sample, we have a set of $2L + 1$ independent components

$$\mathbf{A} = \{(\mathbf{Z}_r^1)^T, (\mathbf{Z}_d^1)^T, \ldots, (\mathbf{Z}_r^L)^T, (\mathbf{Z}_d^L)^T, (\mathbf{Y}^L)^T\}^T, \tag{11}$$

where $\mathbf{Y}^L = \begin{bmatrix} \mathbf{Y}_r^L \\ \mathbf{Y}_d^L \end{bmatrix}$ is the concatenation of last layer's common components.

3.4 Optimization Algorithm

The proposed formulation of cost function (8) is not a convex function of training parameters. Therefore, optimization of the learning parameters is not feasible in a single step. We iteratively optimize subsets of the parameters while keeping others fixed to achieve a suboptimal solution which is already shown effective in different applications [77].

Specifically, the learning parameters of each layer can be divided into two subsets. First are the ones defined for projection and reconstruction of the shared components $\mathbf{Y}_.$, and second consists of similar parameters for individual component $\mathbf{Z}_.$. These two sets are

$$\Omega_Y = \{\mathbf{W}_r, \mathbf{W}_d, \mathbf{Q}_r, \mathbf{Q}_d, \mathbf{b}_{Y_r}, \mathbf{b}_{Y_d}, \mathbf{b}_{\widetilde{X}_r}, \mathbf{b}_{\widetilde{X}_d}\} \tag{12}$$

$$\Omega_Z = \{\mathbf{V}_r, \mathbf{V}_d, \mathbf{U}_r, \mathbf{U}_d, \mathbf{b}_{Z_r}, \mathbf{b}_{Z_d}, \mathbf{b}_{\underset{\sim}{X_r}}, \mathbf{b}_{\underset{\sim}{X_d}}\}. \tag{13}$$

Now, to optimize the overall cost, we first fix Ω_Z (except $\mathbf{b}_{\underset{\sim}{X}}$) and minimize the cost function (8) regarding Ω_Y. Then fix parameters of Ω_Y (except $\mathbf{b}_{\underset{\sim}{X}}$) and optimize regarding Ω_Z and repeat this iteratively to converge into a suboptimal point.

In our implementation, all the optimization steps are done by "L-BFGS" algorithm using off-the-shelf "minFunc" software [78].

4 Structured Sparsity Learning Machine

Previous shared-specific analysis steps were all unsupervised and applied just based on the mutual characteristics of the two modalities. As a result, the factorized features of each component are not guaranteed to be equally discriminative for the following classification step. Hence we adopt the structured sparsity regularization method of [24], [26] aiming to select a number of components/layers sparsely to achieve more robust classification. Since the features of each component are highly correlated, our structured sparsity regularizer bounds the weights of the features inside each component to become activated or deactivated together.

Mathematically, we want to learn a linear projection matrix $\mathbf{B}$ to project our hierarchically factorized features $\mathbf{A}$ (see Equation (11)), to a class assignment matrix $\mathbf{F}$ defined as

$$f_i^j = \begin{cases} 1 & \text{if } j^{th} \text{ sample belongs to the } i^{th}\text{class} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

so that $\mathbf{A}^T\mathbf{B}$ would be as close as possible to $\mathbf{F}$.

Each column of $\mathbf{A}$ consists of $2L+1$ components of features for each training sample. We use the notation $\mathbf{A}^G$ to denote the rows of $\mathbf{A}$ which include the features of component G. Variable G can take values between 1 and $2L+1$ or their corresponding component labels. Correspondingly, columns of $\mathbf{B}$ have the same structure, and we denote the Gth component's parameters as $\mathbf{B}^G$. We refer to the ith column of $\mathbf{B}$ as $\mathbf{b}_i$ which is the projection to our binary classifier for the ith action. Finally $\mathbf{b}_i^G$ refers to the ith column of $\mathbf{B}^G$.

Our classifier is formulated as another optimization problem with the cost function below

$$\begin{aligned} \mathbf{B}^* = \underset{\mathbf{B}}{argmin} & \left\|\mathbf{A}^T\mathbf{B} - \mathbf{F}\right\|_F^2 + \gamma_E \|\mathbf{B}\|_{G_E} \\ & + \gamma_L \|\mathbf{B}\|_{G_L} + \gamma_W \|\mathbf{B}\|_F. \end{aligned} \tag{15}$$

Component-wise regularizer norm, $\|\mathbf{B}\|_{G_E}$, groups the weights of each component by applying a ℓ_2 norm. Then applies the component selection by a ℓ_1 norm over the ℓ_2 values of all components. Mathematically

$$\begin{aligned} \|\mathbf{B}\|_{G_E} &= \sum_{i=1}^{c} \sum_{G=1}^{2L+1} \left\|\mathbf{b}_i^G\right\|_2 \\ &= \sum_{i=1}^{c} \sum_{j=1}^{L} \left(\left\|\mathbf{b}_i^{Z_r^j}\right\|_2 + \left\|\mathbf{b}_i^{Z_d^j}\right\|_2 \right) \\ &\quad + \sum_{i=1}^{c} \left\|\mathbf{b}_i^{Y^L}\right\|_2, \end{aligned} \tag{16}$$

where c is the number of class labels.

This mixed norm dictates the component-wise weight learning regarding their discriminative strength for each action class. Since it applies ℓ_2 norm inside the components and ℓ_1 norm between them, it regularizes the weights within each component, while sparsely selects discriminative components for different classes.

On the other hand, a layer-wise group selection can also be beneficial, because discriminative features may become factorized in some layers of our hierarchical deep network. Based on this intuition, we apply another group sparsity mixed norm to enforce layer selection. Similar to G_E norm, our layer selection norm (G_L) groups the learning parameters corresponding to the components of each layer of the network, and applies ℓ_1 sparsity between them

$$\|\mathbf{B}\|_{G_L} = \sum_{i=1}^{c} \sum_{j=1}^{L} \left\| \begin{bmatrix} \mathbf{b}_i^{Z_r^j} \\ \mathbf{b}_i^{Z_d^j} \end{bmatrix} \right\|_2 + \sum_{i=1}^{c} \left\|\mathbf{b}_i^{Y^L}\right\|_2. \tag{17}$$

The last norm in (15) is a general weight decay regularizer to prevent the entire classifier from overfitting.

Similar to previous section, this optimization is also done using "L-BFGS" algorithm. Upon training the classifier and finding the optimal $\mathbf{B}^*$, we classify each testing sample with exemplar features $\mathbf{a}_q$ as

$$h(\mathbf{a}_q) = \underset{c}{argmax} \langle \mathbf{a}_q, \mathbf{b}_c^* \rangle. \tag{18}$$

5 CCA-RICA Factorization as a Baseline Method

As a baseline to the proposed method to perform the shared-specific analysis of the RGB+D inputs, we combined canonical correlation analysis [13], [14] and reconstruction independent component analysis (RICA) [68], to extract correlated and independent components of input features. In this section we describe this baseline method.

We use the notation $\mathbf{X}_r$ to represent input local RGB features, and $\mathbf{X}_d$ for corresponding local depth features. We define the linear projections of the two input features as

$$\mathbf{Y}_r = \mathbf{W}_{r,c}\mathbf{X}_r, \quad \mathbf{Y}_d = \mathbf{W}_{d,c}\mathbf{X}_d, \tag{19}$$

and to make them maximally correlated we maximize

$$\begin{aligned} \underset{\mathbf{w}_{r,c}^j, \mathbf{w}_{d,c}^j}{maximize}\ & Corr(\mathbf{Y}_r^j, \mathbf{Y}_d^j) \\ & = Corr(\mathbf{w}_{r,c}^j \mathbf{X}_r, \mathbf{w}_{d,c}^j \mathbf{X}_d), \end{aligned} \tag{20}$$

in which superscript j refers to the jth row of the corresponding matrices.

Canonical correlation analysis [13], [14] solves this analytically as an eigenproblem, in which each eigenvector gives one row of the projection and altogether provides the full projection matrices which lead to the maximum correlation between them.

Based on our intuition about insufficiency of shared components for recognition tasks, in the second step, we fix correlation projections ($\mathbf{W}_{r,c}, \mathbf{W}_{d,c}$) and apply a reconstruction independent component analysis formulation [68], to

extract modality-specific components for RGB and depth separately

$$\mathbf{Z}_r = \mathbf{W}_{r,i}\mathbf{X}_r, \quad \mathbf{Z}_d = \mathbf{W}_{d,i}\mathbf{X}_d. \tag{21}$$

For RGB features we optimize

$$\begin{aligned} &\underset{\mathbf{w}_{r,i}}{mininize}\frac{\lambda}{m}\left\|\widetilde{\mathbf{X}}_r - \mathbf{X}_r\right\|_F^2 + \sum_j \left\|\mathbf{W}_{r,i}^j \mathbf{X}_r\right\|_1 \\ &where\ \widetilde{\mathbf{X}}_r = \left[\mathbf{W}_{r,c}^T, \mathbf{W}_{r,i}^T\right]\begin{bmatrix}\mathbf{W}_{r,c}\\ \mathbf{W}_{r,i}\end{bmatrix}\mathbf{X}_r. \end{aligned} \tag{22}$$

Similarly for depth features we optimize

$$\begin{aligned} &\underset{\mathbf{w}_{d,i}}{mininize}\frac{\lambda}{m}\left\|\widetilde{\mathbf{X}}_d - \mathbf{X}_d\right\|_F^2 + \sum_j \left\|\mathbf{W}_{d,i}^j \mathbf{X}_d\right\|_1 \\ &where\ \widetilde{\mathbf{X}}_d = \left[\mathbf{W}_{d,c}^T, \mathbf{W}_{d,i}^T\right]\begin{bmatrix}\mathbf{W}_{d,c}\\ \mathbf{W}_{d,i}\end{bmatrix}\mathbf{X}_d. \end{aligned} \tag{23}$$

Upon convergence of (22) and (23), the RGB+D features of each trajectory (k) can be represented as a quadruple: $\{\mathbf{Z}_r(k), \mathbf{Y}_r(k), \mathbf{Y}_d(k), \mathbf{Z}_d(k)\}$.

6 Experiments

This section presents our experimental setup and the results of the proposed methods on three RGB+D action recognition datasets.

6.1 Experimental Setup

The proposed methods are evaluated on five RGB+D action recognition datasets. All these datasets are collected using the Microsoft Kinect sensor in an indoor environment [79]. This sensor captures RGB videos and depth map sequences, and locates the 3D positions of 20 body joints of actors in the scene.

In our experiments, we try to use features which encode information regarding all the available modalities. From RGB videos, we extract dense trajectories [2] and use HOG, HOF, MBHX, and MBHY features as trajectory descriptors. To encode the global representation of samples based on their trajectories, we use VQ with 2K codewords, for each descriptor. The final representation of each sample video, is the concatenated max-pooled codes of the four descriptors, over 3 levels of the temporal pyramid. For depth sequences, we use the histograms of oriented 4D normals (HON4D) features [6]. To explore different setups on each dataset, we extract this feature in different settings. We describe the details in following sections.

Since the RGB and depth sequences are not fully aligned and not synced in most of the datasets (all the evaluated ones in this paper, except RGBD-HuDaAct), convolutional cubes have to be large enough so that they mostly cover the same parts of the video between the two modalities. To apply the convolutional network, we consider four temporal quarters of the videos. In this way, each input sample has four temporal segments in our convolutional network and the factorized components of all these segments, together with holistic features of the entire sample are considered as the inputs of the stacked network.

To cover various aspects of RGB+D motion and appearances of input samples, we used a combination of different features. For depth channel, we extract Fourier coefficients of the joint locations and local occupancy pattern (LOP) features [80], histogram of oriented 4D normals (HON4D) [6], dynamic skeletons (DS), and dynamic depth patterns (DDP) [49]. From RGB videos, we extract dynamic color patterns (DCP) [49] and dense trajectory features [2].

For depth-based input, we use Fourier coefficients of the joint locations and local occupancy pattern features [80]. The size of $\mathbf{X}_.$, $\mathbf{Y}_.$, and $\mathbf{Z}_.$ vectors is fixed as 100 for local features and 200 for holistic and stacked networks in our experiments. On each of the experiments, the optimal values of gammas in SSLM are found via leave-one-sample-out cross-validation over training samples.

To show the effectiveness of our method, we compare it with two baseline methods below:

Baseline Method 1. Descriptor level fusion. In this method, we concatenate all the input RGB and depth-based features and train a linear SVM for classification.

Baseline Method 2. Kernel level combination. For this baseline method, we calculate the RBF kernel matrices based on all the input RGB and depth-based features and combine them linearly to classify in the form of multi-kernel SVM. We find the weights of kernels via a brute force search in a cross validation setting using training samples [81].

In the following tables, we report the results of our method in two settings:

DSSCA Kernel is the kernel combination of the hierarchically factorized components of our shared-specific analysis network.

DSSCA SSLM. Refers to the proposed structured sparsity learning machine based on the hierarchically factorized components described in Section 4.

It is worth mentioning, there are more than 40 datasets specifically for 3D human action recognition. The survey of Zhang et al. [82] provided a great coverage over the current datasets and discussed their characteristics in different aspects, as well as the best performing methods for each dataset.

6.2 Online RGBD Action Dataset

Online RGBD action dataset [12] is a RGB+D benchmark for action recognition. Unlike most of the other RGB+D benchmarks, this dataset is collected in different locations and provides a cross-environment evaluation setting. It includes samples of seven daily action classes: *drinking*, *eating*, *using laptop*, *reading cellphone*, *making phone call*, *reading book*, and *using remote*. For the recognition task, it provides videos of 24 actors. Each actor performs each of the actions twice. Overall, this dataset include 336 RGB+D video samples. Three different recognition scenarios are defined on this dataset. The first and second scenarios are cross-subject tests. In the first scenario, the first 8 actors are assigned for training and the second eight actors are for testing. The samples of the second scenario are the same as the first one but training and testing samples are swapped. The third scenario is a cross-environment setting. The videos of the third eight actors are collected in another location and are considered as test data. The other 16 actors' videos are used for training. The first and second scenarios are cross-subject and the third is a cross-environment evaluation.

Table 1 compares the results of the deep shared-specific component analysis (DSSCA) and structure sparsity

TABLE 1
Comparison of the Results of Our Methods with the Baselines in Online RGBD Action Dataset

Eval. Dataset	Baseline Method 1	Baseline Method 2	DSSCA Kernel	DSSCA SSLM
Online S1	86.6%	91.1%	92.9%	**95.5%**
Online S2	85.6%	91.0%	91.9%	**93.7%**
Online S3	73.0%	80.2%	82.0%	**83.8%**

S1, S2, and S3 refers to the three different scenarios of the Online RGBD Action dataset. First column shows the performance of descriptor concatenation on all RGB+D input features. Second column reports the accuracy of the kernel combination on the same set of features. Third column shows the result of our correlation-independence analysis. It employs a kernel combination for classification. Last column reports the accuracy of proposed structured sparsity learning machine.

learning machine (SSLM), with baseline methods on this dataset. The results of this experiment show our DSSCA network successfully decompose input features into a more powerful representation which leads into a clear improvement on the classification performance. They also show our SSLM can select the discriminative components and layers and learns a better classifier.

We also compare different structures of our DSSCA network. For each scenario, we report the performance of three structures. "Holistic" refers to the three-layer deep network applied on holistic features. "Local" is the two-layer convolutional network applied on local features. "Stacked local +holistic" is the stacked local and holistic networks, as illustrated in Fig. 3. The results are reported in Table 2. We conclude that the local and holistic features are complementary and applying stacked local+holistic network can improve the final classification accuracy.

In our third experiment on this dataset, performance of the proposed networks is compared with a similar network without modality-specific components. The reference network acts similarly to traditional CCA methods. We compare these two networks on the "local" network of third scenario. The result is shown in Table 3. We can see including independent components is beneficial and improves the accuracy. Performance of the network with these components is clearly higher. The second observation is our method improves the performance more significantly by having multiple layers. Without having the modality-specific components, the values of common components can not change much, on higher layers. This shows our proposed structure is suitable for cascading more layers and decomposing the features layer by layer.

TABLE 2
Performance Comparison for Holistic Network, Local Network, and Stacked Local+Holistic (Fig. 3) Networks on Online RGBD Action Datasets

Evaluation Dataset	Network Structure	DSSCA Kernel	DSSCA SSLM
Online RGB+D Action S1	Holistic	90.2%	92.0%
Online RGB+D Action S1	Local	92.9%	93.8%
Online RGB+D Action S1	Stacked Local+Holistic	92.9%	**95.5%**
Online RGB+D Action S2	Holistic	87.4%	91.0%
Online RGB+D Action S2	Local	88.3%	89.2%
Online RGB+D Action S2	Stacked Local+Holistic	91.9%	**93.7%**
Online RGB+D Action S3	Holistic	79.3%	82.0%
Online RGB+D Action S3	Local	75.7%	77.5%
Online RGB+D Action S3	Stacked Local+Holistic	82.0%	**83.8%**

Reported are the results of our method using kernel combination and SSLM.

TABLE 3
Comparison with a Correlation Network (without Modality-Specific Components) on the Online RGBD Action Dataset, Local Network, Scenario 3

Network Description	Layer 1 SSLM	2 Layers SSLM
Local Without **Z**	73.0%	73.9%
Local With **Z**	76.6%	**77.5%**

Without **Z** *components, the network is limited to the shared ones and acts similar to CCA.*

Table 4 compares our results with the state-of-the-art method on this dataset. Due to the recency of this dataset, only two other works reported results on this dataset. As shown, our method outperforms their results with a large margin, which demonstrates the importance of RGB+D fusion for action recognition as well as the effectiveness of our proposed method for this task.

6.3 MSR-DailyActivity3D Dataset

MSR-DailyActivity dataset [80] is among the most challenging RGB+D benchmarks for action recognition, which has a high level of intra-class variation and a large number of action classes. It provides 320 RGB+D samples, from 16 classes of daily activities: *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,* and *sit down*. Each action is done by 10 actors, twice by each actor. The standard evaluation on this dataset is defined on a cross-subject setting: first five subjects are used for training and others for testing. Results of the experiments on this benchmark are reported in Tables 5 and 6.

Table 7 also shows the accuracy comparison between the proposed method and the state-of-the-art results reported on this benchmark, in which we reduced the error rate by more than 40 percent compared to the best reported results so far. This shows our RGB+D analysis method can effectively improve the performance of the action recognition system.

TABLE 4
Performance Comparison of Proposed DSSCA with the State-of-the-Art Results on Online RGBD Action Dataset

Methods	Setup	Accuracy
HOSM [83]	Same environment	49.5%
Orderlet [12]	Same environment	71.4%
Meng et al. [43]	Same environment	75.8%
Proposed DSSCA-SSLM	Same environment	**94.6%**
HOSM [83]	Cross environment	50.9%
Orderlet [12]	Cross environment	66.1%
Proposed DSSCA-SSLM	Cross environment	**83.8%**

Same environment setup is the average of S1 and S2 scenarios, and cross environment setup is the same as S3 scenario.

TABLE 5
Comparison of the Results of Our Methods with the Baselines in MSR-DailyActivity3D Dataset

Eval. Dataset	Baseline Method 1	Baseline Method 2	DSSCA Kernel	DSSCA SSLM
Daily	91.9%	94.4%	96.3%	**97.5%**

TABLE 6
Performance Comparison for Holistic Network, Local Network, and Stacked Local+Holistic (Fig. 3) Networks on MSR-DailyActivity3D Dataset

Evaluation Dataset	Network Structure	DSSCA Kernel	DSSCA SSLM
MSR Daily Activity 3D	Holistic	95.0%	96.3%
MSR Daily Activity 3D	Local	95.0%	96.9%
MSR Daily Activity 3D	Stacked Local+Holistic	96.3%	**97.5%**

Reported are the results of our method using kernel combination and SSLM.

TABLE 7
Performance Comparison of the Proposed Multimodal DSSCA with the State-of-the-Art Methods on MSR-DailyActivity Dataset

Method	Accuracy
HoDG-RDF [42]	74.5%
Bag-of-FLPs [35]	78.8%
HON4D [6]	80.0%
SSFF [56]	81.9%
ToSP [48]	84.4%
RGGP [47]	85.6%
Actionlet [9]	85.8%
SVN [84]	86.3%
BHIM [50]	86.9%
DCSF+Joint [11]	88.2%
MMTW [10]	88.8%
HOPC [28]	88.8%
Depth Fusion [45]	88.8%
MMMP [41]	91.3%
DL-GSGC [29]	95.0%
JOULE-SVM [49]	95.0%
Range-Sample [27]	95.6%
Proposed DSSCA-SSLM	**97.5%**

6.4 3D Action Pairs Dataset

3D Action Pairs dataset [6] is a less challenging RGB+D dataset for action recognition. This dataset provides six pairs of action classes: *pick up a box/put down a box*, *lift a box/place a box*, *push a chair/pull a chair*, *wear a hat/take off a hat*, *put on a backpack/take off a backpack*, and *stick a poster/remove a poster*. Each pair of the classes have almost the same set of body motions but in different temporal order. Each action class is captured from 10 subjects, each one 3 times. Overall, this dataset includes 360 RGB+D video samples. The first five subjects are kept for testing and others are for training.

Tables 8, 9, and 10 compare the accuracies between the proposed framework, baselines and the state-of-the-art methods reported on this benchmark. Our method ties with two recent works (MMMP [41], and BHIM [50]) in saturating the benchmark by achieving the flawless 100 percent accuracy on this dataset.

TABLE 8
Comparison of the Results of Our Methods with the Baselines in 3D Action Pairs Dataset

Eval. Dataset	Baseline Method 1	Baseline Method 2	DSSCA Kernel	DSSCA SSLM
Pairs	97.7%	98.3%	**100.0%**	**100.0%**

TABLE 9
Performance Comparison for Holistic Network, Local Network, and Stacked Local+Holistic (Fig. 3) Networks on 3D Action Pairs Dataset

Evaluation Dataset	Network Structure	DSSCA Kernel	DSSCA SSLM
3D Action Pairs	Holistic	98.9%	99.4%
3D Action Pairs	Local	99.4%	98.9%
3D Action Pairs	Stacked Local+Holistic	**100.0%**	**100.0%**

Reported are the results of our method using kernel combination and SSLM.

TABLE 10
Performance Comparison of Proposed Multimodal Correlation-Independence Analysis with the State-of-the-Art Methods on 3D Action Pairs Dataset

Method	Accuracy
DHOG [85]	66.11%
Bag-of-FLPs [35]	75.56%
Actionlet [9]	82.22%
HON4D [6]	96.67%
MMTW [10]	97.22%
HOG3D-LLC [86]	98.33%
HOPC [28]	98.33%
SVN [84]	98.89%
MMMP [41]	100.0%
BHIM [50]	100.0%
Proposed DSSCA-SSLM	**100.0%**

6.5 NTU RGB+D Dataset

NTU RGB+D [40] is one of the largest scale benchmark dataset for 3D action recognition. It provided 56,880 RGB+D video samples of 60 distinct actions. The 60 action classes in NTU RGB+D dataset are: *drinking, eating, brushing teeth, brushing hair, dropping, picking up, throwing, sitting down, standing up (from sitting position), clapping, reading, writing, tearing up paper, wearing jacket, taking off jacket, wearing a shoe, taking off a shoe, wearing on glasses, taking off glasses, puting on a hat/cap, taking off a hat/cap, cheering up, hand waving, kicking something, reaching into self pocket, hopping, jumping up, making/answering a phone call, playing with phone, typing, pointing to something, taking selfie, checking time (on watch), rubbing two hands together, bowing, shaking head, wiping face, saluting, putting palms together, crossing hands in front. sneezing/coughing, staggering, falling down, touching head (headache), touching chest (stomachache/heart pain), touching back (back-pain), touching neck (neck-ache), vomiting, fanning self. punching/slapping other person, kicking other person, pushing other person, patting other's back, pointing to the other person, hugging, giving something to other person, touching other person's pocket, handshaking, walking towards each other, and walking apart from each other.*

TABLE 11
Comparison of the Result of Our Method with the Baseline for the Cross-Subject Evaluation Criteria of NTU RGB+D Dataset

Eval. Dataset	Baseline Method 1	DSSCA SSLM
NTU RGB+D	59.7%	**74.9%**

Unlike other evaluated datasets, NTU RGB+D is collected by Microsoft Kinect v.2. Therefore, its skeletal data includes more body joints and is more accurate. For our experiments in this section, we limit the depth-based features to Fourier temporal pyramids over skeletons, HON4D and LOP. For RGB-based inputs we use the same set of features used for the other datasets.

This dataset suggested two evaluation criteria, cross-subject and cross-view. For the cross-view evaluation, our set of RGB based features perform very poorly and could not contribute powerful enough in the proposed multimodal analysis. Therefore, we evaluate the proposed DSSCA-SSLM framework only on the cross-subject criterion of this dataset.

Due to the large size of training video samples in this dataset, evaluation of the kernel-based methods (both baseline method 2 and DSSCA-kernel) were not tractable and we only reported the results for baseline method 1 and DSSCA-SSLM frameworks, as provided in Tables 11 and 12. Table 13 compares the performance of the proposed framework in comparison with other state-of-the-art on this benchmark.

6.6 RGBD-HuDaAct Dataset

RGBD-HuDaAct [46] is a large size benchmarks for human daily action recognition in RGB+D. This dataset includes 1189 RGB+D video sequences from 13 action classes: *exit the room*, *make a phone call*, *get up from bed*, *go to bed*, *sit down*, *mop floor*, *stand up*, *eat meal*, *put on jacket*, *drink water*, *enter room*, *take off jacket*, and *background activity*. The standard evaluation on this dataset is defined on a leave-one-subject-out cross-validation setting. In our experiments we follow the evaluation setup described in [46].

6.6.1 Atomic Local Level Feature Analysis

Unlike most of the other datasets, this benchmark provides fully synchronized and aligned set of RGB and depth videos. This important characteristic enables us to apply the atomic level of analysis on local RGB and depth features within the video samples.

TABLE 12
Performance Comparison for Holistic Network, Local Network, and Stacked Local+Holistic (Fig. 3) Networks on the Cross-Subject Evaluation Criteria of NTU RGB+D Dataset

Evaluation Dataset	Network Structure	DSSCA SSLM
NTU RGB+D	Holistic	70.4%
NTU RGB+D	Local	66.4%
NTU RGB+D	Stacked Local+Holistic	**74.9%**

Reported are the results of our method using SSLM.

TABLE 13
Performance Comparison of Proposed Multimodal Correlation-Independence Analysis with the State-of-the-Art Methods on the Cross-Subject Evaluation Criteria of NTU RGB+D Dataset

Method	Cross-Subject Accuracy
HOG^2 [87]	32.24%
Super Normal Vector [84]	31.82%
HON4D [6]	30.56%
Lie Group [8]	50.08%
Skeletal Quads [30]	38.62%
FTP Dynamic Skeletons [49]	60.23%
HBRNN-L [38]	59.07%
P-LSTM [40]	62.93%
ST-LSTM [65]	69.20%
Proposed DSSCA - SSLM	**74.86%**

As our atomic local level features, we extract the tracked dense trajectories [2] in RGB sequences and their HOG, HOF, MBHX, and MBHY descriptors from both modalities.

To evaluate the effectiveness of the proposed RGB+D analysis, we apply a single layer SSCA to decompose RGB and depth descriptors of the trajectories to their correlated and independent components. For training stage, we sample a set of 40K trajectories from training set. The output of the analysis, which are four factorized components for each trajectory are clustered separately by K-Means with codebook size 1 K. LLC coding [89] and BOF framework are applied on the codes of all the trajectories from each RGB +D video sample to extract their global representations.

In the final step, a linear SVM is used as the action classifier trained on the extracted global representations of the action video samples.

We evaluated the performance of canonical correlation analysis method also. As a better baseline, we also evaluated added independent components. In our implementation of the CCA-RICA method (Section 5) we used the provided codes by the authors of [90] for CCA and [68] for RICA.

All the optimizations in our experiments, are done using "L-BFGS" algorithm. We use the off-the-shelf "minFunc" software released by [78].

Table 16 shows the results of all the experiments described in this section and compares them with other state-of-the-art methods.

At first, we evaluated the performance of correlated components of CCA without any modality specific features, which achieves 93.9 percent outperforming all the reported results on this benchmark. Compared to the accuracy of RGB+D linear coding [53], which has the most similar pipeline of action recognition to ours, CCA components shows about two percents improvement. This approves the robustness of shared components and their advantage over using a simple combination of features from the two modalities.

In the next step, we apply RICA to extract modality-specific components for RGB and depth local features. Adding specific components improves the accuracy of the classification by 2.5 more percents. This supports our argument about the importance of modality-specific components and their discriminative strengths for action classification. The confusion matrix for this method is illustrated in Fig. 4. The majority of the misclassification are caused by the

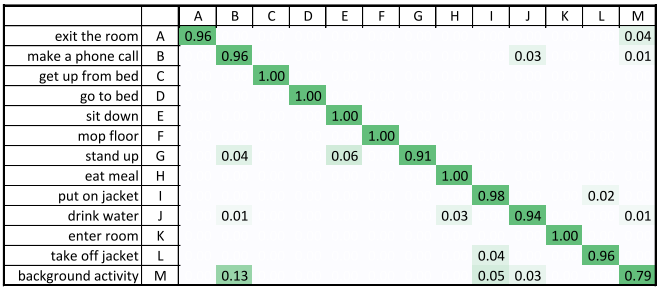


Fig. 4. Confusion matrix for CCA-RICA method on atomic local level features RGBD-HuDaAct dataset. Ground truth action labels are on rows and detections are on columns of the grid.

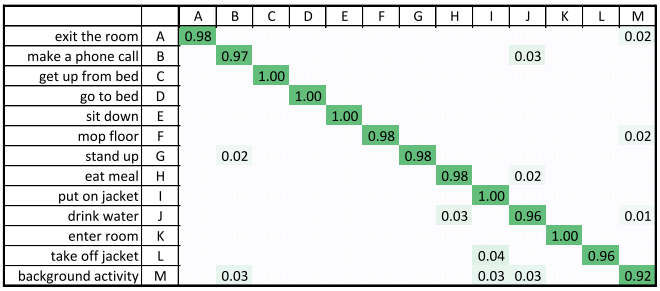


Fig. 5. Confusion matrix for SSCA method on atomic local level features of RGBD-HuDaAct dataset. Ground truth action labels are on rows and detections are on columns of the grid.

TABLE 14
Comparison of the Results of Our Methods with the Baselines on RGBD-HuDaAct Dataset

Eval. Dataset	Baseline Method 1	Baseline Method 2	DSSCA Kernel	DSSCA SSLM
HuDaAct	95.1%	97.6%	98.3%	**99.0%**

First column shows the performance of descriptor concatenation on all RGB+D input features. Second column reports the accuracy of the kernel combination on the same set of features. Third column shows the result of our correlation-independence analysis. It employs a kernel combination for classification. Last column reports the accuracy of proposed structured sparsity learning machine.

background activity class. This class contains samples of random motion and other simple activities which are not covered by other 12 classes, like walking around or stay seated without much of motion. Therefore it is inevitable to have some confusion between this class with classes which contain very small amount of clear motion e.g., making a phone call. Similar action classes with reverse temporal order are also mixed up, e.g., sit down and stand up, or put on jacket and take off jacket classes have the same appearance within individual frames, and their only differences are the arrangement of frames over time.

Next, we evaluate the proposed SSCA method on this atomic local level. SSCA outperforms all other techniques by performing 97.9 percent of correct classification and achieves the state-of-the-art accuracy on this dataset. Compared to CCA-RICA method, SSCA improves the error rate by more than 40 percent which is a notable improvement. The confusion matrix of this experiment is also reported in Fig. 5. Compared to the mixed-up cases of the CCA-RICA method (Fig. 4), the confusion patterns are similar but furthered improved.

TABLE 15
Performance Comparison for Holistic Network, Local Network, and Stacked Local+Holistic Networks on RGBD-HuDaAct Dataset

Scenario Number	Network Structure	DSSCA Kernel	DSSCA SSLM
HuDaAct	Holistic	98.3%	**99.0%**
HuDaAct	Local	98.7%	98.7%
HuDaAct	Stacked Local+Holistic	98.3%	**99.0%**

Reported are the results of our method using kernel combination and SSLM.

TABLE 16
Performance Comparison on RGBD-HuDaAct Dataset

Method	Accuracy
3D-MHIs [46]	70.5%
iM^{2}EDM [88]	76.8%
MF-HMM [55]	78.6%
DLMC-STIPs [46]	81.5%
DIMC-STIPs [52]	87.7%
STIP HOGHOF+LDP [51]	89.1%
Part-based BOW-Pyramid [54]	91.7%
RGB+D Linear Coding [53]	92.0%
CCA (Atomic Level)	93.9%
CCA-RICA (Atomic Level)	96.4%
Proposed Single Unit SSCA (Atomic Level)	**97.9%**
Proposed DSSCA-SSLM (Global Level)	**99.0%**

6.6.2 *Global Level Feature Analysis*

Similar to other datasets reported in the paper, we perform the proposed RGB+D analysis on the global representations extracted from input samples. For RGB signals, the features are HOG, HOF, MBHX, and MBHY descriptors of dense trajectories [2], followed by a K-means clustering and locality-constrained linear coding (LLC) [89] to calculate their global representations as bags-of-features. For depth, we extract HON4D features [6] for holistic and local depth based features. The results of this experiment are reported in Tables 14 and 15 in a similar evaluation setup to other datasets.

As can be seen in Table 16, applying DSSCA analysis in a deep and stacked framework outperforms all the current methods as well as the atomic local level analysis, and achieved the outstanding performance of 99.0 percent on this benchmark, which shows more than 50 percent improvement on the error rate compared to the atomic local level SSCA analysis.

Other reported results are also in accord with our results on other datasets and approve our arguments about the effectiveness of the the proposed framework.

6.7 Comparison with Single Modality

In Table 17, we compare our method with baseline method 2, based on single modality features. Since each modality also has holistic and multiple local features, we perform baseline kernel combination to produce the results. For a fair comparison, we use kernel combination for classification based on our factorized components. It is not surprising to observe our method outperforms the baseline, since ours integrates RGB and depth information effectively.

TABLE 17
Comparison Between Our Method and Baseline Method 2 on Single Modality RGB and Depth Based Input Features, on All the Datasets

Method	MSR Daily Activity 3D	3D Action Pairs	RGBD HuDaAct	Online RGBD S1	Online RGBD S2	Online RGBD S3
Baseline 2 on RGB-based Local+Holistic	89.4%	97.7%	95.2%	81.3%	85.6%	75.7%
Baseline 2 on depth-based Local+Holistic	92.5%	97.7%	79.1%	85.7%	84.7%	66.7%
Ours Kernel	**96.3%**	**100.0%**	**98.3%**	**92.9%**	**91.9%**	**82.0%**

6.8 Analysis of Component Contributions in the Classifier

Table 18 shows the proportion of the weights assigned by SSLM to the factorized components of the stacked local +holistic networks. The weights of $\mathbf{Y}^3$ are relatively high, which supports our initial argument about robustness and discriminative properties of the shared factorized components. The Z components of the both modalities in all three layers also gain weights, which shows they also carry informative features and are complementary for the classification. The reported values in this table shows how discriminative are the factorized features inside each of these components. As can be seen, some components achieve very low (close to zero) values. They hold important components in the distribution of the input multimodal data regardless of the action labels. However, regarding the action classification task, they don't have considerable correlation with action labels and cannot contribute very much in classification, so they gain very low weights.

7 Conclusion

This paper presents a new deep learning framework for a hierarchical shared-specific component factorization, to analyze RGB+D features of human action videos. Each layer of the proposed network is an autoencoder based component factorization unit, which decomposes its multimodal input features into common and modality-specific parts. We further extended our deep factorization framework by applying it in a convolutional setting.

In addition, we proposed a structured sparsity based classifier which utilizes mixed norms to apply component and layer selection for a proper fusion of decomposed feature components.

Provided experimental results on five RGB+D action recognition datasets show the strength of our deep shared-specific component analysis and the proposed structured sparsity learning machine by achieving the state-of-the-art performances on all the reported benchmarks.

TABLE 18
Proportion of the Weights to Factorized Components in SSLM Classifier for Online RGBD, MSR-DailyActivity3D, and 3D Action Pairs Datasets

Dataset	$\mathbf{Z}_r^1$	$\mathbf{Z}_r^2$	$\mathbf{Z}_r^3$	$\mathbf{Y}^3$	$\mathbf{Z}_d^3$	$\mathbf{Z}_d^2$	$\mathbf{Z}_d^1$
Online S1	0.12	0.13	0.18	0.20	0.13	0.05	0.18
Online S2	0.29	0.06	0.03	0.42	0.06	0.11	0.03
Online S3	0.14	0.12	0.06	0.26	0.13	0.00	0.28
Daily	0.22	0.05	0.07	0.23	0.08	0.08	0.28
Pairs	0.06	0.02	0.16	0.42	0.01	0.03	0.29

Reported values are the ℓ_2 norms of all the corresponding weights to each of the components, learned by SSLM on the stacked local+holistic networks.

Acknowledgments

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme. Gang Wang is the corresponding author.

References

[1] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318–1334, Oct. 2013.

[2] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3551–3558.

[3] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *CoRR*, vol. abs/1405.4506, 2014, http://arxiv.org/abs/1405.4506

[4] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. 27th Int. Conf. Advances Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[5] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. 27th Int. Conf. Advances Neural Inform. Process. Syst.*, 2014, pp. 2366–2374.

[6] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 716–723.

[7] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 742–757.

[8] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a Lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 588–595.

[9] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 914–927, May 2014.

[10] J. Wang and Y. Wu, "Learning maximum margin temporal warping for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2688–2695.

[11] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2834–2841.

[12] G. Yu, Z. Liu, and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 50–65.

[13] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, pp. 2639–2664, 2004.

[14] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[15] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *Int. J. Neural Syst.*, vol. 10, pp. 365–377, 2000.

[16] Z. Cai, L. Wang, X. Peng, and Y. Qiao, "Multi-view super vector for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 596–603.

[17] M. Salzmann, C. H. Ek, R. Urtasun, and T. Darrell, "Factorized orthogonal latent spaces," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 701–708.

[18] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[19] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *J. Mach. Learn. Res.*, vol. 15, pp. 2949–2980, 2014.

[20] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[21] Y. Jia, M. Salzmann, and T. Darrell, "Factorized latent spaces with structured sparsity," in *Proc. Advances Neural Inf. Process. Syst.*, 2010, pp. 982–990.

[22] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," 2005.

[23] H. Wang, F. Nie, W. Cai, and H. Huang, "Semi-supervised robust dictionary learning via efficient $l_{2,0^+}$-norms minimization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1145–1152.

[24] H. Wang, F. Nie, and H. Huang, "Multi-view clustering and feature learning via structured sparsity," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. III-352–III-360.

[25] H. Wang, F. Nie, and H. Huang, "Robust and discriminative self-taught learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 298–306.

[26] H. Wang, F. Nie, H. Huang, and C. Ding, "Heterogeneous visual features fusion via sparse multimodal machine," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3097–3102.

[27] C. Lu, J. Jia, and C.-K. Tang, "Range-sample depth feature for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 772–779.

[28] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2430–2443, Dec. 2016.

[29] J. Luo, W. Wang, and H. Qi, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1809–1816.

[30] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4513–4518.

[31] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, pp. 107–123, 2005.

[32] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Trans. Human Mach. Syst.*, vol. 46, no. 4, pp. 498–509, Aug. 2015.

[33] H. Rahmani and A. Mian, "Learning a non-linear knowledge transfer model for cross-view action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2458–2466.

[34] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 99, p. 1, 2017, doi: 10.1109/TPAMI.2017.2691768.

[35] P. Wang, W. Li, P. Ogunbona, Z. Gao, and H. Zhang, "Mining mid-level features for action recognition based on effective skeleton representation," in *Proc. Int. Conf. Digit. Image Comput.: Techn. Appl.*, 2014, pp. 1–8.

[36] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4041–4049.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[38] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1110–1118.

[39] W. Zhu, et al., "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3697–3703.

[40] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1010–1019.

[41] A. Shahroudy, T. T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2123–2129, Oct. 2016.

[42] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Real time action recognition using histograms of depth gradients and random decision forests," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2014, pp. 626–633.

[43] M. Meng, H. Drira, M. Daoudi, and J. Boonaert, "Human-object interaction recognition by learning the distances between the object and the skeleton joints," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, May 2015, vol. 7, pp. 1–6.

[44] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *CoRR*, vol. abs/1601.01006, 2016, http://arxiv.org/abs/1601.01006

[45] Y. Zhu, W. Chen, and G. Guo, "Fusing multiple features for depth-based action recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 6, 2015, Art. no. 18.

[46] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 1147–1153.

[47] L. Liu and L. Shao, "Learning discriminative representations from RGB-D video data," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1493–1500.

[48] Y. Song, S. Liu, and J. Tang, "Describing trajectory of surface patch for human action recognition on RGB and depth videos," *IEEE Signal Process. Lett.*, vol. 22, no. 4, pp. 426–429, Apr. 2015.

[49] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 5344–5352.

[50] Y. Kong and Y. Fu, "Bilinear heterogeneous information machine for RGB-D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1054–1062.

[51] Y. Zhao, Z. Liu, L. Yang, and H. Cheng, "Combing RGB and depth map features for human activity recognition," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2012, pp. 1–4.

[52] Z. Y. Zhao Runlin, "Depth induced feature representation for 4D human activity recognition," *Comput. Model. New Technol.*, vol. 18 (12C), pp. 419–423, 2014.

[53] H. Liu, M. Yuan, and F. Sun, "RGB-D action recognition using linear coding," *Neurocomputing*, vol. 149, pp. 79–85, 2015.

[54] J.-S. Tsai, Y.-P. Hsu, C. Liu, and L.-C. Fu, "An efficient part-based approach to action recognition from RGB-D video with bow-pyramid representation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 2234–2239.

[55] D. Kosmopoulos, P. Doliotis, V. Athitsos, and I. Maglogiannis, "Fusion of color and depth video for human behavior recognition in an assistive environment," in *Proc. 1st Int. Conf. Distrib. Ambient Pervasive Interactions*, 2013, pp. 42–51.

[56] A. Shahroudy, G. Wang, and T.-T. Ng, "Multi-modal feature fusion for action recognition in RGB-D sequences," in *Proc. Int. Symp. Commun. Control Signal Process.*, 2014, pp. 1–4.

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014, http://arxiv.org/abs/1409.1556

[59] C. Szegedy, et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[60] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, and G. Wang, "Recent advances in convolutional neural networks," *CoRR*, vol. abs/1512.07108, 2015, http://arxiv.org/abs/1512.07108

[61] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4694–4702.

[62] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4305–4314.

[63] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "DevNet: A deep event network for multimedia event detection and evidence recounting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2568–2577.

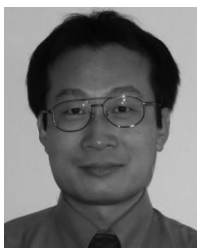
[64] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497.

[65] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 816–833.

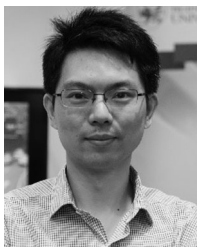
[66] H. Rahmani and A. Mian, "3D action recognition from novel viewpoints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1506–1515.
[67] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, pp. 60–79, 2013.
[68] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Proc. 24th Int. Conf. Advances Neural Inf. Process. Syst.*, 2011, pp. 1017–1025.
[69] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Proc. Int. Conf. Advances Neural Inf. Process. Syst.*, 2008, pp. 873–880.
[70] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," in *Proc. Advances Neural Inf. Process. Syst.*, 2007, pp. 1137–1144.
[71] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.
[72] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, 1–8.
[73] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2007, pp. 153–160.
[74] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, 2006.
[75] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 609–616.
[76] Q. Le, W. Zou, S. Yeung, and A. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3361–3368.
[77] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1794–1801.
[78] M. Schmidt, "Minfunc," 2005. [Online]. Available: http://www.cs.ubc.ca/schmidtm/Software/minFunc.html
[79] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, Feb. 2012.
[80] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1290–1297.
[81] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2007, pp. 401–408.
[82] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "RGB-D-based action recognition datasets: A survey," arXiv, 2016. [Online]. Available: http://arxiv.org/abs/1601.05511
[83] W. Ding, K. Liu, F. Cheng, and J. Zhang, "Learning hierarchical spatio-temporal pattern for human activity prediction," *J. Visual Commun. Image Representation*, vol. 35, pp. 103–111, 2016.
[84] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 804–811.
[85] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 1057–1060.
[86] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Action classification with locality-constrained linear coding," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 3511–3516.
[87] E. Ohn-Bar and M. Trivedi, "Joint angles similarities and hog^2 for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2013, pp. 465–470.
[88] S. Chatzis, "Infinite Markov-switching maximum entropy discrimination machines," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 729–737.
[89] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3360–3367.
[90] M. Borga, "Canonical correlation: A tutorial," *Online Tutorial*, 2001. [Online]. Available: http://people.imt.liu.se/magnus/cca

Amir Shahroudy received the BSc degree in computer engineering from Isfahan University of Technology, Iran, in 2004 and the MSc degree in artificial intelligence from Sharif University of Technology, Iran, in 2006. He is currently a member of the research teams in the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, Singapore, and the Institute for Infocomm Research ($\mathrm{I}^2\mathrm{R}$), A*STAR, Singapore. He is working toward the PhD degree. His research interests include video analysis, human action recognition, 3D vision, and deep learning. He is a student member of the IEEE.

Tian-Tsong Ng received the MPhil degree in signal processing from Cambridge University, in 2001 and the PhD degree in electrical engineering from Columbia University, in 2007. He is currently a deputy department head in the Institute for Infocomm Research. His research focuses on digital image forensics, computer vision, and computational photography. He won the Microsoft Best Student Paper Award at ACM Multimedia Conference in 2005. He is a Commonwealth Scholar and an A*STAR Overseas Graduate Scholar. He is a member of the IEEE.



Yihong Gong received the BS, MS, and PhD degrees in electrical and electronic engineering from the University of Tokyo, in 1987, 1989, and 1992, respectively. He then joined Nanyang Technological University of Singapore, where he worked as an assistant professor in the School of Electrical and Electronic Engineering for four years. From 1996 to 1998, he worked for the Robotics Institute, Carnegie Mellon University as a project scientist. In 1999, he joined NEC Laboratories America, and worked as a group leader, department head and site manager at the Cupertino Branch of the Labs. In 2012, he joined Xi'an Jiaotong University in China, and became a distinguished professor of the National Thousand Talents Program, the chief scientist and vice director of the National Engineering Laboratory for Visual Information Processing. His research interests include computer vision, pattern recognition, machine learning, and multimedia content analysis. He is a senior member of the IEEE.



Gang Wang received the BEng degree in electrical engineering from the Harbin Institute of Technology and the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign. He is an associate professor in the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore. He had a joint appointment at the Advanced Digital Science Center (operated by UIUC) as a research scientist from 2010 to 2014. His research interests include deep learning, scene parsing, object recognition, and action analysis. He is selected as a MIT Technology Review innovator under 35 for Southeast Asia, Australia, New Zealand, and Taiwan. He is also a recipient of Harriett & Robert Perry Fellowship, CS/AI award, best paper awards from PREMIA (Pattern Recognition and Machine Intelligence Association) and top 10 percent paper awards from MMSP. He is an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and an area chair of ICCV 2017. He is a member of the IEEE.