Chapter 10
RGBD-HuDaAct: A Color-Depth Video Database for Human Daily Activity Recognition

Bingbing Ni, Gang Wang, and Pierre Moulin

Abstract In this chapter, we present a home-monitoring oriented human activity recognition benchmark database, based on the combination of a color video camera and a depth sensor. Our contributions are two-fold: (1) We have created a human activity video database named RGBD-HuDaAct, which contains synchronized color-depth video streams, for the task of human daily activity recognition. This database aims at encouraging research in human activity recognition based on multi-modal video data (color plus depth). (2) We have designed two multi-modality fusion schemes which naturally combine color and depth information from two state-of-the-art feature representation methods for action recognition, namely, spatio-temporal interest points (STIPs) and motion history images (MHIs). These depth-extended feature representation methods are evaluated comprehensively, and superior recognition performance related to their uni-modal (color only) counterparts is demonstrated.

10.1 Introduction

Automatic recognition and analysis of human daily activities (e.g., *go to bed*, *mop the floor*, *eat meal*, etc.) is helpful in a variety of applications, e.g., to facilitate effective delivery of health and medical services to isolated, elderly people. In general, video-based human activity recognition has been an active research topic in computer vision over the last decade. However, the inherent limitations of standard sensing devices restrict previous methods [2, 4, 10, 23] to recognition and analysis of lateral motions. However, human bodies and motions are 3-dimensional, and

B. Ni (✉) · G. Wang
Advanced Digital Sciences Center, Singapore, Singapore
e-mail: bingbing.ni@adsc.com.sg

G. Wang
e-mail: gang.wang@adsc.com.sg

P. Moulin
University of Illinois at Urbana Champaign, Urbana, IL 61801, USA
e-mail: moulin@ifp.uiuc.edu

A. Fossati et al. (eds.), *Consumer Depth Cameras for Computer Vision*,
Advances in Computer Vision and Pattern Recognition,
DOI 10.1007/978-1-4471-4640-7_10, © Springer-Verlag London 2013

so the information loss in the depth channel could cause significant degradation in recognition performance. The recent emergence of Microsoft Kinect depth sensors has made it feasible and economically sound to capture in real-time not only the color images, but also depth maps with appropriate spatial resolution (640×480 in pixel) and amplitude accuracy ($\leq$1 cm accuracy). Both 3-dimensional scene structure information and the 3-dimensional motion information can be extracted. Therefore the motion ambiguity of the color camera resulting from the projection of the 3-dimensional motion onto the 2-dimensional image plane can be circumvented.

To date, very few databases provide joint color and depth data for human activity recognition. To encourage such research, we have constructed a video database named **RGBD-HuDaAct** for human activities captured with a RGB-D (i.e., *color plus depth*) sensor. This database is available upon request to the first author. Though the database is developed under the application scenario of daily activity recognition, it could be used as a common test bed for general activity recognition.

Although it is widely believed that combining color and depth provides complementary information, to our knowledge, no studies have yet shown how much gain (in terms of recognition accuracy) could be obtained by exploring the additional depth modality. To demonstrate the capability of the depth information, we develop two color-depth fusion schemes for feature representation from the most representative feature representation methods in human action recognition. Specifically, we first extend the spatio-temporal interest points methods (STIPs) into a depth-layered multi-channel representation; then, we augment the motion history images (MHIs) with two depth-change induced motion history channels. Extensive experimental results demonstrate the superior performance gained by fusing color and depth information for human activity recognition.

The rest of this chapter is organized as follows: Sect. 10.2 gives a brief review of feature representation methods in activity recognition literature. A detailed introduction to the color-depth human daily activity video database is given in Sect. 10.3. The proposed color-depth fusion schemes for activity feature representation are described in Sect. 10.4. Comprehensive experimental evaluations are given in Sect. 10.5 and Sect. 10.6 draws the conclusion and presents possible directions for future work.

10.2 Related Works

Many feature representation methods have been developed for recognizing activities (actions) from video sequences based on color cameras. Sequences of human silhouettes are utilized to model both spatial and temporal characteristics of human actions. In [4], silhouettes are temporally accumulated to form motion energy images (MEIs) and motion history images (MHIs). Seven Hu moments [14] are extracted from both MEIs and MHIs to serve as action descriptors. Davis and Tyagi [8] use Gaussian mixture models (GMM) to capture the distribution of the moments of silhouette sequences. Several other approaches utilize motion flow patterns to represent human actions. Typically, optical flows [11] are calculated for the

entire image by matching consecutive video frames. Then the motion patterns [10] or the estimated motion parameters [2] are used for action representation. However, ambiguity arises when the real-world 3-dimensional motion is projected onto the 2-dimensional image plane.

Recently, a series of spatio-temporal interest points (STIPs)-based methods have been proposed, which achieve state-of-the-art performances in activity recognition. These methods include Harris3D [18], HOG3D [15] and Cuboid [9]. Although slightly different from each other, these methods share the common feature extraction and representation framework, which involves detecting local extrema of the image gradients and describing the point using histogram of oriented gradients (HOG) [7] and histogram of optic flows (HOF).

The first work using RGB-D sensor for activity recognition is [20]. In [20], a bag of $3D$ points (BOPs) are efficiently sampled from the depth map and Gaussian mixture models are used to model the human postures. This method yields superior results over the conventional method which uses $2D$ silhouettes. However, it has several limitations: (1) Instead of direct utilization of the 3-dimensional motion information, it uses 2-dimensional projections of key poses, which could essentially lead to sub-optimal feature representations; (2) only depth information is used for recognition while color information is completely ignored; however, color and depth information are rather complementary than exclusive.

More recently, Sung et al. [26] directly use skeleton motion data extracted from Kinect SDK for activity representation; however, this method cannot be applied when skeleton data cannot be reliably obtained.

10.3 RGBD-HuDaAct: Color-Depth Human Daily Activity Database

10.3.1 Related Video Databases

A summarization of the existing video activity benchmark databases is given in Table 10.1. **KTH** [25] a**nd Weizmann** [3] **Databases**: These databases aim at simple action recognition, including: *walking*, *jogging*, *running*, *hand-waving*, etc. However, the simplicity of the action categories as well as the clean backgrounds make the recognition tasks easy. As the reported accuracies on both databases approach 94.53 % [16] and 100 % [3, 12], respectively, they are no longer considered as good benchmarks. Instead, the RGBD-HuDaAct aims at realistic human daily activities, which are challenging for recognition tasks. **Movie Action Database** [22]: This database is widely used for activity recognition in movies. Given the large variations of the visual contents and the camera movements, this database is challenging. Note that although some of its activity categories overlap with the RGBD-HuDaAct database, the two databases focus on different applications, i.e., the former deals with movie actions under uncontrolled environment with moving cameras, while the latter is for daily activity monitoring under fixed environment and camera settings. **Sports Event Databases** [21, 24]: The UCF sports event database [24] and

Table 10.1 Comparisons of the RGBD-HuDaAct database over other benchmark activity databases

Database	Modality	Resolution	Sample #	Category descriptions
KTH [25]	RGB	160 × 120	2391	6 classes: walking, jogging, running, etc.
Weizmann [3]	RGB	180 × 144	90	10 classes: run, walk, skip, jumping-jack, side, etc.
Hollywood2 [22]	RGB	600 × 450	3669	12 classes: answering the phone, driving car, eating, etc.
UCF Sports [24]	RGB	720 × 480	184	10 classes: swinging, golf swinging, walking, etc.
UCF YouTube [21]	RGB	320 × 240	3040	11 classes: basketball shooting, biking, diving, etc.
MSR Action3D [20]	Depth	320 × 240	4020	20 classes: high arm wave, hand catch, forward punch, etc.
Indoor Activity [26]	RGB-Depth	640 × 480	NA	12 classes: cooking, writing, working on computer, etc.
RGBD-HuDaAct	RGB-Depth	640 × 480	1189	12 classes (plus background activity): drink water, eat meal, phone call, etc.

the UCF YouTube sports database [21] consist of a set of actions collected for various sports events which are typically obtained from websites including BBC Motion gallery, GettyImages, and YouTube.com. These two databases are very challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, and illumination condition, etc. While these two databases consider only outdoor sports, the daily activities in the RGBD-HuDaAct database are all indoor. **MSR Action3D Database** [20]: The only existing depth sensor-based action database is collected by Li et al. [20], which aims at recognizing actions (gestures) in game interaction. However, this database only contains depth maps without corresponding color images. In contrast, the RGBD-HuDaAct database contains synchronized and registered color-depth videos. Used for gesture recognition, this database contains only atomic actions such as hand wave, punch, etc. In contrast, our database aims at higher level human behavior such as mopping the floor, eating meal, etc. **Indoor Kinect Activity Database** [26]: Very recently, Sung et al. [26] use Kinect sensor to construct and indoor (e.g., office, kitchen, bedroom, bathroom, and living room) activity dataset for the task of activity detection, which includes four subjects and 12 activity categories. In addition to RGB-D images, the database also provides skeleton motion data. Most of their categories do not overlap with ours. To have more inter-personal variations, the number of subjects participating our data collection (i.e., 30) is much larger than theirs.

Differently from these databases, our motivation is driven by the application of assisted living in health-care. Monitoring the daily activities of senior citizens has recently become an urgent demand due to the aging population problem. There only exists a very recent video database for senior home monitoring [6], however, it does

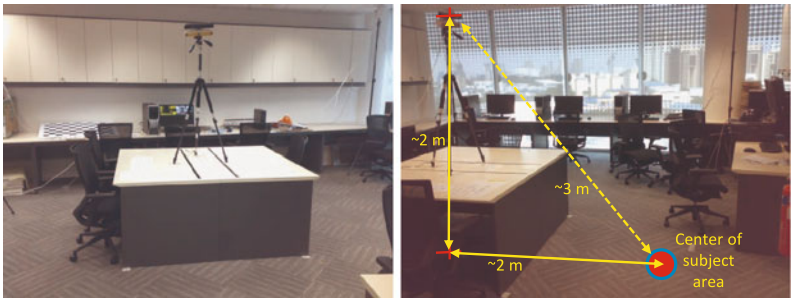

Fig. 10.1 The Kinect camera setup. (*Left*) The video capture environment. (*Right*) The geometric configuration of the Kinect camera

not utilize the depth modality. In contrast, the RGBD-HuDaAct database contains synchronized color and depth videos, which are more suitable for 24 hours monitoring, since the depth sensor also works without visible lighting.

10.3.2 Database Construction

We utilize the recently released Microsoft Kinect sensor to construct the RGBD-HuDaAct video database, collected in a lab environment, which is illustrated in Fig. 10.1. There are minor variations in the camera position and orientation due to repeated mountings of the camera. From Fig. 10.1, it can be noted that the horizontal and vertical distances from the camera to the center of the scene under capture are about 2 and 2 meters, respectively and the average depth of the human subject in the scene is about 3 meters (i.e., which is the optimal operation range of the depth camera). This geometric setting is appropriate for home or hospital ward monitoring. The resolutions of both color image and depth map are 640×480 in pixel. The color image is of 24-bit RGB values; and each depth pixel is a 16-bit integer. Both sequences are synchronized and the frame rates are 30 frames per second (fps). The color and depth frames are stereo-calibrated using the standard stereo-calibration method with a chessboard pattern object available in OpenCV (four corners of the chessboard object are used as corresponding points for depth calibration, as in [1]). We repeat the camera calibration procedure at the beginning of each video capture session and the camera is fixed throughout the session. The database can be downloaded at: http://adsc.illinois.edu/research/ADSC-RGBD-dataset-download-instructions.pdf.

10.3.3 Database Statistics

We are interested in 12 categories of human daily activities motivated by the definitions provided by health-care professionals [17] for *Activity of Daily Living (ADL)*,

which includes: *make a phone call, mop the floor, enter the room, exit the room, go to bed, get up, eat meal, drink water, sit down, stand up, take off the jacket and put on the jacket*. These defined activities are directly corresponding to the ADL category: *using the telephone, maintaining the home, eating, transferring, dressing*, respectively (note that other ADL categories such as *toileting, bathing, managing finances, shopping* are not suitable for visual recognition). We also have a category named as *background activity* that contains different types of random activity. We invited 30 student volunteers to perform these daily activities, which are organized into 14 video capture sessions. The subjects were asked to perform each activity 2–4 times. Finally, we captured about 5,000,000 frames (approximately 46 hours long) for a total of 1189 labeled video samples. Each video sample spans about 30–150 seconds. Note that the size of our database is still growing to include more activity classes and video samples.

Two example frames from each activity category are illustrated in Fig. 10.2, in terms of both color (left) and depth (right) frames. We can make two observations from Fig. 10.2: (1) There exist distinctive depth layers for the moving human body parts in different activities, which implies that incorporating the depth layer information could bring additional discriminating capability for activity feature representation; (2) there exist rich intra-class variations for each activity category.

For example, for the activities *make a phone call* and *drink water*, the subject could be either standing still or sitting on the chair and either hand could be used for phone answering and water drinking. As another example, for the activities *put on the jacket* and *take off the jacket*, different persons have their own styles of performing these actions and they might be facing or not facing the camera. These variations make our database more realistic and challenging.

Note that although the background of the current database is of limited variations and only a single subject is present (i.e., compared to the movie action or YouTube databases), we must emphasize that for the application of indoor home monitoring, using a fixed camera and the current background environment are very typical. One limitation of the current sensor is that the operation range is fixed at about 3 meters and the camera view angle is also fixed. However, in real applications, the actions can occur at any distance with different view angles. Therefore, we are currently collecting more data with various distance ranges and view angles. Also the effective range of the Kinect is limited within 6 meters, and we are currently investigating a multiple-Kinect setup to cover the whole space.

10.4 Color-Depth Fusion for Activity Recognition

In this section, we introduce two feature representation methods for fusing color and depth information for activity recognition, which are straightforwardly developed from two state-of-the-art action representation methods, i.e., spatial-temporal interest points (STIPs) and motion history images (MHIs). On the one hand, we derive a **Depth-Layered Multi-Channel STIPs** (DLMC-STIPs) framework which

Fig. 10.2 Example color and depth frames from each activity category. Note for the depth map, *brighter pixels* mean larger depth values. *Some black regions* correspond to depth measurement errors due to surface reflections, i.e., the PC screen

divides the spatio-temporal interest points into several depth-layered channels, and then STIPs within different channels are pooled independently, resulting in a multiple depth-channel histogram representation. On the other hand, we propose a **3-Dimensional Motion History Images** (3D-MHIs) approach which equips the conventional motion history images (MHIs) with two additional channels encoding the motion recency history in the depth-changing directions. In the experiments, these two color-depth-based feature representation methods are comprehensively evaluated over their color-only counterparts. It is demonstrated that by modeling the 3-dimensional spatial structure of the detected spatio-temporal feature points as well as the 3-dimensional motion history of the human subjects, the discriminating capabilities of the features are boosted.

10.4.1 Depth-Layered Multi-channel STIPs (DLMC-STIPs)

Spatio-temporal interest points (STIPs) are widely used for action recognition. The most representative versions of STIPs employ the Harris3D detector, which was proposed by Laptev and Lindeberg in [18]. The Harris3D detector is a space-time extension of the 2-dimensional Harris detector [13]. At each space-time video point, a spatio-temporal second-moment matrix is computed as $\mu(.;\sigma,\tau) = g(.;s\sigma,s\tau) * (\Delta V(.;\sigma,\tau))(\Delta V(.;\sigma,\tau))^T$ (i.e., V is the video volume), in terms of different spatial and temporal scale values $s\sigma, s\tau$. Namely, space-time gradients ΔV are

computed and smoothed by a separate Gaussian smoothing function $g(.; s\sigma, s\tau)$. The detected locations of space-time interest points are given by local extrema of $H = \det(\mu) - \kappa\,\mathrm{trace}^3(\mu)$, in terms of both spatial and scale space. To characterize local shapes and motions, histograms of oriented gradients (HOG) and histograms of optic flows (HOF) are calculated within the space-time neighborhoods of the detected interest points, see [18]. The HOG and HOF feature descriptors are first quantized into visual words and then each video sequence is represented as a bag of such visual words [27] (i.e., as a histogram vector over the visual word vocabulary).

However, the human subject is in essence a 3-dimensional structure and the detected spatio-temporal feature points are associated with local motions taking place at different 3-dimensional locations; however, the previous pooling methods of STIPs can only utilize this spatial information up to 2-dimensional, i.e., feature poolings are performed within each x–y–t sub-volume, and the spatial information along the depth direction is totally lost. The availability of depth map enables us to recover this lost information. The most straightforward way to utilize the spatial information along the depth direction is to perform the feature pooling by dividing the entire scene into different depth layers, and form a multi-channel STIPs histogram. This basic idea is similar with the space partition in [18], where STIPs are spatially pooled within each x–y–t sub-volume, i.e., the entire 3-dimensional space-time video volume is divided into several x–y–t sub-volumes. Our proposed framework is named as **Depth-Layered Multi-Channel STIPs (DLMC-STIPs)**, which is formulated as follows.

Each video sample V could be represented as a set of (N) STIP feature descriptors (i.e., HOG and HOF), which is denoted $V = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. Each STIP feature descriptor is denoted $\mathbf{x}_i = (x, y, z, t, \sigma, \mathbf{x}_{\mathrm{HOG}}^T, \mathbf{x}_{\mathrm{HOF}}^T)^T$. Here, x, y, z, t, σ represent the $3D$ coordinate (x, y, z), temporal index and the scale of the detected feature point, respectively. $\mathbf{x}_{\mathrm{HOG}}$ and $\mathbf{x}_{\mathrm{HOF}}$ are the $72D$ HOG and $90D$ HOF feature vectors, respectively. We first perform unsupervised clustering on the set of HOG and HOF feature descriptors to construct a visual word vocabulary (codebook). We denote the visual codebook encoded vector (by nearest visual word assignment according to the Euclidean distance) of the feature descriptor $\mathbf{x}_i$ as $\mathbf{v}_i$, i.e., $\mathbf{v}_i$ is a K-dimensional (K is the codebook size) assignment vector with one of the element as 1 and the others as 0s. Then the histogram representation $\mathbf{h}$ for the video sample V is given by

$$\mathbf{h} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i. \tag{10.1}$$

This aggregation process is usually referred as *feature pooling*, i.e., aggregating the set of local features into a global representation vector.

We can also incorporate the spatial information in the feature pooling process. In [18], the entire 3-dimensional space-time volumes are divided into several x–y–t sub-volumes and pooling is performed within each sub-volume. Then the pooled histogram vectors from all the sub-volumes are concatenated to form a multi-channel representation. When the depth value of each detected STIP point is avail-

able, we can also form depth-layered multi-channel representations. Namely, we introduce a set of (M) depth layers $L_1^z = [z_1^l, z_1^u]$, $L_2^z = [z_2^l, z_2^u], \ldots, L_M^z = [z_M^l, z_M^u]$, with lower and upper boundaries denoted as z_m^l and z_m^u for the mth depth layer. Then, we form multi-channel histograms $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M$, as

$$\mathbf{h}_m = \frac{1}{N} \sum_{z(\mathbf{x}_i) \in L_m^z} \mathbf{v}_i, \quad \forall m = 1, 2, \ldots, M. \tag{10.2}$$

These multiple channel histograms could be concatenated into an $M \times K$-dimensional feature vector $\mathbf{h} = (\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_M^T)^T$, as the input to the classification framework, e.g., support vector machines. The distance metric for calculating the kernel matrix could be χ^2 distance. Moreover, we can also use the spatial pyramid matching kernel (SPM) proposed in [19] to better explore the spatial information given in the depth axis. An illustration of the DLMC-STIPs generation process is given in Fig. 10.3. Note the following. (1) The DLMC-STIPs method is not fully $4D$ representation, since the interest point detection and local volume representation are both performed in the x–y–t space. However, improvement has been observed when the local features are not distinctive with this naive extension (see Sect. 10.5). We believe this trial idea (together with the database) will inspire the research community to develop more sophisticated approaches which represent activities in a fully $4D$ manner. (2) The DLMC-STIPs framework does not explicitly model the motion along the depth axis, and a 3D-MHIs approach which explicitly models the 3-dimensional motion is introduced in the next subsection.

10.4.2 3-Dimensional Motion History Images (3D-MHIs)

Another widely used feature representation method for action classification is motion history images (MHIs) developed by Bobick and Davis [4], which is capable of encoding the dynamics of a sequence of moving human silhouettes. In an MHI, each pixel intensity is a function of the motion recency at that location, where brighter value corresponds to more recent motion. This single image contains the discriminative information for determining how a person has moved (spatially and temporally) during the action. Denoting $I(\mathbf{x}, \mathbf{y}, t)$ as an image sequence, each pixel intensity value in an MHI is a function H^I of the temporal history of motion at that point, namely:

$$H_\tau^I(x, y, t) = \begin{cases} \tau, & \text{if } |I(x, y, t) - I(x, y, t-1)| > \delta I_{th} \\ \max(0, H_\tau^I(x, y, t-1) - 1), & \text{else.} \end{cases} \tag{10.3}$$

Here τ is the longest time window we want the system to consider and δI_{th} is the threshold value for generating the mask for the region of motion. The result is a scalar-valued image where brighter pixels indicate more recent motion. Statistical descriptions of the motion history images are then computed based on seven Hu

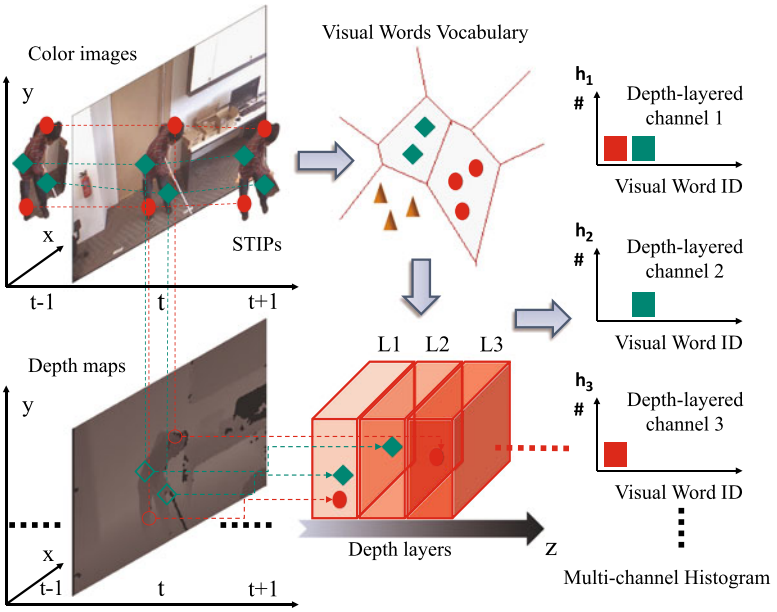

Fig. 10.3 A diagram of the generation process of DLMC-STIPs representation

moment-based features [14], which are known to yield reasonable shape discrimination in a translation- and scale-invariant manner.

However, using only RGB camera, MHIs can only encode the history of motion induced by the lateral component of the scene motion parallel to the image plane. With the additional depth sensor, we can now develop an extended framework which is capable of encoding the motion history along the depth-changing directions. In particular, we propose two depth-change induced motion history images named as DMHIs. DMHIs contain forward-DMHIs (fDMHIs) which encode the forward motion history (increase of depth) and backward-DMHIs (bDMHIs) which encode the backward motion history (decrease of depth). To generate fDMHIs, the following process is adopted:

$$H_\tau^{fD}(x,y,t) = \begin{cases} \tau, & \text{if } (D(x,y,t) - D(x,y,t-1)) > \delta D_{th} \\ \max(0, H_\tau^{fD}(x,y,t-1) - 1), & \text{else.} \end{cases} \tag{10.4}$$

Here, H_τ^{fD} denotes the forward motion history image and $D(x,y,t)$ denotes the depth sequence. δD_{th} is the threshold value for generating the mask for the region of forward motion. The backward-DMHI (i.e., H_τ^{bD}) is generated in a similar way with the thresholding function replaced by $(D(x,y,t) - D(x,y,t-1)) < -\delta D_{th}$. The conventional MHIs are combined with fDMHIs and bDMHIs to represent 3-dimensional motion history and we denote the combined feature representation as **3D-MHIs**. To represent each action video, similar to MHIs, Hu moments are calculated for all three channels (i.e., MHIs, fDMHIs and bDMHIs) and are concate-

Fig. 10.4 Illustration of the MHI, fDMHI and bDMHI in a *sit down* sequence

nated to form a representation vector. An example 3D-MHI is illustrated in Fig. 10.4 in the context of a *sit down* sequence. From Fig. 10.4, we notice obvious motion patterns in fDMHI in contrast to bDMHI, which indicates the subject is moving away from the camera. This example implies that by using fDMHIs and bDMHIs, we can distinguish different actions which present similar motion patterns in the $x-y$ directions but with distinctive motion patterns in the depth-changing directions.

10.5 Experimental Evaluations

10.5.1 Evaluation Schemes

In this work, we use 59 % (i.e., by random sampling a fixed number of samples from each category) of the video samples in the RGBD-HuDaAct database for experiment. The subset we use in the experiments include 18 subjects with nine capture sessions, yielding a total of 702 video samples belonging to 13 activity classes, including the *background activity* videos which are added to the existing 12 activity classes to test how algorithms can recognize the specified activities from some random daily activities such as walk around, stand still, pick-up some object, etc.

To test the generalization capability of the methods for novel input, we use the leave-one-subject-out (LOSO) scheme for algorithmic evaluations. In each run, we choose the samples from one subject as the testing samples, and the remaining samples from the database serve as the training samples. The overall recognition performance is calculated by gathering the results from all training-testing runs.

The evaluation results are reported in terms of classification accuracy as well as class confusion matrix. We regard our human daily activity recognition problem as a multi-class classification problem and each video sample has one and only one activity label (i.e., out of 13 classes). For the LOSO scheme, the classification accuracy is given by the ratio of the correctly classified testing samples over the total number of testing samples, by gathering the classification results from all testing runs. In our experiments, the class confusion matrix C is a 13×13 matrix where each element C_{ij} denotes how many testing samples of the ith class are classified into the jth class. Larger values for the diagonal elements and smaller values for the off-diagonal elements indicate better discriminating capability.

Table 10.2 Comparisons of the classification accuracies (%) for STIPs and DLMC-STIPs under different experimental settings

Setting	$K = 128$	$K = 256$	$K = 512$
STIPs (χ^2)	68.95	76.78	79.77
DLMC-STIPs (χ^2, $M = 2$)	72.43	77.10	79.91
DLMC-STIPs (χ^2, $M = 4$)	74.22	77.91	79.23
DLMC-STIPs (χ^2, $M = 8$)	76.64	79.49	79.49
DLMC-STIPs (SPM)	**77.64**	**81.05**	**81.48**

Prior to feature extraction, we down-sample the original color and depth video sequences in both spatial and temporal dimensions by a factor of 2, yielding 320×240 pixels and 15 fps video samples (i.e., this setting is similar with [20]). We use support vector machines (SVM) [5] (*one-against-one* scheme for multi-class classification) for all classification tasks with different kernels. The penalty parameter $\mathcal{C}$ of SVM is optimized by cross-validation. The bandwidth parameters for χ^2 and RBF kernels are set as the average of the squared distances (χ^2 and Euclidean, respectively) of the training sample pairs.

10.5.2 DLMC-STIPs vs. STIPs

We compare the classification performances between the proposed DLMC-STIPs and the conventional STIPs. We perform K-means clustering to the set of HOG + HOF descriptors, which yields codebooks with size K. We vary the value of K as 128, 256 and 512 for more comprehensive evaluations. For the conventional STIPs, a K-dimensional histogram vector is calculated for representing each video sequence. Note that in order to better reveal the discriminating capability gained by depth-layered multi-channel representation, we fix the setting of other configurations as simple as possible, i.e., we do not partition the STIPs into different x–y–t sub-volume as in [18]. Obviously, space partition in terms of x–y–t for both methods could bring more discriminative information on an equal basis. For DLMC-STIPs, we divide the depth axis into $M = 2, 4, 8$ equally spaced layers according to the depth value distributions of the SITPs. As both DLMC-STIPs and STIPs are histogram-based representations, we use χ^2 distance for calculating the kernel matrix. We also explore the spatial pyramid matching kernel (SPM) [19] for DLMC-STIPs representations with $l = 3$ depth spatial levels. Various classification accuracies under different parameter combinations are given in Table 10.2. We also illustrate the class confusion matrices for both methods in Fig. 10.5, at the setting of $K = 256$.

It can be observed from Table 10.2 that by using depth-layered multi-channel histogram representation, the classification accuracies can be improved consistently; also, by using the spatial pyramid matching kernel (SPM), the classification performances can be further boosted.

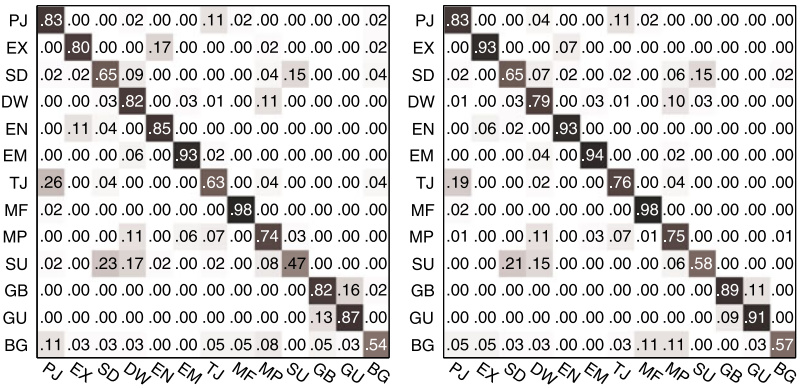

Fig. 10.5 Class confusion matrices for STIPs (*left*) and DLMC-STIPs (*right*, SPM kernel) under the setting of $K = 256$. For better view, we use two characters to represent each activity category, i.e., PJ: *put on the jacket*, TJ: *take off the jacket*, EN:*enter the room*, EX: *exit the room*, SD: *sit down*, SU: *stand up*, DW: *drink water*, EM: *eat meal*, MF: *mop the floor*, MP: *make a phone call*, GB: *go to bed*, GU: *get up* and BG: *background activity*

Table 10.3 Comparisons of the classification accuracies (%) for MHIs and 3D-MHIs under different experimental settings

Kernel	MHIs	fDMHIs + bDMHIs	3D-MHIs
Linear	34.19	68.66	**70.51**
RBF	37.18	66.81	**69.66**

10.5.3 3D-MHIs vs. MHIs

We also compare the classification performances between the proposed 3D-MHIs and the conventional MHIs. For both methods, the τ value is chosen by cross-validations. We further normalize the 3D-MHIs and MHIs by multiplying a scale factor $\frac{1}{\tau}$ to achieve scale invariance. Note that the original implementation of MHIs as in [4] uses a multiple view configuration. In this work, however, we use a single view instead. For SVM classification, we explore both the linear kernel and the RBF kernel, and the classification results are given in Table 10.3. We again show the class confusion matrices for both methods in Fig. 10.6, for the case of linear SVM.

From Table 10.3 and Fig. 10.6, it is noted obviously that by adding the two depth-changing induced motion history images, the discriminating capability of the feature representation is significantly boosted (by nearly 30 %). Furthermore, from Fig. 10.6, we see that the activity *enter the room* is quite easy to confuse with the activities *exit the room* and *mop the floor* due to their similar lateral motion patterns; however, by using 3D-MHIs, these ambiguities are significantly eliminated, since both *enter the room* and *exit the room* include abundant and distinctive depth-changing information.

We also compare the best results obtained from our RGB-D-based methods, i.e., 3D-MHIs and DLMC-STIPs (SPM) with the state-of-the-art action recognition

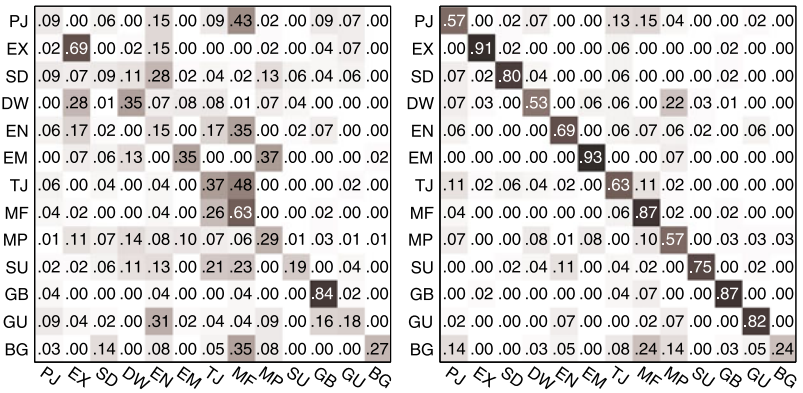

Fig. 10.6 Class confusion matrices for MHIs (*left*) and 3D-MHIs (*right*), at the setting of linear SVM

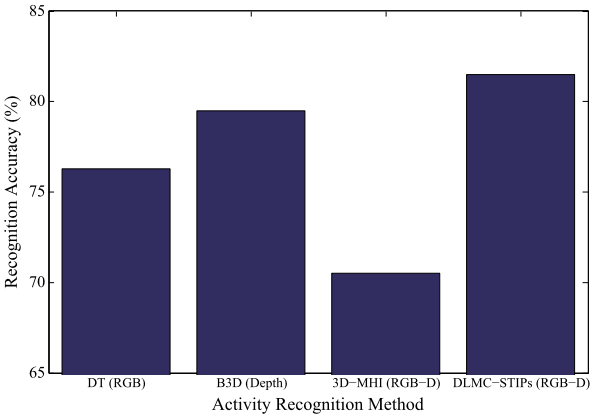

Fig. 10.7 Comparison of recognition accuracies using DT, B3D, 3D-MHIs and DLMC-STIPs (SPM)

methods using RGB images, e.g., dense trajectories (DT) [28] and depth images, e.g., bag of 3D points (B3D) [20]. The related parameters for these comparing methods (e.g., trajectory length, number of visual words of trajectory descriptors, number of mixtures and number of states for bag of 3D points method, the 3D points sampling rate) are tuned optimally on a validation subset. The comparison is illustrated in Fig. 10.7. We can see that fusion RGB and depth information (DLMC-STIPs (SPM)) outperforms single modality-based methods.

10.6 Conclusions

In this work, we introduced a publicly available color-depth video database for human daily activity recognition. We also presented two fusion schemes combining color and depth modalities for action representation, which have shown superior

recognition performances over their color-only counterparts. We hope this database could serve as a benchmark test bed of color-depth-based algorithms for home monitoring oriented activity recognition. In the future, we will extend the current database with actions captured from different distance ranges and view angles.

Acknowledgements This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

References

1. http://nicolas.burrus.name/index.php/research/kinectcalibration
2. Black, M.J., Yacoob, Y., Jepson, A.D., Fleet, D.J.: Learning parameterized models of image motion. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 561–567 (1997)
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: International Conference on Computer Vision, pp. 1395–1402 (2005)
4. Bobick, A., Davis, J.: The representation and recognition of action using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2001)
5. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011)
6. Cheng, H., Liu, Z., Zhao, Y., Ye, G.: Real world activity summary for senior home monitoring. In: IEEE International Conference on Multimedia and Expo (2011)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
8. Davis, J.W., Tyagi, A.: Minimal-latency human action recognition using reliable-inference. Image Vis. Comput. **24**(5), 455–472 (2006)
9. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (2005)
10. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: International Conference on Computer Vision (2003)
11. Fleet, J.L.B.D.J., Beauchemin, S.S.: Performance of optical flow techniques. Int. J. Comput. Vis. **12**(1), 43–77 (1994)
12. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. **29**(12), 2247–2253 (2007)
13. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, pp. 147–151 (1998)
14. Hu, M.: Visual pattern recognition by moment invariants. IRE Trans. Inf. Theory **8**(2), 179–187 (1962)
15. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d gradients. In: British Machine Vision Conference (2008)
16. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition
17. Krapp, K.: Activities of Daily Living Evaluation. Encyclopedia of Nursing and Allied Health (2002)
18. Laptev, I., Lindeberg, T.: Space-time interest points. In: IEEE International Conference on Computer Vision (2003)

19. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition (2006)
20. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: IEEE Conference on Computer Vision and Pattern Recognition—Workshop on Human Communicative Behavior Analysis (2010)
21. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
22. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
23. Ni, B., Yan, S., Kassim, A.: Recognizing human group activities with localized causalities. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
24. Rodriguez, M., Ahmed, J., Shah, M.: Action MACH: a spatio-temporal maximum average correlation height filter for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
25. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: IEEE International Conference on Pattern Recognition (2004)
26. Sung, J., Ponce, C., Selman, B., Saxena, A.: Human activity detection from RGBD images. In: AAAI Workshop on Pattern, Activity and Intent Recognition (2011)
27. Ullah, M.M., Parizi, S.N., Laptev, I.: Improving bag-of-features action recognition with non-local cues. In: British Machine Vision Conference (2010)
28. Wang, H., Kläser, A., Schmid, C., Cheng-Lin, L.: Action recognition by dense trajectories. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3169–3176 (2011)