The Visual Computer (2021) 37:1263–1276
https://doi.org/10.1007/s00371-020-01864-y

ORIGINAL ARTICLE

Real-time multimodal ADL recognition using convolution neural networks

Danushka Madhuranga[1] · **Rivindu Madushan**[1] · **Chathuranga Siriwardane**[1] · **Kutila Gunasekera**[1]

Published online: 12 June 2020

Abstract
Activities of daily living (ADLs) are the activities which humans perform every day of their lives. Walking, sleeping, eating, drinking and sleeping are examples for ADLs. Compared to RGB videos, depth video-based activity recognition is less intrusive and eliminates many privacy concerns, which are crucial for applications such as life-logging and ambient assisted living systems. Existing methods rely on handcrafted features for depth video classification and ignore the importance of audio stream. In this paper, we propose an ADL recognition system that relies on both audio and depth modalities. We propose to adopt popular convolutional neural network (CNN) architectures used for RGB video analysis to classify depth videos. The adaption poses two challenges: (1) depth data are much nosier and (2) our depth dataset is much smaller compared RGB video datasets. To tackle those challenges, we extract silhouettes from depth data prior to model training and alter deep networks to be shallower. As per our knowledge, we used CNN to segment silhouettes from depth images and fused depth data with audio data to recognize ADLs for the first time. We further extended the proposed techniques to build a real-time ADL recognition system.

Keywords Activity recognition · Depth images · Video classification · Data fusion · Silhouette extraction

1 Introduction

Recognizing activities of daily living (ADLs) is a crucial function in ambient assisted living (AAL) systems [20] and elderly-care systems [5] as it helps to monitor activities and detect accidents such as falling. Recently, ADL recognition using depth image sequences has become popular [4, 27, 47] because it helps to overcome issues in RGB videos such as being affected by ambient lights limiting performance at night and exposing sensitive information that could lead to privacy violations. In depth images, a pixel value represents the distance between the corresponding physical point and the camera. Depth images do not contain fined-grain information that reveals identity of people directly. While there are less intrusive alternatives for action recognition such as WiFi signal analysis [1, 42] and pressure sensors [6], they are limited by the number of different ADL-related activities that can be recognized accurately. Humans get assistance from modalities other than vision (e.g., hearing and smell) to identify events around them. Motivated by this, we combine audio with depth data to recognize ADLs.

Most of the works on depth image-based activity recognition are based on handcrafted features such as super normal vectors (SNV) [47] and spatiotemporal interest points (STIP) [46]. Designing handcrafted features is time-consuming and problem specific, and the features tend to lose information-related to subtle movements which are crucial in recognizing activities of daily living. In RGB video-based activity recognition, convolutional neural networks (CNNs) are used as an accurate alternative to handcrafted features [11, 15, 35, 44]. CNNs can be trained with raw images, and the networks extract important features on their own. This also allows to transfer learned knowledge across datasets as feature extraction, and selection is dynamic in neural networks. Therefore, we adopt deep CNNs used for RGB video modeling

✉ Danushka Madhuranga
danushka.15@cse.mrt.ac.lk

Rivindu Madushan
rivindum.15@cse.mrt.ac.lk

Chathuranga Siriwardane
chathuranga.15@cse.mrt.ac.lk

Kutila Gunasekera
kutila@cse.mrt.ac.lk

[1] Department of Computer Science and Engineering, University of Moratuwa, Katubedda, Sri Lanka

to recognize activities in depth videos. The adoption is not trivial. There are two major challenges: 1) Depth images are noisier compared to RGB images and 2) there are no large datasets available for depth videos. We propose two solutions to overcome these challenges: a) we train the network on extracted human silhouettes instead of raw depth images and b) change network architecture to be shallower. Removing the background helps the network to focus on human body parts when learning features.

Image processing techniques such as thresholding [4] and connected component labeling [17] have been the basis in existing works for silhouette extraction from depth images. But fully convolutional neural networks (FCN) have shown to be effective in segmenting RGB and medical images [31, 39]. Hence, we apply FCN for silhouette extraction.

CNNs are very effective at extracting spatial features [18, 36]. But videos contain temporal information as well. There are three popular approaches in applying CNNs to perform video analytics: 1) stack images into a single image and perform 2D convolution [35], 2) perform 3D convolution over the image sequence [15, 38] and 3) combine recurrent neural networks (RNN) with 2D convolution to extract spatiotemporal information [11, 13, 15, 44, 45]. The latter approaches have shown to be more effective compared to the image stacking method as 2D CNNs struggle to preserve temporal information throughout the network. We employs both of the last two approaches to classify depth videos. In 3D CNN networks, the number of frames in each video clip must be the same. Long short-term memory (LSTM) networks have become the most popular variant of RNNs in modeling videos [2, 13, 44]. CNN–LSTM combination provides more flexibility compared to 3D CNNs.

Fusing multiple streams, such as audio and video, for activity recognition can be performed at results level [35, 43, 44] or at feature level [8, 14, 28, 45]. Features contain more information than results. Hence, feature-level fusion facilitates better understanding about the correlation of different streams. Furthermore, feature-level fusion can be performed using either a separate network/framework [8, 28, 45] or by creating a combined network that can be trained in an end-to-end manner [14]. We evaluated both feature-level and result-level fusion in combining audio and depth streams.

Contribution of this paper can be summarized as follows:

- We propose to use convolutional neural networks to segment silhouettes from the depth images.
- We propose to adopt deep learning networks used for RGB video analysis to recognize activities from depth videos.
- We fuse depth and audio data through an end-to-end trainable neural network to recognize ADLs.
- We develop a real-time activity recognition system based on the above trained network.

The rest of this paper is structured as follows. Related work on ADL recognition, silhouette extraction and fusing multiple data streams is described in Sect. 2. Section 3 explains the multimodal dataset we created for evaluation. Our solution approach is described in detail in Sect. 4. Experiments, results and discussion are covered in Sect. 5 and the paper concluded in Sect. 6.

2 Related work

2.1 ADL and other activity recognition

There are many existing activity recognition systems focusing on ADLs and other activities such as sport-related activities [16, 44]. Focusing on the sensors and data used for recognition, there are works based on wearable sensors such as accelerometer and gyroscope [26, 30], WiFi reflecting signal data such as channel state information (CSI) [1, 42], pressure sensors [6], depth camera [4, 27, 41, 46–48], audio [24, 33, 34] and RGB cameras [11, 35]. Non-intrusiveness and preserving privacy are particularly important in ADL recognition as one of the main use cases of ADL recognition is elderly care. While WiFi [1, 42] and pressure sensor-based [6] recognition approaches are among the least intrusive approaches, they are limited by the number of different activities that can be recognized. Arshad et al. [1] and Wang et al. [42] recognize three and eight activities, respectively, whereas Yang et al. [47] identify 16 different activities using a depth camera. Wearables, on the other hand, can accurately recognize more activities but tend to be more intrusive. People may not prefer to wear such devices throughout the day. RGB cameras reveal sensitive privacy concerning information and suffer from sensitivity to illumination variances. Hence, they are not suitable for 24×7 operation. Audio-, depth- and RGB camera-based activity recognition is extensively covered in the following subsections.

2.2 Depth-based activity recognition

Most of the popular existing approaches for recognition of human activities and gestures using depth image sequences include two major steps: extracting features from raw depth sequences and classifying them using a machine learning algorithm. These approaches employ handcrafted features such as super normal vectors (SNVs)—a complex representation of the hypersurface normals that span across space and time [47], histograms of depth values [4], histograms of oriented 4D normals (HON4D) [27], variants of spatiotemporal interest points (STIP) such as depth STIP (DSTIP) [47], and histogram of oriented gradients (HOG) [48]. Support vector machine (SVM) is the most popular classification algorithm [4, 27, 46–48].

There have been few attempts to recognize activities from depth sequences using CNNs. For instance, in [41], authors represented the whole depth sequence in a single frame by merging depth motion maps in a weighted manner, converted the map into a color image using pseudo coloring and classified using a CNN. But again, the input for the machine learning model is a handcrafted feature. A key problem in these representations is the loss of information related to subtle movements which are vital in recognizing daily activities. In addition, selection of the handcrafted features tends to be task specific. Hence, the learning cannot be transferred across different situations.

2.3 RGB video-based activity recognition

Historically, neural networks were used mostly for modeling low-dimensional data because the number of parameters in typical feed forward neural networks increases with the dimensions of the input data. Thus, neural networks were not popular for analysis of images and videos. Instead, handcrafted features such as STIP [10] were commonly used for analysis. But with the introduction of CNNs, part of this problem was eliminated. Note that, since most of the CNN architectures for classification employs fully connected layers at the last stages, a set of parameters scale with the size of the input. One major advantage of CNNs is that either very minimal preprocessing or no preprocessing is required. Also the learned features are general allowing to transfer learned knowledge across datasets. Recently, following the success of different CNN architectures in the ImageNet competition [18, 36], CNNs are widely adopted for image analysis by the deep learning community.

Inspired by the ability of CNNs to extract features from images, researchers started to adopt CNNs to video analysis problems such as action and gesture recognition. ImageNet CNN architectures can effectively extract spatial features [18, 36], but action recognition requires capturing both spatial and temporal features from the image sequences. Different approaches have been proposed to extract temporal features. One approach is to stack optical flow information in each time step as a separate channel and use a 2D CNN for learning [35]. In this representation, each input is a single image, but the motion information is embedded in channels. When using 2D convolution, all the individual channels get summarized into a single channel after the very first convolution occurs. Thus, temporal information may not be preserved throughout the network. To preserve temporal information throughout the network, 3D convolution was used to extract spatiotemporal features from videos [15, 38]. One main problem in using 3D convolution is that all the video clips need to contain the same number of frames. In addition, the number of parameters in the model increases with the number of frames as well.

Recurrent neural networks (RNNs) have shown to be effective in modeling temporal information. Thus, another approach was proposed to rely on RNNs to extract temporal information from image sequences [2, 11, 13, 44, 45]. Unlike in 3D CNNs, RNNs allow the number of frames for each sequence to vary. In addition, the number of parameters do not scale with the length of the sequence. Long short-term memory (LSTM) has been the most popular RNN architecture as it addresses the issue of the vanishing gradient in typical RNNs. The LSTM models were trained on the sequences of extracted spatial features using both handcrafted approaches [13] and neural networks [2, 11, 44, 45]. In [11] and [44], 2D CNNs were used to extract spatial features, while in [45] and [2], 3D CNNs were used. The LSTM layer generates a prediction for each time step. Donahue et al. [11] averaged the prediction over time steps. Others [2, 44, 45] consider only the prediction in final time steps. The predictions of early time steps have higher chance of being incorrect because the model has seen only a part of the action. Thus, if the final prediction is calculated as the average prediction of each time step, the final accuracy can suffer from early miss-predictions.

2.4 Silhouette extraction

Silhouette extraction is a key step in classification of activities using depth image sequences. Unlike RGB images, depth images contain a large amount of noisy pixels. In addition, depth images lack important features such as facial and body details which aid in the identification of presence of humans in color images. Thus, in order to develop an algorithm which can work independent of the environment, segmentation and the extraction of the human silhouette from the background environment are important. Existing research relies on image processing techniques for human silhouette extraction from depth images [4, 17, 40].

In [4], Biswas and Basu isolate the human from the background by an auto thresholding technique. They assume the human is the closest object to the camera and threshold the depth image after the first spike in the depth histogram. Clearly, this approach fails when the human is not the closest object in the scene. In [40], the researchers present an approach capable of detecting multiple humans. This consists of four major steps: floor removal, region growing, object clustering for identification of similar regions and estimating the background using a mixture of Gaussian clusters. Finally, moving objects are recognized as human silhouettes. Although this is a better approach to identify moving humans, it fails when the human subject is not moving for a long time, and when the human is in tight contact with other objects, for example, sleeping on a bed or sitting on a couch, since the region growing and object clustering method will provide false positives in these cases. [17]

presents a connected component labeling and temporal depth difference-based human silhouette extraction methodology. This approach is also built upon motion of objects and is not a suitable method to detect humans having less motion between frames.

Recently, fully convolutional neural networks (FCNs) have been deployed for segmenting RGB and various medical images [3, 31] inspired by the success of ImageNet models [18, 36]. FCNs contain no fully connected layers. Both the input and the output of the FCNs are images. FCNs are trained by providing the original image as the inputs and the expected segmented image as the output. First FCNs downsample the image to understand the image content and later upsample the feature map to create the final segmented image. In UNet [31], upsampling is performed by a special type of convolution known as up-convolution, while SegNet [3] upsamples by reverse max-pooling. Both UNet and SegNet follows the 3×3 convolution and 2×2 max-pooling as in [36]. UNet has shown to be successful in segmenting medical images with fewer training data.

2.5 Audio-based activity recognition

Sound is one of the main sensory information humans use to perceive their surroundings. There are many researches which have studied on recognizing ADL sound events. Acoustic event recognition in bathroom [5], kitchen [37] and other indoor contexts [24] is some examples for ADL recognition systems using only acoustic data. Inbuilt mobile phone microphones [24] and dedicated acoustic sensors [37] have been used to obtain inputs to above ADL recognition systems.

Audio signals are usually preprocessed by constant length frame (window) blocking and constant time shift (hop) between adjacent frames [24]. Audio feature extraction is performed on the resulting set of audio frames.

Mel frequency cepstral coefficients (MFCC) features [23] and short-time Fourier transform (STFT) [5] features have been widely used in extracting features from acoustic data to use with machine learning algorithms. Chroma vectors, zero crossing rate (ZCR) and audio signal energy features can also be used to represent acoustic data. There are some researches in which the mentioned features have been used as a sequence of features [34]. Hidden Markov models (HMMs) [24] and CNN models [33] have been commonly used in sound classification.

2.6 Fusing multiple data streams

It has been shown that it is possible to gain performance improvements using multiple data streams in recognition tasks [8, 14, 28, 35, 43–45]. Multiple streams could be extracted from different sensors such as one from microphone and the other from camera [8, 14, 28, 43, 44] or from the same sensor. For example, in [35, 44], and [45], the optical flow extracted from the image sequence is considered a separate data stream. There are mainly two stages where fusion of data streams can be performed. They can be fused at feature level or at prediction level known as early fusion and late fusion, respectively.

In late fusion, there are independent models to predict results from each data stream whose results are fused by a late-fusion mechanism at the end. In [35], late fusion was implemented with simple averaging of two predictions and using SVM. Wu et al. [43] fused predictions using SVM and fuzzy integrals. In [44], Wu et el. concatenated predictions into a vector and multiplied by a weight matrix to get the final prediction. The weight matrix was trained using logistic regression with a sophisticated regularization framework. Unlike the prior approach [35], latter approaches [43, 44] consider the relationships between classes in the fusion in addition to considering the relationship between the predictions of the different models. Relationships between classes are important as they can prevent miss-classifications. For example, if the two highest probable classes (Walking and Eating) in a prediction have similar probabilities, we can select the class (Eating) which contain more similar movements to the third highest scoring class (Drinking water).

For early fusion, features are extracted as the initial step. Pieropan et al. in [28] and Cristani et el. in [8] fused audiovisual data using hidden Markov model and *K*-nearest neighbor, respectively. In both methods, handcrafted features were used to describe visual information. But Jiang et al. [45] extracted visual features using CNNs and employed a sophisticated regularized neural network to aggregate the extracted features. In [14], Hou et al. proposed an end-to-end trainable network for speech enhancement. The end-to-end trainable network combines feature extraction and fusion together. In this approach, audio data were represented by logarithmic power spectrum. Only the mouth part was extracted from the images and color channels were normalized. The network commences with two CNNs for processing audio and visual data separately and ends with a fully connected feed-forward network with inputs as the merged outputs of prior models. The combined learning in [14] enables even for weights in feature extracting layers to learn from other streams. More recently, attention mechanism has been used when combining different networks to fuse multimodal sensory data which allocates weights to networks according their importance [19].

3 Dataset

There are several benchmark datasets [25, 47] available for action recognition with depth data. But none of those has an audio stream to allow classifying activities using a

combination of audio and depth data. Therefore, we created a new comprehensive dataset to classify ADL using multiple data streams. This dataset has seven streams of data including an RGB video stream, depth video stream, skeleton stream, silhouette stream and three audio streams.

In the collection of data, we used a Microsoft Kinect v1 device [12] along with a couple of Movo USB microphones [22]. The RGB and depth video streams were captured using the Microsoft Kinect device. The RGB camera had a resolution of 640×480 pixels, frame rate of 30 fps and a field of view of 62×48.6 degrees. The Kinect depth camera had a resolution of 320×240 pixels a field of view of 58.5×46.6 degrees. The two USB microphones and the four-array microphone of the Kinect device were used to capture audio streams. These three microphones were placed at different distances from the action: closer, far away and in between.

For the collection of data, we developed a data collection platform in the Python programming language. To access the Kinect device drivers, we used the Microsoft Kinect for Windows SDK v1.8 [21] and the OpenNI/NITE library [29] was used to capture the RGB, depth, skeleton and silhouette data streams.

The dataset consists of 24 segmented activities of daily living performed by 17 actors. The 24 activities were designed based on the researches done by healthcare professionals. These activities are: making a phone call, clapping, drinking, eating, entering from door, exiting from door, falling, lying down, opening pill container, picking object, reading, sit still, sitting down, sleeping, standing up, sweeping, using laptop, using phone, wake up, walking, washing hand, watching TV, water pouring and writing. The activities, making a phone call, clapping, drinking, eating, falling, sit still, sitting down, standing up', using laptop, and using phone, were performed twice, while the activity walking was performed three times by each actor. Repeated performances are attributed to different postures or different environment settings in which activity was performed. All the other activities were performed only once. Each actor has about 36–38 samples.

The activities were performed in five different environment settings designed to model a bedroom, an office room and a corridor. Other than these segmented and labeled activities, we collected a separate set of activities continuously performed by six subjects. Each subject performs about 30–300 s covering all the above activities. The continuous samples can be used as a testing set for real-time activity recognition.

The dataset along with the data collection tool is now publicly available and can be found in https://github.com/RivinduM/Mora-ADL.

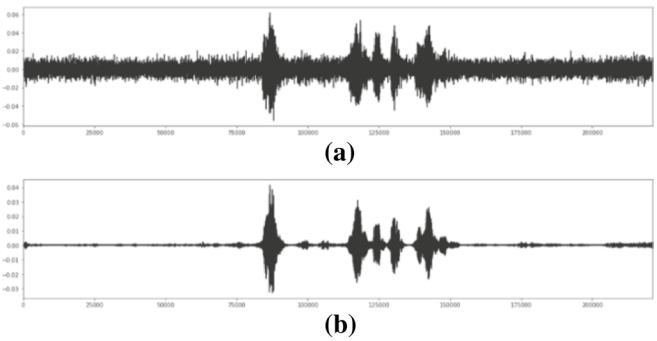


Fig. 1 Waveforms of **a** original audio clip and **b** noise-reduced audio clip

4 Methods

Audio and depth sequences are preprocessed separately. In depth preprocessing, we extract silhouettes from depth frames, while we perform noise suppression in audio preprocessing. Separate networks are developed to extract features from depth and audio streams. An end-to-end trainable network combining the audio and depth processing networks fuses audio and depth sequences for ADL recognition. In the final subsection, we describe the real-time ADL recognition system.

4.1 Audio preprocessing

Audio files obtained from the collected dataset contain background noise. Thus, noise reduction is done before extracting features from audio files. To reduce noise, first, a window of background noise was selected from the original audio clip. Using the noise threshold of the noise window, the original audio clip's noise was suppressed. An implementation [32] of the mentioned noise reduction procedure was used to automate the noise reduction. The noise window was selected as the leading 25000 samples of each audio clip. Figure 1 shows waveforms of a *making a phone call* acoustic event before and after the noise reduction.

Constant length frame blocking was applied to the noise-reduced audio signals. A constant frame length (window length) of 512 and a constant time shift (hop length) of 256 were used in the frame blocking.

4.2 Audio feature extraction

The feature extraction was performed on the constant length frame blocks. First, discrete short-time Fourier transform (STFT) with 512 fast Fourier transform (FFT) coefficient was applied. The STFT features were further reduced by obtaining the absolute and the mean on the constant length

frame blocks. In order to make the features independent from amplitude of the original sound, min–max normalization is applied to the resulting features. Thus, an acoustic event is represented by 127 frequency bins equally distributed from 0 to 22050 Hz. Figure 2 shows the plotted features (normalized frequency bins values) for falling and sweeping acoustic events.

Since we describe audio data using the mean of the STFT values, representation is one dimensional. Hence, the audio data are processed with a regular four-layer feed-forward neural network for extraction of high-level features and classification.

4.3 Silhouette segmentation

Our dataset has only 622 samples of audio and depth image sequences where different actors perform daily activities, whereas RGB action recognition datasets usually contain around 10,000 samples. Moreover, depth images contain large amount of noisy pixels. Thus, directly applying deep learning techniques used to model RGB videos to model depth data is not effective. And it leads to over-fitting. In order to help the models to learn accurate discriminative features, we first extract the human silhouettes from the depth images.

In our dataset, we saved the silhouette stream extracted from the depth stream by the OPENNI/NITE platform [29]. The silhouette stream contains accurately extracted silhouettes in 74.95% of the videos. But in almost all the videos where the activity involves close interactions with other objects (e.g., lying on a bed), the NITE platform fails to detect the human. Figure 3 shows such a scenario. Thus, it is not possible to rely on the silhouettes given by NITE for activity recognition. But since 74.95% of the silhouettes are accurate, the dataset can be used to train a different machine learning model to segment the human from the background.

First, we manually tagged all the video samples with proper silhouettes to avoid training the model with incorrectly segmented images. There were 466 properly segmented videos. Each video was recorded in 480×640 resolution and 30 frames per second. Since human movements are slow, there are no significant variations between consecutive frames recorded with such high frame rates. Thus, each video was subsampled to 15 frames. Furthermore, videos were down-sampled to 240×320 resolution in order to reduce the memory requirement and training time. To avoid overfitting, a mirror image was created from each image. UNet [31] architecture was selected to build the segmenting model. This is because UNet has been effective in segmenting medical images with small number of training samples. Our dataset is also relatively small in size, and depth images

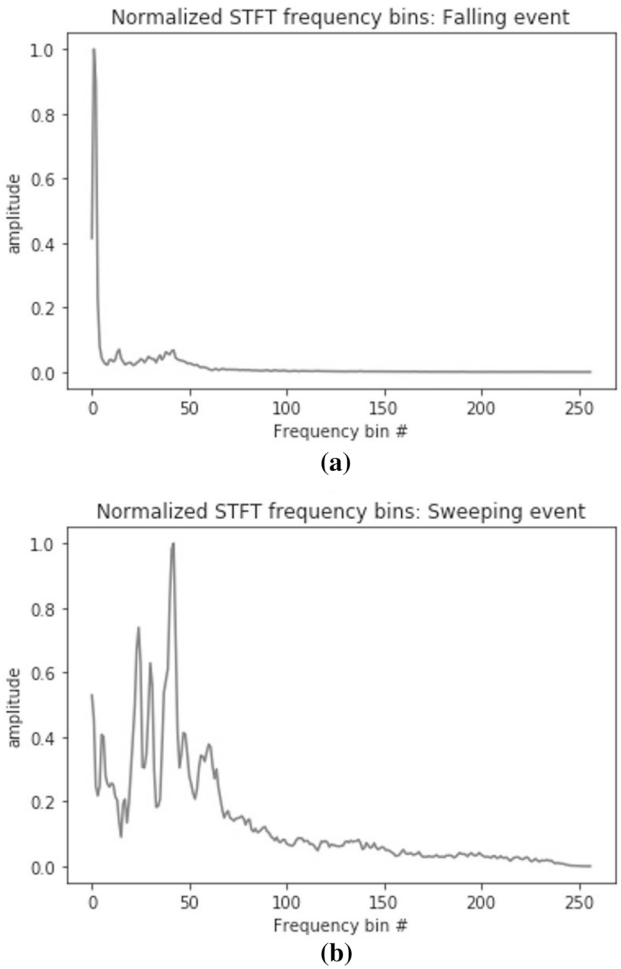


Fig. 2 Normalized frequency bins of acoustic events: **a** falling and **b** sweeping

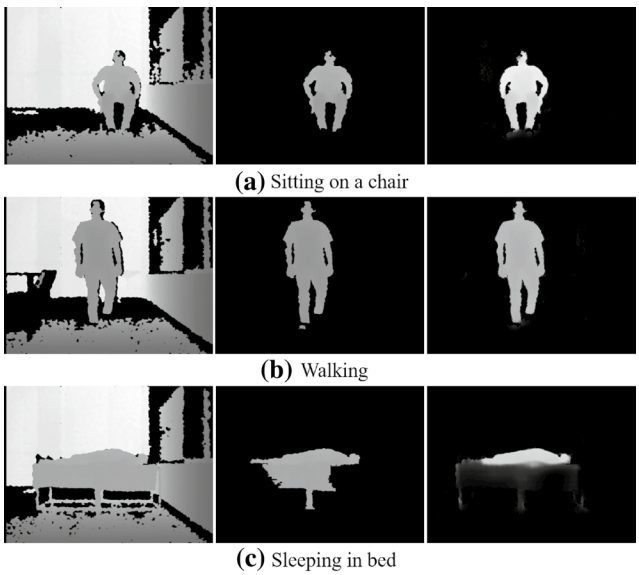


Fig. 3 Comparison of silhouettes. From left to right: raw depth image, NITE silhouette and our silhouette shown

also contain only one information channel similar to many medical images. Figure 3 shows the extracted silhouettes using the trained model.

4.4 Feature extraction from silhouette sequences

We tried two approaches to extract features from silhouette data. In one approach, we rely on a 3D convolutional neural network to extract both spatial and temporal information from the silhouette stream. In the other approach, we employ LSTM to model temporal information while relying on a 2D CNN for spatial information.

4.4.1 3D CNN approach

In this approach, we developed a CNN that employs 3D convolution to extract discriminative information related to activities of daily living from the silhouette streams. The 3D CNN contains three convolutional layers each followed by a rectified linear unit (ReLU) activation. Additionally, there are max-pooling layers after each convolution layer. The architecture is similar to that in [15]. But the filter sizes are set to $3 \times 3 \times 3$ with $2 \times 2 \times 2$ stride. $3 \times 3 \times 3$ filters with $2 \times 2 \times 2$ stride have the same effect as $7 \times 7 \times 3$ filters but require less memory. Also the final 2D CNN layer in [15] is replaced by a 3D layer as it yields improved accuracy.

4.4.2 CNN–LSTM combination

In this CNN–LSTM approach, a network was developed following the architectures of [11, 44] and [45] to extract spatiotemporal information from the silhouette stream. The silhouette stream is processed by a series of 2D convolution and max-pooling layers at the beginning of the network. Unlike in the previous network (3D CNN), the CNN layers in this network process only one image at a time. At the end of the CNN layers, the output is flattened to a vector and given as the input to an LSTM layer. Only the final output of the LSTM layer is considered for further processing. Processing outputs from earlier frames are not necessary since the last output is supposed to contain the important information extracted from all previous time steps. The output of the LSTM layer represents the whole sequence in a single vector. From here onward, the network contains only fully connected layers.

The CNN layers in [11] follow the architectures of AlexNet [18], while [44] and [45] follow the VGG architecture with 19 layers [36]. Both architectures are deep. To avoid overfitting, all of them pre-trained the CNN networks with the ImageNet ILSVRC-2012 dataset which has over 1 million images. Since we are modeling depth data, it is not a viable option. Also, we have less video samples compared to typical RGB video datasets. In addition, memory requirement for training increases with the depth. Considering these limitations, we adopted the shallowest network in VGG which has eight layers of CNN [36] and stripped off the last four layers. Results were observed by removing layers one by one from the back of the eight-layer network. Best results were achieved with only four remaining. Further changes were introduced to the CNN–LSTM networks in [11, 44] and [45]. The number of parameters in the LSTM layer is given by $4 \times m \times (n + m + 1)$ where n and m are the size of the input and output, respectively. The number parameters scale linearly with the size of the input. It is desirable to keep the model simple to avoid overfitting. To reduce the size of the input, we introduced a dense layer (feed-forward layer) in between the CNN and LSTM. The introduced dense layer scales down the output of the CNN layer. The LSTM has 160 hidden units. The number was determined empirically.

4.5 Fusing audio and depth features

We developed a combined end-to-end trainable network similar to [14]. A combined end-to-end trainable network allows weights of the networks modeling both modes to be trained together effectively capturing correlations among modes. The output of audio processing neural network is merged with the output of CNN–LSTM network. The merged results are further processed with two more neural layers to derive the final prediction from the last softmax layer.

First, we developed two independent audio–depth networks one employing 3D CNN and the other using CNN–LSTM to model the silhouette stream. The 3D CNN-only model and the CNN–LSTM-only model can be recovered by removing the CNN–LSTM component (second block) and the 3D CNN component (first block) from the combined model in Fig. 4, respectively. To take advantage of both 3D CNN-based temporal modeling and LSTM-based temporal modeling, we further developed a network that utilizes both 3D CNN and CNN–LSTM to process the silhouettes as shown in Fig. 4.

4.6 Real-time activity recognition

After developing and training networks to recognize activities of daily living in segmented clips of audio and silhouette sequences, the trained models were used to develop a real-time activity recognition system. The system relies on the sliding window approach. The sliding window represents a particular time period. For each sliding window, we take the set of depth frames and the audio clip belonging to the time period and obtain the prediction for them. The prediction is an N-long array where N is the number of activity classes and each element in the array represents the probability of sliding window belonging to each activity. A frame in the depth sequence can be in multiple sliding windows. If the hop length is 1 and the

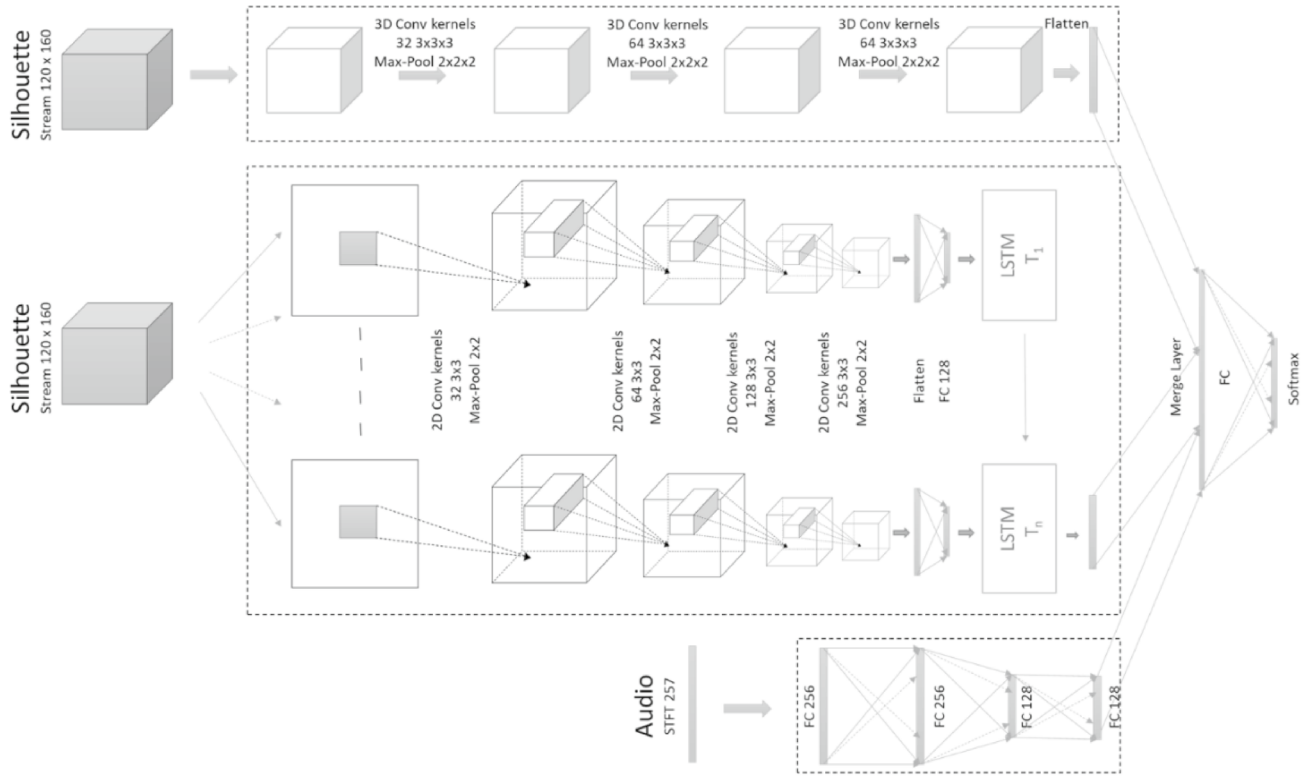


Fig. 4 Architecture of the proposed combined (3D-CNN and CNN–LSTM) end-to-end trainable network

sliding window contains M number of frames, each frame is in M number of sliding windows. As a result, there are M numbers of predictions associated with each frame. Hence, it is desirable to combine all these predictions to obtain the final prediction for a depth frame. But the relatedness of these predictions could vary depending on the position of the frame in the sliding window. Therefore, simply averaging the predictions is not sufficient. Let a depth frame be DF. In the first sliding window that contain DF, DF is at the end of the sliding window. Similarly, in the second sliding window, DF is the second last frame. It is more likely that the predictions of sliding windows where DF is at a position closer to the middle of the window are more relevant to DF. Hence, combining predictions for a frame was calculated in a weighted manner where prediction with the frame closer to the middle of the window gets higher weights. When no activity is predicted with a probability higher than a given threshold, it is considered as an unidentified activity. Figure 5 shows the real-time activity recognition process.

5 Evaluation

5.1 Experimental setup

Currently, there is no ADL or other activity dataset with both audio and depth streams nor any work that recognize activities fusing audio and depth data. Hence, our evaluation takes the form of an ablation study.

Our dataset is composed of actions performed by 17 actors. Samples from three actors (close to 20% of the dataset) were allocated to the test set, and the samples from remaining 14 actors were allocated to the training set. The train/test split of actors is identical for both the silhouette extraction network and the ADL recognition networks, but the selection of actions and samples was differed. This is because we only selected samples with properly segmented silhouettes from the NiTe platform to train the silhouette segmentation network. Since some actions, lying down, sleeping, and waking up to be specific, did not contain any samples with properly segmented silhouettes, they were excluded from both the training and quantitative testing sets. Later, the samples without properly segmented silhouettes were also used for qualitative testing.

All ADL recognition networks, 3D CNN, CNN–LSTM and 3D CNN–CNN–LSTM combined, are trained in an end-to-end manner. Adam optimizer is used as it can adjust the learning rate throughout the training process. Initial learning rate is set to 0.001, while β_1 and β_2 set to 0.9 and 0.999, respectively. Random dropout with a rate of 50% was introduced after each layer. Figure 6 presents the variation of the validation and training loss of the CNN–LSTM depth–audio network with the epochs.

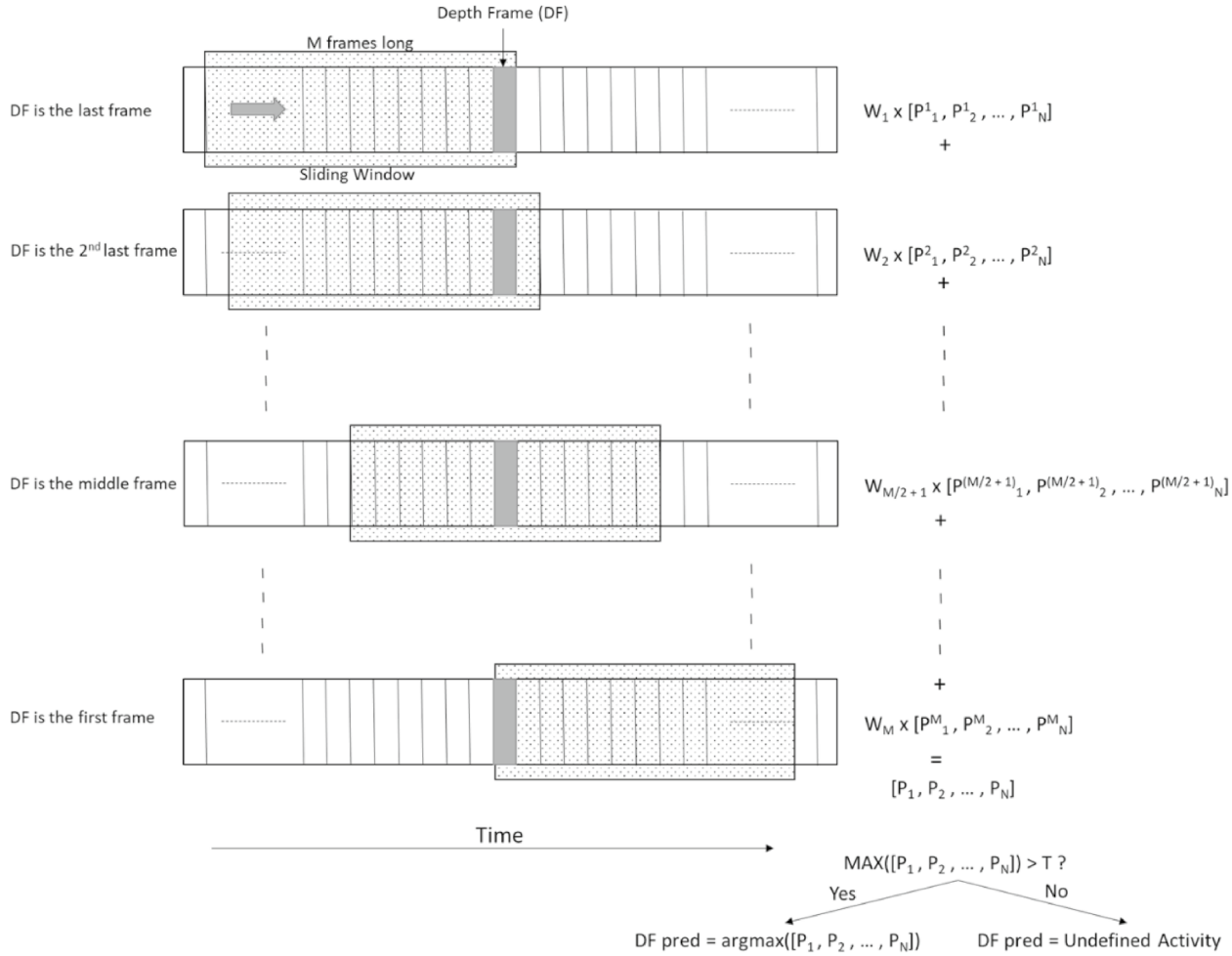


Fig. 5 Calculation of the prediction of a frame in the Real-time ADL recognition system

In addition to fusing audio and depth data in the mentioned three end-to-end trainable networks, we developed separate networks for depth and audio. The network for audio is obtained by removing the silhouettes processing network part and connecting the audio processing network part directly to the final layer avoiding the merge layer. Similarly, networks were derived to model silhouettes. As mentioned in 2.6, in RGB video-based activity recognition, results of separate audio and video classifying networks are combined using SVM [35, 43] and neural networks [45]. Similarly, the predictions of our separate networks were fused using SVM and neural networks. The results are compared in Table 1.

In the real-time activity recognition system, first, the prediction of a frame was derived as the prediction of the sliding window which centers the frame. Secondly, it was calculated as the unweighted summation of the predictions of all its sliding windows. Finally, we implemented the proposed weighted summation approach. Prior two approaches are compared with the proposed (final) approach. When implementing the real-time activity recognition system, we selected the CNN–LSTM network with audio to predict activities. All the silhouette–audio models have similar performance. But the CNN–LSTM network can handle variable length videos. The sliding window spans across ten silhouettes and operates with a stride of 1. All the models were implemented using Keras deep learning framework [7]. The experiments on real-time activity recognition were conducted using a commodity laptop operating on Windows 10 with Intel(R) Core(TM) i5-5200U CPU@2.20 GHz processor, 4 GB of memory, and GeForce 920M GPU (2 GB memory).

5.2 Results and discussion

5.2.1 Silhouette segmentation

The network for silhouette segmentation was trained with the loss function as the average binary cross-entropy loss over each pixel in all the images. The model yields a loss of 0.0074 and achieves an accuracy of 99.8%. The accuracy

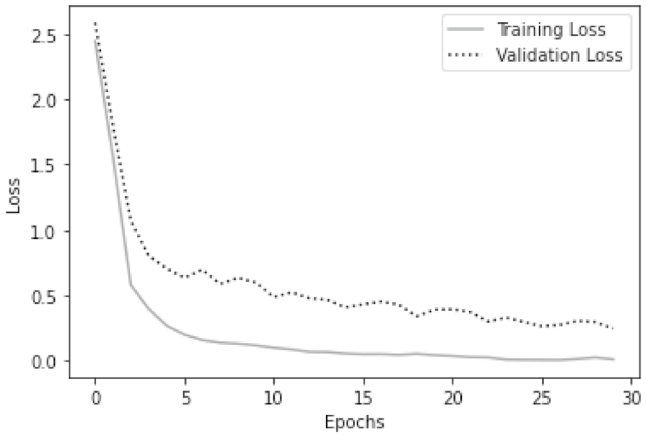


Fig. 6 Training and validation loss with epochs

is the average number of pixels that were labeled correctly. As mentioned before, the silhouette dataset contains both properly segmented silhouettes and poorly segmented silhouettes. These results were obtained by training and testing with properly segmented silhouettes and their respective raw depth images. The model delivers strong results. The training and testing scenarios cover humans performing a wide range of activities. Thus, the silhouettes are in different static and dynamic postures as shown in Fig. 3. We further evaluated the performance of the model with the depth images that have poorly segmented silhouettes in the dataset. Since the targets (the silhouettes given by the NiTe platform) are not accurate, manual qualitative inspections were carried out. And the model produces properly segmented silhouettes in most of the cases. Moreover, it is noteworthy to mention that there were no properly segmented silhouettes in training data where the subjects perform actions: sleeping, waking up and lying down on a bed. But the model was able to accurately segment the silhouette in those scenarios

Table 1 Network comparison

	Accuracy
Audio only	29.54
3D CNN	69.54
CNN–LSTM	74.49
3D CNN + Audio (NN fusion)	77.42
3D CNN + Audio (SVM fusion)	75.19
3D CNN + Audio (Combined model)	84.83
CNN–LSTM + Audio (NN fusion)	83.33
CNN–LSTM + Audio (SVM fusion)	83.87
CNN–LSTM + Audio (Combined model)	86.67
3D CNN + CNN–LSTM + Audio (NN fusion)	81.72
3D CNN + CNN–LSTM + Audio (SVM fusion)	79.58
3D CNN + CNN–LSTM + Audio (Combined model)	88.20

outperforming the NiTe platform. An example case is shown in Fig. 3 where our model accurately segments the silhouette, while the silhouette segmented by the NiTe contains a part of the bed as well. The results indicate that the model has learned the semantic concept of the human accurately.

5.2.2 ADL recognition

The ADL recognition accuracy of the audio-only model is included in the first entry of Table 1. The next two results present the results obtained from network that relies only on the silhouette stream to recognize activities. All the other results are from classifying activities with both audio and silhouette stream. It is clearly visible that the audio–depth fusion is capable of improving the ability to recognize activities of daily living. The accuracy difference between the best performing silhouette-only approach and the silhouette–audio approach is 13.71%.

It is also visible that the CNN–LSTM combination has outperformed the 3D CNN network in each case. This is due to the ability of LSTM to model sequences over CNNs. The flow of an action differs from one actor to the other. For example, consider the activity of sitting on a chair. There are three sub-activities in this activity as standing still, sitting down and sitting still. Time spent on each sub-action varies for different people. When training, 3D CNN applies filters for blocks of frames separately such as the first three frames separately and next three frames separately. When the time duration varies for sub-actions, parts of the sub-actions might move across frame blocks and make it difficult for the 3D CNN to learn temporal variation accurately.

When fusing the results obtained from different models, the neural network-based approach has slightly better results compared to the SVM-based approach. But the end-to-end trainable combined network has outperformed both other approaches by significant margins. In the NN-based fusion approach and the SVM-based fusion approach, we perform the fusion of results. In the combined network approach, results are yielded at the end. Hence, the fusion considers features which are in higher dimensions. Features contain more information about a data sequence than results. Also, the model weights are trained together. This facilitates a better understanding of the correlation between data streams.

Figure 7 presents the confusion matrix related to the results obtained from the best performing model (combined model with 3D CNN, CNN–LSTM and audio). The leftmost column indicates the actual action, and other columns indicate the predicted action. A cell represents which percentage of an activity indicated by the row is classified as the activity indicated by the column. Out of 20 activities, 11 activities yield 100% accuracy. Walking is also identified with very high (94%) accuracy. Washing hand is the most

misclassified activity. In the dataset, the sink is placed at the furthest corner in the room from the camera location. Hence, the resolution of the human in recorded depth images is smaller and the pixels are nosier. Depth data become noisier as a point moves away from the camera. Thus, the portion of correct depth pixels representing the human silhouette is at the minimum in this activity. This could be the reason for the activity being misclassified more often. The second most misclassified activity is eating. It has been misclassified as making a call which contains similar hand movement as eating and as sit still which is a position in which eating is performed. Also, it is observed that making a call is also being misclassified as eating and sit still being misclassified as eating. We note the recent work by Das et al. [9] as a possible approach to improve accuracy in distinguishing between similar activities.

A key difference in the proposed ADL recognition network to those of RGB-based activity recognition networks is that the proposed network is shallower. While keeping the depth of the CNN network short has reduced the potential for overfitting with a relatively small number of samples, it could be the main factor that limits the accuracy at high eighties. This is because shallower networks have less filters and limited capacity to learn complex features.

5.3 Real-time activity recognition

Real-time activity recognition needs to segment the silhouettes from depth images, extract STFT features from the audio and get the prediction for the sliding window. The time taken by the silhouette segmenting model to process a single depth image is 55ms. The activity recognition model takes 84ms to predict a sliding window with 15 depth images. Time taken to extract STFT features is negligible compared to the other operations. As a result, the system can process seven depth frames per second.

Since the continuous dataset is not tagged at the frame level, qualitative performance inspections were carried out manually. When only a single sliding window is considered to calculate the prediction of a frame, the predictions tend to show sudden variations which are incorrect. In both unweighted and weighted summation, predictions are stable and more accurate. This is because sudden variations are damped by predictions of other sliding windows.

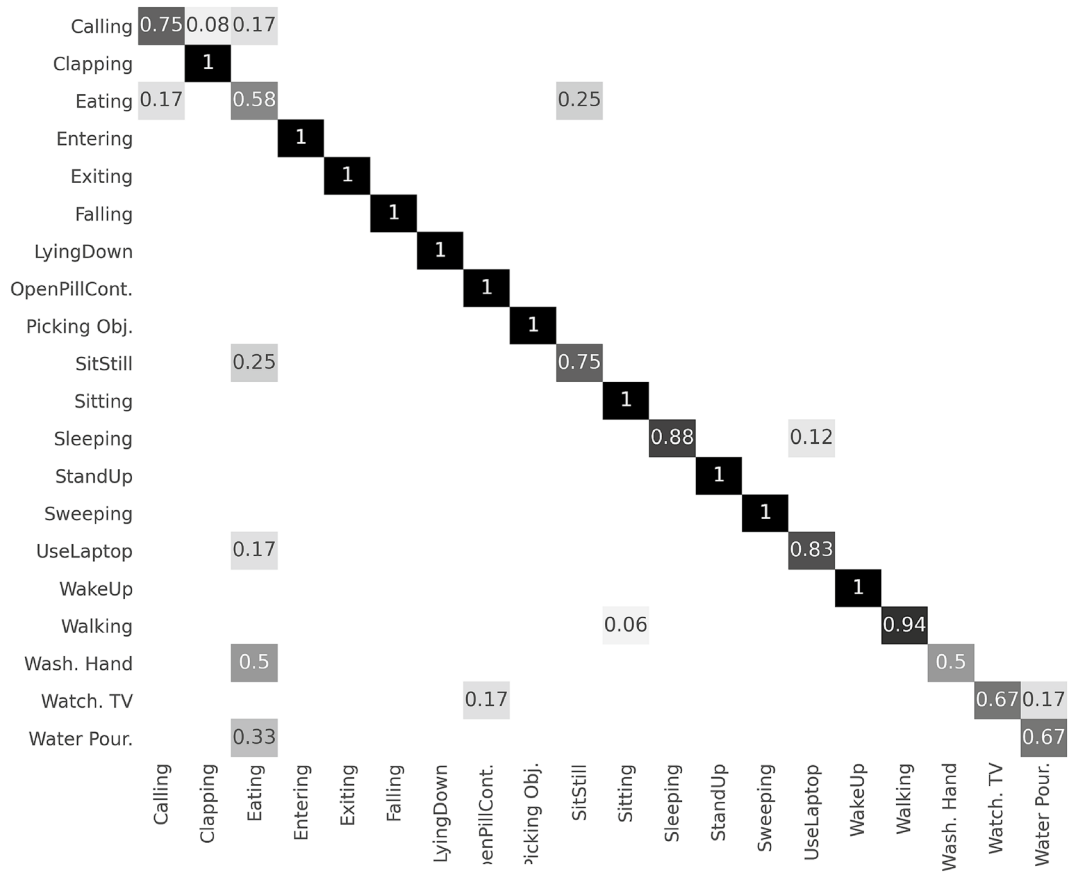


Fig. 7 Confusion matrix for combined model (with 3D-CNN, CNN–LSTM and audio). Actual action given in leftmost column with prediction given in subsequent columns

Moreover, the unweighted summation makes more errors at the edges of an action compared to the weighted summation approach.

6 Conclusion and future work

In this paper, we employed convolution neural networks for silhouette segmentation in depth images and demonstrated its effectiveness with strong results. We adopted convolutional neural network-based approaches used for RGB video analysis to recognize activities of daily living using depth data. Our dataset is smaller compared to most datasets used for RGB video classification. Two strategies were used to address the issue: 1) use of extracted silhouettes instead of raw depth images to train the models and 2) use of shallower networks. We fused audio data stream with depth data and showed that the audio–depth fusion contributes to a significant improvement in recognition accuracy. We further extended the developed models to build a real-time ADL recognition system which is capable of recognizing ADL on a commodity laptop computer.

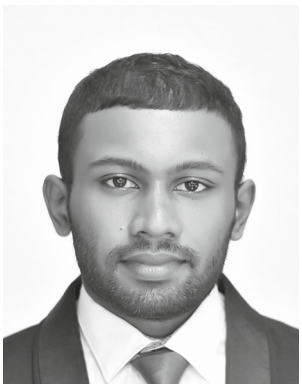
Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Arshad, S., Feng, C., Liu, Y., Hu, Y., Yu, R., Zhou, S., Li, H.: Wi-chase: a wifi based human activity recognition system for sensorless environments. In: 2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1–6 (2017)
2. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential deep learning for human action recognition. In: Salah, A.A., Lepri, B. (eds.) Human Behavior Understanding, pp. 29–39. Springer, Berlin (2011)
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017). https://doi.org/10.1109/TPAMI.2016.2644615
4. Biswas, K.K., Basu, S.K.: Gesture recognition using microsoft kinect®. In: The 5th International Conference on Automation, Robotics and Applications, pp. 100–103 (2011). https://doi.org/10.1109/ICARA.2011.6144864
5. Chen, J., Kam, A.H., Zhang, J., Liu, N., Shue, L.: Bathroom activity monitoring based on sound. In: Gellersen, H.W., Want, R., Schmidt, A. (eds.) Pervasive Computing, pp. 47–61. Springer, Berlin (2005)
6. Cheng, J., Sundholm, M., Zhou, B., Hirsch, M., Lukowicz, P.: Smart-surface: large scale textile pressure sensors arrays for activity recognition. Pervasive Mob. Comput. **30**, 97–112 (2016). https://doi.org/10.1016/j.pmcj.2016.01.007
7. Chollet, F., et al.: Keras. (2015). https://keras.io
8. Cristani, M., Bicego, M., Murino, V.: Audio-visual event recognition in surveillance video sequences. IEEE Trans. Multimed. **9**(2), 257–267 (2007). https://doi.org/10.1109/TMM.2006.886263
9. Das, S., Thonnat, M., Bremond, F.F.: Looking deeper into time for activities of daily living recognition. In: WACV 2020-IEEE Winter Conference on Applications of Computer Vision. Snowmass village, Colorado, United States (2020). https://hal.inria.fr/hal-02368366
10. Das Dawn, D., Shaikh, S.H.: A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector. Vis. Comput. **32**, 289–306 (2016). https://doi.org/10.1007/s00371-015-1066-2
11. Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 677–691 (2017). https://doi.org/10.1109/TPAMI.2016.2599174
12. Gasparrini, S., Cippitelli, E., Spinsante, S., Gambi, E.: A depth-based fall detection system using a Kinect® sensor. Sensors (Switzerland) **14**(2), 2756–2775 (2014). https://doi.org/10.3390/s140202756
13. Grushin, A., Monner, D.D., Reggia, J.A., Mishra, A.: Robust human action recognition via long short-term memory. In: The 2013 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2013). https://doi.org/10.1109/IJCNN.2013.6706797
14. Hou, J.C., Wang, S.S., Lai, Y.H., Tsao, Y., Chang, H.W., Wang, H.m.: Audio-visual speech enhancement using multimodal deep convolutional neural networks. IEEE Transactions on Emerging Topics in Computational Intelligence **2** (2018). https://doi.org/10.1109/TETCI.2017.2784878
15. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013). https://doi.org/10.1109/TPAMI.2012.59
16. Jiang, X., Zhong, F., Peng, Q., Qin, X.: Online robust action recognition based on a hierarchical model. Vis. Comput. **30**, 1021–1033 (2014). https://doi.org/10.1007/s00371-014-0923-8
17. Kamal, S., Jalal, A., Kim, D.: Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified hmm. J. Electr. Eng. Technol. **11**(6), 1857–1862 (2016). https://doi.org/10.5370/jeet.2016.11.6.1857
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017). https://doi.org/10.1145/3065386
19. Ma, H., Li, W., Zhang, X., Gao, S., Lu, S.: Attnsense: Multi-level attention mechanism for multimodal human activity recognition. pp. 3109–3115. International Joint Conferences on Artificial Intelligence Organization (2019). https://doi.org/10.24963/ijcai.2019/431
20. Mainetti, L., Manco, L., Patrono, L., Secco, A., Sergi, I., Vergallo, R.: An ambient assisted living system for elderly assistance applications. In: 2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1–6 (2016)
21. Microsoft Corporation: Kinect - Windows app development. https://developer.microsoft.com/en-us/windows/kinect
22. Movo Photo Corporation: MOVO USB computer Lavalier microphone (20'ft cord). https://www.movophoto.com/products/movo-m1-usb-lavalier-lapel-condenser-computer-microphone
23. Muda, L., Begam, M., Elamvazuthi, I.: Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. J. Comput. **2**, 34–41 (2010). https://doi.org/10.5120/20312-2362
24. Naronglerdrit, P., Mporas, I., Sotudeh, R.: Monitoring of indoors human activities using mobile phone audio recordings. In: Proceedings-2017 IEEE 13th International Colloquium on Signal Processing and its Applications, CSPA 2017 (2017). https://doi.org/10.1109/CSPA.2017.8064918

25. Ni, B., Wang, G., Moulin, P.: RGBD-HuDaAct: a color-depth video database for human daily activity recognition BT - consumer depth cameras for computer vision. 2011 IEEE International Conference on Computer Vision Workshops pp. 1147–1153 (2011). https://doi.org/10.1109/ICCVW.2011.6130379
26. Ordóñez, F.J., Roggen, D.: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors **16**, 115 (2016). https://doi.org/10.3390/s16010115
27. Oreifej, O., Liu, Z.: Hon4d: histogram of oriented 4d normals for activity recognition from depth sequences. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 716–723 (2013). https://doi.org/10.1109/CVPR.2013.98
28. Pieropan, A., Salvi, G., Pauwels, K., Kjellström, H.: Audio–visual classification and detection of human manipulation actions. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3045–3052 (2014). https://doi.org/10.1109/IROS.2014.6942983
29. PrimeSense: prime sensortm nite 1.3 algorithms notes-version 1.0 (2010)
30. Ronao, C., Cho, S.B.: Human activity recognition with smartphone sensors using deep learning neural networks. Expert Syst. Appl. **59**, 235–244 (2016). https://doi.org/10.1016/j.eswa.2016.04.032
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. (2015). ArXiv arXiv:abs/1505.04597
32. Sainburg, T.: noisereduce · PyPI. https://pypi.org/project/noisereduce/
33. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. **24**(3), 279–283 (2017). https://doi.org/10.1109/LSP.2017.2657381
34. Siantikos, G., Giannakopoulos, T., Konstantopoulos, S.: A low-cost approach for detecting activities of daily living using audio information: a use case on bathroom activity monitoring. In: International Conference on Information and Communication Technologies for Ageing Well and e-Health (2016)
35. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. (2014) ArXiv arXiv:abs/1406.2199
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR pp. 1–14 (2014). arXiv:1409.1556
37. Stork, J.A., Spinello, L., Silva, J., Arras, K.O.: Audio-based human activity recognition using non-markovian ensemble voting. In: 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, pp. 509–514 (2012). https://doi.org/10.1109/ROMAN.2012.6343802
38. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4489–4497 (2015). https://doi.org/10.1109/ICCV.2015.510
39. Tran, P.V.: A fully convolutional neural network for cardiac segmentation in short-axis mri. (2016) CoRR arXiv:abs/1604.00494
40. Trinh, L.A., Thang, N.D., Tran, H.H., Hung, T.C.: Human extraction from a sequence of depth images using segmentation and foreground detection. Proceedings of the 5th Symposium on Information and Communication Technology-SoICT **14**, (2014). https://doi.org/10.1145/2676585.2676624
41. Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C., Ogunbona, P.O.: Action recognition from depth maps using deep convolutional neural networks. IEEE Trans. Hum. Mach. Syst. **46**(4), 498–509 (2016). https://doi.org/10.1109/THMS.2015.2504550
42. Wang, W., Liu, A.X., Shahzad, M., Ling, K., Lu, S.: Device-free human activity recognition using commercial wifi devices. IEEE J. Sel. Areas Commun. **35**(5), 1118–1131 (2017)
43. Wu, Q., Wang, Z., Deng, F., Chi, Z., Feng, D.D.: Realistic human action recognition with multimodal feature selection and fusion. IEEE Trans. Syst. Man Cybern. Syst. **43**(4), 875–885 (2013). https://doi.org/10.1109/TSMCA.2012.2226575
44. Wu, Z., Jiang, Y.G., Wang, X., Ye, H., Xue, X.: Multi-stream multi-class fusion of deep networks for video classification. In: Proceedings of the 24th ACM International Conference on Multimedia, MM '16, pp. 791–800. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2964284.2964328
45. Wu, Z., Wang, X., Jiang, Y.G., Ye, H., Xue, X.: Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In: Proceedings of the 23rd ACM International Conference on Multimedia, MM '15, pp. 461–470. ACM, New York, NY, USA (2015). https://doi.org/10.1145/2733373.2806222
46. Xia, L., Aggarwal, J.K.: Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 2834–2841 (2013). https://doi.org/10.1109/CVPR.2013.365
47. Yang, X., Tian, Y.: Super normal vector for activity recognition using depth sequences. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14, pp. 804–811. IEEE Computer Society, Washington, DC, USA (2014). https://doi.org/10.1109/CVPR.2014.108
48. Yang, X., Zhang, C., Tian, Y.: Recognizing actions using depth motion maps-based histograms of oriented gradients. In: Proceedings of the 20th ACM International Conference on Multimedia, MM '12, pp. 1057–1060. ACM, New York, NY, USA (2012). https://doi.org/10.1145/2393347.2396382



Danushka Madhuranga is a final year undergraduate at Computer Science & Engineering Department of the University of Moratuwa, Sri Lanka. His research interests include computer vision, machine learning and information retrieval.

Rivindu Madushan is an undergraduate at Computer Science & Engineering Department of the University of Moratuwa, Sri Lanka. His research interests include computer vision and machine learning.

Chathuranga Siriwardane is an undergraduate at Computer Science & Engineering Department of the University of Moratuwa, Sri Lanka. His research interests include signal processing, computer vision and machine learning.

Kutila Gunasekera received a bachelor's degree in Computer Science & Engineering and a master's degree in Computer Science from the University of Moratuwa, Sri Lanka. He received his PhD from Monash University, Australia. Currently, he works as a Senior Lecturer at the Department of Computer Science & Engineering at the University of Moratuwa. His research interests include pervasive computing, Internet of things and software engineering.