DMMs-Based Multiple Features Fusion for Human Action Recognition

Mohammad Farhad Bulbul, Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, China

Yunsheng Jiang, Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, China

Jinwen Ma, Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, China

ABSTRACT

The emerging cost-effective depth sensors have facilitated the action recognition task significantly. In this paper, the authors address the action recognition problem using depth video sequences combining three discriminative features. More specifically, the authors generate three Depth Motion Maps (DMMs) over the entire video sequence corresponding to the front, side, and top projection views. Contourlet-based Histogram of Oriented Gradients (CT-HOG), Local Binary Patterns (LBP), and Edge Oriented Histograms (EOH) are then computed from the DMMs. To merge these features, the authors consider decision-level fusion, where a soft decision-fusion rule, Logarithmic Opinion Pool (LOGP), is used to combine the classification outcomes from multiple classifiers each with an individual set of features. Experimental results on two datasets reveal that the fusion scheme achieves superior action recognition performance over the situations when using each feature individually.

Keywords: *Action Recognition, Depth Motion Maps, Edge Oriented Histograms, Kernel-based Extreme Learning Machine, Local Binary Patterns*

INTRODUCTION

Automatic human action/gesture recognition is an active research topic in the area of computer vision. Researchers are fueled by the increasing number of real-world applications including autonomous visual surveillance, video retrieval, human-computer interaction, health care, sports training, etc.(e.g., C. Chen, Liu, Jafari, & Kehtarnavaz, 2014a; C. Chen, Kehtarnavaz, & Jafari, 2014b). Human action recognition is very challenging due to the significant variations in human body sizes, appearances, postures, motions, clothing, camera motions, viewing angles, illumination changes, etc. Moreover, the complexity grows due to that the same action is performed differently by different persons, even for same person at different times.

DOI: 10.4018/IJMDEM.2015100102

A large portion of researchers have addressed this problem by using features extracted from 2D intensity images (Chaaraoui, Climent-Pérez, & Flórez-Revuelta, 2012; Poppe, 2010; Wiliem, Madasu, Boles, & Yarlagadda, 2010; H. Wang & Schmid, 2013). However, the 2D intensity images captured by the conventional RGB video cameras do not have enough information to perform the comprehensive analysis. Moreover, they are sensitive to lighting condition, and the process of identifying key points depends on the object texture instead of object geometry (L. Chen, Wei, & Ferryman, 2013). On the other hand, these intensity images have many obstacles to perform robust computer vision tasks such as background subtraction and object segmentation.

Recently, with the availability of low-cost depth cameras (e.g., Microsoft Kinect), some of the difficulties for intensity images have been alleviated. The outputs of depth cameras are called depth images (which are sometimes mentioned as depth maps or depth frames according to context). Depth images preserve the depth information corresponding to the distances from the surface of scene objects to the viewpoint (Shotton, et al., 2013). The pixels in a depth image indicate calibrated depths (i.e., depths in a scale) in the scene, instead of intensity or color. This depth information achieves an additional robustness to color information due to its invariant to illumination and textures changes (Zhu & Pun, 2013). Moreover, the depth data captures the 3D structure of the scene as well as the 3D motion of the subjects/objects in the scene. Therefore, depth cameras show many advantages over the conventional intensity cameras, such as working under low light conditions and even in darkness, estimating calibrated depth, being steady to color and texture variations, and giving solution of the silhouette problem in posture (Shotton, et al., 2013). They also remove many ambiguities in computer vision tasks like background subtraction and object segmentation.

This paper proposes an effective action recognition framework by fusing the outcomes of multiple classifiers, each of which has an individual features set. This type of fusion is essential, as often a single kind of features or feature-level fusion may not exhibit enough discriminatory power. Therefore, we combine the classification decisions from classifiers with three types of features that extracted from Depth Motion Maps (DMMs) (C. Chen, Liu, & Kehtarnavaz, 2013): i) Contourlet-based Histogram of Oriented Gradients (CT-HOG) (Farhad, Jiang, & Ma, 2015a), ii) Local Binary Patterns (LBP) (Ojala, Pietikäinen, & Mäenpää, 2002) and iii) Edge Oriented Histograms (EOH) (Conaire). More specifically, we first represent an action video sequence with three DMMs (see Section 3 for more details). Then, CT-HOG, LBP and EOH are computed on each DMM separately. Finally, three feature sets are fed into three Kernel-based Extreme Learning Machine (KELM) (Huang, Zhu, & Siew, 2006) classifiers to provide the probability outputs for each action. The obtained probability outputs are merged using Logarithmic Opinion Pool (LOGP) (Benediktsson & Sveinsson, 2003) and Majority Voting (MV) (Lam & Suen, 1997) decision rules to label the query sample. Overall, the decision-level fusion operates on probability outputs and fuses multiple decisions into a joint one.

The main contributions of this paper are summarized as follows:

1. We compute three feature descriptors employing DMMs, CT-HOG, LBP and EOH. DMMs are utilized to capture specific appearances and shapes in a depth video sequence. Then, CT-HOG, LBP and EOH are employed on DMMs to obtain contour, texture and edge features respectively. Here, all these features lead us to achieve a compact representation of DMMs and enhance discriminatory power for the recognition algorithm.
2. Decision-level fusion is employed to the extracted compact features. In the decision-level fusion, more than one decision strategies are utilized to merge the probability outputs of each classification.

3. The proposed features and the classification framework have been tested extensively on two publicly available human action datasets, MSR-Action3D (Li et al., 2010) and UTD-MAD (C. Chen, Jafari, & Kehtarnavaz, 2015b). The experimental results demonstrated that the proposed action recognition method achieved superior performance over several state-of-the-art methods.

The rest of this paper is organized as follows: In Section 2, are view on related work is stated. The details of CT-HOG, LBP and EOH feature descriptors and KELM classifiers are described in Section 3. DMMs-based multiple features pooling and decision-level fusion are stated in Section 4. In Section 5, the experimental results on two standard datasets are reported and compared. Finally, the conclusion appears in Section 6.

RELATED WORK

Action recognition using depth images has drawn much more attraction to the computer vision community after the proliferation of the depth sensors. In this context, we broadly divide the action recognition approaches into four categories: depth image-based approaches, skeleton joints based approaches, depth and color images (fusion) based approaches, and depth/color images and skeleton joints (fusion) based approaches. Here, our discussion is restricted with several well-known approaches. See the work of Aggarwal and Ryoo (2011) for the comprehensive reviews of the previous studies.

In the first category, some researchers have focused on exploiting silhouette and edge pixels as discriminative information. For example, a collection-of-3D-points feature was proposed for action recognition from depth video sequences, where the 3D points were sampled from the silhouettes of the depth images (Li, Zhang, & Liu, 2010). An action graph was then employed for their classification framework, where each action was encoded in one or multiple paths in the action graph. The nodes of the action graph were used to represent the salient postures. However, the method suffers from two main disadvantages: the loss of spatial context information between interest points as well as the high computational cost due to having 3D points sampling scheme. In addition, due to noise and occlusions in the depth images, the side and top view of the silhouettes was noisy. Thus, it was very difficult to get a robust sampling scheme for the interest points describing the geometry and motion variations between different subjects.

On the other hand, silhouettes were generated in 3-dimensional space utilizing space-time occupancy patterns (Vieira, Nascimento, Oliveira, Liu, & Campos, 2012). In a cell structured spatiotemporal depth volume, a filled cell was labeled by 1, an unfilled by 0 and a partially-filled cell by a fraction. All the fully and partially cells were distinguished based on an ad hoc parameter. Instead of using simple occupancy patterns, a vector consisting of ..Haar features on a uniform grid in the 4-dimensional volume was considered by J. Wang, Liu, Chorowski, Chen, and Wu (2012a). In this approach, LDA and SVM were used for the detection of discriminative feature positions and action classification respectively. However, high computational complexity was the barrier for the both methods.

Yang, Zhang, and Tian (2012b) used Depth Motion Maps (DMMs)-based Histograms of Oriented Gradients (HOG) features and Support Vector Machine (SVM) classifier for human action recognition. More specifically, the depth frames were projected onto three orthogonal Cartesian planes and the entire depth video sequence was accumulated generating three DMMs similar to the Motion History Images (MHI) (Bobick & Davis, 2001). Then, HOG features are computed for each DMM. The concatenation of the three HOG feature vectors was used as the

input feature vector to a linear SVM classifier. The computational cost of this approach was relatively low as HOG was computed for DMMs. C. Chen et al. (2013) presented a computationally efficient solution for human action recognition problem using modified DMMs descriptors and l_2 -collaborative representation classifier (l_2 -CRC). The proposed approach was able to achieve real-time action recognition. Later, compact texture features using DMMs and LBP were extracted in (C. Chen, Jafari, & Kehtarnavaz, 2015a). This feature exhibits higher discriminatory power than the features used in (Yang et al., 2012b; C. Chen et al., 2013). Furthermore, we proposed an effective method by using HOG features from DMMs-based contourlet sub-bands (Farhad et al., 2015a). We also enhanced the discriminatory power of the feature representation through the fusion approach of the DMMs-based texture and edge features (Farhad, Jiang, & Ma, 2015b).

In the second category, some algorithms have been built discovering the correlation between action categories and body-part joints from the depth images. For instance, Yang and Tian (2012a) computed pairwise 3-dimensional joint position differences for each depth frame and temporal differences across images to describe human actions. However, the recognition accuracy of this approach was not high as 3-dimensional joints were not good enough to capture all the discriminative information. In the same year, this approach was extended by J. Wang, Liu, Wu, and Yuan (2012b) using the depth histogram-based features. These features were obtained from a specific domain around each joint in each depth frame. Low-frequency Fourier components were treated as temporal dimension features over the temporal dimension. An SVM was employed to establish a discriminative set of joints.

Joint positions could be represented compactly as reported by Xia, Chen, and Aggarwal (2012). In specific, they proposed a new feature called Histogram of 3D Joints Locations (HOJ3D), which essentially encodes spatial occupancy information relative to the skeleton root. They applied linear discriminant analysis to reduce feature dimensionality and Hidden Markov Model (HMM) to model the dynamics and action recognition.

To optimize skeleton joints features, a genetic-based evolutionary algorithm was proposed (Chaaraoui, Padilla-López, Climent-Pérez, &Flórez-Revuelta, 2014). The topological structure for the skeleton was considered to improve the performance of the algorithm. Basically, they took into account a binary vector where each gene described the further consideration or not of a special feature. They employed filter and wrapper model for its implementation. But, the high computational cost of the approach and early convergence created some drawbacks of this approach. In the wrapper-based evolutionary approach, the fitness calculation of an important number of solutions was required to reach the final solution. Here, the involvement of the calculation of a single fitness with a complete training and recognition process resulted in considerable time for the whole evolution. In addition, early convergence occurs while the evolutionary search was accumulated in a local minimum and a good solution was not achieved. However, for the applications of those skeleton-based algorithms, it's required to have available skeleton information beforehand.

In the third category, Ni, Wang, and Moulin (2013) introduced two feature extraction methods for fusing color and depth information. These feature extraction schemes were developed based on two state-of-the-art action representation approaches: Spatial-temporal Interest Points (STIPs) and Motion History Images (MHIs). Specifically, on the one hand, they derived a framework (named as depth-layered multi-channel STIPs) to partition the STIPs into several depth-layered channels. These STIPs within different channels are pooled independently to obtain a multiple depth channel histogram representation. On the other hand, three-dimensional MHIs were used to equip the conventional MHIs with two additional channels.

In the last category, the discriminative features from depth/color images as well as from 3D joint positions were fused to boost the discriminating capabilities of the algorithm. Rahmani, Mahmood, Huynh, and Mian (2014) incorporated four features from depth videos and the 3D joint positions: the 4D depth, depth gradient histograms from depth videos and joint displacement histograms, joint movement occupancy volumes from 3D joint positions. Random decision forest was used for feature pruning and classification. In order to improve recognition accuracy, some researchers combine features from RGB and depth video sequences. Furthermore, 3D joint features are also associated with spatio-temporal features to classify actions more accurately (Luo, Wang, & Qi, 2014). The spatio-temporal features were extracted from the RGB video sequence using center-symmetric motion local ternary pattern.

PRELIMINARIES

In this section, we first study on descriptors that are used here to represent an action video sequence. Then, a comprehensive overview of KELM classifier is presented. However, the next section shows the implementation of those descriptors in this paper.

DMMs Computation

Yang et al. (2012b) first introduced the so-called DMMs to represent human actions by stacking motion energy of each depth frame in a depth video sequence. In fact, the accumulated motion energy corresponding to an action category generates specific appearances and shapes on DMMs. However, the concept of the DMMs was modified by C. Chen et al. (2013) and it's used in this work due to its computational simplicity. Concretely, we consider a depth video sequence with M depth frames. Each depth frame in the video can generate three 2D projected maps by projecting the frame onto three orthogonal Cartesian planes. These projections are taken from the front (f), side (s) and top (t) projection views (see Figure 1). The 2D projected maps corresponding to the projection views are labeled by map_f, map_s and map_t. The summation of all the absolute differences between two consecutive projected maps for a specific projection view produces a single DMM. As a result, we obtain three DMMs for the three projection views, which are denoted as DMM_f, DMM_s and DMM_t. The generation of DMMs can be expressed in mathematical form as follows:

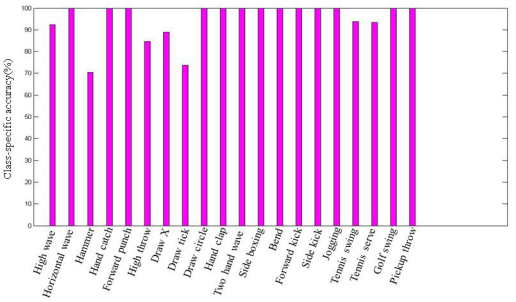
$$DMM_p = \sum_{k=1}^{M-1} \left| map_v^{k+1} - map_v^k \right|, \tag{1}$$

where k is the index for a depth video frame, $v \in \{f, s, t\}$ and map_v^k is the projection of the kth frame under a projection view $v \in \{f, s, t\}$. The non-zero region of each DMM is cropped (setting a suitable bounding box) and is considered as the final DMM.

There are two main advantages of using DMMs. The 4D information of body shape and motion in depth maps are encoded in three projected maps through the DMMs representation of the video sequence. On the other hand, DMMs mechanism reduces data of the depth sequence to just three 2D maps, which alleviates the computation cost considerably.

An example of DMM_f generation for a *two hand wave* depth sequence is shown in Figure 2. In the figure, there is an action video sequence consists of six depth frames. The front projection views of these video frames produce the equal number of projected maps. Then, there are five absolute differences are gained by considering consecutive difference between two pro-

Figure 1. Projection views of a depth video frame



jected maps. The summation of all the absolute differences generate the DMM_f of the video sequence.

CT-HOG Features

In our previous work, we computed CT-HOG features on DMMs (Farhad et al., 2015a) to capture human contour compactly. Contourlet (Do & Vetterli, 2005) was applied on DMMs to remove noise and enhance the shape information, i.e. to enhance DMMs. Since the computational complexity of the algorithm increases by employing all the contourlet sub-bands corresponding to each DMM, the low-frequency sub-band is coupled with the high-frequency sub-bands obtained from the first-level contourlet decomposition (if the decomposition-level is more than one). The experimentations were also carried out using high-frequency sub-bands from other decomposition-levels, but the promising result was occurred for high-frequency sub-bands from the first-level. Overall, five sub-bands (one low-frequency sub-band and four high-frequency sub-bands) were selected for each DMM. Since there are three DMMs for each video sequence, fifteen sub-bands were used to describe a depth video. To build a compact feature representation of the fifteen sub-bands, we employed HOG on overlapped-structured sub-bands to achieve computational efficiency as well as good classification performance. See our paper for more details (Farhad et al., 2015a).

Figure 2. The DMM_f of a two hand wave depth video sequence

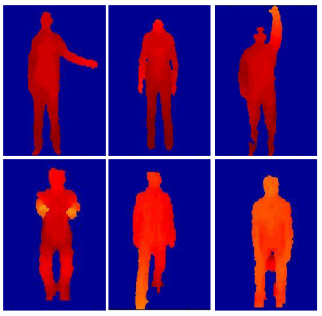
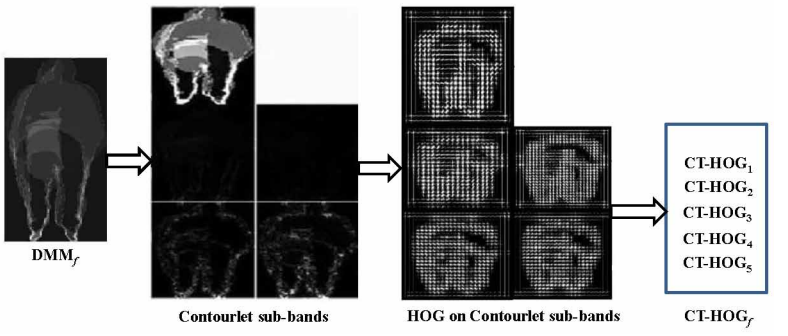


Figure 3. LBP-label generation of a pixel p_c for the neighborhood $(8,1)$



LBP Features

The LPB (Ojala et al., 2002) operator is an effective gray scale and rotation invariant operator that can be used to capture texture information of an image. To extract texture information, the LBP-coded image corresponding to the raw-image is first calculated. In the LBP-coded image, original pixels are labeled with decimal numbers that encode local texture information.

Let p_c be a pixel in an image, whose neighbors are equally spaced on a circle with center $p_c(0,0)$ and radius $r(r>0)$. If there are n neighbors p_j $(j=0,\ldots,n-1)$, the coordinates of the neighbors are $(-r\sin(2\pi j/\mathrm{n}), \mathrm{rcos}(2\pi j/n))$. An example of a neighbor set and the pixel labeling process is shown in Figure 3 for $n=8$ and $r=1$. The thresholding of the neighbors $\{p_0,\ldots,p_{n-1}\}$ with the center pixel p_c generates the n-bit binary number (i.e., LBP). Notice that, in the thresholding, if a neighbor is greater than or equal to the center pixel, this neighbor will be assigned a value of 1 and otherwise 0. The decimal form of the binary numbers is used to label the center pixel or the candidate pixel p_c. Thus, pixels in the image can be labeled using decimal form and the resulted image is denoted as LBP-coded image. Histograms features from the LBP-coded image are calculated to capture the texture information compactly.

EOH Features

The EOH (Conaire) feature descriptordescribes the edge-based shape information from an image. First, the image is subjected to filtering for noise removal. There are broadly used filters (e.g., Gaussian filter and Median filter) that are utilized to remove noise. In fact, this type of typical preprocessing step enhances the result of further processing of the image in the subsequent steps. Then, edges (4 directional and one non-directional edge) in the image are detected through an edge detection algorithm. Actually, the edge detection approachsignificantly reduces the amount of data and filters out the useless information in an image while preserving its structural shapes. Finally, EOH features are computed from the edge image. The implementation of the EOH feature extraction algorithm is comprehensively described in the discriminative features pooling section.

KELM Classifier

Extreme Learning Machine (ELM) was developed based on single-hidden-layer feed-forward neural networks (Huang et al., 2006). Recently, Extreme Learning Machine (KELM) (Huang,

Zhou, Ding, & Zhang, 2012) has been utilized by extending explicit activation functions in ELM to implicit mapping functions, which have exhibited a better generalization capability and stability than ELM (Li, Chen, Su, & Du, 2015).

Let us consider a dataset with C classes. The class label of a sample in the dataset can be defined as $y_i \in \{0,1\}$, where $i = 1, 2, \cdots, C$. The sample belongs to the ith class if $y_i = 1$. Assume there are m training samples $\{x_j, y_j\}_{j=1}^{m}$, where $\mathbf{x}_j \in \mathbb{R}^D$ and $y_j \in \mathbb{R}^C$, the output function of a single-hidden-layer feed-forward neural network with N hidden nodes can be presented as

$$h_N(\mathbf{x}_j) = \sum_{k=}^{N} \pm_k f(\mathbf{w}_k . \mathbf{x}_j + e_k) = \mathbf{y}_j, j = 1,2,\ldots,m \tag{2}$$

where $f(.)$ denotes the nonlinear activation function, $\mathbf{w_k} \in \mathbb{R}^D$ and $\pm_k \in \mathbb{R}^C$ are the weight vectors connecting the kth hidden node to the input and output nodes respectively, and e_k is the bias for the kth hidden node. Considering all the m equations, Equation 2 can be figured out as

$$F\alpha = Y, \tag{3}$$

where $\alpha = [\alpha_1^T, \ldots, \alpha_m^T]^T \in \mathbb{R}^{N \times C}$, $Y = [y_1^T, \ldots, y_m^T]^T \in \mathbb{R}^{m \times C}$, and F denotes the hidden layer output matrix of the neural network, which is written as

$$F = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_m) \end{bmatrix} = \begin{bmatrix} f(w_1 . x_1 + e_1) & \cdots & f(w_N . x_1 + e_N) \\ \vdots & \ddots & \vdots \\ f(w_1 . x_m + e_1) & \cdots & f(w_N . x_m + e_N) \end{bmatrix} \tag{4}$$

Here, $f(x_j) = [f(w_1 . x_j + e_1), \ldots, f(w_N . x_j + e_N)]$ is the output of the hidden nodes for the input $\mathbf{x}_j$. As $N \ll m$ (N is the number of hidden nodes and m is the number of training samples) happens frequently, the smallest norm least-squares solution (Huang et al., 2006) of Equation 3 can be considered, *i.e.*

$$\alpha,, = F^{\dagger} Y \tag{5}$$

where $F^{\dagger}$ denotes the Moore-Penrose generalized inverse of matrix F, $F^{\dagger} = F^T (FF^T)^{-1}$. To gain a better stability and generalization $\frac{1}{\rho}(\rho > 0)$ is simply added to the diagonal of FF^T. Hence, Equation 2 (i.e., the output function) can be expressed as

$$h_N(x_j) = f(x_j)\alpha = f(x_j)F^T(\frac{I}{\rho} + FF^T)^{-1}Y \tag{6}$$

In ELM, a kernel matrix can be used as follows (if the feature mapping [INSERT FIGURE 001] is unknown):

$$\Omega_{ELM} = FF^T : \Omega_{ELM_{j,s}} = f(x_j).f(x_s) = K(x_j, x_s) \tag{7}$$

As a result, the output function for the kernel matrix-based ELM (KELM) is represented by

$$h_N(x_j) = \begin{bmatrix} K(x_j, x_1) \\ \vdots \\ K(x_j, x_m) \end{bmatrix} \left(\frac{I}{\rho} + \Omega_{ELM} \right)^{-1} Y \tag{8}$$

The label of a query sample x_t is obtained according to

$$y_t = \underset{j=1,2,\ldots,C}{\arg\max}\ \mathrm{h}_N(\mathbf{x}_t)_j \tag{9}$$

where $h_N(x_t)_j$ is the jth output of $h_N(x_t) = [h_N(x_t)_1, \ldots, h_N(x_t)_C]^T$.

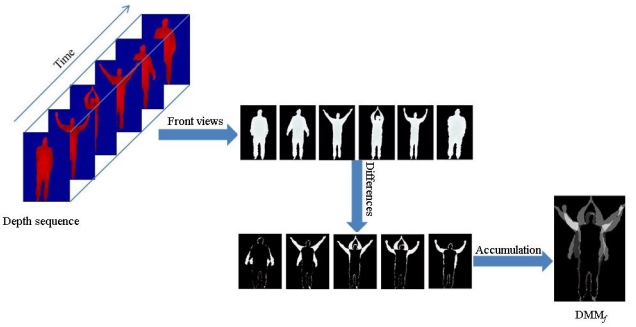
PROPOSED FUSION METHOD

In our approach, DMMs-based CT-HOG, LBP and EOH features are extracted to represent a depth video sequence from different perspectives. To fuse these features, decision-level fusion is considered. It should be noted that feature-level fusion could be employed here. But, the feature-level fusion mechanism is incompatible of multiple feature sets and large dimensionality although it is straightforward. Therefore, we consider decision-level fusion, where the results from a classifier ensemble are combined through Logarithmic Opinion Pool (LOGP) (Benediktsson et al., 2003) and Majority Voting (MV) (Lam & Suen, 1997)decision algorithms as both are simple and effective. In this section, we describe DMMs-based multiple features pooling and its decision-level fusion as well. The decision-level fusion based on LOGP is stated here comprehensively and the discussion on MV-based fusion is skipped as it's well-known. Our proposed fusion approach is shown in Figure 7.

Discriminative Features Pooling

For each depth video sequence, we extract the CT-HOG, LBP and EOH features from the three DMMs respectively (see Figure 7). In this context, we figure out these features for a single DMM as an example. To extract the CT-HOG features from a DMM, contourlet transform is employed on the DMM and several contourlet sub-bands are selected using the concept as described by Farhad et al. (2015a). Then, HOG features (Junior, Delgado, Goncalves, & Nunes, 2009) are computed on these selected sub-bands. The HOG features corresponding to all the sub-bands are concatenated to represent the contour feature from the DMM. The CT-HOG feature vectors from the three DMMs are labeled as $CT-HOG_f$, $CT-HOG_s$ and $CT-HOG_t$. An example of the CT-HOG feature extraction scheme is illustrated in Figure 4.

Figure 4. CT-HOG feature from DMM_f of a bend action sequence



In the LBP feature extraction procedures, the corresponding LBP-coded image of the DMM is split into overlapped blocks and histograms are computed block by block. The uniform pattern (Ojala et al., 2002) is used in this paper to calculate the histogram features. The LBP-histograms for all the blocks are merged to construct a feature vector. The LBP-histograms feature vectors are named as LBP_f, LBP_s and LBP_t corresponding to three DMMs. Figure 5 shows an example of LBP-histogram features.

For the EOH features, we use the Gaussian filter kernel to remove the noise from each DMM. The Canny edge detection algorithm (Canny,1986) is used to detect the edges of the filtered DMM. The Canny operator is employed here as it is optimal and widely used as detection algorithm in research. On the other hand, an optimal edge detection technique is able to mark real edges as many as possible. The DMM is divided into non-overlapped blocks. Then, EOH are computed from each block. The EOH corresponding to all the blocks of the DMM are concatenated to form a feature vector. The EOH features from DMM_f, DMM_s and DMM_t are indicated by EOH_f, EOH_s and EOH_t respectively. An example of EOH generation is shown in Figure 6.

Decision Fusion Using LOGP

In the KELM, the output function (i.e., $h_N(\mathbf{x}_j)$ in Equation 8) is used to estimates the accuracy of the output label. Therefore, the posterior probabilities are estimated through the decision function. In this case, all the posterior probabilities increase proportionally with the higher

Figure 5. LBP histogram feature from a DMM

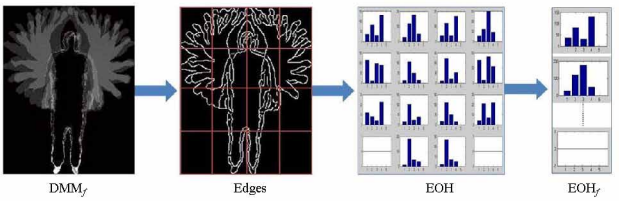
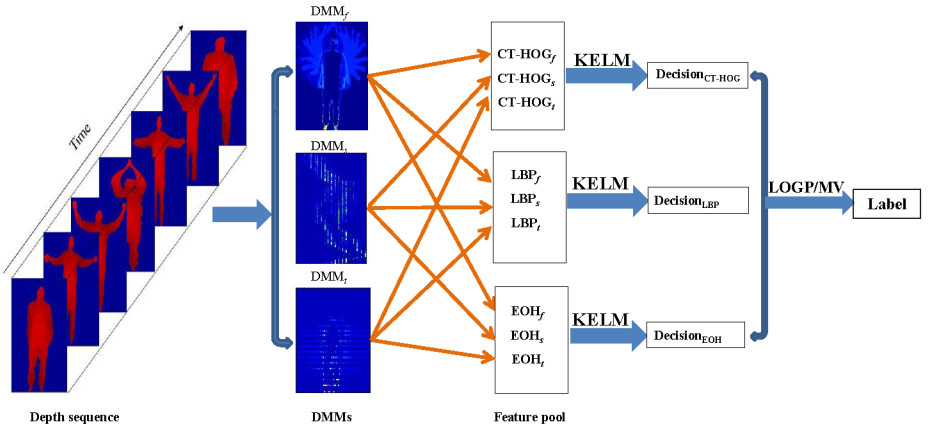


Figure 6. EOH feature extraction scheme



values of the decision function (Platt, 1999). Hence, $h_N(\mathbf{x})$ is scaled to [0, 1].Then, the posterior probabilities are approximated employing the Platt's empirical analysis as follows:

$$\mathrm{p}(y_i \mid \mathbf{x}) = \frac{1}{1+\exp(Ah_N(\mathbf{x})_i + B)} \tag{10}$$

To simplify the form of Equation (10), A and B are set to $A=-1$ and $B=0$.

In the LOGP strategy, these approximated posterior probabilities are utilized to estimate a global membership function as

$$P(y_i \mid x) = \prod_{q=1}^{L} p_q(y_i \mid x)^{\beta_q} \tag{11}$$

Or

$$\log P(y_i \mid x) = \sum_{q=1}^{L} \beta_q p_q(y_i \mid x) \tag{12}$$

Figure 7. Overview of the proposed fusion approach

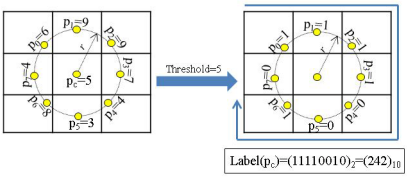
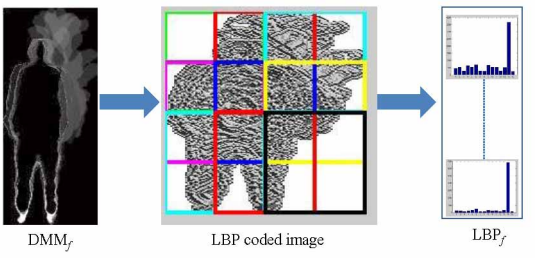


Figure 8. Sample depth images from MSR-Action3D dataset (top) and UTD-MHAD dataset (bottom)



where L stands for the number of classifier and $\{\beta_q\}_{q=1}^{L}$ are classifier weights. For simplicity, we use the uniform weights, i.e., $1/L$.

The single outcome is determined as follows:

$$y^* = \arg\max_{i=1,2,\ldots,C} \mathrm{P}(\mathrm{y}_i \mid \mathbf{x}) \tag{13}$$

EXPERIMENTAL RESULTS

We evaluate our recognition method on two standard public domain datasets: MSR-Action3D dataset (Li et al., 2010) and UTD-MHAD (C. Chen et al., 2015b) dataset. These datasets provide sequences of depth maps captured by commercial depth cameras. An example of depth maps from these datasets is shown in Figure 8. In our work, the Radial Basis Function (RBF) kernel is considered in KELM.

Evaluation on MSR-Action3D Dataset

The MSR-Action3D dataset (Li et al., 2010) consist of 20 different action categories performed by 10 subjects: *high wave, horizontal wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, and pick up throw*. Each action was performed two or three times by each subject facing to the depth camera during performance. Overall, the dataset includes 557 segmented depth sequences. Recognition task using this dataset is very challenging as it contains many actions with similar appearance (e.g., *draw x* and *draw tick*). In order to facilitate a fair evaluation of the method, the same experimental setup by J. Wang et al. (2012a) is followed. A total of 20 actions are employed and half of the total subjects $(1,3,5,7,9)$ are utilized for training and the rest ones for testing.

For CT-HOG feature extraction, all the DMMs are first resized to 256×256 according to our empirical analysis (Farad et al. 2015). Then, four-level contourlet decomposition with for directional sub-bands, i.e., $[2,2,2,2]$ is applied on each DMM. The HOG features on contourlet sub-bands are computed using 7×7 blocks and 8 bins (Junior et al., 2009). In LBP, m and r are set to $m = 4$ and $r = 1$ in terms of using 5-fold cross-validation accuracy. To compute LBP histogram features DMM_f, DMM_s and DMM_t are split into 4×2, 4×3 and 3×2 over-

lapped blocks respectively (C. Chen et al. 2015a). Here, the overlap between two blocks is taken to be one half of the block size. Besides, DMM_f, DMM_s and DMM_t are divided into 4×4 non-overlapped blocks to get EOH features (Farhad et al. 2015b). In our experiment, The RBF kernel parameters for the KELM classifier are chosen using 5 -fold cross-validation

We compare our method with the existing methods in Table 1. As can be seen from this table, our decision-level fusion approach (based on both the fusion algorithms) achieves superior performance over the listed methods as well as our other methods. The class-specific accuracies for the decision-level fusion are presented in Figure 9. Notice that 13 out of 20 actions in the MSR-Action3D dataset are classified with 100% classification accuracy. The classification accuracies of the remaining 7 actions are pretty good, where 3 actions are recognized with above 90% accuracy. Overall, the proposed approach achieves considerable recognition accuracies for all the actions except *hammer* and *draw x*.

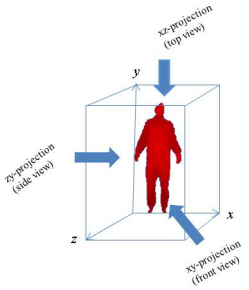
UTD-MHAD DATASET

The UTD-MHAD (C. Chen et al., 2015b) dataset consists of 27 different actions performed by 8 subjects (4 females and 4 males). Each action was repeated 4 times by each performer. There are 861 action sequences after removing 3 corrupted data sequences. The 27 actions are: *right arm swiping to the left, right arm swiping to the right, right hand wave, two hand front clap, right arm throw, cross arms in the chest, basketball shoot, right hand draw x, right hand draw circle (clockwise), right hand draw circle (counter clockwise), draw triangle, bowling (right hand), front boxing, baseball swing from right, tennis right hand forehand swing, arm curl (two arms), tennis serve, two hand push, right hand knock on door, right hand catch an object, right hand pick up and throw, jogging in place, walking in place, sit to stand, stand to sit, forward lunge (left foot forward), squat (two arms stretch out)*. The actions were captured using a Microsoft Kinect camera and a wearable inertial sensor. For the actions 1 through 21, the inertial sensor was placed on the subject's right wrist while for the actions 22 through 27, the inertial sensor was placed on the subject's right thigh. It can be seen from the list of actions, there are a comprehensive set of human actions in the dataset. For example, sport actions (e.g., bowling),

Table 1. Comparison of recognition accuracies on MSR-Action3D dataset

Method		Accuracy (%)
Yang et al. (2012b)		85.5
J. Wang et al. (2012a)		86.5
Oreifej & Liu (2013)		88.9
Xia & Aggarwal (2013)		89.3
C. Chen et al. (2015a)		93.0
Vieira et al. (2012)		84.8
Yang et al.(2012a)		82.3
Ours	DMMs-based CT-HOG	89.7
	DMMs-based LBP	91.9
	DMMs-based EOH	83.2
	Decision-level fusion(MV)	91.9
	Decision-level fusion(LOGP)	**94.9**

Figure 9. Class-specific accuracy for MSR-Action3D dataset (Using LOGP-based decision-level fusion)



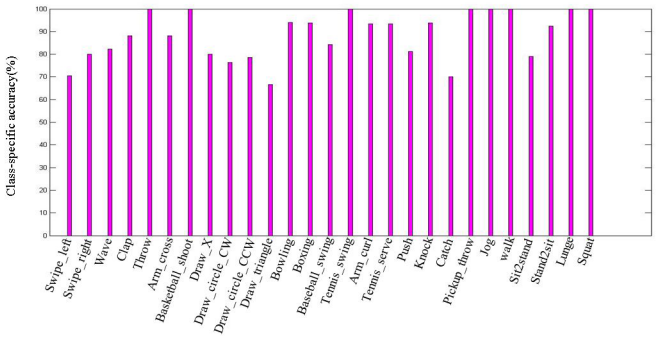
hand gestures (e.g., draw x), daily activities (e.g., knock on door), and training exercises (e.g., arm curl).

For the UTD-MHAD dataset, all the parameters are set as MSR-Action3D dataset except m and r. Here, $m = 6$ and $r = 5$ are set following the same technique. The classification results are stated in Table 2. In the table, the decision-level fusion scheme (either using LOGP or MV) recognizes actions with higher recognition accuracy. The individual classification accuracy is shown in Figure 10.

Table 2. Comparison of recognition accuracies on UTD-MHAD dataset

Method		Accuracy (%)
C. Chen et al. (2015b)		79.1
Ours	DMM-CT-HOG	83.5
	DMM-LBP	84.2
	DMM-EOH	75.3
	Decision-level fusion(MV)	85.3
	Decision-level fusion (LOGP)	**88.4**

Figure 10. Class-specific accuracy for UTD-MHAD dataset (Using LOGP-based decision-level fusion)



CONCLUSION

We have established an effective action recognition method of fusing classification outcomes from multiple classifiers with different kinds of features. Three feature descriptors, Contourlet-based Histogram of Oriented Gradients (CT-HOG), Local Binary Patterns (LBP) and Edge Oriented Histograms (EOH) are computed on depth motion maps. All of them are fed into different Kernel-based Extreme Learning Machine (KELM) classifier and their probability outputs are merged using soft decision rules, Logarithmic Opinion Pool (LOGP)and Majority Voting (MV), to assign a class label of the unknown sample. We carry out experiments on two standard datasets and compare with other existing methods as well as results using DMMs-based CT-HOG, LBP and EOH features. Experimental results demonstrate that the human actions can be recognized more accurately while we use LOGP-based decision-level fusion approach rather than employing individual feature descriptor.

ACKNOWLEDGEMENT

This work is supported by the National Science Foundation of China under Grant 61171138.

REFERENCES

Aggarwal, J., & Ryoo, M. (2011). Human Activity Analysis: A Review. *ACM Computing Surveys*, *43*(3), 1–43. doi:10.1145/1922649.1922653

Benediktssono, J. A., & Sveinsson, J. (2003). Multisource Remote Sensing Data Classification Based on Consensus and Pruning. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(4), 932–936. doi:10.1109/TGRS.2003.812000

Bobick, A., & Davis, J. (2001). The Recognition of Human Movement Using Temporal Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(3), 257–267. doi:10.1109/34.910878

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *8*(6), 679–698. doi:10.1109/TPAMI.1986.4767851 PMID:21869365

Chaaraoui, A. A., Climent-Pérez, P., & Flórez-Revuelta, F. (2012, September). A Review on Vision Techniques Applied to Human Behaviour Analysis for Ambient-Assisted Living. *International Journal of Expert Systems with Applications*, *39*(12), 10873–10888. doi:10.1016/j.eswa.2012.03.005

Chaaraoui, A. A., Padilla-López, J. R., Climent-Pérez, P., & Flórez-Revuelta, F. (2014). Evolutionary joint selection to improve human action recognition with RGB-D devices. *Journal of Expert Systems with Applications*, *41*(3), 786–794. doi:10.1016/j.eswa.2013.08.009

Chen, C., Jafari, R., & Kehtarnavaz, N. (2015a). Action Recognition from Depth Sequences Using Depth Motion Maps-Based Local Binary Patterns. *In Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, (pp. 1092-1099). Waikoloa Beach, HI. doi:10.1109/WACV.2015.150

Chen, C., Jafari, R., & Kehtarnavaz, N. (2015b). UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor. *In Proceedings of the International Conference on Image Processing*. Quebec city, Canada.

Chen, C., Kehtarnavaz, N., & Jafari, R. (2014b). A Medication Adherence Monitoring System for Pill Bottles based on a Wearable Inertial Sensor. *In Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 4983–4986). Chicago,IL. doi:10.1109/EMBC.2014.6944743

Chen, C., Liu, K., Jafari, R., & Kehtarnavaz, N. (2014a). Home-Based Senior Fitness Test Measurement System Using Collaborative Inertial and Depth Sensors. *In Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 4135-4138). doi:10.1109/EMBC.2014.6944534

Chen, C., Liu, K., & Kehtarnavaz, N. (2013). Real-Time Human Action Recognition Based on Depth Motion Maps. *Journal of Real-Time Image Processing*. doi:10.1007/s11554-013-0370-1

Chen, L., Wei, H., & Ferryman, J. (2013). A Survey of Human Motion Analysis Using Depth Imagery. *Pattern Recognition Letters*, *34*(15), 1995–2006. doi:10.1016/j.patrec.2013.02.006

Conaire, C. Ó. (n.d.). *Computer Vision Source Code*. Retrieved from http://clickdamage.com/sourcecode/index.php: http://clickdamage.com/sourcecode/code/edgeOrientationHistogram.m

Do, M. N., & Vetterli, M. (2005). The Contourlet Transform: An Efficient Directional Multiresolution Image Representation. *IEEE Transactions on Image Processing*, *14*(12), 2091–2106. doi:10.1109/TIP.2005.859376 PMID:16370462

Farhad, M., Jiang, Y., & Ma, J. (2015a). Human Action Recognition Based on DMMs, HOGs and Contourlet Transform. *In Proceedings of the IEEE International Conference on Multimedia Big Data*, (pp. 389–394). Beijing, China.

Farhad, M., Jiang, Y., & Ma, J. (2015b). Real-Time Human Action Recognition Using DMMs-Based LBP and EOH Feautres. *In Proceedings of the International Conference on Intelligent Computing*. Fuzhou, China.

Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme Learning Machine for Regression and Multiclass Classification. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, *42*(2), 513–529. doi:10.1109/TSMCB.2011.2168604 PMID:21984515

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme Learning Machine: Theory and Applications. *Journal of Neurocomputing*, *70*(1-3), 489–501. doi:10.1016/j.neucom.2005.12.126

Junior, O. L., Delgado, D., Goncalves, V., & Nunes, U. (2009). Trainable Classifier-Fusion Schemes: an Application to Pedestrian Detection. *IEEE International Conference on Intelligent Transportation Systems*, (pp. 432–443). St. Louis. doi:10.1109/ITSC.2009.5309700

Lam, L., & Suen, C. Y. (1997). Application of Majority Voting to Pattern Recognition: An analysis of Its Behaviour and Performance. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans*, *27*(5), 553–568. doi:10.1109/3468.618255

Li, W., Chen, C., Su, H., & Du, Q. (2015). Local Binary Patterns and Extreme Learning Machine for Hyperspectral Imagery Classification. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(7), 3681–3693. doi:10.1109/TGRS.2014.2381602

Li, W., Zhang, Z., & Liu, Z. (2010). Action Recognition Based on a Bag of 3D Points. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (pp. 9-14). San Francisco,CA. doi:10.1109/CVPRW.2010.5543273

Luo, J., Wang, W., & Qi, H. (2014). Spatio-Temporal Feature Extraction and Representation for RGB-D Human Action Recognition. *Pattern Recognition Letters*, *50*, 139–148. doi:10.1016/j.patrec.2014.03.024

Ni, B., Wang, G., & Moulin, P. (2013). Rgbd-Hudaact: A Color-Depth Video Database for Human Daily Activity Recognition. *In Proceedings of the Consumer Depth Cameras for Computer Vision* (pp. 193-208). Springer London. doi:10.1007/978-1-4471-4640-7_10

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987. doi:10.1109/TPAMI.2002.1017623

Oreifej, O., & Liu, Z. (2013). HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. *In Proceedings of the Conference on Computer Vision and Pattern Recognition*, (pp. 716-723). doi:10.1109/CVPR.2013.98

Platt, J. (1999). *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. Cambridge, MA: MIT Press.

Poppe, R. (2010, June). A Survey on Vision-Based Human Action Recognition. *Journal on Image and Vision Computing*, *28*(6), 976–990. doi:10.1016/j.imavis.2009.11.014

Rahmani, H., Mahmood, A., Huynh, D. Q., & Mian, A. (2014). Real-Time Action Recognition Using Histograms of Depth Gradients and Random Decision Forests. *In Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, (pp. 626-633). doi:10.1109/WACV.2014.6836044

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., & Blake, A. et al. (2013, January). Real-Time Human Pose Recognition in Parts from Single Depth Images. *Communications of the ACM*, *56*(1), 116–124. doi:10.1145/2398356.2398381

Vieira, A. W., Nascimento, E. R., Oliveira, G. L., Liu, Z., & Campos, M. M. (2012). STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences. *In Proceedings of the Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, (pp. 252-259). Buenos Aires,Argentina.

Wang, H., & Schmid, C. (2013). Action Recognition with Improved Trajectories. *In Proceedings of the IEEE International Conference on Computer Vision*, (pp. 3551-3558). Sydney, Australia.

Wang, J., Liu, Z., Chorowski, J., Chen, Z., & Wu, Y. (2012a). Robust 3D Action Recognition with Random Occupancy Ptterns. *In Proceedings of the European Conference on Computer Vision*, (pp. 872-885). Florence,Italy.

Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012b). Mining Actionlet Ensemble for Action Recognition with Depth Cameras. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1290-1297). Providence, RI. doi:10.1109/CVPR.2012.6247813

Wiliem, A., Madasu, V., Boles, W., & Yarlagadda, P. (2010). An Update-Describe Approach for Human Action Recognition in Surveillance Video. *In Proceedings of the International Conference on Digital Image Computing: Techniques and Applications*, (pp. 270-275). Sydney, Australia. doi:10.1109/DICTA.2010.55

Xia, L., & Aggarwal, J. (2013). Spatio-Temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera. *In Proceedings of the Conference on Computer Vision and Pattern Recognition*, (pp. 2834-2841). doi:10.1109/CVPR.2013.365

Xia, L., Chen, C.-C., & Aggarwal, J. (2012). View Invariant Human Action Recognition Using Histograms of 3D Joints. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 20-27). doi:10.1109/CVPRW.2012.6239233

Yang, X., & Tian, Y. (2012a). EigenJoints-Based Action Recognition Using Naïve-Bayes-Nearest-Neighbor. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (pp. 14-19). Province, RI.

Yang, X., Zhang, C., & Tian, Y. (2012b). Recognizing Actions Using Depth Motion Maps-Based Histograms of Oriented Gradients. *In Proceedings of the 20th ACM International Conference on Multimedia*, (pp. 1057-1060). New York, USA. doi:10.1145/2393347.2396382

Zhu, H.-M., & Pun, C.-M. (2013). Human Action Recognition with Skeletal Information from Depth Camera. *In Proceedings of the IEEE International Conference Information and Automation*, (pp. 1082 –1085). Yinchuan, China. doi:10.1109/ICInfA.2013.6720456