Multi-Modal Human Action Recognition With Sub-Action Exploiting and Class-Privacy Preserved Collaborative Representation Learning

CHENGWU LIANG1,2, DEYIN LIU3,4, LIN QI1, AND LING GUAN3, (Fellow, IEEE)
1School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China
2School of Electrical and Control Engineering, Henan University of Urban Construction, Pingdingshan 467036, China
3Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada
4School of Information Technology and Electrical Engineering, The University of Queensland (UQ), Brisbane, QLD 4072, Australia
Corresponding author: Deyin Liu (iedyzzu@outlook.com)

This work was supported in part by the Key Program of National Natural Science Foundation of China (NSFC) under Grant 61331021, in part by the Program of NSFC under Grant U1804152, and in part by the Canada Research Chair Program.

ABSTRACT Multimodal human action recognition with depth sensors has drawn wide attention, due to its potential applications such as health-care monitoring, smart buildings/home, intelligent transportation, and security surveillance. As one of the obstacles of robust action recognition, sub-actions sharing, especially among similar action categories, makes human action recognition more challenging. This paper proposes a segmental architecture to exploit the relations of sub-actions, jointly with heterogeneous information fusion and Class-privacy Preserved Collaborative Representation (CPPCR) for multi-modal human action recognition. Specifically, a segmental architecture is proposed based on the normalized action motion energy. It models long-range temporal structure over video sequences to better distinguish the similar actions bearing sub-action sharing phenomenon. The sub-action based depth motion and skeleton features are then extracted and fused. Moreover, by introducing within-class local consistency into Collaborative Representation (CR) coding, CPPCR is proposed to address the challenging sub-action sharing phenomenon, learning the high-level discriminative representation. Experiments on four datasets demonstrate the effectiveness of the proposed method.

INDEX TERMS Action recognition, feature fusion, class-privacy preserved, sub-action sharing.

I. INTRODUCTION

"Human action or activity recognition has played significant roles in many potential applications, including security surveillance, human-computer interaction (HCI), health monitoring and intelligent transportation [1]–[6]. For instance, in healthcare environments, by monitoring the behavior of people and recognizing human activities, the activity habits and patterns of people can be understood. Thus correct emergency decisions can be made so that a healthier and more secure living environment can be created for the community. Human action recognition involves specific tasks such as action detection, localization and action recognition from different data modalities with RGB, Depth, infrared or inertial sensors. Normally, action recognition is to classify a video or data sequence into one of the pre-defined action categories, whereas action detection is to determine the presence of the interested action in continuous untrimmed data streams. Action localization aims to find the potential proposals which contain certain human movements, i.e., the time and area that an action of interest happens.

For an intelligent machine to achieve the level of action recognition like humans do, the first is the representation capability, i.e., the ability to perceive the informative observations (features) from multi-modal data. Based on these observations, a feature space is learned with a good capacity. This feature space receives and stores distinct characteristics of the objects of interest, which needs representation learning methods. Multimodal observations may facilitate the level of capacity of receiving useful information and the level of receptivity for impressions.

Earlier action or activity recognition researches focus more on the using of RGB video captured by conventional RGB cameras [8], [9]. The limitations of using RGB cameras is
that conventional RGB images lack 3D action data, which is regarded as one critical clue to improve the recognition performance. The advancement of sensor technology makes it possible to sense 3D action data by new depth sensors. In addition, 3D depth images provide a way to acquire 3D skeleton of a person for better action or activity recognition. For human action recognition, RGB-Depth video, and the skeleton pose sequence enrich the representation space of human motions. Figure 1 shows the RGB-Depth and the skeleton samples. Depth sensors-based action recognition has been studied for years [10]–[13], which is the focus of this paper.

Depth sensor-based human action recognition provides new opportunities whereas faces some challenges. Firstly, one important observation of action datasets is that different actions have irrelevant/similar actions and therefore certain actions from different classes share similar or even same characteristics. As shown in Figure 1, in SBU action dataset [7], sub-actions are shared between different action categories, especially among the similar categories “Punching”, “Hugging” and “Departing”. This is truly ubiquitous in different action datasets. This common phenomenon makes the human action recognition confusing and challenging. In another MSR Action 3D dataset [14], the existing literatures [10], [15]–[17] show that the lower accuracy always happened among three very similar actions, No.4 action “Hand catch”, No.7 action “Draw x and” No.9 action “Draw circle”. The reason is that they share sub-actions, leading to low recognition performance.

Secondly, as shown in Figure 1 (the last row), in the SBU-Kinect interaction dataset [7], the ground truth action category “Hugging” is similar and could be confused with “Approaching” and “Departing”. Action category “Pushing” is a composition of sub-actions from “Kicking”, “Approaching” and “Departing”. Consequently, other samples from “class i” may be represented “collaboratively” with the assistance of samples from other classes, as demonstrated in [18]. Therefore, with the potential assistance of samples from other classes, samples from distinct classes may be “collaboratively” expressed. Thus the feature of a testing video can be coded collaboratively on the globally shared dictionary (i.e., the entire dictionary constructed from the samples of all categories).

Thirdly, the feature space constructed from multi-modal sources is high-dimensional and not uniform distribution. The action performing subjects have their own uncontrolled freedom and behavior habits. This results in larger within-class variations, making human action recognition more difficult than image classification tasks. Sparse representation (SR) has been successful for RGB-based action recognition [3], [19], [20]. The key assumptions of SR are: the features or representation of each category of training data is sufficient enough to span a separable subspace; and the training data are collected carefully, making the extracted feature space distribute uniformly. These preconditions limit their generalization to video analytic tasks. For example the gaming action dataset, UTD-MHAD-Kinect V2 [21] is a multi-modal dataset. It is a typical small-training sample-size dataset, likely causing unacceptable representation errors and unstable classification results when applying strong supervised approaches.

In this paper, to increase the capacity of perceiving information, two heterogeneous low-level features are extracted from depth and skeleton modalities, respectively. Then
Canonical Correlations Analysis (CCA) is utilized for features correlation analysis, providing us with compact and shared mid-level heterogeneous features. The ubiquitous sub-action sharing challenge is regarded as an opportunity and is exploited by the proposed sub-action segmentation method and Class-privacy Preserved Collaborative Representation (CPPCR) learning method. In CCA feature space, CPPCR integrates the low-dimensional manifold (local consistency) into the collaborative sub-action learning process (globality), obtaining the final high-level discriminative representation. CPPCR not only leverages the collaborative representation to address the challenge of sub-action sharing phenomenon among different classes globally, but also preserves the expected local geometric structures of action classes. The analytical solution of CPPCR makes the computation efficient and avoids being trapped in local optima.

The main contributions of the proposed method are summarized as follows:

1) We propose an energy guided sub-action segmentation method based on which the input activity is decomposed into an unfixed number of temporally segmented sub-activities. Accordingly the depth features are extracted based on the new energy-guided sub-actions. In addition, guided by the proposed motion energy function generated from depth modality, the “Cross-modality Parameters Transferring” transfers the sub-action segmentation parameters into the synchronous skeleton modality for heterogeneous feature representation.

2) The proposed CPPCR is demonstrated to be an effective scheme for addressing sub-action sharing problem. CPPCR integrates local consistency into the collaborative sub-action learning process, alleviating the adverse effect caused by sub-action sharing, which leads to a final high-level discriminative representation. CPPCR demonstrates not only performance improvement over the CR learning process, but also efficient computation with the closed-form solution.

3) For human action recognition aiming to address the sub-action sharing phenomenon, the proposed framework demonstrates an effective Statistics Machine Learning (SML) based pipeline. It could be easily extended to DNN framework or hybrid framework combining SML and Deep Neural Network (DNN), if the hand-crafted feature extraction components are replaced by DNNs followed by either a DNN-based or a SML and DNN combined classification module.

In the following, Section II reviews the related works. Section III presents the proposed adaptive energy guided sub-action segmentation method, heterogeneous feature extraction and fusion. Section IV introduces the CPPCR method for addressing the sub-action sharing challenge. Experiments and analyses are conducted in Section V, and subsequently a conclusion is summarized in Section VI.

II. RELATED WORK
According to the feature extraction for action recognition, existing methods can be categorized into hand-crafted feature based and deep learning based. The relevant fusion and representation learning methods are also reviewed.

A. HAND-CRAFTED FEATURE BASED METHODS

1) DEPTH MAP FEATURES
From depth sensors, hand-crafted features, such as bag of 3D points [14], depth surface normal feature super normal vector (SNV) [22], histograms of oriented principle components (HOPC) [10], depth motion maps (DMM) [23], spatio-temporal depth cuboid [24] motion history and statistical metrics are extracted for action recognition. Histograms of oriented 4D normals (HON4D) [25] takes the 3D depth data as an opportunity to construct the 4D normals of body parts and uses the statistical information as data representation. Similar to the motion history images (MHIs) and motion-energy images (MEIs) [26] which are successful for RGB based action recognition, Depth Motion Maps (DMM) [23] with depth sensor [27] aims to model human body shape and motion’s history information. Generated from depth cloud points, Histograms of oriented principle components (HOPC) [10] feature uses the information of eigenvectors of its support cuboid, for view-invariant action recognition. Super normal vector (SNV) [22] is constructed from depth cloud points, Histograms of oriented principle components (HOPC) [10] feature uses the information of eigenvectors of its support cuboid, for view-invariant action recognition. Super normal vector (SNV) [22] is constructed from depth cloud points.

2) SKELETON FEATURES
As a distinct modality, skeleton data is heterogeneous with depth and provides us with 3D human skeletal joints. Therefore skeleton data features can be extracted from 3D skeleton data, such as skeletal quads feature [28], eigenjoints feature [29], 3DMTM-PHOG [30], active skeleton feature [13], body-pose feature [7] poselet mining, [31] joint trajectory maps, [11], covariance 3D Joint [32], skeleton optical spectra [33]. Skeletal quads feature [28] is proposed for 3D action recognition. By encoding the skeleton limbs into states via Markov random field, active skeleton representation [13] is aggregated for characterizing human actions. Based on the differences of skeleton joints of static poses and the dynamic poses over time, the eigenjoints feature [29] is used successfully for skeleton based action recognition. Skeletal representation by curved manifold Lie Group [34] is a novel method that models the 3D joint points of human bodies via 3D geometric algebra. In [33], skeleton optical spectra is proposed, in which the skeleton data are rendered into color images and then CNNs are used to learn the features for action recognition. By dividing 3D points into 4D grids, Vieira et al. [35] employed occupancy patterns to describe 4D grids spatially and temporally. Actionlet ensemble model [15] extracts local features in the neighborhood area of skeleton joints. For view-variant action recognition, Rahmani and
Mian [36] proposed a nonlinear model, transforming data from distinct views into a canonical view. However, as shown in Figure 1, non-accurate skeleton poses and sub-action sharing phenomenon make action analysis more challenging.

**B. DEEP LEARNING BASED METHODS**

With the evolution of neural networks, bidirectional Recurrent Neural Networks (RNN) [16], structured Convolutional Neural Networks (CNNs) [37], cross-modality feature analysis [12], Graph Convolutional Networks (GNN) [38], [39], attention-based long short-term memory (LSTM) [40], [41] and Spatio-temporal attention network [42] are proposed for action or video analysis tasks. By designing a general attention neural cell, spatio-temporal attention network with heterogeneous data is proposed for traditional RGB action recognition. Wang et al. [43] proposed to consider attention based deep 3D CNN features with LSTM for action recognition. Amir et al. [44] presented a spatio temporal LSTM networks for 3D action recognition. Zhu et al. utilized LSTM with Co-occurrence scheme [45] and achieved good performance. In order to address the challenge of action representations with view variations, two view adaptive neural networks [46] were combined for high performance skeleton-based action recognition. SkeletonNet method [47] transforms the features of skeleton frames into images and feeds them into the proposed deep learning framework. Similarly, in [17], the authors designed an RNN driven by privileged information (PI) for action recognition. From depth modality, three kinds of dynamic depth image features using rank pooling are generated [5], and then are fed into CNNs for action recognition.

However, deep neural network-based methods [5], [6], [17], [48], [49] for depth-action recognition are not popular as RGB-action recognition. One reason is deep learning based methods are data-driven, requiring big data with labels. The depth datasets are of relatively small or medium size so that the data-driven models are weakened, at the risk of over-fitting. By the augmented skeleton data, multiview LSTM fusion model with attend scheme [6] was proposed for skeletal action recognition. Liu et al. [48] constructed a 3D-based CNN (3DCNN) to learn the depth features. Then a hand-crafted skeleton joint based feature is fused with these 3DCNN learned depth features. DMM feature is weighted hierarchically, then is fed into three channel deep CNN [49] on small training datasets. To overcome the drawback of the less color or pixels of depth maps, methods [5], [17] were proposed.

Although end-to-end deep learning has many advantages, AI systems built through it often show the following fatal weaknesses: inexplicability, vulnerability (robustness is poor), easy to be deceived and attacked, and need a lot of data. These weaknesses make it possible to be used only in limited scenarios, such as complete information, deterministic information, static (or evolving according to deterministic laws) environment, and limited domains. Therefore, creating explainable, credible and robustness theories and methods is necessary for complex applications.

**C. INFORMATION FUSION AND REPRESENTATION LEARNING METHODS**

Fusing information from multiple modalities is useful for performance improvement. Fusion strategy can be performed at data-level, feature/representation-level and score/decision-level [8]. Each fusion category has its own cons and pros, and the selection of the fusion method is generally dependent on the types of features and data sources. Score-level fusion requires no post-processing or dimension reduction, and is independent on the types and lengths of different multi-modal features. However, score-level fusion has the main drawbacks: (1) Independent classification decisions that relate to each sensing modality, need to be combined via some soft rule for the final decision; (2) For n different modalities, the decision-level fusion needs to train n separated classifiers, resulting in more parameters and time consumption especially when using deep learning classifiers; (3) Decision scores are obtained from data streams separately, therefore the correlation between different modalities is largely lost when fusion takes place at the decision level.

In contrast, in the practice for multi-modal human action recognition system, concurrent data from multiple sources is a good clue to collect sufficient amount of information for making improved decisions. Therefore SML based feature-level fusion projects the features concurrently collected from multiple sensors to a new space by vigorous mathematical transformation to optimize information representation for high quality decision making.

The earlier works mainly focused on feature fusion for action recognition [50]–[57]. In [50], a multi-modal learning framework was proposed to fuse depth and skeleton-based features. Feature-level fusion [51] of depth features and skeleton joints based on random forests were proposed by the rule of Winner-Take-All. By extending CCA model [58], heterogeneous domain adaptation by $\ell_1$ regularized CCA [52] was proposed to exploit the correlation subspace, for cross-view action recognition. Recent work [56] a 3D CNN fusion strategy is proposed by combing the softmax scores for action recognition with arbitrary length. In [57], RGB and depth futures are fused for RGB-D videos based action recognition.

On the other hand, representation learning could provide us with stronger discriminative power [42], [59]–[61]. Collaborative Representation (CR) [62] has been demonstrated to be effective for face recognition and is fast as it has a closed-form solution. Kernel collaborative representation (KCR) [63] and discriminative collaborative representation (DCR) [64] were proposed and then based on them dictionary learning and discriminant projection methods were designed to determine appropriate features. Discriminative compact representation [18], KCR with locality constrained dictionary (KCRC-LCD) [60] and locality-constrained collaborative representation (LCCR) [59] were proposed to extend collaborative representation for face recognition.
However, more complicated fusion methods increase dramatically the computational cost. Simpler but efficient fusion methods are preferred for action recognition. Moreover, CR-based methods have a weakness that they rely heavily on a “good” dictionary (properly controlled) which is derived from the training dataset. This shortcoming limits the application of the CR model in action recognition.

III. PROPOSED HETEROGENEOUS FEATURES AND FUSION

For depth sensor based human action datasets, skeleton data are not always accurate, as shown in the third row of Figure 1. Global depth feature DMMs [23] which is extracted from the entire videos contains more long-term temporal information of human movement while less short-term information. Speed variations will directly affect the appearances of DMM feature since it is extracted from inter-frame motions. Moreover, there are large intra-class variations since depth sequences are generated at different speeds, i.e., the depth sequences have different lengths, as the cases demonstrated in Figure 3 (a). These observations and ubiquitous sub-action sharing prompt us to propose an adaptive energy guided sub-action segmentation method. It discovers sub-actions automatically in diverse action video instances.

A. ENERGY-ORIENTED SUB-ACTION SEGMENTATION

As introduced in Section I, the sub-action sharing phenomenon within distinct action categories is existed, decreasing the recognition performance, which can be addressed by exploiting the relations of shared sub-actions. When dividing an action sequence, an intuitive segmentation strategy is to divide the action sequence into segments with the same length directly in the time axis. Thus each segment is a sub-action of equal length, which is called as “time-oriented” segmentation. In contrast, in order to express the dynamic information such as the speed variations of human motion over time, we propose to segment each video into temporal sub-actions according to the motion energy function, so that sub-actions with different lengths can be obtained, which is called as “energy-oriented” segmentation.

Assuming there is an action sequence with \( N \) depth maps, we first project each depth map with three-dimensional information onto three orthogonal Cartesian planes, each of which corresponds to a perspective of the 3D space. The three planes are denoted as \( v \in \{\text{front, side, top}\} \). The difference between two consecutive projected maps on three views is then thresholded to generate a binary map. Then the accumulated motion energy on the i-th frame, \( E(i) \), is defined as:

\[
E(i) = \sum_{v=1}^{3} \sum_{j=1}^{i-1} \left( \text{sum}(|F_{v}^{j+1} - F_{v}^{j}| \geq \theta) \right),
\]

where \( F_{v}^{j+1} \) is the \( j+1 \)-th depth frame on view \( v \in \{\text{front, side, top}\} \) from the depth modality, \( \text{sum}(.) \) returns the number of non-zero element in a binary map, \( \theta \) is the threshold. Since the motion energy function is accumulated, it starts from the first \( i \) video frames. The motion energy of a frame reflects current frame’s relative motion status and location with respect to the entire activity. Based on this method, a video is divided adaptively into sub-segments of unequal length globally, effectively capturing the motion’s temporal orders.

The video is segmented based on equal division of the normalized motion energy, where each segment has the same percentage energy. As shown in Figure 2, the total motion energy \( E(N) \) of an action video with \( N \) depth frames is normalized to one. Thus we divide this normalized energy into a set of segments whose corresponding indices of frames are used to partition a video. In the example of Figure 2, the video is segmented into \( P \) parts based on equal division of the normalized motion energy. Each segment accounts for approximately \( 1/P \) of the total energy. For easy computation, \( P \) is set to be the power of 2 and \( P = 2^{\text{Scale} - 1} \) where the \( \text{Scale} \) is temporal pyramid scale parameter. Thus the frame indexes for sub-action segmentation is obtained if \( \text{Scale} \) and \( P \) are ready. For instance, as illustrated in Fig. 2 (b)
when $Scale$ is set to be 3, thus $P = 4$. We will have the entire video, i.e., $\{T_0, T_4\}$ in the first scale; 2 sub-actions, i.e., $\{T_0, T_2\}$ and $\{T_2, T_4\}$ in the second scale; and 4 sub-actions, i.e., $\{T_0, T_1\}$, $\{T_1, T_2\}$, $\{T_2, T_3\}$ and $\{T_3, T_4\}$ in the third scale where $Ti$ corresponds to different frame index number. As a result we will have $\sum_{s=3}^{Scale=3} 2^{s-1} = 1 + 2 + 4 = 7$ sub-actions in three temporal scale on three views. Parameter $Scale$, $P$ in Fig. 2 were determined experimentally, i.e., we evaluated the recognition rates versus different values of $P$, while fixing the others, and chose the best ones.

**B. ENERGY-ORIENTED DEPTH FEATURE EXTRACTION**

Depth maps have more information and geometric shape in the third depth dimension, which can be used as an important clue to describe the shape of human motion. But it contains structural and background noises, as shown in Figure 3.

Suppose that $start_i$ denotes the beginning frame indexes of $i$-th parts. The segmentation parameters should satisfy $i \in \{1 : P\}$. DMMs feature [23] of each perspective is extracted and they are concatenated for each sub-action as follows:

$$DMM^i_v = \sum_{f=start_i}^{N_i+start_i} \left( |map^i_v f+1 - map^i_v f| \geq \varepsilon \right),$$

$$DMM^i = [DMM^i_{\text{front}} T, DMM^i_{\text{side}} T, DMM^i_{\text{top}} T]$$

(2)

where $N_i$ is the number of video frames of the $i$-th sub-action, and $f$ is frame index. $T$ means matrix transposition and $\varepsilon$ is the background noise threshold. In the experiment $\varepsilon = 50$. The symbols $F^i_v f+1$ and $map^i_v f+1$, in Subsection III-A and Subsection III-B, are slightly different. $F^i_v f+1$ is the $(j+1)$th depth frame on view and $map^i v f+1$ is the $(j+1)$th depth map on view $v \in \{\text{front}, \text{side}, \text{top}\}$ which is resized after using bounding box to extract the foreground.

**C. SKELETON FEATURE EXTRACTION BY CROSS-MODALITY PARAMETERS TRANSFERRING**

3D depth image sequences lack the global semantic dynamic of the entire action videos. To compensate for the insufficient information therein, the skeleton features that characterize the relationship of skeleton joints are extracted.

The depth sensors capture the skeleton and depth data simultaneously and synchronously. Thus the motion energy function, generated from depth modality, characterizes the execution speed of actions in depth and skeleton data simultaneously. This motivates us to transfer the action segmentation parameters, generated from depth modality, into skeleton modality to divide the skeleton sequence.

In skeleton modality, based on the sub-sequences, Dynamic skeleton (DS) features [15] are extracted by computing the relative positions between each pair of trajectories. Then Fourier features and its gradient information (i.e., DSG) are further extracted and concatenated together as introduced in [15]. From the result of this practice, gradient information could depict the velocity change of the action’s motion.

**D. HETEROGENEOUS INFORMATION FUSION**

Here, the goal of heterogeneous information fusion is to analyze and exploit the relations between heterogeneous feature sets. Thus the obtained representation is more discriminative than any of the input ones. The features extracted above always have high dimensionality with a certain redundancy. Therefore, these redundant features should be fused, compressed or refined. With the advantages of analytic capabilities of machine learning, features of high dimensionality can be analyzed efficiently to extract meaningful information, forming compressed and discriminative representations. Feature level fusion is perceived as simpler, more effective and meaningful than the other levels of fusion [65]. In this paper, the Canonical Correlation Analysis (CCA) is adopted.
to reduce the dimension of high dimensional features. CCA incorporates the vector associations into the correlation analysis of the feature sets, maximizing the correlation across the two feature sets.

Given two feature matrices $X \in \mathbb{R}^{p \times n}$ and $Y \in \mathbb{R}^{q \times n}$, which are generated from $n$ training samples. The feature vectors are from depth ($p$ dimension) and skeleton ($q$ dimension) modalities respectively. Within-set covariance matrices of $X$ and $Y$ are defined as $S_{xx} \in \mathbb{R}^{p \times p}$ and $S_{yy} \in \mathbb{R}^{q \times q}$ respectively. Between-set covariance matrix of $X$ and $Y$ is defined as $S_{xy} \in \mathbb{R}^{p \times q}$ and $S_{yx} = S_{yx}^T$. The aim of CCA is to maximize the pair-wise correlations of $X$ and $Y$, resulting in the transformation matrices $W_x$ and $W_y$. By these transformation matrices, linear combinations $X^* = W_x^T X$ and $Y^* = W_y^T Y$ are obtained by solving the eigenvalue equations:

$$
\begin{align*}
S_{xx}^{-1} & S_{xy}^{-1} S_{yx} \hat{W}_x = R_1^2 \hat{W}_x, \\
S_{yy}^{-1} & S_{yx}^{-1} S_{xy} \hat{W}_y = R_2^2 \hat{W}_y,
\end{align*}
$$

(3)

where $\hat{W}_x$ and $\hat{W}_y$ are the eigenvectors and $R^2$ is the diagonal matrix of eigenvalues or squares of the canonical correlations. Two typical feature fusion methods using CCA are: serial feature fusion (CCA-serial) and parallel feature fusion (CCA-parallel):

$$
Z_1 = \left( X^* Y^* \right) = \left( W_x^T X \right) \left( W_y^T Y \right)^T = \left( W_x \right)^T \left( X \right) \left( Y \right)^T
$$

(4)

and

$$
Z_2 = X^* + Y^* = W_x^T X + W_y^T Y = \left( W_x \right)^T \left( X \right) \left( Y \right)^T
$$

(5)

where $Z_1$ and $Z_2$ are called the canonical correlation discriminant features. $X^*$ and $Y^* \in \mathbb{R}^{d \times n}$ are known as canonical variates and have two useful qualities: they have nonzero correlation only on their corresponding indices, and are uncorrelated within each feature set.

For instance, in SBU interaction dataset, there are $n = 196$ training samples, the dimension of depth features is $p = 47376$, the dimension of skeleton features is $q = 54810$. Using feature-level fusion method, the dimension of the computed CCA feature space is $d = 195$, so that the final dimension of fused features is either $d = 195$ by CCA-parallel summation or $d = 390$ by CCA-serial concatenation.

IV. THE PROPOSED CLASS-PRIVACY PRESERVED COLLABORATIVE REPRESENTATION (CPPCR) ACTION RECOGNITION METHOD

Sub-actions sharing among action categories is a challenge for action recognition. The global scheme in CR-based learning is designed to leverage the shared feature subspace. However, the locality of the action category is a strong clue to recognize actions. We propose to make the class-privacy property preserved, and select the linear combination of nearby characteristics/sub-actions, favoring class-preserved locality (which preserves some expected local geometric structures) even though the testing sample can be described by another classes’ few far characteristics/sub-actions, which is called Class-privacy Preserved Collaborative Representation.

A. PRELIMINARY: CR-BASED LEARNING AND CLASSIFICATION

In popular low-dimensional manifold models [19], [53], for each feature space, one feature vector is represented by the linear combination of a few representative points. However, in these models the characteristics of the testing data can only be represented by the learned characteristics of one class from the training samples. As introduced in Section I and shown in Figure 4(b) and (c), sub-actions sharing is ubiquitous throughout the different interaction categories. The sub-actions hugging are shared between “Hugging”, “Approaching” and “Departing”, which are different interaction categories.

This paper regards the negative sub-action sharing challenge as a positive chance, and ingeniously transforms challenge into opportunity in the proposed CPPCR method. As introduced in Section I, from Immanuel Kant’s statement, the sub-actions sharing challenge is also a chance since sub-actions from other classes are helpful to represent the testing sample. If all the other sub-actions’ extracted features are used as possible training features (samples) for representing each sub-action, we can not only significantly improve the ability of learning features (knowledge representation), but also mitigate sub-action share challenges.

Let $D = \left[ D_1, D_2, \ldots, D_i, \ldots, D_C \right]$ be the dictionary, which has $C$ human action categories. Each sub-dictionary $D_i$ is associated with the $i$ action category, and each column of $D_i$ represents the fused feature set of training samples from class $i$. The fused feature set is obtained via CCA from heterogeneous data. Let $y$ be the fused feature of testing sample.

Firstly, in CR learning method [62], the columns of $D$ is normalized to have unit norm. Then $y$ is represented on $D$ collaboratively and globally using the $\ell_2$-minimization Lagrangian formulation:

$$
\hat{a} = \arg \min_a \left\{ \| y - D \hat{a} \|_2^2 + \lambda \| a \|_2 \right\}
$$

(6)

where $\lambda$ is a Lagrangian scalar parameter to balance the residual of representation function and the regularization term.

Secondly, by computing the representation residuals $e_i = \| y - D_i \hat{a} \|_2 / \left\| \hat{a} \right\|_2$, the class label is determined via $\text{class}(y) = \arg \min_i \{ e_i \}$.

B. CPPCR FOR ACTION RECOGNITION

CR [62] favors the global relationship and encodes the testing sample as a linear combination of sub-actions of training samples from all categories. However, the locality of the action category is a strong clue to recognize actions. In other words, we tend to select the linear combination of neighboring characteristics/sub-actions, rather than the global relationship. Class-preserved local geometric structures is favored
in CPPCR, even though the testing sample can be described by another few far characteristics/sub-actions. By integrating this class-preserving locality to CR [62], we will improve the discrimination ability of the feature representation.

In this section, by adding a class-privacy preserved locality constraint $C_{pp} = (\sum_{y_i \in N_K(y)} \|y_i - Da\|_2^2)/K$ to Eq. (6), the objective function is:

$$\min_\alpha (1 - \gamma) \|y - Da\|_2^2 + \lambda \|\alpha\|_2$$

$$+ \frac{\gamma}{K} \sum_{y_N \in N_K(y)} \|y_N - Da\|_2^2$$  (7)

where $N_K(y)$ are the top-$K$ nearest training samples of the testing sample $y$. And $\lambda$ and $\gamma$ are regularization parameters to balance the reconstruction residual term, the sparsity term and the locality constraint term. The neighborhood local region is determined by some distance metrics, such as Spearman, Cityblock and Euclidean distance by KNN searching. Under this locality constraint, the closer the distance is, the greater the contribution will be.

By computing the derivative with respect to $\alpha$ of Eq. (7) and letting it be zero, we have the solution as follows:

$$\hat{\alpha} = \left(D^TD + \lambda I\right)^{-1}D^T \left\{ (1 - \gamma)y + \frac{\gamma}{K} \sum_{y_N \in N_K(y)} \|y_N - Da\|_2^2 \right\}$$  (8)

In the training phase, the term$(D^TD + \lambda I)^{-1}D^T$ can be precomputed. This precomputed term is independent of the testing sample, depending only on the training data.

In the recognition phase, given a testing sample $y$, the two low-level heterogeneous features are extracted and then fused by CCA to get the mid-level feature. Moreover, the neighborhood region $N_K(y)$ is determined by KNN searching from the training set, then the $(1 - \gamma)y + \frac{\gamma}{K} \sum_{y_N \in N_K(y)} \|y_N - Da\|_2^2$ is calculated and multiplied by the pre-calculated projection matrix. Finally the regularized representation residuals $e_i(y)$ is computed by

$$e_i(y) = \|y - D\hat{\alpha}_i\|_2 / \|\hat{\alpha}_i\|_2$$  (9)

and the action category label can be predicted via

$$label(y) = \arg\min_i \{e_i(y)\}$$  (10)

The proposed CPPCR algorithm for action recognition is concluded in Algorithm 1.

V. EXPERIMENTS AND PERFORMANCE EVALUATION

To evaluate the proposed method for multi-modal action recognition, we first conduct an ablation study. Then extensive experiments on four public datasets are conducted, including one interaction dataset [7] which has two-person interactions, and three action datasets [14], [21], [66] which include single person actions.

A. ABLATION EVALUATION

Firstly, we evaluated the energy guided sub-action segmentation for feature extraction and fusion strategies. The ablation evaluations are conducted on the SBU-Kinect interaction dataset [7]. As shown in Table 1, the performance of the proposed energy-guided sub-action segmentation method is superior to that of the time-guided method. Secondly, for individual heterogeneous features, the depth feature has better performance than the skeleton feature since skeleton data sometimes contains incorrectly tracked skeleton poses. For single feature, the energy-guided depth feature has the highest recognition accuracy of 87.69%.

For feature fusion strategies, the CCA-serial fusion strategy performs better than the CCA-parallel one. As a result,
Algorithm 1 CPPCR Algorithm for Action Recognition

1: **Training phase and inputs:**

   (1) A heterogeneous feature matrix (dictionary) $D = [D_1, D_2, \ldots, D_i, \ldots, D_K]$ constructed by Section III, containing the extracted heterogeneous (depth and skeleton) feature of all training samples. $D_i$ is the heterogeneous feature set of training video samples from class $i$.

   (2) The heterogeneous feature vectors from the testing sample $y$.

   (3) The regularization parameters $\lambda$ and $\gamma$.

   (4) The parameter $K$, which is the size of the top-K local neighborhood region.

2: **Pre-calculation process:** Calculate the projection matrix $(D^TD + \lambda I)^{-1}D^T$.

3: **Testing phase:** For each testing action sample $y$:

   (1) Use KNN searching, from the training feature set $D$, to determine the neighborhood region $N_K(y)$.

   (2) Calculate the class-privacy preserved constraint item $\frac{K}{K} \sum_{y \in N_K(y)} \|y - D\alpha\|_2^2$.

   (3) Balance the importance between the testing sample and its local neighborhood region samples by $(1 - \gamma)\|y\|_{2} + \frac{K}{K} \sum_{y \in N_K(y)} \|y - D\alpha\|_2^2$.

4: **function** CPPCR($P$, $y$, $\gamma$, $\lambda$, $K$) // Where $P$ – projection matrix; $y$ – the test sample; $\gamma$, $\lambda$, $K$ – the parameter

   (1) Calculate the CPPCR solution via Eq.(8).

   (2) Calculate the regularized representation residuals via Eq.(9).

   (3) Inference the label via Eq.(10).

5: **end function**

### TABLE 1. Contributions of heterogeneous feature and fusion strategies, evaluated on the SBU Interaction dataset [7].

| Features and Fusion Strategies | Accuracy (%) |
|-------------------------------|--------------|
| Time-guided Skeleton Feature  | 80.0         |
| Time-guided Depth Feature     | 83.08        |
| Time-guided Features via CCA-parallel Fusion | 90.77     |
| Time-guided Features via CCA-serial Fusion | 90.77     |
| Energy-guided Skeleton Feature | 84.56       |
| Energy-guided Depth Feature   | 87.69        |
| Energy-guided Features via CCA-parallel fusion | 92.31   |
| Energy-guided Features via CCA-serial fusion | **95.39** |

The CCA-serial fusion of two heterogeneous energy-guided features has the highest recognition accuracy of 95.39%. This demonstrates that the appropriate feature fusion methods effectively preserve the complementary information of heterogeneous data.

### B. PARAMETERS EVALUATION

The key parameters of CPPPCR are investigated, in terms of analyzing the recognition rates iteratively. Table 2 reports the recognition accuracies versus the variant values of parameter $K$. From the evaluations, it’s observed that the action performances are better when $K = 3$ and the metric distance is chosen to be Spearman metric. From the experimental evaluations, the parameter $K$, $\lambda$ and $\gamma$ are set as $K = 3$, $\lambda = 0.01$, $\gamma = 0.2$ respectively in all the following experiments.

### C. SBU INTERACTION DATASET AND PERFORMANCE EVALUATION

The SBU Interaction dataset [7] is a multi-modal dataset with 8 interaction activities. It was collected from 7 participants, providing synchronized RGB video, depth map and skeleton pose modalities. It consists of 230 video sequence samples from 8 interaction categories. In most scenarios, interactions are performed when one person is acting and the other person is reacting.

The challenges of this data set include: (1) human motion categories are the interaction between two persons; (2) in most interactions, one person is acting while the other is responding, and most of the interactions are associated with the security surveillance, healthcare monitoring, smart buildings/home applications; (3) as illustrated in Figure 1, most of these action categories are social behaviors; (4) most categories are non-periodic human-to-human interactions, containing sub-actions and comparable physical movements. Same settings in [7] is followed, where a standard 5-fold cross-validation scheme is employed.

The result of the proposed method is 95.39%, as shown in Table 3. It is observed that the performance of the proposed method is better than body pose feature with libSVM method [7], privileged information-based RNNs method [17], representation learning of temporal dynamics by RNNs [16], deep structured model [37] and co-occurrence LSTM model [45]. This indicates that the proposed method is effective for person-to-person interaction recognition.

The confusion matrix is a useful tool to show the recognition accuracies of each class and the confusion percentage between distinct categories, which is for analyzing the detailed recognition results. Confusion matrix of SBU interaction dataset is illustrated in Figure 5. We can see that most interaction categories are recognized correctly.
FIGURE 5. Confusion matrix of the SBU Kinect Interaction dataset by the proposed method.

TABLE 3. Performance (%) Evaluation on SBU-Kinect Interaction dataset. Note that “D” represents the Depth feature, “S” represents the Skeleton feature, “A” represents the Accelerometer feature.

| Methods | Accuracy | Modality |
|---------|----------|----------|
| Multiple Instance Learning Roost [7] | 91.1 | Skeleton |
| Skeleton Pose [7] | 87.6 | Skeleton |
| Body-Pose Feature [7] | 91.1 | Skeleton |
| Poselet Mining [31] | 86.9 | Skeleton |
| Privileged Information-Based RNNs [17] | 89.2 | D+S |
| Deep Structured Data-level Fusion [37] | 93.4 | RGB+D |
| ST-LSTM + Trust Gates [44] | 93.3 | skeleton |
| LST + Co-occurrence+ Dropout [45] | 90.41 | Skeleton |
| Hierarchical RNN [16] (reported in [45]) | 80.35 | Skeleton |
| SkeletonNet (Skeleton + CNN) [47] | 93.47 | Skeleton |
| Global Context-Aware + LSTM [67] | 94.1 | Skeleton |
| LSTM-Multimodal Feature Fusion [6] | 95.0 | Skeleton |
| Decision-level fusion with VGG-F [68] | 96.26 | RGB+D+S |
| Multi-stream LSTM Feature Fusion [41] | 92.5 | Skeleton |

The Proposed Method 95.39 D+S

D. MSR ACTION 3D DATASET AND PERFORMANCE EVALUATION

The famous MSR Action 3D dataset [14] has 20 categories and is collected by 10 subjects. Cross-subject experiment settings were adopted as in [14], where the data of 5 subjects are used for model training and the remaining data are used for recognition.

The proposed method has the recognition accuracy 93.82%, as shown in Table 4, better than methods [10], [23], [25] and most existing methods compared. It’s observed that the proposed method is comparable to statistical method [22], heterogeneous features fusion method [54] and deep learning methods [16], [47], [67]. In addition, sub-segmentation by cross-modality parameter transferring is effective from the results. For feature fusion strategies, the CCA-serial fusion contributes more than CCA-parallel strategy. This indicates that the proposed method effectively preserves the spatio-temporal information of the two heterogeneous features, outputting high-level discriminative action representation.

E. UTD-MHAD DATASET AND PERFORMANCE EVALUATION

Multi-modal dataset UTD-MHAD [66] is collected by depth sensor and wearable inertial sensor. It consists of 27 action categories, 4 modalities, and each modality has 861 samples.

F. UTD-MHAD-KINECT V2 DATASET AND PERFORMANCE EVALUATION

UTD-MHAD-Kinect V2 [21] is a multi-modal dataset, which contains heterogeneous data from depth sensor and inertial...
TABLE 5. Comparisons on the UTD-MHAD dataset. Note that “D”, “S”, “A” and “G” represent Depth, Skeleton, Accelerometer and Gyroscope feature respectively.

| Methods                        | Accuracy | Modality |
|--------------------------------|----------|----------|
| BIC-KSCD [71]                  | 76.19    | Skeleton |
| Multi-modal decision-level fusion [66] | 79.10    | D+A      |
| Deep Decision-Level Fusion [72] | 89.2     | D+G+A    |
| Decision-level fusion with VGG-P [68] | 94.6     | RGB+D+S  |
| Covariance 3D Joint [32]       | 85.58    | Skeleton |
| Skeleton Optical Spectra+CNN [33] | 86.97    | Skeleton |
| Joint Trajectory Maps+CNN [11] | 87.90    | Skeleton |
| Three Sensors Feature-Level Fusion [73] | 84.89    | RGB+D+S  |
| Two Sensors Feature-Level Fusion [70] | 89.3     | RGB+D    |
| Two Sensors Feature-Level Fusion [70] | 91.6     | G+A      |
| Two Sensors Feature-Level Fusion [70] | 93.7     | D+G      |
| Two Sensors Feature-Level Fusion [70] | 94.8     | D+A      |
| Four Sensors Feature-Level Fusion [70] | 98.2     | RGB+D+G+A|
| Time-guided Skeleton Feature   | 84.9     | Skeleton |
| Time-guided Depth Feature      | 76.2     | Depth    |
| Time-guided Fusion via CCA-parallel | 87.0     | D+S      |
| Time-guided Fusion via CCA-serial | 90.7     | D+S      |
| Energy-guided Skeleton Feature | 85.6     | Skeleton |
| Energy-guided Depth Feature    | 82.8     | Depth    |
| Energy-guided Fusion via CCA-parallel | 91.2     | D+S      |
| The Proposed Method            | 94.2     | D+S      |

sensor. It consists of 1200 sequences from three modalities. It has 10 action categories and is performed by 3 female and 3 male subjects. Same experimental setting of [21] is adopted in this paper.

Firstly, the recognition accuracy of single heterogeneous feature and fusion methods are investigated. The performances of the four features, skeleton feature guided by time, depth feature guided by time, skeleton feature guided by energy and depth feature guided by energy are first derived. From Table 6, it can be seen that the proposed Energy-oriented method improves the performance. For single features, the Energy-oriented depth feature has higher recognition accuracy than that of Time-oriented. The performance of the depth feature proposed in this paper is better than that of the skeleton feature.

The results demonstrate that the proposed CPPCR achieves 91.5% which is higher than Multimodal Hybrid Centroid CCA, Multimodal Centroid CCA and MCCA by 1.5%, 3.5% and 9%, respectively. For feature fusion strategies, CCA-serial fusion contributes more than CCA-parallel strategy. CCA-serial fusion of two heterogeneous Energy-oriented features has the highest recognition accuracy of 90.0%, further improved to 91.5% by the proposed representation learning method. This indicates that the proposed CPPCR preserves the spatiotemporal information of actions, outputting high-level discriminative action representation.

G. QUALITATIVE ANALYSIS OF THE PROPOSED METHOD

Here, we did the qualitative analysis of the proposed method, including: (1) Whether the two heterogeneous features have complementary characteristics; (2) the relationship between feature dimensions and recognition accuracies after the representation learning method CPPCR. The experiment was conducted on SBU Kinect Interaction dataset to do qualitative analysis.

We evaluated the recognition accuracies versus feature dimensions by CCA-serial fusion method. The dimensions of the proposed two heterogeneous features are reduced to 195 from dimensions 47376 and 54810, respectively. As shown in Figure 6, on SBU-Kinect Interaction dataset, the recognition accuracies increase drastically when the feature dimensions are greater than 195, indicating that in the new CPPCR feature space, the proposed two heterogeneous features are highly complementary with each other. Furthermore, the accuracies are stable if the dimensions are higher than 230. The highest accuracy is 95.39% when the dimensions = [298, 309, 310, 311, 312]. It should be noted that 195 is precisely the feature dimension from the first data modality, skeletal sequence data. This shows that the features extracted from the two modalities are highly complementary with each other in the new CPPCR feature space. They are compact and high-level features for action representation. Fusing them together contributes to performance improvement.

H. DISCUSSIONS

On SBU Kinect interaction dataset, it is observed that the proposed method is better than numerous methods compared [16], [17], [37], [44], [45], [47], [67], as shown in Table 3. The proposed method is worse than feature fusion methods [70] using multi-modal data from RGB, depth and inertial sensors (accelerometer and gyroscope) on dataset UTD-MHAD. On MSRAction 3D dataset shown in Table 4, the performance is worse than the hierarchical skeleton...
model [17] whereas yields better performance than active skeleton representation [13] and decision-level fusion [69]. There are several main possible reasons:

1) The scale of the datasets. The proposed method achieves the superior performance compared to most state-of-the-art on small or medium scale datasets. According to the development rule of artificial intelligence, deep neural networks are data-driven methods, requiring a large amount of training data. Statistical machine learning can be applied to databases of different scales.

2) Ubiquitous sub-action sharing challenge. As demonstrated in Figure 4 (b) and (c), sub-actions sharing are ubiquitous throughout the different interaction categories. The sub-action shuffling are shared between “Hugging” and “Pushing”, which are different interaction categories. This paper regards the negative sub-action sharing challenge as a positive chance, and ingeniously transforms risk into opportunity in the proposed CPPCR method. As introduced in Section I, from Immanuel Kant’s statement, the sub-actions sharing challenge is also a chance since sub-actions from other classes are helpful to represent the testing sample. If the features extracted from all the other sub-actions are used as possible training features (samples) for distinct sub-action, we can not only significantly improve the ability to learn features (knowledge representation), but also mitigate sub-action sharing challenges.

Non-accurate skeleton data. The methods [6], [13], [16], [17], [37], [44], [45], [47], [67] used skeleton features. However, the estimated skeleton joints sometimes are not accurate because of the body parts occlusion and missing fragment, as illustrated in Figure 1 (a).

The sub-action sharing challenge and the experimental results demonstrate that it is a novel choice to employ the proposed heterogeneous feature fusion method with CPPCR learning.

VI. CONCLUSION

There are many challenges in human action recognition based on RGB-Depth sensors, among which ubiquitous sub-action sharing phenomenon (especially among the similar categories) is a critical one. To this end, sub-action segmentation based on equal motion energy and class-privacy preserved collaborative representation (CPPCR) learning are proposed to jointly explore/address the long-range temporal dynamic structure involved in the actions/interactions. The action motion energy is computed in the depth modality and accordingly the action videos are segmented into sub-actions based on equal motion energy division. Then the action segmentation parameters are transferred from depth to the temporally synchronous skeleton modality, thus two heterogeneous features are extracted respectively and fused. In addition, the proposed CPPCR takes the negative sub-action sharing challenge as a positive opportunity, addressing the sub-action sharing challenge.

The experimental results on four datasets consistently demonstrate the effectiveness of the proposed method. Qualitative analysis of the two features, as shown in Figure 6 and 7, illustrates that in the learned CPPCR feature space depth and skeleton features are complementary with each other, fusing them leads to superior performance than using either of the two individually.

REFERENCES

[1] L. Wang, D. Q. Huynh, and P. Koniusz, “A comparative review of recent kinect-based action recognition algorithms,” IEEE Trans. Image Process., vol. 29, pp. 15–28, 2020.

[2] J. K. Aggarwal and M. S. Ryoo, “Human activity analysis: A review,” ACM Comput. Surv., vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011.

[3] T. Guha and R. K. Ward, “Learning sparse representations for human action recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 8, pp. 1576–1588, Aug. 2012.

[4] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhang, “Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks,” IEEE Trans. Multimedia, vol. 20, no. 9, pp. 2330–2343, Sep. 2018.

[5] P. Wang, W. Li, Z. Gao, C. Tang, and P. O. Ogunbona, “Depth pooling based large-scale 3-D action recognition with convolutional neural networks,” IEEE Trans. Multimedia, vol. 20, no. 5, pp. 1051–1061, May 2018.

[6] Z. Fan, X. Zhao, T. Lin, and H. Su, “Attention-based multiview re-observation fusion network for skeletal action recognition,” IEEE Trans. Multimedia, vol. 21, no. 2, pp. 363–374, Feb. 2019.

[7] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, “Two-person interaction detection using body-pose features and multiple instance learning,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2012, pp. 28–35.

[8] X. Peng, L. Wang, X. Wang, and Y. Qiao, “Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice,” Comput. Vis. Image Understand., vol. 150, pp. 109–125, Sep. 2016.

[9] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks for action recognition in videos,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[10] H. Rahmani, A. Mahmood, D. Huyhn, 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.

[11] P. Wang, W. Li, C. Li, and Y. Hou, “Action recognition based on joint trajectory maps with convolutional neural networks,” Knowl.-Based Syst., vol. 158, pp. 43–53, Oct. 2018.

[12] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang, “Deep multimodal feature analysis for action recognition in RGB-D videos,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 5, pp. 1045–1058, May 2018.
[61] C. Liang, L. Qi, Y. He, and L. Guan, “3D human action recognition using a single depth feature and locality-constrained affine subspace coding,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 10, pp. 2920–2932, Oct. 2018.

[62] L. Zhang, M. Yang, and X. Feng, “Sparse representation or collaborative representation: Which helps face recognition?” in Proc. Int. Conf. Comput. Vis., Nov. 2011, pp. 471–478.

[63] G. Zhang, H. Sun, G. Xia, and Q. Sun, “Kernel collaborative representation based dictionary learning and discriminative projection,” Neurocomputing, vol. 207, pp. 300–309, Sep. 2016.

[64] Y. Wu, W. Li, M. Mukunoki, M. Mino, and S. Lao, Discriminative Collaborative Representation for Classification. Cham, Switzerland: Springer, 2015, pp. 205–221.

[65] A. H. Gunati, N. D. Elmadany, and Y. He, “Feature-level and decision-level fusion of noncoincidentally sampled sensors for land mine detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 577–589, Jun. 2001.

[66] C. Chen, R. Jafari, and N. Kehtarnavaz, “UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2015, pp. 168–172.

[67] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, “Skeleton-based human action recognition with global context-aware attention LSTM networks,” IEEE Trans. Image Process., vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

[68] P. Khare, P. Kumar, and J. Imran, “Combining CNN streams of RGB-D and skeletal data for human activity recognition,” Pattern Recognit. Lett., vol. 115, pp. 107–116, Nov. 2018.

[69] C. Chen, R. Jafari, and N. Kehtarnavaz, “Action recognition from depth sequences using depth motion maps-based local binary patterns,” in Proc. IEEE Winter Conf. Appl. Comput. Vis., Jan. 2015, pp. 1092–1099.

[70] M. Ehatisham-Ul-Haq, A. Javed, M. A. Azam, H. M. A. Malik, A. Irtaza, I. H. Lee, and M. T. Mahmood, “Robust human activity recognition using multimodal feature-level fusion,” IEEE Access, vol. 7, pp. 60736–60751, 2019.

[71] L. Zhou, W. Li, Y. Zhang, P. Ogunbona, D. T. Nguyen, and H. Zhang, “Discriminative key pose extraction using extended LC-KSVD for action recognition,” in Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA), Nov. 2014, pp. 1–8.

[72] N. Dawar, S. Ostadabbas, and N. Kehtarnavaz, “Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition,” IEEE Sensors Lett., vol. 3, no. 1, pp. 1–4, Jan. 2019.

[73] E. Escobedo and G. Camara, “A new approach for dynamic gesture recognition using skeleton trajectory representation and histograms of cumulative magnitudes,” in Proc. 29th SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI), Oct. 2016, pp. 209–216.

[74] N. E. D. Elmadany, Y. He, and L. Guan, “Multimodal learning for human action recognition via bimodal/multimodal hybrid centroid canonical correlation analysis,” IEEE Trans. Multimedia, vol. 21, no. 5, pp. 1317–1331, May 2019.

CHENGWU LIANG received the M.A.Sc. degree in information and communication engineering from the University of Electronic Science and Technology of China, Chengdu, China. He is currently pursuing the Ph.D. degree with Zhengzhou University, China. He is a Lecturer with the Henan University of Urban Construction. He was a Visiting Student with Ryerson University, Toronto, ON, Canada. His current research interests include machine learning, computer vision, and statistical pattern recognition, especially video understanding. He was a recipient of the Top five Papers Award of 2017 IEEE International Conference on Visual Communication and Image Processing.

DEYIN LIU received the bachelor’s degree from Zhengzhou University, Zhengzhou, China, where he is currently pursuing the Ph.D. degree. He is also a Visiting Joint Ph.D. Candidate with The University of Queensland, Australia. His major research interests include video analysis, computer vision, pattern recognition, and machine learning.

LIN QI received the B.Sc. degree in radio engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, the M.A.Sc. degree in computer science from Zhengzhou University, Zhengzhou, China, and the Ph.D. degree in information and communication engineering from the Beijing Institute of Technology, Beijing, China. He is currently a Professor with the School of Information Engineering, Zhengzhou University. His current research interests include image or video analysis and processing, pattern recognition, and signal detection and estimation.

LING GUAN (Fellow, IEEE) received the Ph.D. degree in electrical engineering from The University of British Columbia, Vancouver, BC, Canada, in 1989. He is currently a Professor with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada. He also held visiting positions at British Telecom, in 1994, Tokyo Institute of Technology, in 1999, Princeton University, in 2000, National ICT Australia, in 2007, Hong Kong Polytechnic University, from 2008 to 2009, and Microsoft Research Asia, in 2002, 2009, and 2017. He has published extensively in multimedia processing and communications, human-centered computing, machine learning, adaptive image and signal processing, and more recently, multimedia computing in the immersive environment. He is an IEEE Circuits and System Society Distinguished Lecturer and an Elected Member of the Canadian Academy of Engineering. He was a recipient of the 2014 IEEE Canada C.C. Gotlieb Computer Medal and the 2005 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS Best Paper Award. In 2001, he was appointed as a Tier I Canada Research Chair in Multimedia and Computer Technology.

* * *

VOLUME 8, 2020

39933