RGBD-HuDaAct: A Color-Depth Video Database For Human Daily Activity Recognition

Bingbing Ni
Advanced Digital Sciences Center
Singapore 138632
bingbing.ni@adsc.com.sg

Gang Wang
Advanced Digital Sciences Center
Singapore 138632
gang.wang@adsc.com.sg

Pierre Moulin
UIUC
IL 61820-5711 USA
moulin@ifp.uiuc.edu

Abstract

*In this paper, 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 publicly releasable human activity video database (*i.e., *named as **RGBD-HuDaAct**), which contains synchronized color-depth video streams, for the task of human daily activity recognition. This database aims at encouraging more research efforts on human activity recognition based on multi-modality sensor combination (*e.g., *color plus depth). 2) Two multi-modality fusion schemes, which naturally combine color and depth information, have been developed from two state-of-the-art feature representation methods for action recognition,* i.e., *spatio-temporal interest points (STIPs) and motion history images (MHIs). These depth-extended feature representation methods are evaluated comprehensively and superior recognition performances over their uni-modality (*e.g., *color only) counterparts are demonstrated.*

1. Introduction

Being able to recognize and analyze human daily activities (*e.g.*, *go to bed*, *mop the floor* and *eat meal* etc.) in a low cost and intelligent way (*e.g.*, vision-based) for elderly people living-alone is essential for further providing them with appropriate health and medical services [1]. Video-based (color camera) Human activity (action) recognition has been an active research topic in computer vision over the last decade. However, the inherent limitation of the sensing device (*i.e.*, color camera) restricts previous methods [5, 11, 3, 24] to be only capable of describing lateral motions. As human bodies and motions are in essence three-dimensional, the information loss in the depth channel could cause significant degradation of the representation and discriminating capability for these feature representations. Recent emergence of depth sensor (*e.g.*, Microsoft Kinect) has made it feasible and economically sound to capture in real-time not only the color images, but also depth maps with appropriate resolution (*e.g.*, 640×480 in pixel) and accuracy (*e.g.*, $\leq 1cm$). It can provide three-dimensional structure information of the scene as well as the three-dimensional motion information of the subjects/objects in the scene. Therefore the motion ambiguity of the color camera, *i.e.*, projection of the three-dimensional motion onto the two-dimensional image plane, could be bypassed.

However, there exist very few works or benchmark databases that utilize the color-depth sensor combination for human activity recognition. To encourage more research efforts, in this work, we collect a public releasable video database (named as **RGBD-HuDaAct**) for human activities, under the RGB-D setting, *i.e.*, *color plus depth*. 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 tasks.

To demonstrate the capability of the depth information, in this work, two color-depth fusion schemes for feature representation are developed 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 multichannel representation; then, we augment the motion history images (MHIs) with two depth change induced motion history channels. Extensive experimental results well demonstrate the superior performances gained by fusing color and depth information for human activity recognition.

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 [5], silhouettes are temporally accumulated to form motion energy images

(MEIs) and motion history images (MHIs). Seven Hu moments [15] are extracted from both MEIs and MHIs to serve as action descriptors. Davis and Yyagi [9] 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 [12] are calculated for the entire image by matching consecutive video frames. Then the motion patterns [11] or the estimated motion parameters [3] are used for action representation. However, ambiguity arises when the real-world three-dimensional motion is projected onto the two-dimensional image plane.

Recently, a series of spatio-temporal interest points (STIPs) based methods have been proposed, which achieve the state-of-the-art performances in activity recognition. These methods include Harris3D [19], HOG3D [16] and Cuboid [10]. Although slightly different from each other, these methods share the common feature extraction and representation framework, which involves detecting local extremes of the image gradients and describing the point using histogram of oriented gradients (HOG) [8] and histogram of optic flows (HOF).

The first work using RGBD sensor for activity recognition is [21]. In [21], 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 three-dimensional motion information, it uses two-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. [27] 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.

3. RGBD-HuDaAct: Color-Depth Human Daily Activity Database

3.1. Related Video Databases

A summarization of the existing video activity benchmark databases is given in Table 1. **KTH and Weizmann 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% [17] and 100% [4, 13], 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:** 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:** The UCF sports event database [25] and the UCF YouTube sports database [22] 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:** The only existing depth sensor based action database is collected by Li et al. [21], 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. **Indoor Kinect Activity Database:** Very recently, Sung et al. [27] 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 4 subjects and 12 activity categories. In addition to RGBD images, the database also provides skeleton motion data. Most of their categories do not overlap with ours.

Different with these databases, our motivation is driven by the application of assisted living in heath-care. Monitoring the daily activities of senior citizens has recently become a urgent demand due to the aging population problem. There only exists a very recent video database for senior home monitoring [7], however, it does not utilize the depth modality. In contrast, the RGBD-HuDaAct database contains the synchronized color and depth videos, which is more suitable for 24 hours monitoring, since the depth sensor also works without visible lighting.

3.2. Database Construction

We utilize the recent released Microsoft Kinect sensor to construct the RGBD-HuDaAct video database, collected in a lab environment.There are minor variations in the camera position and orientation due to repeated mountings of the camera. The horizontal and vertical distances from the camera to the center of scene under capture are about 2 and 2 meters, respectively and the average depth of the human

Database	Modality	Resolution	Sample#	Category Descriptions
KTH [26]	RGB	160 × 120	2391	6 classes: walking, jogging, running, etc.
Weizmann [4]	RGB	180 × 144	90	10 classes: run, walk, skip, jumping-jack, side, etc.
Hollywood2 [23]	RGB	600 × 450 (in average)	3669	12 classes: answering the phone, driving car, eating, etc.
UCF Sports [25]	RGB	720 × 480	184	10 classes: swinging, golf swinging, walking, etc.
UCF YouTube [22]	RGB	320 × 240	3040	11 classes: basketball shooting, biking, diving, etc.
MSR Action3D [21]	Depth	320 × 240	4020	20 classes: high arm wave, hand catch, forward punch, etc.
Indoor Activity [27]	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, make a phone call, etc.

Table 1. Comparisons of the RGBD-HuDaAct database over other benchmark activity databases.

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 an 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 [2]). We repeat the camera calibration procedure at the beginning of each video capture session and the camera is fixed throughout the session.

3.3. Database Statistics

We are interested in 12 categories of human daily activities, motivated by the definitions provided by health-care professionals [18], including: *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*. We also have a category named as *background activity* that contains different types of random activities. We invite 30 student volunteers to perform these daily activities, which are organized into 14 video capture sessions. The subjects are asked to perform each activity $2-4$ times. Finally, we capture 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.

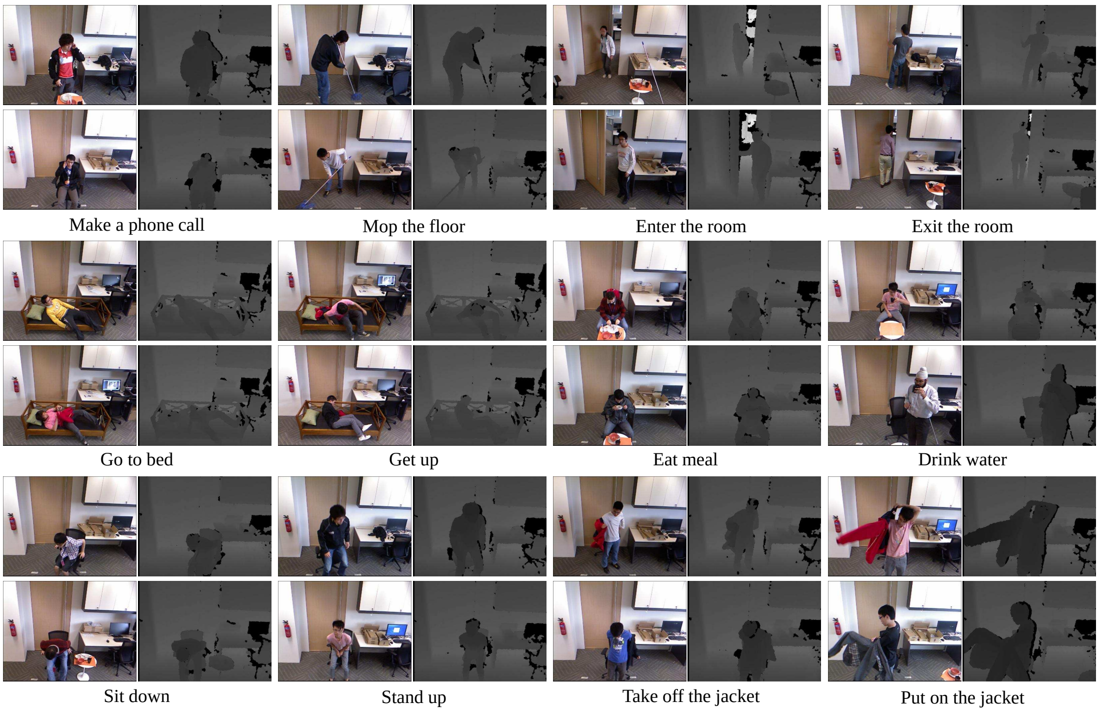
Two example frames from each activity category are illustrated in Figure 1, in terms of both color (left) and depth (right) frames. We can make two observations from Figure 1: 1) There exist distinctive depth layers for the moving human body parts in different activities, which implies that incorporating 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. Note that although the background of the current database is of limited variations and only single subject is present (*i.e.*, compared with those 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 is very typical.

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 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 **Three-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 three-dimensional spatial structure of the detected spatio-temporal feature points as well as the three-dimensional motion history of the human subjects, the discriminating capabilities of the features are boosted.

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 is proposed by Laptev and Lindeberg in [19]. Harris3D detector is a space-time extension of the two-dimensional Harris detector [14]. At each space-time video point, a spatio-temporal second-moment matrix is computed as:



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

Put on the jacket

Figure 1. 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.

$\mu(.;\sigma,\tau) = g(.;s\sigma,s\tau)*(\Delta V(.;\sigma,\tau))(\Delta V(.;\sigma,\tau))^T$ (*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 extremes of $H = det(\mu) - \kappa trace^3(\mu)$, 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 [19]. 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 [28] (*i.e.*, as a histogram vector over the visual word vocabulary).

However, the human subject is in essence a three-dimensional structure and the detected spatio-temporal feature points are associated with local motions taking place at different three-dimensional locations; however, the previous pooling methods of STIPs can only utilize this spatial information up to two-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 [19], where STIPs are spatially pooled within each $x-y-t$ sub-volume, *i.e.*, the entire three-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 as $V = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$. Each STIP feature descriptor is denoted as $\mathbf{x}_i = (x, y, z, t, \sigma, \mathbf{x}_{HOG}^T, \mathbf{x}_{HOF}^T)^T$. Here, x, y, z, t, σ represent the $3D$ coordinate (x, y, z), temporal index and the scale of the detected feature point, respectively. $\mathbf{x}_{HOG}$ and $\mathbf{x}_{HOF}$ 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,

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

This aggregation process is usually referred as *feature pool-*

ing, *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 [19], the entire three-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 available, we can also form depth-layered multi-channel representations. Namely, we introduce a set of (M) depth layers $L_1^z = [z_1^l, z_1^u], L_2^z = [z_2^l, z_2^u], \cdots, L_M^z = [z_M^l, z_M^u]$, with lower and upper boundaries denoted as z_m^l and z_m^u for the $m-th$ depth layer. Then, we can form multi-channel histograms $\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_M$, as:

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

These multiple channel histograms could be concatenated into an $M \times K$-dimensional feature vector $\mathbf{h} = (\mathbf{h}_1^T, \mathbf{h}_2^T, \cdots, \mathbf{h}_M^T)^T$, 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 [20] to better explore the spatial information given in the depth axis. An illustration of the DLMC-STIPs generation process is given in Figure 2. Note that: 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 Section 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 three-dimensional motion is introduced in the next subsection.

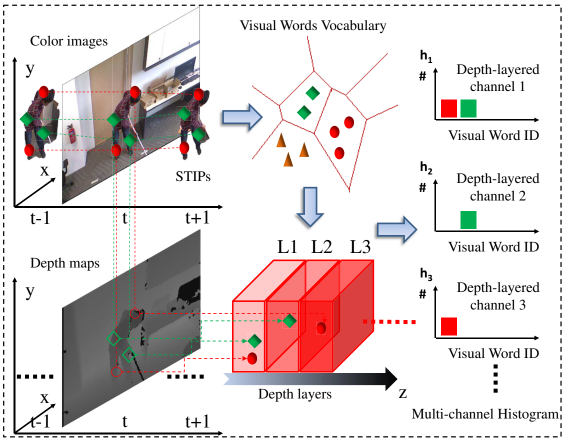


Figure 2. A diagram of the generation process of DLMC-STIPs representation.

4.2. Three-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 [5], 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:

$$H_\tau^I(x, y, t) = \begin{cases} \tau, if |I(x, y, t) - I(x, y, t-1)| > \delta I_{th} \\ \max(0, H_\tau^I(x, y, t-1) - 1), else. \end{cases} \quad (3)$$

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 moment-based features [15], 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,

$$H_\tau^{fD}(x, y, t) = \begin{cases} \tau, if (D(x, y, t) - D(x, y, t-1)) > \delta D_{th} \\ \max(0, H_\tau^{fD}(x, y, t-1) - 1), else. \end{cases} \quad (4)$$

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 three-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 concatenated to form a representation vector. An example 3D-MHI is illustrated in Figure 3 in the context of a *sit down* sequence. From Figure 3, we can 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.

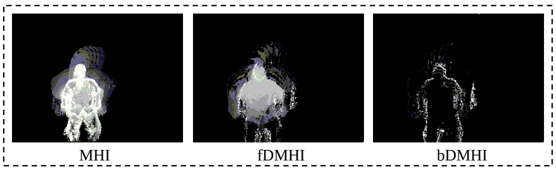


Figure 3. Illustration of the MHI, fDMHI and bDMHI in a *sit down* sequence.

5. Experimental Evaluations

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 9 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 inputs, 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 i-th class are classified into the j-th class. Larger values for the diagonal elements and smaller values for the off-diagonal elements indicate better discriminating capability.

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**

Table 2. Comparisons of the classification accuracies (%) for STIPs and DLMC-STIPs under different experimental settings.

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 [21]). We use support vector machines (SVM) [6] (*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.

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 [19]. 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) [20] for DLMC-STIPs representations with $l = 3$ depth spatial levels. Various classification accuracies under different parameter combinations are given in Table 2. We also illustrate the class confusion matrices for both methods in Figure 4, at the setting of $K = 256$.

It can be observed from Table 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.

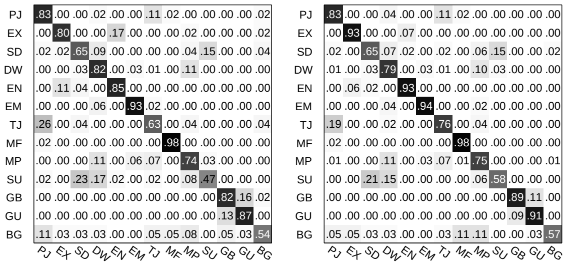


Figure 4. Class confusion matrices for STIPs (left) and DLMC-STIPs (right, SPM kernel) under the setting of $K = 256$. For better view, we use two characters to represent each activity category, *i.e.*, PJ: *put on the jacket*, TJ: *take off the jacket*, EN:*enter the room*, EX: *exit the room*, SD: *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*.

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

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

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 [5] 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 3. We again show the class confusion matrices for both methods in Figure 5, for the case of linear SVM.

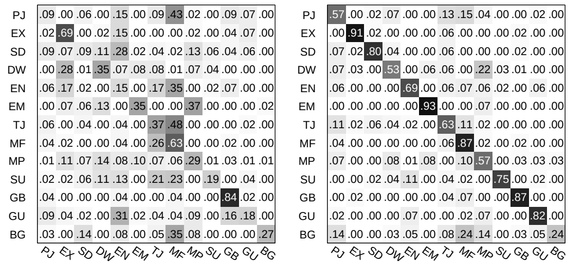


Figure 5. Class confusion matrices for MHIs (left) and 3D-MHIs (right), at the setting of linear SVM.

From Table 3 and Figure 5, 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 Figure 5, we can 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.

References

[1] http://en.wikipedia.org/wiki/activities_of_daily_living. 1

[2] http://nicolas.burrus.name/index.php/research/kinectcalibration. 3

[3] M. J. Black, Y. Yacoob, A. D. Jepson, and D. J. Fleet. Learning parameterized models of image motion. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 561–567, 1997. 1, 2

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *IEEE International Conference on Computer Vision*, pages 1395–1402, 2005. 2, 3

[5] 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

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. 6

[7] H. Cheng, Z. Liu, Y. Zhao, and G. Ye. Real world activity summary for senior home monitoring. In *IEEE International Conference on Multimedia and Expo*, 2011. 2

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2005. 2

[9] J. W. Davis and A. Tyagi. Minimal-latency human action recognition using reliable-inference. *Image and Vision Computing*, 24(5):455–472, 2006. 2

[10] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005. 2

[11] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision*, 2003. 1, 2

[12] J. L. B. D. J. Fleet and S. S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994. 2

[13] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007. 2

[14] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1998. 3

[15] M. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187, 1962. 2, 5

[16] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d gradients. In *The British Machine Vision Conference*, 2008. 2

[17] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition.* 2

[18] K. Krapp. Activities of daily living evaluation. *Encyclopedia of Nursing and Allied Health*, 2002. 3

[19] I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE International Conference on Computer Vision*, 2003. 2, 3, 4, 5, 6

[20] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2006. 5, 6

[21] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *CVPR workshop*, 2010. 2, 3, 6

[22] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos ”in the wild”. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2, 3

[23] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 3

[24] B. Ni, S. Yan, and A. Kassim. Recognizing human group activities with localized causalities. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009. 1

[25] M. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 2, 3

[26] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *IEEE International Conference on Pattern Recognition*, 2004. 3

[27] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgbd images. In *AAAI workshop on Pattern, Activity and Intent Recognition*, 2011. 2, 3

[28] M. M. Ullah, S. N. Parizi, and I. Laptev. Improving bag-of-features action recognition with non-local cues. In *The British Machine Vision Conference*, 2010. 4