J Electr Eng Technol Vol. 9, No. 2: 739-748, 2014
http://dx.doi.org/10.5370/JEET.2014.9.2.739

ISSN(Print) 1975-0102
ISSN(Online) 2093-7423

# Human Action Recognition Via Multi-modality Information

**Zan Gao*, Jian-ming Song*, Hua Zhang*, An-An Liu†, Yan-bing Xue* and Guang-ping Xu***

**Abstract** – In this paper, we propose pyramid appearance and global structure action descriptors on both RGB and depth motion history images and a model-free method for human action recognition. In proposed algorithm, we firstly construct motion history image for both RGB and depth channels, at the same time, depth information is employed to filter RGB information, after that, different action descriptors are extracted from depth and RGB MHIs to represent these actions, and then multimodality information collaborative representation and recognition model, in which multi-modality information are put into object function naturally, and information fusion and action recognition also be done together, is proposed to classify human actions. To demonstrate the superiority of the proposed method, we evaluate it on MSR Action3D and DHA datasets, the well-known dataset for human action recognition. Large scale experiment shows our descriptors are robust, stable and efficient, when comparing with the-state-of-the-art algorithms, the performances of our descriptors are better than that of them, further, the performance of combined descriptors is much better than just using sole descriptor. What is more, our proposed model outperforms the state-of-the-art methods on both MSR Action3D and DHA datasets.

**Keywords**: Action recognition, Multi-modality, Feature fusion, RGB, Depth, MMCRR, DMHI, RDMHI, RDMHI-AHB, RDMHI-Gist

## 1. Introduction

Human action recognition is a hot research topic in the field of computer vision and machine learning in the recent years and has been widely applied in many domains, such as visual surveillance, human computer interaction and video retrieval etc. In the past decades, researcher mainly focused on video sources captured by traditional RGB cameras, and the-state-of-art methods [1-6] can obtain satisfying performance on well-known benchmarks, such as Weizmann [2] and KTH [4]. Motion history image (MHI) [1-2] and spatio-temporal interest points methods [3-6] have been extensively used for action representation. In order to precisely capture the appearance and motion of the target, we often need to segment the target from the background. Similarly, when we extract spatio-temporal interest points, some of them might locate in the background regions, which will be noise for the representation of the target. Therefore, researchers often need the contour of the target to filter noisy interest points. However, in traditional RGB videos, it is extremely challenging to quickly and reliably detect, segment and track human body especially with low illumination and consequently most of state-of-art approaches would fail.

As the imaging technique advances, e.g. the launch of Microsoft Kinect, it has become possible to capture both color images and depth maps in real time with RGB-D sensors. Fig. 1 shows the inner structure of Kinect sensor, and Fig. 2 shows the outputs of Kinect sensor in which RGB image and depth map are given respectively. From Fig.2, it is intuitive that we can much more easily obtain silhouette of one person with the depth map, which would provide rich shape information for human action recognition. Many researchers have been engaged in leveraging depth information to identify actions. Space-time volume in depth and simple descriptors [7] were extracted for action representation, and then approximate string matching (ASM) was utilized for classification. Li et al. [8] proposed a bag-of-3D-points or 3D silhouettes method to represent postures by sampling 3D points from depth maps, and then an action graph was then adopted to model the sampled 3D points for action recognition.

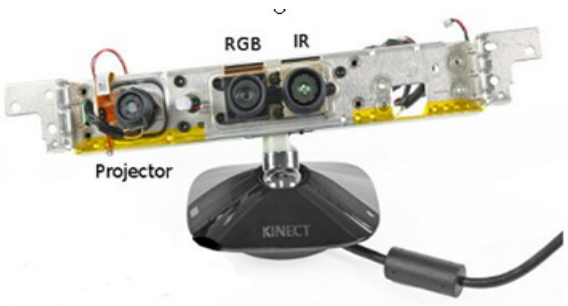



**Fig.1.** Kinect consists of Infrared (IR) projector, RGB camera and IR camera

† Corresponding Author: School of Electronic Information Engineering, Tianjin University, P.R. China. (anan0422@gmail.com)
* School of Computer and Communication Engineering, Tianjin University of Technology, P.R. China. (zangaonsh4522@gmail.com)
Received: August 23, 2013; Accepted: November 4, 2013

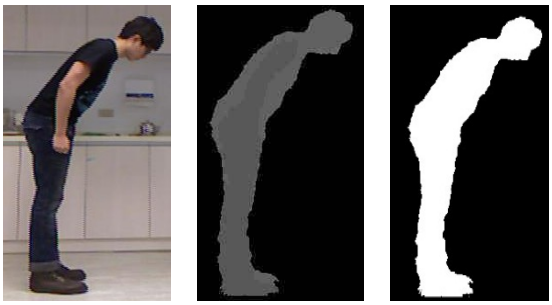


**Fig.2.** from left to right: RGB image, depth map, and 0-1 mask of human silhouettes of bending

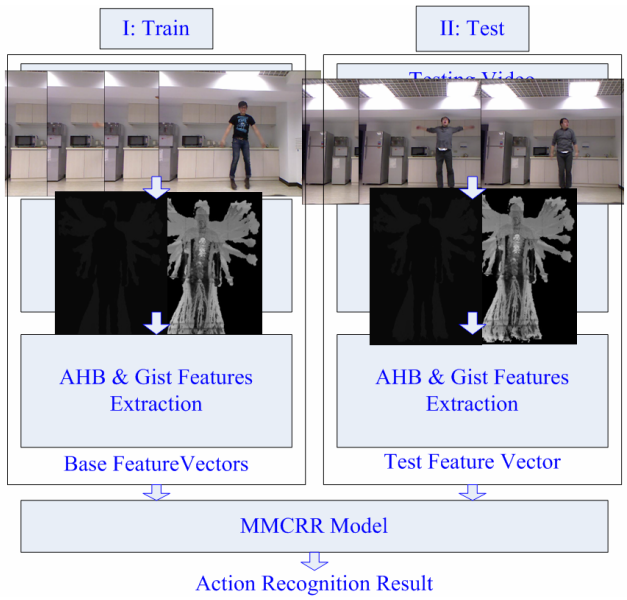



**Fig.3.** General framework of the proposed approach

These descriptors usually focus on depth channel captured by Kinect, and ignore RGB information. However, RGB and depth information represent the target object by two different modalities, and consequently may have certain complementary characteristics each other. Therefore, it is important to fuse both these information for discriminative feature representation and action recognition. In this paper, we propose pyramid appearance and global structure descriptors on both RGB and depth motion history images for human action representation. Furthermore, we propose multi-modality information collaborative representation and recognition model for the fusion of multiple sources. The generalization ability of MMCRR can be easily extended by simply adding new action video as bases and information fusion and recognition can be realized naturally in our model. Fig.3 displays the general framework of the proposed method.

The rest of the paper is structured as follows. In Section 2 the related work is introduced, and then human action representation is detailed. In Section 4 we present the multi-modality information collaborate representation and recognition model. The experimental evaluation and discussion are illustrated in Section 5 and conclusions are given in Section 6.

## 2. Related Work

According to the sensor types, the previous works can be divided into two classes: action recognition based on RGB camera and action recognition based on depth camera. In the past decades, most of researchers mainly focused on analyzing video sources captured by RGB cameras, and the state-of-the-art schemes could be roughly divided into three classes: 1) human silhouette-based method: Modeling the human body by a convex polygon-based star figure and gaussian mixture model was proposed [9]; Bobick and Davis [1] adopted temporal template-- motion energy image (MEI) and motion history image (MHI) to describe the actions and then the similarity matching schemes were adopted to classify actions; Gorelick et al. [2] regarded human actions as three dimensional shapes induced by the silhouettes in the space time volume, and then utilized properties of the solution to the poisson equation to extract space-time features such as local space-time saliency, action dynamics, shape structure and orientation; 2) temporal inference model-based method: Wang et al. [11] utilized hidden conditional random fields (HCRF) to take advantage of the spatiotemporal context to model the latent dynamics of human behavior. Wang and Mori [12-13] developed HCRF and max-margin hidden conditional random fields (MM-HCRF) to recognize actions; Shi et al. [14] proposed semi-markov model to discover the local structure during the complicated action sequence; 3) spatio-temporal interest point-based method: Various robust spatio-temporal interest point detectors and descriptors were proposed in [3-6, 16-18], where the authors extracted spatio-temporal interest points based on different rules, and then adopted a bag-of-word (BoW) scheme to represent and normalize these points, after that, SVM classification model was employed for action modeling. For example, Laptev et al. [5] proposed a 3D Harris corner detector by extending 2D Harris corner detectors, to detect compact and distinctive interest points, which have high intensity variations in both temporal and spatial dimensions. However, experimental results discovered that the variations in all three dimensions were very rigorous and thus very few interest points were produced. Dollar et al. [3] utilized temporal context to detect periodic frequency components by employing Gabor filters and therefore could find much more interest points than the approach in [5]. Although they declared that points with strong periodic responses usually included enough discriminated characteristics, it was still very difficult for them to represent complex actions since human motions are very complicated. Chen et al. [6] proposed MoSIFT detector and descriptor, in which the famous local interest point algorithm was adopted to

detect visually remarkable areas in the spatial domain, and then these candidate interest points were reserved with the motion constraints.

Although these algorithms have obtained good performance, the previous related works mainly focused on analyzing video sources captured by RGB camera. Recently, many researchers are paying attention to action recognition with depth information [7-8, 20, 22]. For example, Lin et al. [7] proposed several kinds of descriptors based on the constructed space-time volume on depth modality, and then employed approximate string matching for classification. Wang et al. [20] employed both the depth and skeleton information to construct the actionlet ensemble model for action recognition. Li et al. [8] proposed a bag-of-3D-point to represent postures by sampling 3D points from depth maps, and then employed action graph to model these points to realize action recognition. Their experimental results on MSR Action3D dataset demonstrated that 3D silhouettes from depth sequence are much helpful for action recognition than 2D silhouettes. Megavannan et al. [22] extracted motion history images for representation and then trained SVM classifier for decision. Different from these model-based methods, some authors have proposed a model-free method for human action recognition via sparse representation [23-25]. Sparse representation and classification (SRC) [26] has been proposed firstly for face recognition, in which SRC codes a testing sample as a sparse linear combination of all the training samples, and then classifies the testing sample by evaluating which class leads to the minimum representation error. Similar to SRC, the philosophy of the proposed method in [23-25] is to decompose each video sample containing one kind of human actions as a $\ell_1$ sparse linear combination of several video samples containing multiple kinds of human actions. The main idea is that the coefficients in such a sparse decomposition reflect the point's neighborhood structure, providing better similarity measures among the decomposed data point and the rest of the data points. After that, Zhang et al. [27] discussed the role of $\ell_1$ -norm and $\ell_2$ -norm respectively, and then concluded that the sparsity in SRC was not so important, and collaborative representation and classification (CRC) played much more important roles.

## 3. Low Level Multi-modality Human Action Representation

In order to encode the dynamics of human silhouettes during one action, RGB and depth motion history image is constructed firstly, and then two kinds of action descriptors are extracted.

### 3.1 Low level RGB motion history image

Motion history images (called **MHI**) [1], in which the dynamic silhouettes are accumulated and encoded, has been widely used to describe human motion. However, it is necessary to segment targets from the background for MHI representation, which itself is nontrivial and difficult. Thus, in the construction of **MHI**, we often make no discrimination for all pixels in RGB image which will have negative impacts for the discriminative ability of **MHI**. The second column in Fig. 4 shows the sample of traditional MHI. It is obvious that MHI is not clear and distinctive enough to describe the motion shape. Fortunately, depth information is very useful to detect the target, and we can adopt depth information to detect the target and filter those static noise pixels, thus, we can more easily capture the target motion process.

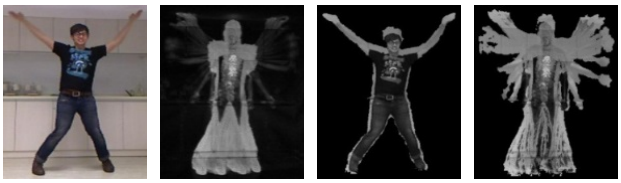


**Fig. 4.** from left to right columns: RGB Image, traditional MHI, RGB filtered by depth and RDMHI of jacking action respectively

In details, RGB image is first pixelwisely multiplied by the corresponding depth image, and then the maximum and minimum values of each pixel in video sequence are calculated. Human motion history image on RGB modality, (called **RDMHI)** is obtained by the subtraction between maximum and minimum image pixels. The third and fourth columns in Fig.4 show their results. The definition of the processing is given as follows:

$$rd(i,j,t) = r(i,j,t) * d(i,j,t), t \in [1...N] \quad (1)$$

$$RD_{\max}(i,j) = \max\{rd(i,j,t) : rd(i,j,t) \neq 0, t \in [1...N]\} \quad (2)$$

$$RD_{\min}(i,j) = \min\{rd(i,j,t) : rd(i,j,t) \neq 0, t \in [1...N]\} \quad (3)$$

$$RDMHI(i,j) = RD_{\max}(i,j) - RD_{\min}(i,j) \quad (4)$$

where $i$ and $j$ is the horizontal and vertical pixel index, $t$ is frame index, and $N$ is the total frame number of an action sequence, $d(i,j,t)$ and $r(i,j,t)$ is the pixel depth value and RGB value in $t$ frame respectively. $r(i,j,t)$ is filtered by $d(i,j,t)$, and then $rd(i,j,t)$ is produced. $RD_{\max}(i,j)$ and $RD_{\min}(i,j)$ are maximum and minimum images respectively, and $RDMHI(i,j)$ is RGB motion history image filtered by depth information.

### 3.2 Low level depth motion history image

Depth information is usually very helpful for foreground segmentation because the object often move within a certain distance to the camera. Therefore a suitable threshold depth value can be set to remove the background in depth image. We proposed depth motion history image for depth motion image sequence to represent the spatial

**Fig. 5** From left to right: RGB Image, Depth Map and DMHI of jacking action

and temporal information about an action.

Supposed there is a depth image sequence $S = \{f_i\}_{i=1}^{N}$ and then maximum and minimum motion energy for each pixel in this sequence is calculated, after that the difference between maximum and minimum images is computed. The details are given as follows:

$$DMHI(i,j) = d_{\max}(i,j) - d_{\min}(i,j) \quad (5)$$

$$d_{\max}(i,j) = \max\{d(i,j,t), d(i,j,t) \neq 0\ t \in (1 \sim N)\ N \geq 2\} \quad (6)$$

$$d_{\min}(i,j) = \min\{d(i,j,t), d(i,j,t) \neq 0\ , t \in (1 \sim N), N \geq 2\} \quad (7)$$

where $i$ and $j$ is pixel index, $t$ is frame index, and $N$ is the total frame number of an action sequence, $d(i,j,t)$ is pixel depth value in $t$ frame. **DMHI** image can not only conveys important shape and motion characteristics of a human movement, but also filter most of background pixels. Fig. 5 shows the corresponding RGB, depth map and depth motion history images respectively.

### 3.3 Low level visual representation

After obtaining **DMHI** and **RDMHI**, we need to design discriminative descriptors to represent the actions. In this paper, we propose two kinds of visual features, representing the characteristics of appearance and structure in pyramid manner.

#### *3.3.1 DMHI-AHB and RDMHI-AHB Descriptors*

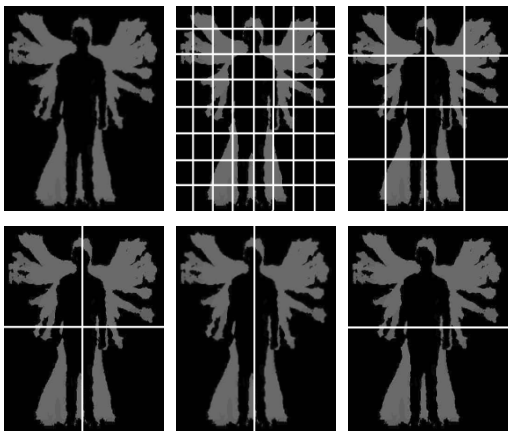


**Fig. 6.** Sample of the construction of pyramid structure in **DMHI**

We first extract a rectangular bounding box of one action in **DMHI** and **RDMHI** images, and then divide the bounding box in pyramid manner into 8*8, 4*4, 2*2, 2*1 and 1*2 blocks respectively. Average value in each hierarchical block (abbreviated as A**HB**) is calculated for local appearance representation and finally concatenated for the representation of **DMHI** and **RDMHI. T**hese descriptors are separately named as **DMHI-AHB** and **RDMHI-AHB** descriptors respectively. Fig. 6 shows one sample of the construction of pyramid structure in **DMHI**.

#### *3.3.2 DMHI-Gist and RDMHI-Gist Descriptors*

The orientation and scale information are helpful for feature representation because people often perform the same actions with different orientations and different distances. Though a lot of perceptual experiments, Oliva *et al.* [28] proposed that perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) were very important for scene representation and they also employed the different scales and orients filter to compute these perceptual dimensions, in which each dimension depicted a meaningful property of the space of the scene. Motivated by this work, we regard that Gist descriptor is helpful for our task. In fact, **DMHI** and **RDMHI** images can be considered as different scenes in which naturalness, openness, roughness, expansion and ruggedness are very different. For example, the roughness dimension of two hands waving will be much bigger than that of one hand waving, the openness dimension of running will be larger than that of jumping. Thus, after obtaining **DMHI** and **RDMHI**, the Gist descriptor is adopted, called **DMHI-GIST** and **RDMHI-GIST** descriptors respectively. In these descriptors, four scales and eight orients filter are employed, and then each filtered **DMHI** and **RDMHI** is divided into 4*4 blocks, and average value of each block is computed, thus, the vector of **DMHI-GIST** and **RDMHI-GIST** descriptors are 512 dimensions.

Since **DMHI-Gist** and **DMHI-AHB** describe local appearance and global structure respectively, they are complementary to each other and consequently **DMHI-Gist** descriptor can be combined with **DMHI-AHB** descriptor by feature concatenation end by end to form **DMHI-AHB-DMHI-Gist** descriptor, and the similar scheme is employed to other descriptors.

## 4. Multi-modality Information Collaborative Representation and Recognition Model

Since we can simultaneously obtain RGB and depth information, thus, we not only need to construct robust and efficient descriptors, but also need to consider how to combine these features for model learning. Sparse signal representation has been proven to be an extremely powerful tool for acquiring, representing, and compressing

high-dimensional signals. This success is mainly due to the fact that important classes of signals such as audio and images have naturally sparse representations with respect to fixed bases (i.e., Fourier, Wavelet), or concatenations of such bases. Moreover, efficient and provably effective algorithms based on convex optimization or greedy pursuit, are available for computing such representations with high fidelity. Thus, motivated by this, we propose multi-modality information collaborative representation and recognition (**MMCRR**) model, which not only inherits the advantage of **CRC**, but also puts feature fusion and recognition into the object function. By this way, we can mine the relationship between color and depth information naturally. Supposed there are $k$ action classes and $n$ training samples, and let $A=[A_1, A_2...A_k]$ be the set of training samples where $A_i$ is the subset of training samples from the *i-th* action class, at the same time, we also assume that $G_i$ and $H_i$ are two different modality features extracted from the *ith* training sample, and $F_i$ is their fused feature. $D_D=[G_1, G_2...G_n]$, $D_R=[H_1, H_2...H_n]$ and $D_C=[F_1, F_2...F_n]$ are the dictionary of depth, color and their fused feature respectively.

Given a test sample $y$, we first extract color and depth modality features, marked as $y_d$ and $y_r$ respectively, and their fused feature is named $y_c$. Denote by $x$ the coding coefficient of $y$ over $D_R$, $D_D$ or $D_C$, such that $y_d \approx D_D * x$. The proposed **MMCRR** model can be formulated as:

$$\hat{x}=\arg\min_x \left\{ \begin{matrix} \lambda_1 \| y_d - D_D x \|_2^2 + \lambda_2 \| y_r - D_R x \|_2^2 \\ +\lambda_3 \| y_c - D_C x \|_2^2 + \lambda_4 \| x \|_2 \end{matrix} \right\} \quad (8)$$

where $\hat{x}=[\hat{x}_1, \hat{x}_2, ...\hat{x}_n]$, $\hat{x}_i$ is the coefficient vector associated with class $i$, and $\lambda_i$ is a scalar constant which controls the weight of each term. In the object function, it is constructed by four stems: I) Fidelity term: The top three stems are adopted to measure the fitting error in different aspects respectively. They can naturally mine respective advantages of different modality features, and enforce the test sample reconstruction with linear combination. II) Collaborative term: the last stem is utilized to induce the collaborative representation. When all four terms are employed, the reconstruction coefficients in $\hat{x}$ corresponding to the class associated with the test sample are non-zero, but the other coefficients are almost zero, thus, it will be very useful to classify. As the object function in Eq. (8) is convex and differentiable in $\hat{x}$. Except for turning to any gradient decent algorithm for optimization, we derive the analytical solution as:

$$\begin{aligned} \hat{x} = & [\lambda_1 D_D^T D_D + \lambda_2 D_R^T D_R + \lambda_3 D_C^T D_C + \lambda_4 I]^{-1} \\ & * [\lambda_1 D_D^T y_d + \lambda_2 D_R^T y_r + \lambda_3 D_C^T y_c] \end{aligned} \quad (9)$$

Thus, in the computing of $\hat{x}$, we just need matrix operation, which can be completed in real time. After collaborative representation, the reconstruction error $e_i$ associated with *ith* class can be calculated with $\hat{x}_i$ the coefficient vector associated with this class, and the formulation in (10) can be further used for human action recognition by choosing the action class with the minimum $e_i$.

$$\begin{aligned} & identify(l) = \arg\min_i (e_i) \\ & e_i = \lambda_1 \| y_d - D_D \hat{x} \|_2^2 + \lambda_2 \| y_r - D_R \hat{x} \|_2^2 \\ & \quad + \lambda_3 \| y_c - D_C \hat{x} \|_2^2 + \lambda_4 \| \hat{x} \|_2 \end{aligned} \quad (10)$$

## 5. Experimental Evaluation and Analysis

We choose **MSR- Action3D** dataset and **DHA** dataset to evaluate the proposed depth and RGB descriptors using **SVM** and **CRC** model respectively, and then **MMCRR** model for action recognition are assessed. In all experiments, we adopt average accuracy as evaluation criterion. The parameters in CRC and MMCRR model were selected by cross validation within the range of [1, $10^{-1}$, $10^{-2}$, $10^{-3}$ and $10^{-4}$].

### 5.1 Action dataset

**MSRACTION3D *dataset*[8]**：It is an action dataset of depth sequences captured by a depth camera. This dataset contains twenty actions, and each action was performed by ten subjects for three times. The frame rate is 15 frames per second and the resolution is 640×480. In total, the dataset has 23797 frames of depth map for 402 action samples.

In order to facilitate a fair comparison, we follow the same experimental settings as [8] to split 20 action categories into three subsets, and the details are given in Table 1. For each subset, there are three different test schemes, i.e. Test One (One), Test Two (Two), and Cross Subject Test ( CrSub ), whose details can be found in [8].

***DHA dataset:*** In this dataset [7], it contains 17 action categories: (1) bend, (2) jack, (3) jump, (4) one-hand-wave, (5) pjump, (6)run, (7)side, (8) skip, (9) two-hand-wave, (10) walk, (11)clap-front, (12) arm-swing, (13) kick-leg, (14) pitch, (15) swing, (16) boxing and (17) tai-chi, and each action was performed by 21 people (12 males and 9 females), such that there are totally 357 videos in DHA dataset, each with both the color and depth data recorded.

**Table 1.** Three subsets of actions in MSRAction3d dataset

| AS1 | AS2 | AS3 |
|---|---|---|
| horizontal arm wave | high arm wave | high throw |
| hammer | hand catch | forward kick |
| forward punch | draw x | side kick |
| high throw | draw tick | jogging |
| hand clap | draw circle | tennis swing |
| bend | two hand wave | tennis serve |
| tennis serve | forward kick | golf swing |
| pick up & throw | side-boxing | pick up & throw |

Although the background in DHA dataset is relative clean, there are some similar actions which will be very difficult to recognize. In addition, Lin *et al.* [7] divided **DHA** dataset into two parts, but in our previous work [29], we had discussed the evaluation protocol of action recognition, and the leave-one-person-out method should be much more reasonable. Thus, we will utilize the leave-one-person-out protocol, and the average accuracy is employed.

### 5.2 Performance evaluation of depth descriptors

We will evaluate our depth descriptors on **MSRACTION3D** dataset and **DHA** dataset with different classification models respectively. We also compare their performances with translation, scale and orientation invariant Hu moments and Gabor feature. In order to compare fairly, all experimental settings are the same. In order to describe conveniently, when **7 Hu moments**, **Gabor**, **AHB** and **Gist** features are extracted from **DMHI**, we marked them as **DMHI-Hu**, **DMHI-Gabor**, **DMHI-AHB** and **DMHI-Gist** respectively. Their performances are given in Table 2 and Table 3.

From them, we can see that no matter what kinds of models are used, the performances of **DMHI-AHB** and **DMHI-Gist** descriptors are much better than that of **DMHI-Hu** and **DMHI-Gabor** descriptors on all three different datasets. Especially, **DMHI-Gist** descriptor achieves the best performance when **CRC** and **SVM** models are employed.

**Table 2.** Performance comparison on MSRAction3d dataset when SVM model and different descriptors are employed

| SVM (RBF) | DMHI-Hu | DMHI-Gabor | DMHI-AHB |
|---|---|---|---|
| AS1One | 0.65 | 0.66 | 0.90 |
| AS2One | 0.61 | 0.50 | 0.91 |
| AS3One | 0.65 | 0.59 | 0.87 |
| **AverOne** | **0.64** | **0.58** | **0.89** |
| AS1Two | 0.72 | 0.76 | 0.96 |
| AS2Two | 0.64 | 0.76 | 0.93 |
| AS3Two | 0.72 | 0.73 | 0.95 |
| **AverTwo** | **0.69** | **0.75** | **0.95** |
| AS1CrSub | 0.61 | 0.51 | 0.69 |
| AS2CrSub | 0.41 | 0.50 | 0.80 |
| AS3CrSub | 0.51 | 0.53 | 0.71 |
| **AverCrSub** | **0.51** | **0.51** | **0.73** |

**Table 3.** Performance comparison on MSRAction3d dataset when CRC model and different descriptors are adopted

| CRC | DMHI-Hu | DMHI-Gabor | DMHI-AHB | DMHI-Gist |
|---|---|---|---|---|
| AS1One | 0.14 | 0.56 | 0.85 | 0.90 |
| AS2One | 0.13 | 0.57 | 0.86 | 0.95 |
| AS3One | 0.13 | 0.51 | 0.84 | 0.90 |
| AverOne | **0.13** | **0.55** | **0.85** | **0.92** |
| AS1Two | 0.17 | 0.78 | 0.93 | 0.99 |
| AS2Two | 0.13 | 0.74 | 0.90 | 0.97 |
| AS3Two | 0.11 | 0.72 | 0.94 | 0.98 |
| AverTwo | **0.14** | **0.75** | **0.93** | **0.98** |
| AS1CrSub | 0.20 | 0.50 | 0.82 | 0.85 |
| AS2CrSub | 0.13 | 0.44 | 0.72 | 0.73 |
| AS3CrSub | 0.23 | 0.56 | 0.88 | 0.88 |
| AverCrSub | **0.187** | **0.50** | **0.807** | **0.82** |

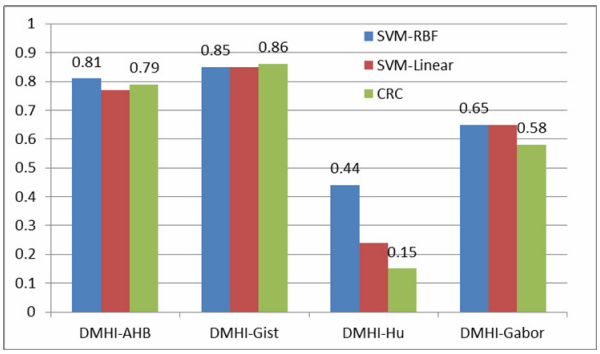



**Fig. 7.** Performance comparison on DHA dataset when different descriptors and different models are used

In addition, we also evaluate our proposed descriptors further on **DHA** dataset, and its performance is shown in Fig. 7. From it, we can know that when **DMHI-AHB** descriptor is extracted, its performance reaches about 80%. When **DMHI-Gist** descriptor is used, its accuracy achieves about 85%. Further, the performances of **DMHI-AHB** and **DMHI-Gist** descriptors are also much better than **DMHI-Hu** and **DMHI-Gabor** descriptors regardless of what kind of different models are trained.

### 5.3 Performance evaluation of RGB descriptor

When obtaining RGB **MHI**, we often need to detect and segment the foreground target for each frame, as the background change or camera move will affect the target motion process. However, in real conditions, it will be very difficult for us to detect and segment the target. Lucky, with the launch of Microsoft Kinect, it will be possible for us to make it true, as Kinect can provides additional skeleton point information and depth information, thus, we can much more easily detect and segment the target according to skeleton point information and depth information. In order to prove it, we make the experiments on DHA dataset to compare the Gist features extracted from RGB MHI (RMHI-GIST) and RDMHI (RDMHI-GIST). Table 4 shows that when depth information is used to obtain the foreground target, the performance of Gist descriptor is 0.894 otherwise its performance just reaches 0.675. Therefore, the depth information is very helpful.

In order to further assess our proposed descriptors, SVM model with different kernels and CRC model are estimated. At the same time, we also compare with the-state-of-art schemes, and their performances are provided in Fig. 8. From it, we can understand that **RDMHI-AHB** and **RDMHI-Gist** descriptors are much better than that of **RDMHI-Hu** and **RDMHI-Gabor** descriptors regardless of what kinds of models are engaged.

**Table 4.** Performance comparison on DHA dataset whether the depth is employed or not

| Feature | Accuracy (SVM-RBF) |
|---|---|
| RMHI-GIST | 0.675 |
| RDMHI-GIST | **0.894** |

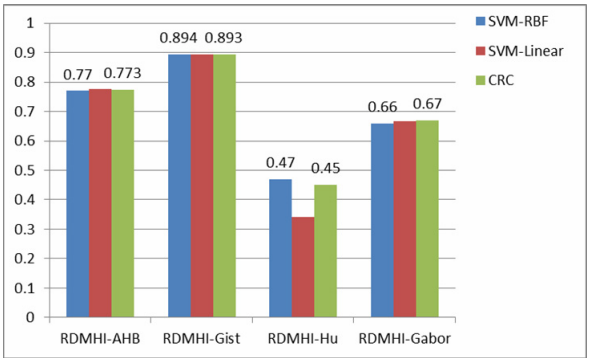



**Fig. 8.** Performance comparison on DHA dataset when different descriptors and models are employed

From above evaluation and analysis, we can conclude that our proposed descriptors can achieve much better performance than the-state-of-the-art schemes in depth and RGB modalities no matter what kind of models are employed. What is more, the performance of RGB information can be improved with the help of depth information.

### 5.4 Performance evaluation of MMCRR model

As different descriptors and different modalities have some complement each other, it would be helpful for action recognition by fusing all of them. In addition, the traditional recognition algorithms do not have good generalization, and when new data are added, the model needs to be retrained. The proposed **MMCRR** model allows more flexibility for modeling since the bases for it can be dynamically changed. Therefore, feature representation, fusion, and classification can be done together.

In order to evaluate the performance of **MMCRR,** we performed the experiments on **MSRACTION3D** and **DHA** datasets, and the direct fusion scheme (two different features are directly concatenated end by end) is also assessed by **SVM** and **CRC** models. In addition, we also compare it with the-state-of-the-art algorithms, and their performances are provided in Table 5 and Fig. 9 respectively. Table 5 shows that when we directly link **DMHI-AHB** and **DMHI-Gist** descriptors and **SVM** and **CRC** models are adopted, their fusion performances can obtain some improvements against **DMHI-AHB** or **DMHI-Gist** descriptors only. For example, the average performances of **DMHI-Gist** descriptor with SVM model on three datasets in Table 2 are 0.9, 0.97 and 0.75

**Table 5.** Performance comparison on MSR Action3d dataset when DMHI-AHB and DMHI-Gist descriptors are employed

| Schemes | Li et al.[8] | SVM | CRC | MMCRR |
|---|---|---|---|---|
| AS1One | 0.90 | 0.92 | 0.93 | 0.94 |
| AS2One | 0.89 | 0.94 | 0.94 | 0.97 |
| AS3One | 0.96 | 0.92 | 0.94 | 0.95 |
| **AverOne** | **0.92** | **0.93** | **0.94** | **0.95** |
| AS1Two | 0.73 | 0.99 | 0.99 | 1.0 |
| AS2Two | 0.93 | 1.00 | 0.96 | 0.97 |
| AS3Two | 0.96 | 0.97 | 0.98 | 0.99 |
| **AverTwo** | **0.87** | **0.99** | **0.98** | **0.99** |
| AS1CrSub | 0.73 | 0.78 | 0.86 | 0.92 |
| AS2CrSub | 0.72 | 0.79 | 0.84 | 0.85 |
| AS3CrSub | 0.79 | 0.77 | 0.82 | 0.93 |
| **AverCrSub** | **0.75** | 0.78 | 0.84 | **0.90** |

respectively, but the average performances of fusion descriptor with SVM model on three datasets in Table 5 are 0.93, 0.99 and 0.78 respectively. The same situation occurs in the **DMHI-Gist** descriptor with **CRC** model. Although the direct feature fusion scheme is useful for action recognition, its improvement is limit. Thus, **MMCRR**, which puts the feature representation, feature fusion and classification together naturally, is leveraged to mine the relationship between different descriptors, and its average performances of **MMCRR** on three datasets in Table 5 are 0.95, 0.99 and 0.90 respectively, whose performance is better than that of **SVM** and **CRC**.

After that, when comparing **MMCRR** to a bag of 3D points [8], its improvement achieves 3%, 12% and 10% respectively. Finally, we also display all the average performance on Cross-Subject dataset by some the-state-of-the-art algorithms, and their performance comparison is given in Fig. 9. From it, we can see that our algorithms obtain the top performance.

Further, when **MMCRR** model is used to fuse these descriptors naturally in the object function, the average performances of these collaborated features are 0.938, 0.9222, 0.928 and 0.952 respectively. When comparing **MMCRR** to SVM model, its improvement reaches 2.6%, 3.2%, 4.8% and 3.2% severally. Similarly, in comparison with **CRC** model, the enhancement of **MMCRR** is about 2%. Finally, though the comparison between **MMCRR** and the-state-of-the-art algorithm (0.87), we can know that our improvement achieves 8.2%. At the same time, the confusion matrix of **MMCRR** and **DMHI-Gist-RDMHI-Gist** is given in Fig. 10**,** and we can see that the accuracies of most actions are above 90% and even 100%, that is to say, we can recognize nine in ten actions.

In total, when different descriptors are linked directly and simply, regardless of what kinds of models are employed, their performances can obtain some improvements when comparing with sole descriptors. Further, when **MMCRR** model is used to fuse and recognize action together, its performance can improve further. In a word, our descriptors and models are robust, stable and efficient.

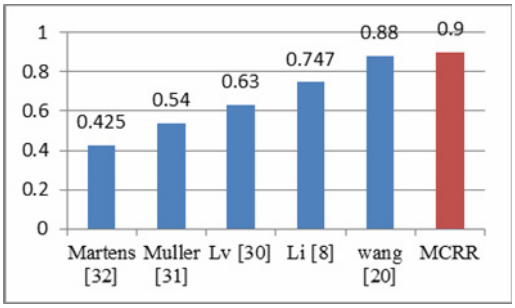



**Fig. 9.** Performance comparison on cross-subject MSRACTION3D dataset

**Table 6.** Performance comparison on DHA dataset

| Descriptors | SVM | CRC | MMCRR |
|---|---|---|---|
| DMHI-Gist-RDMHI-Gist | 0.92 | **0.899** | **0.952** |
| DMHI-AHB-RDMHI-Gist | 0.9 | 0.927 | 0.938 |
| DMHI-Gist-RDMHI-AHB | 0.88 | 0.91 | 0.928 |
| DMHI-AHB-DMHI-Gist | 0.89 | 0.905 | 0.922 |
| DMHI-Gist | 0.85 | 0.86 | / |
| DMHI-AHB | 0.81 | 0.79 | / |
| RDMHI-Gist | 0.89 | 0.88 | / |
| DMHI-AHB | 0.77 | 0.84 | / |
| Lin et al. [7] | | 0.87 | |

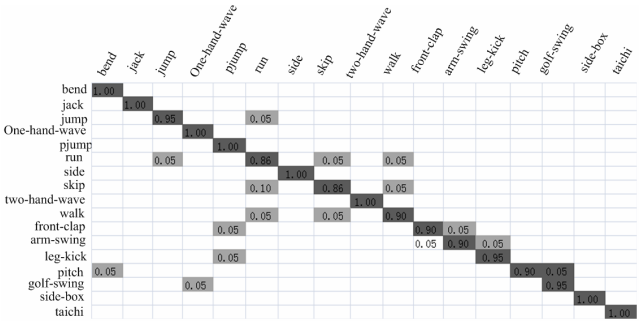



**Fig. 10.** Confusion Matrix of MMCRR and DMHI-Gist-RDMHI-Gist

## 6. Conclusions

In our work, we propose pyramid appearance and global structure action descriptors with both RGB and depth information and multi-modality information collaborative representation and recognition model to recognize human actions. Firstly, we construct the motion history image for both RGB and depth channels. At the same time, depth information is utilized to filter RGB information. And then different action descriptors are extracted from **DMHI** and **RDMHI** to represent these actions, after that, different descriptors are fused and **MMCRR** model are proposed to fuse descriptors and recognize actions. Large-scale comparison experiments on **MSRAction3D** and **DHA** datasets demonstrate that our descriptors are robust, stable and efficient, when comparing with the-state-of-art algorithms, the performance of our descriptors are much better than that of them. Further, the performance of combined descriptors is much better than just using only sole descriptor. What is more, our **MMCRR** model, in which the feature fusion and recognition are put together and naturally, can obtain the top performance on both **MSRACTION3D** and **DHA** datasets, and its best performance can reaches 0.90 and 0.952 respectively.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.61202168, No.61201234, 61100124, 21106095).

## References

[1] A. Bobick and J. Davis. The representation and recognition of action using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3):257-267, 2001. 1, 5, 7.

[2] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. PAMI. 29 (12):2247-2253, 2007.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp 65- 72, 2005.

[4] Schuldt, C. Laptev, and B. I. Caputo. Recognizing human actions: a local SVM approach. ICPR (17), pp 32-36, 2004.

[5] I. Laptev and T. Lindeberg. Space-time interest points. ICCV, pages 432-439, 2003.

[6] M.-y. Chen and A.Hauptmann. MoSIFT: Reocgnizing Human Actions in Surveillance Videos. CMU-CS-09-161, Carnegie Mellon University, 2009.

[7] Yan-Ching Lin, Min-Chun Hua, Wen-Huang Cheng, Yung-Huan Hsieh, Hong-Ming Chen, Human Action Recognition and Retrieval Using Sole Depth Information, ACM MM 2012.

[8] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In Human Communicative Behavior Analysis Workshop (in conjunction with CVPR), 2010. 2, 5, 6.

[9] Chen, D.Y. Efficient polygonal posture representation and action recognition, Electronic Letter, 2011, 47, (2), pp. 101-103.

[10] Kosta, G., Pedro, C., and Benoit, M. Modelization of limb coordination for human action analysis. Proc. IEEE ICIP, Atlanta, CA, USA, 2006, pp. 1765-1768.

[11] Wang, S.B., Quattoni, A., and Morency, L.P., et al.: Hidden conditional random fields for gesture recognition. Proc. IEEE CVPR, New York, NY, USA, 2006, pp. 1521-1527.

[12] Yang Wang, Greg Mori, Max-Margin Hidden Condi-

tional Random Fields for Human Action Recognition, CVPR, 2009
[13] Y. Wang and G. Mori. Learning a discriminative hidden part model for human action recognition, In NIPS 21, 2008.
[14] Qinfeng Shi, Li Wang, Li Cheng, Alex Smola, Discriminative Human Action Segmentation and Recognition using Semi-Markov Model, CVPR,2008.
[15] S.-F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. ICCV, pages 1-8, 2007.
[16] G. Willems, T. Tuytelaars, and L. V. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. ECCV, pages 650-663, 2008.
[17] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. CVPR, pages1-8, 2008.
[18] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. IJCV, 79(3):299-318, 2008.
[19] Yan-Ching Lin, Min-Chun Hua, Wen-Huang Cheng, Yung-Huan Hsieh, Hong-Ming Chen, Human Action Recognition and Retrieval Using Sole Depth Information, ACM MM 2012.
[20] Jiang Wang, Zicheng Liu, Ying Wu, Jusong Yuan, Mining actionlet ensemble for action recognition with depth cameras, in CPRR 2012, pp.1290-1297.
[21] Bingbing Ni, Gang Wang, Pierre Moulin, RGBD-HuDaAct: A Color-Depth Video Database for Human Daily Activity Recognition, ICCV workshop, 2012.
[22] Vennila Megavannan, Bhuvnesh Agarwal R. Venkatesh Babu, Human Action Recognition using Depth Maps, International Conference on Signal Processing and Communications (SPCOM), 2012.
[23] Liu, A. and Han, D. Spatiotemporal Sparsity Induced Similarity Measure for Human Action Recognition. In Proceedings of JDCTA. 2010, 143-149.
[24] Kai Guo, Prakash Ishwar, and Janusz Konrad, Action Recognition Using Sparse Representation on Covariance Manifolds of Optical Flow, 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance, Aug. 29 2010-Sept. 1 2010, pp: 188 - 195.
[25] Changhong Liu, Yang Yang, Yong Chen, Human action recognition using sparse representation, Intelligent Computing and Intelligent Systems, 2009, IEEE International Conference on, 20-22 Nov. 2009, PP. 184-188.
[26] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009.
[27] L. Zhang, M. Yang and X. Feng, "Sparse Representation or Collaborative Representation: Which Helps Face Recognition?" in ICCV 2011.
[28] Oliva A; Torralba A Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, International Journal of Computer Vision, 42(3):145-175, 2001.
[29] Zan Gao, Ming-yu Chen, Alexander G. Hauptmann, Anni Cai, Comparing Evaluation Protocols on the KTH Dataset, International Conference on Pattern Recognition, 2010, pages 88-100.
[30] F. Lv and R. Nevatia. Recognition and Segmentation of 3-D Human Action Using HMM and Multi-class AdaBoost. In ECCV, pages 359-372, 2006. 2,6.
[31] M. Muller and T. R¨oder. Motion templates for automatic classification and retrieval of motion capture data. In Proceedings of the 2006 ACM SIGGRAPH/ Eurographics symposium on Computer animation, pages 137-146. Eurographics Association, 2006. 2, 6, 8
[32] J. Martens and I. Sutskever. Learning Recurrent Neural Networks with Hessian-Free Optimization. In ICML, 2011. 2, 6.

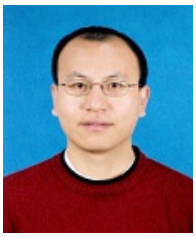
**Zan Gao** is an associate professor in the School of Computer and Communication Engineering, Key Laboratory of Computer Vision and System, Ministry of Education, Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology. From 2009 to 2010, he was a visiting scholar in the School of Computer Science, Carnegie Mellon University, USA. He received his Ph.D degree from Beijin University of Posts and Telecommunications in 2011. His research interests include computer vision, multimedia analysis and retrieval.

**Jian-Ming Song** is pursuing his master degree in the school of Computer and Communication engineering, Tianjin University of Technology. He received his Bachelor degree from Zhengzhou University in 2012. His research interests include computer vision, multimedia analysis and retrieval.

**Hua Zhang** is a professor in the school of Computer and Communication Engineering, Tianjin University of Technology, Tianjin, China. She received her Ph.D degree from Tianjin University in 2008. Her research interests include multimedia analysis and virtual reality.

**An-an Liu** is an associate professor in the school of Electronic Information Engineering, Tianjin University, P.R. China. From 2008 to 2009, he was a visiting scholar in the Robotics Institute, Carnegie Mellon University, USA. His research interests include learning-based computer vision, multimedia analysis and retrieval, biomedical image processing. He is an IEEE member now.

**Yan-Bin Xue** is an associate researcher in the school of Computer and Communication Engineering, Tianjin University of Technology, Tianjin, China. He received his master degree from Tianjin University of Technology in 2005. His research interests include multimedia analysis and computer vision.



**Guang-ping Xu** is an associate professor in the school of Computer and Communication Engineering, Tianjin University of Technology, Tianjin, China. He received his Ph.D and M.S degree from Nankai University in 2009 and 2005 respectively. His research interests include optimal design and performance evaluation of multimedia systems and distributed storage networks.