NAIVE BAYESIAN FUSION FOR ACTION RECOGNITION FROM KINECT

Amel Ben Mahjoub[1], Mohamed Ibn Khedher[2], Mohamed Atri[1] and Mounim A. El Yacoubi[2]

[1]Electronics and Micro-Electronics Laboratory, Faculty of Sciences of Monastir, Monastir University, Tunisia
[2]SAMOVAR, Telecom SudParis, CNRS, University of Paris Saclay, France

ABSTRACT

The recognition of human actions based on three-dimensional depth data has become a very active research field in computer vision. In this paper, we study the fusion at the feature and decision levels for depth data captured by a Kinect camera to improve action recognition. More precisely, from each depth video sequence, we compute Depth Motion Maps (DMM) from three projection views: front, side and top. Then shape and texture features are extracted from the obtained DMMs. These features are based essentially on Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) descriptors. We propose to use two fusion levels. The first is a feature fusion level and is based on the concatenation of HOG and LBP descriptors. The second, a score fusion level, based on the naive-Bayes combination approach, aggregates the scores of three classifiers: a collaborative representation classifier, a sparse representation classifier and a kernel based extreme learning machine classifier. The experimental results conducted on two public datasets, Kinect v2 and UTD-MHAD, show that our approach achieves a high recognition accuracy and outperforms several existing methods.

KEYWORDS

Action recognition, Depth motion maps, Features fusion, Score fusion, Naive Bayesian fusion, RGB-D.

1. INTRODUCTION

The field of action recognition has been considered as an active challenging domain in computer vision research for more than two decades. It is necessary for several applications such as intelligent video surveillance, robot control, video understanding, healthcare, etc. In the past few years, further investigations [1–4] have been initially focused on recognizing actions from RGB video sequences recorded by traditional 2D cameras. Recently, the emergence of low-cost RGB-D cameras, such as Microsoft Kinect v2, has gained much attention in computer vision thanks to its excellent accuracy in action recognition. Kinect v2 provides RGB and depth data modalities. It has been used to improve the performance of human action recognition systems. The rapid development of such cameras has opened the door to a rich representative work [5–10] in learning and recognizing actions based on depth video sequences. Depth maps have various

David C. Wyld et al. (Eds) : CNDC, DPPR, AIAA, WiMNET, ACITY - 2017
pp. 53– 69, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.71606

advantages compared to traditional color videos. First, they are insensitive to change in lighting conditions. Second, they provide a three-Dimensional (3D) structure and shape information that improves the distinguish ability of different poses. These innovations have been behind producing a lot of multimodal datasets dedicated to human action recognition systems. [11] described most RGB-D datasets currently exploited in recognizing actions. Three levels of information fusion have shown an improvement in accuracy: (i) data level, where data from several sensors can be integrated to supply new data; (ii) feature level, where the different feature sets extracted from a data source are fused to create a new fused feature vector; and (iii) decision level, where the fusion of multiple classifiers is used to make the final classification decision.

This paper addresses how to enhance recognition accuracy using feature and score fusion levels. First, three Depth Motion Maps (DMMs) [7] are computed in order to represent each action video sequence. Next, the description of the obtained DMMs is performed on the basis of Histogram of Oriented Gradients (HOG) [12] and Local Binary Patterns (LBP) [13] descriptors that encode contour and texture depth features. The HOG-LBP feature fusion approach is applied to carry out a compact DMM representation. To get action prediction outputs from the feature variables, we train three classifiers: Collaborative Representation Classifier (CRC) [10, 14], Sparse Representation Classifier (SRC) [15, 16] and Kernel based Extreme Learning Machine (KELM) [17]. These techniques are among the most widely used methods in the literature [9,10,14,15,18,19], as they have shown good performances for activity recognition systems, but as far we know, this is the first time that these three classifiers are fused together to classify action. Finally, we consider a Naive-Bayes approach to combine the classification scores, that shows an improvement in the accuracy of human action recognition when tested on publicly available datasets [14] [20]. The naive Bayesian approach is a commonly known methodology for classifier output fusion, is proposed in various works as [21–23]. Our experimental results substantiated that our proposed human action recognition approach performs better than various state-of-the-art methods.

The rest of the paper is organized as follows. In section 2, a state of the art of human action recognition methods is presented. Section 3 describes the DMM as well as our proposed fusion and classification approaches. The experimental results are presented in section 4. Section 5 includes a conclusion and perspectives.

2. STATE OF THE ART

Several recent action recognition approaches have been presented recently [24–26]. Earlier, action recognition data was provided from an RGB camera. However, human activity recognition from color video sequences has many difficulties such as illumination changes and variations in human appearance.

Recently, by the appearance of depth cameras like Microsoft Kinect, several RGB-D-based human action recognition methods have been developed, as reviewed in [27–29]. The Kinect sensor captures data as color and depth information. In the literature, these provided RGB-D data in addition to skeleton joints have been well explored to improve human action recognition.

In [5], the authors define Space-Time Occupancy Patterns (STOP) to represent 3D depth maps. Both space and time axes are divided into several segments to present a 4D grid. An occupancy feature, calculated in each grid cell, represented the number of occupied space-time points. The

occupancy values of all cells formed STOP feature vectors. A nearest neighbor classifier was used to recognize human actions.

Wang and Lie [6] extracted the Random Occupancy Pattern (ROP) from depth sequences by considering a 3D depth sequence as a 4D shape. ROP features were calculated by applying a weighted sampling scheme founded on rejection sampling. An elastic-net regularized model was utilized to choose the most discriminative features to train red a Support Vector Machine (SVM) classifier for action recognition.

In [7], the authors proposed shape and motion-action representations. Each 3D depth frame was projected into three 2D maps and then the difference between two consecutive maps yielded a motion energy. The concatenation of all these motion energies over the video give out the DMM. The HOG was computed from front, side and top DMM maps as a distribution of local intensity gradients. The DMM-HOG features were matched by an SVM classifier for action recognition.

Oreifej and Liu [8] introduced the histogram of oriented 4D normals which was the extension of histograms of oriented 3D normals [30] by appending the time derivative. The depth, time and spatial coordinates of a 4D space were quantified by a regular polychoron, and then each human action was modeled as a distribution of the normal surface orientations.

Moreover, Chen [9] presented a DMM based on an LBP descriptor for human action. The DMM was presented from three projection views to characterize the 3D local motion. The LBP features were extracted from the depth maps to measure the local image texture by encoding each pixel with decimal numbers. All the extracted LBP features from each projected DMM were concatenated to give a single feature vector, used to train a KELM classifier.

Farhad and Jiang in [10] developed a new descriptor that computed HOG features from DMMs based on contourlet sub-bands. A Contourlet Transform (CT) combined the Laplacien pyramid and the directional filter bank technique to decompose the DMMs into low-frequency and high-frequency sub-bands. This method was used to decrease the noise and clearly present the depth shape information at several scales and directions. Afterwards, HOG features were extracted from these DMM contourlet sub-bands. The combination of the histograms obtained from the three depth views provided the final DMM-CT-HOG feature vector, trained by the CRC to classify human action.

The work of [31, 32] was inspired by the success of deep learning in human action recognition. A new deep learning based action recognition framework using depth and skeleton data was defined in [32]. The deep convolutional neural network was used to extract the spatio-temporal features from depth sequences. A jointVector feature was obtained by computing the angle and position between skeleton joint information. A SVM classifier matched high-level and jointVector features separately to get class probability vectors. The fusion of these two kinds of weighted vectors gives final action recognition.

Ivan in [33] introduced an approach of body-pose estimation based on RGB-D video sequences to recognize complex human activities. Two geometric and motion descriptors were applied to each RGB-D datum to encode respectively the spatial configuration and dynamic features of every body-pose. A hierarchy of three levels was modeled to produce the global activities prediction. At the first level, each activity was decomposed into atomic actions. The intermediate level

represented the atomic human action with sparse composition. At a higher level, the complex human activities were described based on spatio-temporal compositions of atomic actions.

Various publications have appeared in the recent years demonstrating the importance of fusion methods in improving the accuracy of action recognition systems. One of the examples for decision-level fusion was presented in [14]. The authors extracted feature vectors from depth, skeleton and inertial data. Next, they matched three CRC classifiers separately to get the video label. Finally, a Logarithmic Opinion Pool (LOGP) was carried out for decision-level fusion. Imran and Kumar in [34] suggested a new human action recognition method based on classification fusion. Four deep convolutional neural networks were utilized to classify the Motion History Image (MHI) vector descriptors from RGB data and three DMM vectors from depth sequences. The fusion of the output scores of these four networks was done using average and product rule approaches. The Dempster-Shafer (DS) method was put forward in [35] for decision level fusion. In the latter work, both depth and inertial data were exploited to extract feature vectors. The authors used the DS technique to fuse the decision outputs of two CRC classifiers. The authors in [18] described three DMMs by CT-HOG, LBP and Edge Oriented Histograms (EOH) feature descriptors. Then, they used three KELM classifiers to make a decision for each feature vector. The LOGP and majority voting methods were proposed to combine the outputs of classifiers in order to improve activity recognition accuracy. A probabilistic classification fusion approach utilizing a Bayes formalism was performed in [23]. Multiple Hidden Markov Models (HMM) classifiers were matched given the features from accelerometer sensors placed on the body. The naive Bayesian fusion technique was executed to concatenate the classification output vectors from all HMMs.

3. PROPOSED APPROACH

In this section, we present our approach which is broadly described in Figure 1. In order to identify the action of a person in the depth video sequence, we firstly extract the shape and texture features using HOG and LBP descriptors. The key idea of our method is to fuse these two kinds of feature vectors before classification. A dimensionality reduction is secondly performed based on the Principal Component Analysis (PCA) technique. Thirdly, the training is done by applying CRC, SRC and KELM classifiers to get different classification score outputs. Finally, we fuse the probabilistic information sources via a Naive-Bayes technique to output the label of the sequence.

4. FEATURE EXTRACTION

Two features are extracted to describe the DMM images : 1) HOG and 2) LBP.

4.1 Depth Motion Map

The DMM [7] is used to characterize the 3D structure and shape information from depth maps. A depth video sequence contains M frames, where each frame is projected over three orthogonal Cartesian planes to build DMM_f , DMM_s and DMM_t images from front, side and top views, respectively. The computation of motion energy is performed by subtracting the consecutive maps from each projected DMM. The sum of motion energy over the video sequence yields the $DMM_{f,s,t}$ as follows:

$$DMM_{f,s,t} = \sum_{i=o}^{M-1} (DMM_{f,s,t}^{i+1} - DMM_{f,s,t}^{i}| > \epsilon) \quad (1)$$

where f,s,t are the front, side and top views and $\in$ 2 is a threshold. For each projected map, the entire sequence frames are not used but just their extracted regions of interest. These extracted foreground DMMs are normalized to generate the final DMM features. Figure2 shows an example of three projected maps of a right hand high wave depth sequence.

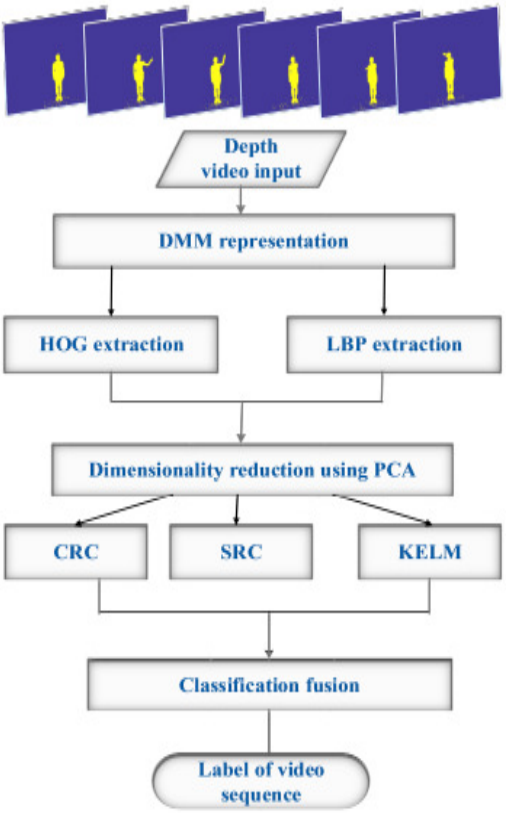

Figure 1. Flowchart of our approach

Figure 2. DMM views: DMM_f, DMM_s and DMM_t

4.2 Histogram of Oriented Gradients

The HOG was introduced in [12] and was used to encode the local appearance and shape on DMM maps [7] with the distribution of local intensity gradients or edge directions. After finding the object using depth information, the idea is to calculate the occurrences of discretized gradient orientations in the depth local region to represent the body shape and motion information. We divide every projected depth map into 8x8 non overlapping cells, where each cell has nine orientation bins. The pixels of these cells throw a weighted vote for an orientation histogram based on the value of the gradient magnitude to yield a histogram of nine gradient directions. It results in three HOG vectors that describe map features from front, side and top DMMs. These vectors are concatenated to produce a 6,588 dimensional final DMM-HOG descriptor for the entire action video sequence.

4.3 Local Binary Patterns

The first LBP encoding schemes were proposed for quantifying the local image contrast. Ojala in [13] extended the LBP to an arbitrary circular derivation to describe the local texture pattern. The Computation of the texture can be carried out by thresholding a neighborhood by the gray value of its center and by assigning decimal numbers to the pixels of the image. Let g_c be a scalar value of the center pixel and g_P(**P=0, ... P-1**) be the gray values of its neighborhood of P which are pixels equally spaced on a circle with a radius R. This circle constitutes a circular symmetric whole of its neighbors. The LBP feature is established by subtracting the g_P neighbors from the center value g_c to produce a P digit binary number converted to a decimal form as follows:

$$LBP_{P,R} = \sum_{p=o}^{P-1} S(g_P - g_c)2^P \tag{2}$$

where P is the number of neighborhood pixels and S(x) is

$$S(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

We then obtain 2^P uniform patterns. The evaluation of a histogram over an image, which represents the frequency of each occurring number, is generated to encode the texture information. In this paper, we use the LBP operator to extract features from the DMM maps as in [9].

4.4 Feature fusion

Recently, the use of information fusion has attracted the attention of researchers in the action recognition domain as a consequence of its greater accuracy. The concatenation of feature vectors to improve the system performance is a simple and traditional method frequently used in the literature. The feature vectors of different kinds have been concatenated together to find a single long feature vector to train the classifier. In [36], Wang and Han employed this fusion approach to combine HOG and LBP feature vectors for human detection. Dimitrovski and Kocev [37] performed the low level feature fusion approach to describe medical images. Several extracted features have been concatenated in a single feature vector before the classification step. Local

feature and boundary based shape features were concatenated in [38] to improve the recognition performance of the object class. In [39], the authors started by extracting the HOG and LBP features. After that, they applied PCA to each of them to reduce the dimension. Finally, the concatenation of the two obtained feature vectors was performed to give the mixed HOG-LBP descriptor.

In our work, we extract HOG and LBP features from the DMM to represent the depth sequence from diverse prospects. We then apply the fusion algorithm based on the PCA to concatenate DMM-HOG and DMM-LBP feature vectors.

5. CLASSIFICATION ALGORITHM

5.1 Collaborative Representation Classifier

The CRC has been employed in various work [10, 14] owing to its performance and efficiency in classification. Given c as the number of classes, we have n training samples belonging to c classes, denoted $\mathbf{X= [X_1, X_2, ..., X_c]} = [x_1, x_2, ...,x_n]$, where $\mathbf{X} \in \mathbf{R}^{d \times n}$ and the total number of training samples is $n= n_1 + n_2 + ... + n_c$, n_k being the number of samples pertaining to class k. The test sample y is described in the CRC as a linear association: $\mathbf{y = \alpha X}$, where $\alpha = [\alpha_1, \ldots, \alpha_c]$ represents the coefficient vector of the corresponding training sample. We then apply the l_2 norm to optimize α by a minimizing formulation as follows:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\alpha} \| \mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\mathbf{L}\|_2^2 \tag{3}$$

Where λ is a parameter of regularization and L is the Tikhonov regularization matrix that represents the distance weighted matrix by giving less weight to the training samples dissimilar from the test samples as follows:

$$\mathbf{L} = \begin{bmatrix} \|\mathbf{y} - \mathbf{x}_1\|_2 & 0 & \ldots & 0 \\ 0 & \|\mathbf{y} - \mathbf{x}_2\|_2 & \ldots & 0 \\ 0 & 0 & \ldots & \|\mathbf{y} - \mathbf{x}_n\|_2 \end{bmatrix} \tag{4}$$

The Tikhonov regularization is used to solve this minimization problem:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{L}\mathbf{L}^T)^{-1}\mathbf{X}^T\mathbf{y} \tag{5}$$

A minimization of the reconstruction error is made to take a classification decision.

$$label(\mathbf{y}) = \arg\min_{k} e_k(\mathbf{y})) = \| \mathbf{y} - \boldsymbol{X_k}\hat{\boldsymbol{\alpha}}_k\|_2 \tag{6}$$

where $k \in [1, 2, ...c]$ and e_k is the residual error.

5.2 Sparse representation

Human action recognition utilizing SRC is inspired by the work of Wright and Yang [15] which used sparse representation to recognize faces. In the SRC, the testing samples are obtained by a sparse linear combination of the training samples. The unknown sample is identified by finding the label with the lowest residual error. Given a matrix X of training samples for c classes and an optional error tolerance $\epsilon > 0$, the test sample of the kth class y is represented from the training set $\mathbf{X}_k$ with the coefficients $\boldsymbol{\alpha}_k$. According to the sparse representation of y in terms of dictionary constructed from training samples of all c classes, we can retrieve α_k as follows :

$$\begin{aligned} &\arg\min_{\boldsymbol{\alpha}} \| \boldsymbol{\alpha} \|_1 \\ &\text{subject to } \| \mathbf{X}\boldsymbol{\alpha} - \mathbf{y} \|_2^2 \leq \epsilon \end{aligned} \tag{7}$$

Subsequently, we classify y for k= 1, 2, ..., c as follows:

$$identity(\mathbf{y}) = \arg\min_k (r_i = \| \mathbf{y} - \mathbf{X}_k \boldsymbol{\alpha}_k \|_2^2) \tag{8}$$

The calculation of identity (y) defines the label of the test sample y from all distinct c classes.

5.3 Kernel based extreme learning machine

KELM was developed in [17] to solve regression and multiclass classification tasks. An extreme learning machine was initially dedicated to match a Single-hidden Layer Feedforward Neural Network (SLFNN) in order to overcome the learning slowness. We have the n training samples $\{\mathbf{x}_i, l_i\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbf{R}^d$ and $l_i \in [1 \ldots c]$ is it corresponding label. $\mathbf{t_i} = [t_{i1}, \ldots, t_{ic}]^T$ is the network target binary vector that denotes the sample belonging, where only one component is non null. For example, if $t_ik = 1$, it implies that the sample belongs to class k. The responses of the SLFNN to $\mathbf{x}_i$, $\mathbf{h}_i = [h_{il}, \ldots, h_{ic}]^T$ is:

$$h_{ik} = \sum_{j=1}^{D} \boldsymbol{\alpha}_{kj} f(\mathbf{w}_j \mathbf{x}_i + e_j) \tag{9}$$

where D is the number of the hidden nodes, f(.) is a linear activation function for the network output layer, $\mathbf{w}_i \in \mathbf{R}^d$ and $\boldsymbol{\alpha}_{ki} \in \mathbf{R}^D$ are the weight vectors that connect ith hidden node to the input and output nodes respectively and e_i is the bias for the ith hidden node. For all n equations we have: $\mathbf{h} = \boldsymbol{\alpha}\ \mathbf{F}$, where $\boldsymbol{\alpha}$ is the network output weights $\in \mathbf{R}^{D \times c}$ and $\mathbf{F}$ is the matrix of hidden layer outputs of all training sets x_i, which is written as:

$$\mathbf{F} = \begin{bmatrix} f(\mathbf{w}_1 \mathbf{x}_1 + e_1) & \ldots & f(\mathbf{w}_D \mathbf{x}_1 + e_D) \\ \ldots & \ldots & \ldots \\ f(\mathbf{w}_1 \mathbf{x}_n + e_1) & \ldots & f(\mathbf{w}_D \mathbf{x}_n + e_D) \end{bmatrix}$$

$T = [\ t_1,..,\ t_n\ ]^T$ is the matrix that contains the network target vectors. The output weights α is analytically computed as:

$$\boldsymbol{\alpha} = \mathbf{F}'\mathbf{T}^T \tag{10}$$

where $\mathbf{F}'$ is the Moore-Penrose generalized inverse of the matrix **F:**

$$\mathbf{F}' = \mathbf{F}(\mathbf{F}\mathbf{F}^T)^{-1} \tag{11}$$

A positive regularization term $1/\rho$ is added to the diagonal elements of $\mathbf{F}\mathbf{F}^T$, so we have:

$$\boldsymbol{\alpha} = (\mathbf{F}\mathbf{F}^T + \frac{\mathbf{I}}{\rho})^{-1}\mathbf{F}\mathbf{T}^T \tag{12}$$

The kernel matrix for the ELM is used as follows:

$$\Omega_{ELM} = \mathbf{F}\mathbf{F}^T: \Omega_{ELMj,s} = f(\mathbf{x}_j).f(\mathbf{x}_i) = K(\mathbf{x}_j, \mathbf{x}_i).$$

Therefore, the output of KELM classifier is :

$$h(\mathbf{x}) = \begin{bmatrix} K(\mathbf{x}, \mathbf{x}_1) \\ ... \\ K(\mathbf{x}, \mathbf{x}_n) \end{bmatrix} (\frac{\mathbf{I}}{\rho} + \Omega_{ELM})^{-1}\mathbf{T}^T \tag{13}$$

where $\mathbf{I} \in \mathbf{R}^{D \times D}$ is an identity matrix .

The predicted class label of the testing sample $\mathbf{y} \in \mathbf{R}^d$ is the index number of the network output node which has the highest value. Considering $\mathbf{f}_k(\mathbf{y})$ as the output function of the kth hidden node, where $\mathbf{f}(\mathbf{y}) = [f_1(\mathbf{y}),\ldots, f_c(\mathbf{y})]T$, the predicted class of y is calculated as :

$$label(\mathbf{y}) = \arg\max_{k=1...c} f_k(\mathbf{y}) \tag{14}$$

5.4 Classifier fusion

The fusion of various classifiers is a known robust technique as it is usually more robust and accurate than a single learner system. For benchmarking, we consider nine fusion approaches. The Majority vote serves in collecting all the votes of the different classifiers and selecting the label that is the most frequently occurring value. The maximum approach chooses the most confident classifier with the highest classification score. The sum function calculates the sum of score output elements of classifiers and outcomes the label with the highest value. The minimum method gives the class which has a minimum objection by different classifiers. The mean of the output classifier scores consists in choosing the label with the highest mean value. The product fusion technique consists in multiplying the vector elements of the classifier score outputs, and the final decision corresponds to the class with the highest probability. The Decision Template (DT) is a simple and robust classifier fusion method that compares the classifier output combination with a representative template for each class. The decision templates are the

averages of all classifier decision outputs during the training step that ties to the belonging of training samples to each class. These templates are later used to output the final class based on similarity measure. The Dempster-Shafer (DS) technique is like the DT and we can place both methods in the same group. The decision templates are generated from the training data to represent the most characteristic decision profile for each class. In the testing step, the DS method compares the DT to give the label with the largest similarity.

Naive-Bayes is a powerful technique for combining the confidence outputs of different classifiers [21, 22]. This method consists in calculating the a posteriori probability of each possible class w_k given the output labels s_i of the different classifiers. We assume that we have c classes and L classifiers D_i match the data $\mathbf{y} \in \mathbf{R^d}$. Each classifier generates a label s_i, $i \in [1,L]$, so we have the output vector $s = [s_1,\ldots s_L]$. We define $p(s_i)$ as the probability that classifier D_i labels $\mathbf{y}$ in class s_i. Naive Bayesian approach computes the a posteriori probability that y is labeled as w_k as follows:

$$p(w_k/\mathbf{s}) = \frac{p(\mathbf{s}/w_k)p(w_k)}{p(\mathbf{s})} \qquad k = 1...c \tag{15}$$

where $p(w_k)$ is the a priori probability of the hypothesis w_k, $p(\mathbf{s} / w_k)$ is the likelihood and $p(\mathbf{s})$ is the evidence used for normalization, which can be neglected. The equation that describes the support for class w_k can be written as :

$$\mu_k(\mathbf{y}) \propto p(w_k) \prod_{i=1}^{L} p(s_i/w_k) \tag{16}$$

The $\mathbf{c^*c}$ confusion matrix CM^i is defined for each classifier D_i where cm^i_{k,s_t} is the number of data elements having a true class w_k , and labeled by D_i as class w_s. Assuming that the training dataset $\mathbf{X}$ contains a total of n elements and n_s elements from class w_s, the probability for class w_k is given by $\frac{n_k}{n}$ and the probability $p(s_i/w_k)$ can be written in the form $\frac{cm^i_{k,s_t}}{n_k}$. In this way and according to (16), we obtain:

$$\mu_k(\mathbf{y}) \propto \frac{1}{n_k^{L-1}} \prod_{i=1}^{L} cm^i_{k,s_t} \tag{17}$$

The highest value $\mu_k(\mathbf{y})$ is used to label y in class w_k

6. EXPERIMENTAL RESULTS

In order to verify the validity of our method, we have carried out experiments based on the two Kinect v2 and UTD-MHAD datasets.

Table 1. UTD-MHAD dataset actions

Arm cross	Arm curl
Baseball swing	Basketball shoot
Bowling	Boxing
Catch	Clap
Draw circle CCW	Draw circle CW
Draw triangle	Draw X
Jog	Knock
Lunge	Pickup and throw
Push	Sit to stand
Squat	Stand to sit
Swipe left	Swipe right
Tennis serve	Tennis swing
Throw	Walk
Wave	

6.1 Kinect v2 dataset

We harnessed a public multimodal dataset defined in [14], with collected data from a Kinect v2 camera and a wearable inertial sensor simultaneously. This database contains three modalities of depth, skeleton joint position and inertial signals. In this paper, we use the depth images modality. As shown in Figure 3, the dataset includes the following ten actions: right hand high wave, right hand catch, right hand high throw, right hand draw x, right hand draw tick, right hand draw circle, right hand horizontal wave, right hand forward punch, right hand hammer and hand clap (two hands). These actions were effected by three female subjects and three male ones and each subject rehearsed the actions five times. Consequently, we obtained 300 depth video sequences in total. We performed the subject-specific experiment as in [14] to divide the data into training and testing steps. For each subject, the first two repetitions were chosen in training, and the remaining repetitions in testing. Table 2 compares our proposed approach with that used in [14], where the authors applied the DMM and CRC techniques to recognize actions from depth samples.

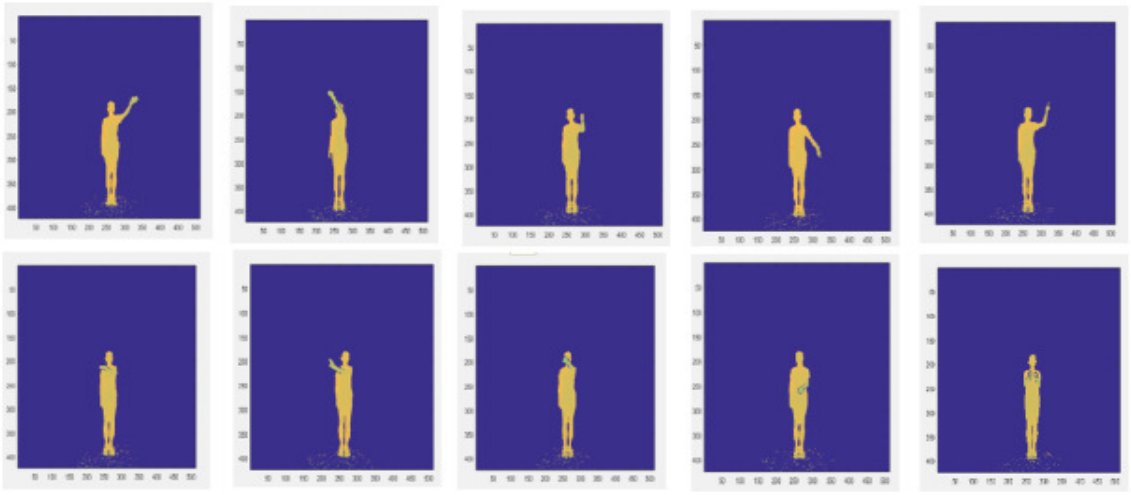

Figure 3. Kinect v2 dataset depth images

6.2 UTD-MHAD dataset

The UTD-MHAD dataset, described in [20], is acquired by a Kinect camera and an inertial sensor that collect multimodal data. UTD-MHAD encloses 27 actions which are mostly an arm or a leg based activity, as shown in Table 1.

Table 2. Results comparison on kinect v2 dataset

Approach	Description	Recognition rate (%)
[14]	DMM	79.4 %
Our method	classification fusion	93.9 %

Table 3. Results comparison on UTD-MHAD dataset for the half-subject experiment

Approach	Description	Recognition rate (%)
[20]	Kinect(only)	66.1
[40]	DMM	73.4
[41]	GF + LF	84.89
[34]	Depth + deep CNN	87.9
[18]	Decision-level fusion (LOGP)	88,4
Our	Classification fusion	90.5

Each action was carried out by four females and four males, and each subject performed four repetitions, so we have 861 video sequences in total after eliminating three corrupted videos. The UTD-MHAD is a multimodal database that includes color and depth videos, skeleton joint positions and inertial sensor signals (acceleration, angular velocity and magnetic strength). We executed two types of experiments on the UTD-MHAD dataset. The first is nominated half-subject where subjects 1, 3, 5 and 7 were used for training and subjects 2, 4, 6 and 8 for testing. The comparison of the existing work with our approach using the UTD-MHAD dataset for the half-subject experiment is illustrated in Table 3. The second experiments are the subject-specific settings in [19]. As each subject performs an action four times, the first two repetitions are used for training and the two remaining repetitions for testing. Table 4 depicts our obtained results compared to the method implemented in [19] based on the Kinect depth feature only.

6.3 Evaluation protocol

The experiment was carried out using a computer i7 3.4GHZ with a RAM of 16 GB.

Table 5 lists the recognition rate results of nine methods for score level fusion of the CRC, the SRC and the KELM. Figure 4 reports the classification results using the SRC, the CRC and the KELM each alone and then the fusion of these three classifiers based on the naive-Bayes fusion method.

Table 4. Results comparison on UTD-MHAD dataset for the subject-specific experiment

Approach	Description	Recognition rate (%)
[19]	Kinect (only)	85.1
Our method	classification fusion	91.6

6.4 Discussion

Recognizing human actions with a good recognition rate is an key computer vision requirement. In our paper, we propose a fusion approach based on Kinect v2 and UTD-MHAD datasets to improve accuracy. In the first step, we start by extracting the HOG and the LBP from the DMM representation of depth video sequences. Then, we concatenate these features using PCA to reduce dimension. Finally, the naive Bayesian approach is applied to fuse the classification score outputs of the CRC, the SRC and the KELM classifiers. As detailed in Table 5, we test different fusion methods on the Kinect v2 and UTD-MHAD datasets. These findings point to the usefulness of naive-Bayes as a score level fusion approach as they give a great recognition rate for both datasets. Figure 4 highlights how important our score level fusion method is for improving the recognition rate compared to approaches using each classifier alone. It is apparent from Table 2 that our method on the Kinect v2 dataset outperforms previous methods [14] by around 15%. The results of the two types of experiments on UTD-MHAD can be seen in Table 4 and Table 3. In Table 4, our technique for the subject-specific experiment demonstrates a clear advantage over the work in [19], which improves the accuracy by 6.5%. We observe from Table 3 that our fused method for the half-subject experiment outperforms previous work based on depth data with an accuracy of 90.5%. We are aware that our work may have two limitations. The first is that we have not considered conjointly the depth skeleton joint position and inertial multimodalities data and the fusion between them which can improve the classification rate. The second negative factor regarding our algorithm is that it has a limit with real-time constraints. An acceleration of our action recognition method can be suggested by using field-programmable gate array or graphics processing unit to speed up the application.Despite this, we can still state that our approach outperforms several previous methods based on only depth data, and it can be employed in computer vision applications.

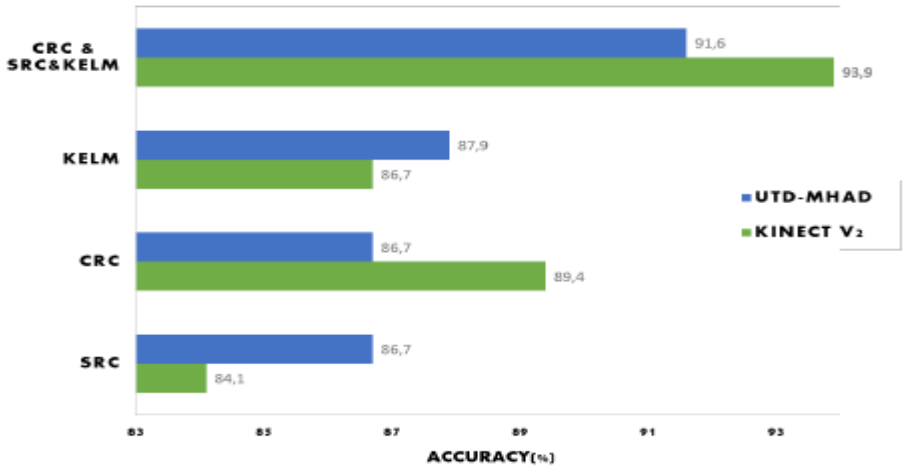

Figure 4. Recognition accuracy of our proposed approach

Table 5. comparison of fused methods of CRC, SRC and KELM classifiers

fused method	accuracy on kinect v2(%)	accuracy on UTD-MHAD (%)
Majority vote	87.8	86.9
Maximum	87.8	86.9
Minimum	88.9	87.9
Sum	86.7	87.4
Average	88.9	87.4
Product	68.9	79.8
Decision template	86.7	89.3
Dempster-shafer	85.5	85
Naive-bayes	93.9	91.6

7. CONCLUSION

This paper has given an account of our proposed probabilistic score level fusion based on the Bayes theorem for human action recognition tested on the Kinect v2 and UTD-MHAD datasets. We have exploited the depth video sequence in these datasets to calculate the DMM. To represent the DMM, we have used the HOG and LBP feature descriptors. The concatenation of the DMM-HOG and the DMM-LBP using the PCA technique has been then performed. Finally, we have applied a naive Bayesian approach to fuse the SRC, CRC and KELM classification scores, that has been shown to outperform different other fusion methods. Our results indicate that our system presents a good recognition accuracy compared to existing work. Future work will focus on multimodalities fusion at data, feature or score levels. We will also develop a co-design architecture to speed up the system.

ACKNOWLEDGEMENT

This work was carried out at the laboratory of Electronics and Microelectronics at the Faculty of Sciences of Monastir, Tunisia and the laboratory of SAMOVAR at Telecom SudParis, France.

REFERENCES

[1] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 1–8, IEEE (2008).

[2] J. Sun, X.Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, “Hierarchical spatio-temporal context modeling for action recognition,” in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2004–2011, IEEE (2009).

[3] M. Selmi, M. A. El-Yacoubi, and B. Dorizzi, “Two-layer discriminative model for human activity recognition,” IET Computer Vision 10(4), 273–278, IET (2016).

[4] M. Selmi, M. El Yacoubi, and B. Dorizzi, "On the sensitivity of spatio-temporal interest points to person identity," in Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium on, 69–72, IEEE (2012).

[5] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," in Iberoamerican Congress on Pattern Recognition, 252–259, Springer (2012).

[6] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," in Computer vision–ECCV 2012, 872–885, Springer (2012).

[7] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in Proceedings of the 20th ACM international conference on Multimedia, 1057–1060, ACM (2012).

[8] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 716–723 (2013).

[9] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on, 1092–1099, IEEE (2015).

[10] M. F. Bulbul, Y. Jiang, and J. Ma, "Human action recognition based on dmms, hogs and contourlet transform," in Multimedia Big Data (BigMM), 2015 IEEE International Conference on, 389–394, IEEE (2015).

[11] J. Zhang,W. Li, P. O. Ogunbona, P.Wang, and C. Tang, "Rgb-d-based action recognition datasets: A survey," Pattern Recognition 60, 86–105, Elsevier (2016).

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 1, 886–893, IEEE (2005).

[13] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Transactions on pattern analysis and machine intelligence 24(7), 971–987, IEEE (2002).

[14] C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, 2712–2716, IEEE (2016).

[15] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," IEEE transactions on pattern analysis and machine intelligence 31(2), 210–227, IEEE (2009).

[16] M. I. Khedher, M. A. El Yacoubi, and B. Dorizzi, "Multi-shot surf-based person re-identification via sparse representation," in Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on, 159–164, IEEE (2013).

[17] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42(2), 513–529, IEEE (2012).

[18] M. F. Bulbul, Y. Jiang, and J. Ma, "Dmms-based multiple features fusion for human action recognition," International Journal of Multimedia Data Engineering and Management (IJMDEM) 6(4), 23–39, IGI Global (2015).

[19] C. Chen, R. Jafari, and N. Kehtarnavaz, "A real-time human action recognition system using depth and inertial sensor fusion," IEEE Sensors Journal 16(3), 773–781, IEEE (2016).

[20] C. Chen, J. Roozbeh, and K. Nasser, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in ICIP, 168–172, IEEE (2015).

[21] H. Altınc¸ay, "On naive bayesian fusion of dependent classifiers," Pattern Recognition Letters 26(15), 2463–2473, Elsevier (2005).

[22] L. I. Kuncheva, "Combining pattern classifiers: methods and algorithms," John Wiley & Sons (2004).

[23] P. Zappi, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Troster, "Activity recognition from on-body sensors by classifier fusion: sensor scalability and robustness," in Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on, 281–286, IEEE (2007).

[24] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," in Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications, 149–187, Springer (2013).

[25] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," Computer vision and image understanding 115(2), 224–241, Elsevier (2011).

[26] M. Selmi and M. A. El-Yacoubi, "Multimodal sequential modeling and recognition of human activities," in International Conference on Computers Helping People with Special Needs, 541–548, Springer (2016).

[27] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," IEEE transactions on cybernetics 43(5), 1318–1334, IEEE (2013).

[28] J. K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," Pattern Recognition Letters 48, 70–80, Elsevier (2014).

[29] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," Multimedia Tools and Applications 76(3), 4405–4425, Springer (2017).

[30] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao, "Histogram of oriented normal vectors for object recognition with a depth sensor," in Asian conference on computer vision, 525–538, Springer (2012).

[31] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, and others, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks.," in AAAI, 2, 8 (2016).

[32] Z. Liu, C. Zhang, and Y. Tian, "3d-based deep convolutional neural network for action recognition with depth sequences," Image and Vision Computing 55, 93–100, Elsevier (2016).

[33] I. Lillo, J. C. Niebles, and A. Soto, "Sparse composition of body poses and atomic actions for human activity recognition in rgb-d videos," Image and Vision Computing 59, 63–75, Elsevier (2017).

[34] J. Imran and P. Kumar, "Human action recognition using rgb-d sensor and deep convolutional neural networks," in Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on, 144–148, IEEE (2016).

[35] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," IEEE Transactions on Human-Machine Systems 45(1), 51–61, IEEE (2015).

[36] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in Computer Vision, 2009 IEEE 12th International Conference on, 32–39, IEEE (2009).

[37] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Dˇzeroski, "Hierarchical annotation of medical images," Pattern Recognition 44(10), 2436–2449, Elsevier (2011).

[38] M. Noridayu, R. A. R. Abdul, R. Ava, and R. Dhanesh, "Feature fusion in improving object class recognition," Citeseer (2012).

[39] N. Manshor, A. R. A. Rahiman, A. Rajeswari, and D. Ramach, "Feature fusion in improving object class recognition," Citeseer (2012).

[40] N. E. D. Elmadany, Y. He, and L. Guan, "Human action recognition using hybrid centroid canonical correlation analysis," in Multimedia (ISM), 2015 IEEE International Symposium on, 205–210, IEEE (2015).

[41] E. Escobedo and G. Camara, "A new approach for dynamic gesture recognition using skeleton trajectory representation and histograms of cumulative magnitudes," in Graphics, Patterns and Images (SIBGRAPI), 2016 29th SIBGRAPI Conference on, 209–216, IEEE (2016).