Vision-Based Multi-Modal Framework for Action Recognition

Beddiar Djamila Romaissa[1],[2], Oussalah Mourad[2] and Nini Brahim[1]
[1]Research Laboratory on Computer Science's Complex Systems
University Laarbi Ben M'hidi, Oum El Bouaghi, Algeria
Email: ad_beddiar@esi.dz , Djamila.Beddiar@oulu.fi
[2]Center for Machine Vision and Signal Analysis,
University of Oulu, Finland

***Abstract*—Human activity recognition plays a central role in the development of intelligent systems for video surveillance, public security, health care and home monitoring, where detection and recognition of activities can improve the quality of life and security of humans. Typically, automated, intuitive and real-time systems are required to recognize human activities and identify accurately unusual behaviors in order to prevent dangerous situations. In this work, we explore the combination of three modalities (RGB, depth and skeleton data) to design a robust multi-modal framework for vision-based human activity recognition. Especially, spatial information, body shape/posture and temporal evolution of actions are highlighted using illustrative representations obtained from a combination of dynamic RGB images, dynamic depth images and skeleton data representations. Therefore, each video is represented with three images that summarize the ongoing action. Our framework takes advantage of transfer learning from pre-trained models to extract significant features from these newly created images. Next, we fuse extracted features using Canonical Correlation Analysis and train a Long Short-Term Memory network to classify actions from visual descriptive images. Experimental results demonstrated the reliability of our feature-fusion framework that allows us to capture highly significant features and enables us to achieve the state-of-the-art performance on the public UTD-MHAD and NTU RGB+D datasets.**

I. Introduction

Human activity recognition (HAR) refers to the process of identification and categorization of a sequence of recorded data from ubiquitous or visual sensors into well-defined basic activities. The identification of a single activity, also called activity detection, consists in temporally localizing the movements of the person in the scene. While the categorization of an activity, known as activity classification, consists of distinguishing the nature of a person movements using some spatial and temporal cues or any other meaningful features that best describe the ongoing actions and assigns it to its corresponding class. It is to note that we use here the terminology of both action and activity to refer to the same concept.

Vision-based HAR has become a very active research topic in computer vision and image processing due to its wide application fields. Roughly speaking, it covers areas of automatic video surveillance, public security, virtual and augmented reality, health care and home monitoring, human-computer interaction and robot learning [1]. In addition, the increasing progress in sensing technologies prompted the emergence of intelligent real-time systems that can potentially impact the development of efficient human activity recognition systems and enhance the quality of life and security of the individuals [2]. Major vision-based HAR works focus on using one single sensor modality to classify activities. This yields some limitations while discriminating complex activities due to environment conditions such as lighting, perspective changes and occlusions [3]. To achieve good results and enable robust HAR systems, it is important to exploit more than one modality and to this end, different fusion strategies are to be explored [1]. In this respect, the current paper attempts to combine three modalities from RGB, depth and skeleton data for vision-based HAR in order to achieve high recognition accuracy. The proposed framework integrates three sets of images created using data acquired from the modalities mentioned above. For RGB and depth data, we use an approximated version of rank-pooling in the same spirit as [4], [5] to create two sets of dynamic images. Each dynamic image summarizes information contained in the video frames into one single visual image. Furthermore, a skeleton data is used to create images that encode the locations of the skeleton joints among the video frames and hence describe the temporal aspect of the action. These newly created images are then employed to enable transfer learning from a pre-trained model in order to extract significant features. A feature-fusion strategy is performed later on using the Canonical Correlation Analysis CCA [6] to create highly discriminative feature vector that combines selective features from the three single feature vectors. Once the underlined unified vector is created, we train a Long Short-Term Memory LSTM network to classify activities from the video sequences. Finally, we test our proposal on the public UTD-MHAD dataset [7] as well as the NTU RGB+D dataset [8] and analyse the obtained results. The findings demonstrate that our approach can achieve high recognition accuracy and compete with the state-of-art HAR approaches. The general overview of our HAR proposed methodology is shown in "Fig. 1". Unlike [9] where 5 CNN streams of front, side and top depth motion maps along with motion history images and skeleton joints clustering are fused. Especially, 2D CNN of stacked dense flow difference images and a bi-gated RNN of augmented skeleton data and a 1D CNN of augmented inertial data are combined [10]. Our main contributions can

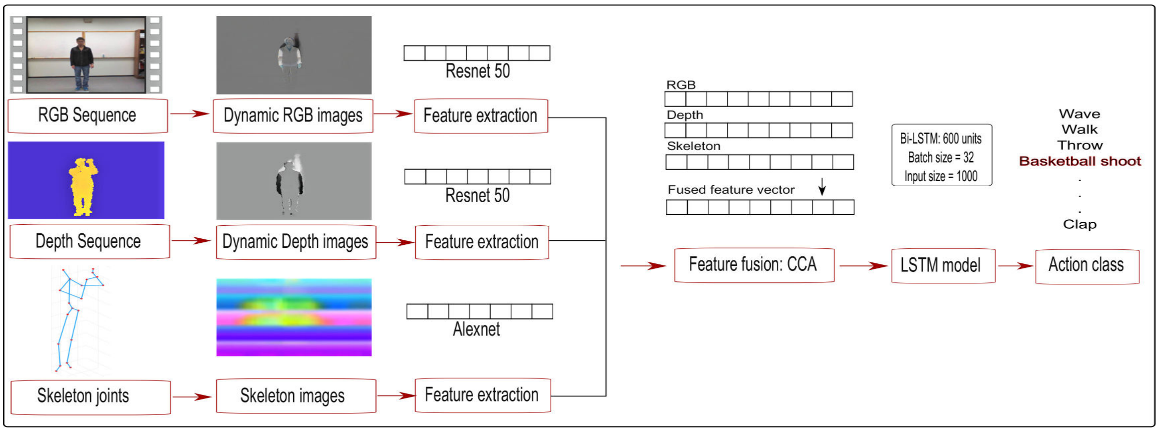

Fig. 1: The general overview of our proposed vision-based multi-modal approach for HAR

be summarized into the following:

- Summarizing RGB and depth videos into dynamic images using an approximate rank pooling method introduced in [4], [5].
- Encoding locations of skeleton joints along the video frames into new representations.
- Extracting new features from RGB and depth dynamic images and skeleton representations using transfer learning from pre-trained models.
- Developing a new feature fusion based strategy using Canonical Correlation Analysis of RGB, depth and skeleton data modalities.
- Developing one bi-directional Long Short Term model for action classification that has for input the resulting features fusion vectors.
- Evaluating our proposed method on two datasets: UTD-MHAD and NTU RGB+D where we performed cross-view evaluation.

In the next section, we will briefly review related works in vision-based HAR where we will discuss video representations, RGB, depth and skeleton-based HAR techniques, rank pooling of videos and multi-modal HAR approaches. Our proposed multi-modal method is described in detail in section III. In Section IV, we evaluate the proposed framework on the public UTD-MHAD and NTU RGB+D datasets and analyse the findings. Finally, we conclude the paper and outline future research directions in the last section.

II. Vision-based Human activity recognition

Human activity recognition (HAR), as an important research direction, has been extensively studied in the literature. Summarizing all the existing HAR systems is quite hard [1] and beyond the scope of this paper. We will restrict in this section to review only vision-based HARs that overlap to some extent with our approach. The existing vision-based approaches can be categorized according to the input data type into many categories among which RGB images, skeleton data and depth images are the most commonly employed ones. It is worth noting that, actually, some works exploit a single modality while many others combine one or two modalities to improve the performance of their HAR systems [1]. On the other hand, the temporal dimension of the action is often taken into account explicitly to enhance the recognition performance. However, many approaches extract spatial features of the image and deal with the temporal variations in the classification stage. In line with our multimodal HAR proposed framework highlighted in the introduction section, we review related work in video representation and action recognition based on the three data modalities: RGB, depth and skeleton data.

Video representations: Features extracted from image sequences expand to variations in action execution, person appearance, shape and motion. These should be sufficiently distinctive to allow distinguishing between different actions. Efficient action representation is the key to yield robust and expressive features. Therefore, many video-based HAR methods are based on video representation to efficiently describe the action. They can be grouped accordingly into two main categories. The first one considers video as a stream of still images or as transitions between frames. The second category represents videos as 3D dimensional volumes. The majority of hand-crafted-based or deep learning-based human activity representations belongs to the first category such as [5], [11]. The popularity of the first category raises from its efficiency and simplicity of use for activity recognition. Moreover, video is represented as a spatial-temporal volume by stacking frames over a given sequence and action recognition is performed based on either spatial or temporal features or both. These features may be texture, color, posture, histograms of optical flow or histograms of oriented gradients. Many authors use spatio-temporal templates and 3D CNNs to learn features from spatio-temporal volumes and capture dynamics. For instance, [12] uses spatio-temporal templates, while [13] uses 3D CNNs for activity recognition based on video volume representations. Furthermore, a multi-view system to understand, in real time, the interactions between the ball and the players based on their respective 3D trajectories was presented by [14].

RGB images for HAR: Many works in the literature rely on the RGB videos to construct robust HAR systems. Several holistic action representations based on RGB images and powerful features are adopted by many authors to describe actions robustly. For instance, [15] combined shape features and optical flow calculated among RGB frames to detect change in motion. These approaches allow us to track the person in the scene and classify the ongoing activity. In addition, [16] proposed a long-term motion descriptor called sequential Deep Trajectory Descriptor (sDTD) which feeds a CNN-RNN network with dense trajectories to learn an effective representation for long-term motion. On the other hand, silhouettes were exploited by [17], [18] after extracting them using background subtraction. Automatic feature extraction from RGB images through deep learning was also suggested in many works. This is justified by the strong ability of CNN networks in dealing with RGB images. The authors of [19] extracted features from video sequences using pre-trained model, then an architecture of Deep Bidirectional LSTM for learning sequence information in the features of video frames was proposed. Similarly, [20] presented a Deep Long-term Recurrent Convolutional Network that deals with spatial and temporal features at once. Deep learning methods represent low level to high level features with multiple layers of the neural networks. Activity classification is therefore performed either using a popular machine learning classifier or using deep learning networks. The above can also be combined as in [21] where a method that integrates graphical models and deep neural networks into a joint framework was presented.

Depth images for HAR: Due to the emergence of depth cameras that can overcome some inherent privacy and limitations issues related to traditional cameras, depth image-based representations for HAR have evolved significantly in recent years. 3D structure of the body can be generated by integrating depth sensors and body tracking, enabling straightforward action recognition. Many limitations related to lighting variations, perspective change, variation in appearance, complex backgrounds and scale variation can be resolved using depth-based HAR methods. Furthermore, the extraction of information content generated by depth images is often straightforward. This includes the body shape information, the silhouette data and the whole image region within camera view. Several depth-based HAR approaches have been suggested in the literature. For instance, a local spatio-temporal descriptor for action recognition from depth video sequences is developed by [22]. It takes into account shape discrimination, motion change and action speed variations to distinguish between different actions. Similarly, [3] proposes a human pose representation model based on deep convolutional neural network CNN. The proposed model maps human poses acquired from several views of depth videos to a view-invariant high-level space. Although, depth image-based HAR has drawn growing interest by providing very promising results, depth-based methods still face difficult issues such as occlusion.

Skeleton-based action recognition: With the quick advent of depth sensors and algorithms of real-time skeleton estimation, many authors have demonstrated that skeleton features are more robust than RGB and depth features. This allows them to take advantage of this type of features for HAR. 3D locations and angles of joints are common features that can be used to build robust skeleton representations for HAR. Various methods based on skeleton analysis and representations of the set of joints for action recognition have been proposed in the literature. For instance, [23] proposes an end-to-end fully connected deep LSTM network for skeleton-based action recognition that relies on co-occurrence features of the skeleton joints. The co-occurrences of the joints are proven to be able to characterize accurately human actions. Similarly, [24] proposes a novel adaptive recurrent neural network (RNN) with LSTM architecture to automatically regulate observation viewpoints during the occurrence of an action. The 3D skeleton newly represented in a new coordinate system is used for accurate action recognition. Moreover, an end-to-end spatial and temporal attention model based on Recurrent Neural Networks (RNNs) and LSTM for HAR from discriminative joints of the skeleton was proposed by [25]. Likewise, authors of [26] presented an enhanced skeleton visualization method for view invariant HAR, where a sequence-based view invariant transform for the skeleton joints is performed and then the newly generated skeleton is visualized as series of enhanced RGB-images that encode spatial and temporal information related to skeleton joints. Finally, features were extracted using a CNN-based model allowing action classification.

Rank pooling videos: Rank pooling in videos was introduced by [5], [11]. It allows to capture the video-wide temporal evolution while preserving actions execution temporal ordering. The authors of [5] proposed to train a linear ranking machine on the video frames and to use its parameters as a new video representation. When trained on different samples of the same action, the authors demonstrated that the ranking machines would have similar ranking functions. [27] extended the rank pooling to encode video sequences at multiple levels recursively where the output of each encoding level is itself the input of the next encoding level in order to capture higher-order dynamics. Similarly, [4] introduced a CNN-based approximated rank pooling approach that allows us to learn dynamic image networks for action recognition.

Multi-modal methods for HAR: Multi-modal data fusion in HAR consists in combining many sensor modalities data in order to increase the robustness and the reliability of the recognition system while reducing single sensor effects such as noise [1]. To achieve this, it is essential to provide a complementary highly discriminative fusion of these modalities. In the literature, many fusion strategies have been employed to efficiently select meaningful information among different combined modalities [1]. Feature-level fusion, through increasing feature-space, projecting on some external frame, or using correlation-like analysis, is one of the best strategies for fusing heterogeneous modalities [1]. For instance, depth data, skeleton information and RGB images provide important complementary features. Indeed, depth data is more robust to illumination changes and scale variation but sensitive to occlusion; while skeleton information is more robust to occlusion effects and RGB image provides fine-grained image segmentation. Many vision-based HAR approaches combine two of these three modalities to improve the recognition accuracy but very few works focused on the combination of all of the three modalities. On the other hand, a robust HAR approach combining skeleton and RGB data streams was presented by [2], although, the authors used decision-level fusion instead of feature-level fusion. To effectively fuse features extracted from several modalities, some works use Canonical Correlation Analysis which allows us to learn from heterogeneous data and afford high linear correlation outputs. For instance, [28] developed a deep canonical correlated analysis to fuse accelerometer and gyroscope data for human activity recognition.

Building on [28], our work combines features extracted from dynamic images and skeleton images using canonical correlation analysis. Dynamic images were calculated from RGB and depth sensors separately, while skeleton images refer to RGB image representation that we derive from skeleton joints information.

III. Proposed method

This paper advocates a feature-level fusion framework for multi-modal human activity recognition using Canonical Correlation Analysis of the three modalities: RGB, depth information and skeleton. For this purpose, we created a set of dynamic images from RGB and depth videos separately. Skeleton visual images were inferred from skeleton joint information. Dynamic images are extracted from the video sequence in a way to capture spatial and temporal information among all frames. Especially, a dynamic image allows us to encode the video sequence robustly and describe the ongoing action in the video. For this purpose, we use an approximate rank pooling method as suggested in [4] to construct dynamic images. Moreover, skeleton images are constructed using 3D locations of the skeleton joints. Once the three sets of images were created, we use transfer learning from a pre-traind model to extract features from these images. Afterwards, we perform a fusion of these features using a feature-level fusion strategy based on Canonical Correlation Analysis. Finally, we train a bi-directional LSTM network to recognize and classify activities in the input video sequences. In summary, our methodology is composed of four steps that we explain in detail in the following subsections.

A. Dynamic image construction for RGB and depth images

Dynamic image (DI) consists of a single image representation of a video sequence, capturing the temporal evolution of ongoing action. DIs can provide simple, powerful and efficient representations that can be used for action recognition. The concept of "Dynamic images" has been presented in [5], [4], [29]. For that, the authors of [4] suggest to use an approximated rank pooling method to construct DIs. They observed that: (1) DIs focus on the motion instead of background pixels which are averaged away, (2) DIs behave differently for actions of different speeds and (3) DIs are reminiscent of some other imaging effects such as blur and panning. Similarly to [4], we use DIs to encode each video into one single image. The latter can provide us useful information on the ongoing action in the scene. We use the proposed approximated rank pooling method to calculate DIs for both RGB and depth video sequences separately. From the above and in the same spirit as [5], the video sequence is presented as a ranking function of its frames as follows:

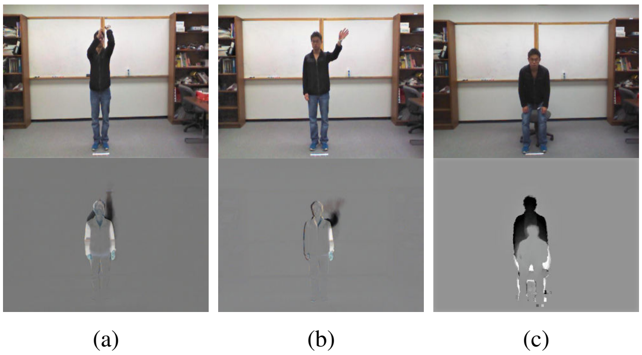

Fig. 2: Samples of RGB video frames from the UTD-MHAD dataset [7] in the first row and their corresponding dynamic RGB images in the second row. Column (a) corresponds to a basketball shoot while the subject is waving and sitting in columns (b) and (c) respectively.

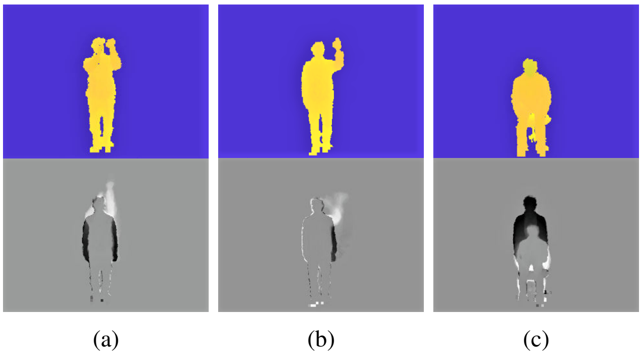

Fig. 3: Samples of Depth video frames from the UTD-MHAD dataset [7] in the first row and their corresponding dynamic Depth images in the second row. Column (a) corresponds to a basketball shoot. In column (b) the subject is waving, and he is sitting in column (c).

We refer to the feature vector extracted from frame I_t by $\psi(I_t)$. So $V_t = \frac{1}{t}\sum_{T=1}^{t}\psi(I_t)$ is the average of the features extracted from frames $\{I_1,I_2 ...,I_t\}$ over time t. The ranking function assigns a score $S(t|d) =< d, V_t >$ to each time increment t, where $d \in \mathbb{R}$ is a vector of parameters.

To reflect the rank of the frames in the video, d is learned as a convex optimization problem using the RankSVM formulation since later times are associated with larger scores, i.e $\forall\{q,t\}$ s.t $q > t \Rightarrow S(q|d) > S(t|d)$.

d^* is the optimizing function to the objective function given in "(2)" and T is the number of frames. We can see from "(1)" that $\rho(I_1, ...I_T; \psi)$ maps T video frames to a single vector d*. This operation of construction of d^* from T frames is called Rank Pooling.

$$d^* = \rho(I_1, ...I_T; \psi) = \underset{d}{argmin} E(d) \tag{1}$$

$$E(d)=\frac{\lambda}{2}||d||^2+\frac{2}{T(T-1)}\sum_{q>t}max\{0,1-S(q|d)+S(t|d)\} \tag{2}$$

This objective function is composed of two terms: the first one corresponds to the usual quadratic regularizer of SVM while the second term serves to count how many pairs q $>$ t are incorrectly ranked by the scoring function. In other words, it counts the number of pairs for which their associated scores are not separated by at least a unit margin.

The vector d^* contains enough information to rank all frames of the video. Similarly to [4], we apply rank pooling directly to RGB frame and depth image pixels. For that, $\psi(I_t)$ performs a component-wise non-linearity such as the square root function. As observed, d^* has the same number of elements as video frames and can therefore be used to represent the video.

Solving "(2)" may be computational expensive. For this purpose, we use approximated rank pooling which gives good results in practice to smooth the computation and make it faster. More specifically, the idea behind the approximated rank pooling is to consider the first step in a gradient-based optimization of "(2)". We then start with $d = \overrightarrow{0}$ and get a first approximated solution by gradient descent:

$d^* = \overrightarrow{0} - \eta \bigtriangledown E(d)|_{d=\overrightarrow{0}} \propto - \bigtriangledown E(d)|_{d=\overrightarrow{0}}$ for $\eta > 0$ where :

$$\bigtriangledown E(\overrightarrow{0}) \propto \sum_{q>t}\bigtriangledown max\{0,1-S(q|d)+S(t|d)\}|_{d=\overrightarrow{0}}$$
$$=\sum_{q>t}\bigtriangledown < d, V_t - V_q > = \sum_{q>t} < V_t - V_q > \tag{3}$$

So, we can extend d^* as follows, where β_t are scalar coefficients.

$$d^* \propto \sum_{q>t} < V_q - V_t > = \sum_{t=1}^{T}\beta_t V_t \tag{4}$$

By expanding the sum $\sum_{q>t} V_q - V_t$, each *Vt* appears (t-1) times with positive sign and (T-t) times with negative sign. Hence, we can deduce that $\beta_t = (t-1)-(T-t) = 2t-T-1$.

Since we already have $V_t = \frac{1}{t}\sum_{T=1}^{t}\psi(I_t)$, d^* can be written as a linear combination of the feature vector $\psi(I_t)$: $d^* \propto \sum_{t=1}^{T}\beta_t V_t = \sum_{t=1}^{T}\alpha_t\psi(I_t)$.

The approximated rank pooling is given such that the operator d^* is reduced to respect "(5)". So, the calculation of DIs consists in accumulating the video frames after being multiplied by α_t while $\alpha_t = 2(T-t+1)-(T+1)(H_T-H_{t-1})$ and $H_t = \sum_{i=1}^{t}\frac{1}{t}$ is the t-th Harmonic number and $H_0 = 0$.

$$\hat{\rho}(I_1,...I_T;\psi)=\sum_{t=1}^{T}\alpha_t\psi(I_t) \tag{5}$$

The vectors d^*_{RGB} and d^*_{Depth} obtained from rank pooling the RGB and depth videos respectively, comprise our DIs which we call DI_{rgb} and DI_{depth}. "Fig. 2" and "Fig. 3" illustrate some examples of RGB, depth images (from UTD-MHAD dataset) and their correponding dynamic RGB and dynamic Depth images, respectively. Columns of both images (in their order of appearance) correspond to basketball shoot, wave and stand to sit activities. We can see from these figures, that dynamic images were able to accurately summarize the execution of each of the activities as still images.

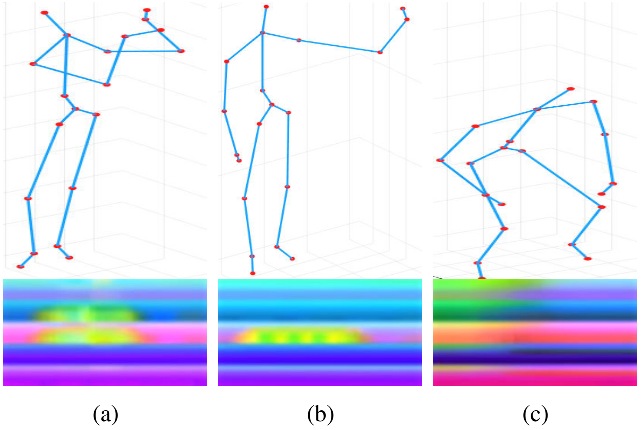
(a) (b) (c)

Fig. 4: Examples of skeleton representation from the UTD-MHAD dataset [7] in the first row and their corresponding skeleton visual images in the second row. Columns (a), (b) and (c) correspond to a basketball shoot, wave, stand to sit activities respectively.

B. Skeleton images from skeleton joints

Human activity recognition from skeleton information have been facing many challenges among which is: how to effectively represent spatio-temporal skeleton sequences? Moreover, retrieving features from RGB images using pre-trained models is giving very promising results in many tasks as well as for human activities recognition. Therefore, to take advantage of these models and HAR from skeleton data, we create images from skeleton sequences, then we extract discriminative features from these images using a pre-trained model. For that, and for each video sequence, we normalize the coordinates of the skeleton joints *(x,y,z)* and use them to create an RGB image which we call I_{skel}. Skeleton image allows us to track changes of each skeleton joint over time and, hence, describe the corresponding activity. "Fig. 4" illustrates some examples of skeleton representations from the UTD-MHAD dataset and their corresponding skeleton images. Columns correspond to basketball shoot, wave and stand to sit activities respectively.

C. Features Extraction using pre-trained models

Due to the large amounts of data needed for training an LSTM network, we extract features from our image sets $\{$DIs$_{rgb}$ and DIs$_{depth}\}$ using the Resnet50 model pretrained on the large Imagenet dataset. Used widely as a backbone for many computer vision tasks, it has been integrated in many HAR approaches as well. It allows us to explore multiple levels of deep features by dint of its stack of layers that is composed of more than 150 layers. In addition, we use Alexnet to extract features from Is$_{skel}$ images set. This feature extraction step is important because it provides us a strong initialisation to our feature fusion strategy compared to a straightforward use of these images. Feature vectors calculated in this step are then fused using Canonical Correlation Analysis which allows us to select meaningful features.

D. Feature Fusion and activity classification

To obtain more discriminative feature vectors from our created representations, we apply a feature fusion on the extracted features from our three sets of images: the dynamic RGB images (DIs_{rgb}), the dynamic depth images (DIs_{depth}) and the skeleton images (Is_{skel}). Our feature fusion method consists of combining feature vectors of the three modalities into one single feature vector. The resulting feature vector is supposed to be more meaningful than each single aforementioned modality related feature vector. For that, and similarly to [6], we use Canonical Correlation Analysis (CCA) which has been widely used for feature fusion.

Let our three feature vectors be $V_x \in \mathrm{I\!R}^{\mathrm{pxn}}$, $V_y \in \mathrm{I\!R}^{\mathrm{qxn}}$ and $V_z \in \mathrm{I\!R}^{\mathrm{rxn}}$ extracted from dynamic RGB images, dynamic depth images and skeleton images respectively. Each of these vectors contain *n* samples. To get a representative feature vector that fuses the three vectors, we apply CCA twice. We apply CCA firstly on two vectors, for example V_x and V_y, we obtain F_1. Then, we apply CCA again on F_1 and V_z.

For each two vectors *X* and *Y*, we calculate the within-sets covariance matrices and the between-set covariance matrix that we call: $S_{xx} \in \mathrm{I\!R}^{\mathrm{pxp}}$, $S_{yy} \in \mathrm{I\!R}^{\mathrm{qxq}}$ and $S_{xy} \in \mathrm{I\!R}^{\mathrm{pxq}}$ respectively and S_{yx} is the transpose of S_{xy}: S_{xy}^T. Next, we create the covariance matrix $S \in \mathrm{I\!R}^{(\mathrm{p+q})\mathrm{x}(\mathrm{p+q})}$ as illustrated below.

$$S = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} = \begin{pmatrix} cov(X) & cov(X,Y) \\ cov(Y,X) & cov(Y) \end{pmatrix} \tag{6}$$

It is observed that understanding the correlations between *X* and *Y* using the covariance matrix *S* is difficult. Therefore, CCA is used to maximize the pairwise correlations given in “(7)” across the two data sets using Lagrange multipliers. Solution to the objective function is given by the optimizer linear combinations X^* and Y^*.

Canonical variates X^* and Y^* are defined as $X^* = W_x^T X$, $Y^* = W_y^T Y$ and $var(X^*) = var(Y^*) = 1$ where $cov(X^*, Y^*)$, $var(X^*)$ and $var(Y^*)$ are calculated using the set of “(8)”.

$$corr(X^*, Y^*) = \frac{cov(X^*, Y^*)}{var(X^*)var(Y^*)} \tag{7}$$

$$\begin{Bmatrix} cov(X^*, Y^*) = W_x^T S_{xy} W_y \\ var(X^*) = W_x^T S_{xx} W_x \\ var(Y^*) = W_y^T S_{yy} W_y \end{Bmatrix} \tag{8}$$

To obtain the transformation matrices W_x and W_y, one should solve the eigenvalue “(9)”. $\widehat{W}_x$ and $\widehat{W}_y$ are the eigenvectors and Λ^2 is the diagonal matrix of eigenvalues or squares of the canonical correlations. For each equation, the number of non-zero eigenvalues (i.e. $\lambda_1 >= \lambda_2 ... >= \lambda_d$) that are sorted in decreasing order is $d = rank(S_{xy} <= min(n, p, q))$. W_x and W_y consist of sorted eigenvectors corresponding to the non-zero eigenvalues.

$$\begin{Bmatrix} S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} \widehat{W}_x = \Lambda^2 \widehat{W}_x, \\ S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} \widehat{W}_y = \Lambda^2 \widehat{W}_y \end{Bmatrix} \tag{9}$$

Hence, the covariance matrix *S* defined above will be of the following form. Let I_d be the identity matrix and $diag(\lambda_1, ..., \lambda_d)$ be the diagonal matrix of the associated eigenvalues.

$$S = \begin{pmatrix} I_d & diag(\lambda_1, ..., \lambda_d) \\ diag(\lambda_1, ..., \lambda_d) & I_d \end{pmatrix} \tag{10}$$

We can observe that X^* and Y^* have non-zero correlation only on their corresponding indices and are therefore uncorrelated within each data set. Finally, we perform feature-level fusion by concatenating the transformed feature vectors. The resulting feature vector F_1 is used to perform another time the feature-level fusion by concatenating the transformed feature vectors F_1^* and Z^* (*Z* corresponds to the third modality feature vector).

$$F_1 = \begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} W_x^T X \\ W_y^T Y \end{pmatrix} = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \tag{11}$$

Once our fused feature vectors are calculated, we perform activity recognition using a bi-directional LSTM network. This allows us to comprehend temporal dynamics encoded by the feature extractor (Resnet50 model for RGB and depth images and Alexnet for skeleton images) into feature maps. These feature maps are fused using CCA and fed to the classifier.

IV. Experimental results

We evaluate our approach on the publicly available datasets UTD-MHAD [7] and NTU RGB+D [8]. In the subsequent sections, we present a brief description of these datasets followed by our experimental results. We compare the recognition performance of each individual sensor modality to the performance of combining each pair of modalities and finally to the performance of fusing the three modalities. We display our performance results in Table I and Table II for UTD-MHAD and NTU RGB+D datasets.

In our LSTM model, we incorporate a bi-directional long short term memory layer. We use 600 hidden units and a feature sequence of 1000 to 1092 length. A mini-batch size of 32 samples is employed to train images of the subset and we calculate the accuracy. For the UTD-MHAD dataset, we create the training and the testing subsets using the same protocol as [7]. Data from the subject numbers 1, 3, 5, 7 were used for training, while data for the subject numbers 2, 4, 6, 8 were used for testing. We report the classification accuracy on the NTU RGB+D dataset by following the action classification evaluation protocol presented in [8]: cross-view evaluation where videos from cameras 2 and 3 are used for training while videos from camera 1 are used for testing.

A. Datasets

a) UTD-MHAD: is a multi-modal dataset [7], composed of four data modalities: RGB videos, depth videos, skeleton joint positions and inertial sensor signals. The dataset includes 861 video sequences and was recorded using a Microsoft Kinect sensor and a wearable inertial sensor in an indoor environment. It consists of 27 different actions performed by 8 subjects. Each of them repeats the same action 4 times.

b) NTU RGB+D: is a large-scale dataset for multi-modal human action recognition. It includes 56880 videos of 60 action classes of 40 subjects recorded in highly variant camera settings, where each action is performed twice. Three Microsoft Kinect v2 sensors were used to collect four data modalities.

B. Results and analysis

First, We calculate the performance accuracy of each single modality. In other words, we compare activity classification from straightforward images towards newly created images (dynamic RGB, depth images and skeleton images). For both situations, we calculate the accuracy of applying LSTM on the extracted features using a pretrained model. Resnet50 and Alexnet are used as feature extractors.

As can be seen from Table I which illustrates the results of uni-modal activity recognition for UTD-MHAD and NTU RGB+D datasets, the accuracy was improved when using our created images. For the UTD-MHAD dataset, skeleton features perform the best accuracy value for both configurations with 74.52% for skeleton joints sequences and 87.43% for skeleton images using Alexnet as feature extractor. Similarly, for the NTU RGB+D dataset, the best accuracy of 49.91% was obtained for skeleton joints sequences while dynamic depth images outperform the dynamic RGB and skeleton images with a value of 51.66%.

TABLE I: Accuracy (%) of activity classification with LSTM of uni-modal features and features extracted (using pre-trained models) from our newly created image representations on the UTD-MHAD and NTU RGB+D datasets.

Uni-modal feature	**UTD-MHAD**	**NTU RGB+D**
RGB	51.35	39.85
Depth	37.45	45.90
Skeletal data	74.52	49.91
Dynamic RGB	72.28	41.53
Dynamic Depth	71.91	51.66
Skeleton images	87.43	50.81

Furthermore, we calculate the recognition accuracy for each pairwise fusion and for the three features fusion. We compare in Table II, the results obtained using a feature-level fusion: the Canonical Correlation Analysis on each set of features. We can see from these tables that, by combining the features from each two sets of images, the recognition accuracy was improved over that using a single modality alone for both datasets. The best results were obtained by fusing dynamic depth and skeleton images as they present complementary temporal features. We achieve for that an accuracy of 97.95% for the UTD-MHAD dataset and 70.85% for the NTU RGB+D dataset.

Fusing the three modalities has as well improved the recognition accuracy over that using single modalities or pairwise modalities. The order of fusing features was also investigated and the results demonstrate that when changing this order, the accuracy is also improved. We obtain an accuracy of 98.88% for fusing RGB and depth dynamic images and then fusing the resulting vector with skeleton images feature vector for the UTD-MHAD dataset and an accuracy of 75.50% for the NTU RGB+D dataset for the same configuration.

Table III presents a comparison of the results of our method to the state-of-the-art on the publicly available UTD-MHAD dataset. We can see that our method outperforms all the previous methods of feature fusion on the UDT-MHAD dataset. Again, Table IV illustrates a comparison of our results with the state-of-the-art results on the NTU RGB+D dataset. Our results are comparable to some of the existing methods such as [30] and [8]. However, we still can improve the results by enhancing the Canonical Correlation Analysis fusion strategy.

TABLE II: Accuracy (%) of activity classification using fusion of multi-modal features extracted (using pre-trained models) from our newly created image representations on the UTD-MHAD dataset and NTU RGB+D dataset respectively (DI refers to dynamic images).

Pairwise Fusion	**UTD-MHAD**	**NTU RGB+D**
DI RGB + DI Depth	85.39	60.42
DI RGB + Skeleton images	93.26	68.62
DI Depth + Skeleton images	**97.95**	**70.85**
By three Fusion		
(DI RGB + DI Depth) + Skeleton images	**98.88**	**75.50**
(DI RGB + Skeleton images) + DI Depth	92.13	73.72
(DI Depth + Skeleton images) + DI RGB	93.26	72.64

TABLE III: Comparison of the proposed method with previous methods on UTD-MHAD Dataset.

Method	Accuracy %
Decision Fusion Using LOGP [31]	88.40
Depth + inertial data fusion + CRC classifier [7]	79.10
5-CNN fusion of skeleton images [9]	95.38
fusion with CCA and KELM [10]	97.91
DI RGB + DI Depth + Skeleton images + LSTM (Ours)	**98.88**

TABLE IV: Comparison of the proposed method with previous methods on NTU RGB+D Dataset.

Method	Accuracy %
Deep RNN [8]	64.09%
Deep LSTM [8]	67.29%
Joint trajectory maps + CNN [30]	75.20%
Part-aware LSTM [8]	70.20%
DI RGB + DI Depth + Skeleton images + LSTM (Ours)	**75.50%**

V. CONCLUSION

We present in this paper a vision-based multi-modality fusion approach for human activity recognition. RGB images, depth images and skeleton joint data are used to construct RGB dynamic images, depth dynamic images and skeleton images, respectively. These constructed visual images are then employed to generate features using pre-trained models that allow us to retrieve meaningful features from the image sets. Afterward, for each video sequence, a feature fusion strategy based on the Canonical Correlation Analysis is carried out to select highly discriminative features from our three feature vectors. The resulting feature fusion vectors are then fed to a bi-directional LSTM network in order to recognize and classify activities. We evaluate our approach on the publicly available UTD-MHAD and NTU RGB+D datasets and record recognition accuracy for each single modality, fusion of each pair of modalities and fusion of three modalities. Our experiments show that the results of our proposed approach can achieve high recognition accuracy and outperform the state-of-the-art results for both datasets. In the future, we can explore other fusion schemes and integrate some data augmentation methods to improve the performance of our proposal. Besides, we believe there is also a room for further improvement on the recognition accuracy achieved by NTU RGB+D dataset throughout a more fine-gained optimization of the parameters of the underlined LSTM model.

Acknowledgment

This work is partly supported by the Algerian Residential Training Program Abroad Outstanding National Program (PNE) that supported the first author stay at University of Oulu and European YougRes project (Ref. 823701), which are gratefully acknowledged.

References

[1] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-Garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Information Fusion*, vol. 46, pp. 147–170, 2019.

[2] A. Franco, A. Magnani, and D. Maio, "A multimodal approach for human activity recognition based on skeleton and rgb data," *Pattern Recognition Letters*, 2020.

[3] H. Rahmani and A. Mian, "3d action recognition from novel viewpoints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1506–1515.

[4] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2799–2813, 2017.

[5] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5378–5387.

[6] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Fully automatic face normalization and single sample face recognition in unconstrained environments," *Expert Systems with Applications*, vol. 47, pp. 23–34, 2016.

[7] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International conference on image processing (ICIP)*. IEEE, 2015, pp. 168–172.

[8] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[9] P. Khaire, J. Imran, and P. Kumar, "Human activity recognition by fusion of rgb, depth, and skeletal data," in *Proceedings of 2nd International Conference on Computer Vision & Image Processing*. Springer, 2018, pp. 409–421.

[10] J. Imran and B. Raman, "Evaluating fusion of rgb-d and inertial sensors for multimodal human action recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 1, pp. 189–208, 2020.

[11] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 773–787, 2016.

[12] H. Sial, M. Yousaf, and F. Hussain, "Spatio-temporal rgbd cuboids feature for human activity recognition," *The Nucleus*, vol. 55, no. 3, pp. 139–149, 2018.

[13] R. Chopra *et al.*, "Activity recognition based on 3d cnn-lstm-assisted approach," *Journal of the Gujarat Research Society*, vol. 21, no. 6, pp. 454–466, 2019.

[14] M. Leo, N. Mosca, P. Spagnolo, P. L. Mazzeo, T. D'Orazio, and A. Distante, "Real-time multiview analysis of soccer matches for understanding interactions between ball and players," in *Proceedings of the 2008 international conference on Content-based image and video retrieval*, 2008, pp. 525–534.

[15] M. H. Kolekar and D. P. Dash, "Hidden markov model based human activity recognition using shape and optical flow based features," in *2016 IEEE Region 10 Conference (TENCON)*. IEEE, 2016, pp. 393–397.

[16] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream cnn," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510–1520, 2017.

[17] K. Gościewska and D. Frejlichowski, "Silhouette-based action recognition using simple shape descriptors," in *International Conference on Computer Vision and Graphics*. Springer, 2018, pp. 413–424.

[18] L. González, S. A. Velastin, and G. Acuna, "Silhouette-based human action recognition with a multi-class support vector machine," 2018.

[19] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017.

[20] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[21] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4772–4781.

[22] C. Chen, M. Liu, H. Liu, B. Zhang, J. Han, and N. Kehtarnavaz, "Multi-temporal depth motion maps-based local binary patterns for 3-d human action recognition," *IEEE Access*, vol. 5, pp. 22 590–22 604, 2017.

[23] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[24] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2117–2126.

[25] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[26] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[27] B. Fernando, P. Anderson, M. Hutter, and S. Gould, "Discriminative hierarchical rank pooling for activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1924–1932.

[28] G. Chetty and M. White, "Body sensor networks for human activity recognition," in *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2016, pp. 660–665.

[29] S. Mukherjee, L. Anvitha, and T. M. Lahari, "Human activity recognition in rgb-d videos by dynamic images," *arXiv preprint arXiv:1807.02947*, 2018.

[30] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 102–106.

[31] 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)*, vol. 6, no. 4, pp. 23–39, 2015.