applied sciences

MDPI

Article

KFSENet: A Key Frame-Based Skeleton Feature Estimation and Action Recognition Network for Improved Robot Vision with Face and Emotion Recognition

Dinh-Son Le [1], Hai-Hong Phan [1], Ha Huy Hung [2], Van-An Tran [1], The-Hung Nguyen [3] and Dinh-Quan Nguyen [2],*

[1] Faculty of Information Technology, Le Quy Don Technical University, 236 Hoang Quoc Viet, Bac Tu Liem, Ha Noi 11900, Vietnam; sonld@lqdtu.edu.vn (D.-S.L.); hongpth@lqdtu.edu.vn (H.-H.P.); tavistu@gmail.com (V.-A.T.)

[2] Faculty of Aerospace Engineering, Le Quy Don Technical University, 236 Hoang Quoc Viet, Bac Tu Liem, Ha Noi 11900, Vietnam; hahuyhung@lqdtu.edu.vn

[3] Faculty of Technical Management, Le Quy Don Technical University, 236 Hoang Quoc Viet, Bac Tu Liem, Ha Noi 11900, Vietnam; hungnt515@gmail.com

* Correspondence: ndquan@lqdtu.edu.vn; Tel.: +84-969-205-305

Abstract: In this paper, we propose an integrated approach to robot vision: a key frame-based skeleton feature estimation and action recognition network (KFSENet) that incorporates action recognition with face and emotion recognition to enable social robots to engage in more personal interactions. Instead of extracting the human skeleton features from the entire video, we propose a key frame-based approach for their extraction using pose estimation models. We select the key frames using the gradient of a proposed total motion metric that is computed using dense optical flow. We use the extracted human skeleton features from the selected key frames to train a deep neural network (i.e., the double-feature double-motion network (DDNet)) for action recognition. The proposed KFSENet utilizes a simpler model to learn and differentiate between the different action classes, is computationally simpler and yields better action recognition performance when compared with existing methods. The use of key frames allows the proposed method to eliminate unnecessary and redundant information, which improves its classification accuracy and decreases its computational cost. The proposed method is tested on both publicly available standard benchmark datasets and self-collected datasets. The performance of the proposed method is compared to existing state-of-the-art methods. Our results indicate that the proposed method yields better performance compared with existing methods. Moreover, our proposed framework integrates face and emotion recognition to enable social robots to engage in more personal interaction with humans.

Keywords: robots; vision; key frame; deep neural network; optical flow; face recognition; action recognition; skeleton features

check for updates

Citation: Le, D.-S.; Phan, H.-H.; Hung, H.H.; Tran, V.-A.; Nguyen, T.-H.; Nguyen, D.-Q. KFSENet: A Key Frame-Based Skeleton Feature Estimation and Action Recognition Network for Improved Robot Vision with Face and Emotion Recognition. *Appl. Sci.* **2022**, *12*, 5455. https://doi.org/10.3390/app12115455

Academic Editor: Athanasios Nikolaidis

Received: 27 March 2022
Accepted: 24 May 2022
Published: 27 May 2022

1. Introduction

Social robotics is an important area of research and an emerging field in robotics as robots are increasingly landing new roles which are mostly of an assistive nature, namely in education and especially for children with special needs or in elderly care [1–4]. Improving the human–robot interaction (HRI) in such contexts would require a robot to be more aware of the human that it is interacting with, and the ability to identify people, detect their actions and recognize their faces and emotions in particular would be required at the very least [3]. Hence, there is a need for an integrated approach to robot vision that incorporates action recognition along with face and emotion recognition to make robots more amiable in their social interactions with their human counterparts.

Human action recognition using video cameras is a challenging task in computer and robot vision. However, it has wide-ranging applications in health care, sports training, physical rehabilitation of the injured, interactive entertainment and video understanding [5]. Action recognition typically requires the processing of huge amounts of information streaming through the input video that must be analyzed in real time, which is a challenging task. One approach to this problem is to create and use larger architectures [6] which are useful in achieving better precision, whereas a second approach involves getting rid of redundant video data and creating a more concise representation of the raw input video, which would result in reduced computational costs for more accurate action recognition in real time [6].

Action recognition involves at least three steps: the processing of the raw video input, followed by the creation of an appropriate representation for the human actions and finally the classification of those representations into the appropriate action categories. The speed and accuracy of human action recognition can be improved by improving any of these steps in terms of computational speed and accuracy. Multiple studies [7–10] have indicated that videos contain temporally redundant data, implying that multiple frames of a video can be easily skipped without losing much information on the actions recorded in that video. Moreover, there are some action classes in the standard benchmark datasets that do not require motion or temporal information for their identification. Hence, with only a few key frames as inputs, these actions can be easily identified.

Human skeletal features (i.e., points that show the relative placement of different parts or joints of the human skeleton), has been one of the most effective and extensively utilized elements to represent various human actions in recent years. These features can be used for the identification of human action when they are analyzed over multiple frames in the video. Thus, action recognition on the basis of human skeletal features is simply a timed motion analysis of human poses [11]. Different input data can be used to gain information about the skeletal joints, such as a simple RGB video, a spatiotemporal joint trajectory obtained through a sensor system, a 3D skeleton derived from image data or a hybrid system that combines any of the preceding input data sources. Many studies have found it to be very effective in simplifying the network architecture required for action recognition [12–14].

In this paper, we propose an integrated framework for robot vision that incorporates action recognition with face and emotion recognition. Our methodology for human action recognition involves the extraction of key frames from the input video based upon the gradient of a proposed motion metric that is calculated using dense optical flow. After the key frames are extracted, pose estimation models are then used to extract human skeleton features, also known as human keypoints, from the extracted key frames. These human keypoints are then used to train a deep neural network, which can then be used to identify different human actions. For the proposed framework, we create an algorithm that uses the gradient of our proposed motion metric to select the most significant frames (i.e., the key frames). Many studies have proposed using deep learning models for the extraction of key frames [15–17]. These solutions, however, have the drawbacks of a huge model size and hence a poor execution speed. In addition, these studies also propose using deep network models for skeleton-based action representation and classification. In order to improve the execution speed of action recognition and perform it in real time, we need to use a rather simple and efficient method to extract key frames from the input video, as is proposed in this study.

In this study, we propose an integrated framework for robot vision that involves action recognition using human skeleton features extracted from the key frames of a video. A new algorithm is proposed that uses the gradient of a motion metric calculated using dense optical flow, as discussed in Section 3.3, to select the key frames. Pose estimation models are applied on the key frames to extract the human skeleton features that are then used to train a classifier for action recognition. Moreover, facial and emotion recognition is also incorporated into the proposed framework for robot vision. The proposed approach for action recognition is thoroughly evaluated using the standard benchmark dataset for action recognition (i.e., the University of Central Florida (UCF) Sports [18]), which includes 10 action classes, and a self-created dataset with 6 action classes. Moreover, the face recognition component of the proposed integrated framework for robot vision is tested on the standard labeled faces in the wild (LFW) [19] and the self-collected HMT datasets. The UCF Sports and LFW datasets are regarded as standard datasets that are used by researchers to benchmark the performance of their algorithms for action and face recognition, respectively. Similarly, we test the emotion recognition component of our proposed framework using the FER2013 dataset [20] for facial emotion recognition, with more than 32,000 images of 48 × 48 pixels each split across 7 emotion categories.

The main contributions of the this study are as follows:

1. A novel integrated approach to robot vision is presented which incorporates action recognition with face and emotion recognition, affording social robots the ability to engage in more personal interaction with humans.
2. A new motion metric is proposed to calculate the total motion between two consecutive frames using dense optical flow. Furthermore, the gradient of this motion metric is proposed to be used for selecting the key frames of a video.
3. A novel pipeline is proposed for real-time action recognition using a key frame-based approach, which involves the extraction of human skeleton features from only the key frames of a video instead of all the frames. This drastically reduces the amount of data by almost 80–90% which must be processed for the extraction of human skeleton features using pose estimation models. Consequently, the dimensionality of the human skeleton feature vectors is also reduced, which improves the classification performance of the deep neural network (i.e., double-feature double-motion network (DDNet)) used for action recognition.

The rest of this paper is organized as follows. Section 2 reviews the related work in human action recognition based on skeleton features and deep learning models. It also highlights the contributions of this work. Section 3 presents the proposed methodology for an integrated approach to robot vision in a detailed manner, while Section 4 describes the datasets and the computational platform used to implement the proposed methodology. Section 5 presents the experimental results and discussions, whereas conclusions are drawn in Section 6.

2. Related Work

This section provides a brief review of the research work related to this study, primarily in relation to three key issues: the representation of features extracted from human skeletal joints, the modeling of temporal dynamics across successive frames in a video sequence for action recognition and approaches for the efficient extraction of key frames from video sequences. Deep learning has automated the process of representation learning in many application areas, especially computer vision [21]. In [22], the authors proposed using a deep convolutional neural network called the high-resolution network (HRNet) for pose tracking on the PoseTrack dataset. HRNet automatically extracts the human skeleton feature vectors for action recognition. Meanwhile, in [23], the authors proposed a multi-level representation of Fisher vectors and other skeleton-based geometric features for the recognition of hand gestures. In [24], the authors presented a novel representation method called *PoTion*, which uses heat maps for human joints to encode the movement of some semantic keypoints.

In [25], the authors asserted that optical flow-based techniques are more effective for action recognition, as they are invariant to appearance, are more accurate at boundaries and yield better action recognition performance even with small displacements. Moreover, the detection and extraction of key frames in videos has been found to be an effective strategy in many application areas, such as action recognition and video classification. Different techniques have been proposed for the extraction of key frames in videos. For instance, Xiang et al. [9] proposed a deep learning-based approach for the detection of key frames in videos. They proposed using a two-stream deep ConvNet for the detection of key frames in a video. Their model is trained to directly predict the locations of key frames in a video. Similarly, in [8], the authors proposed a self-supervised and automatic method to select key frames in a video. In their approach, these authors proposed to combine a two-stream ConvNet with a novel annotation architecture that automatically annotates key frames in a video. Moreover, Gowda et al. [26] proposed another method for filtering key frames in a video, which works by aggregating the video frames instead of looking at them one by one.

Similarly, different approaches have been proposed for the modeling of temporal dynamics across successive frames for action recognition. For instance, in [27], Sawant et al. proposed a systematic approach for the recognition of human activities in real time using OpenPose and a long short-term memory (LSTM) recurrent neural network (RNN). The authors in [28] proposed using a hierarchical RNN for modeling the temporal dynamics associated with human actions across successive frames, whereas in [29] a temporal sliding LSTM network was proposed for the same purpose. Similarly, Yang et al. [30] proposed a fast and lightweight network (i.e., the double-feature double-motion network (DDNet)) for skeleton feature-based action recognition. Moreover, Yasin et al. [7] proposed using a key frame-based approach for 3D action recognition using a deep neural network. In their study, they performed extensive experiments on the motion capture database *HDM05* [31], which contains more than 4 h of recorded motion video, and the Carnegie Mellon motion capture database, which is divided into 6 categories and 23 subcategories [32].

In [33], the authors demonstrated that dense sampling methods are more effective for human action recognition compared with methods that rely on sparse interest point representations, which may overlook some important space structures and hence lead to higher misclassification. Many authors have tried to compensate for the loss of important structural information in methods based on local representations. For instance, in [34], the authors explored the use of spatiotemporal structural information and neighborhood-based features. Similarly, in [35], the authors proposed a local descriptor based on an independent subspace analysis algorithm to learn the spatiotemporal invariant features from unlabeled video data and hence improve the performance of local descriptor-based methods. Meanwhile, in [36], Al Ghamdi et al. presented a space-time extension of the scale-invariant feature transform (SIFT), which was originally constructed using a spatiotemporal difference of Gaussians (DoG). In [37], Wang et al. used dense trajectories and motion boundary descriptors to capture information about local motion in videos. Similarly, in [38], the authors proposed a new framework that integrates motion maps using a deep 3D convolutional neural network and a long short-term memory (LSTM) network for action recognition. Meanwhile, in [39], Kim et al. proposed the aggregation of features from visual attention and pose estimation for action recognition. However, most of these techniques involve the training of an end-to-end network with fully convolutional kernels, which renders these techniques computationally expensive as they boast a large model size. However, applications that are constrained by the availability of data may benefit from shallow learning approaches based on hand-crafted features [35–37].

3. The Proposed Methodology

The main components of the proposed framework for the integrated vision system of a social robot that incorporates action recognition with face and emotion recognition in real time are illustrated in Figure 1.

As shown in Figure 1, the proposed methodology begins by first extracting key frames from the visual input to the robot, which is assumed to be a steady video stream that is coming from a camera installed on the robot. The technique used for the extraction of key frames is discussed in more detail in Section 3.1. Once the key frames are extracted from the input video, then pose estimation models are applied to those frames to extract human skeleton features. These human skeleton features are also called human keypoints, and they represent the coordinates of the two-dimensional skeleton joints of the human objects in the key frames. The techniques used for the extraction of these keypoints are discussed in more detail in Section 3.2. These human keypoints are then used as input to a deep neural network—the double-feature double-motion network (DDNet)—which classifies them into an appropriate action category as discussed in Section 3.3. In addition to action recognition, the proposed framework also involves face and emotion recognition of human objects in the input video, as discussed in Section 3.4. Face and emotion recognition are useful in improving the quality of human–machine interaction and in making it more personal for humans.

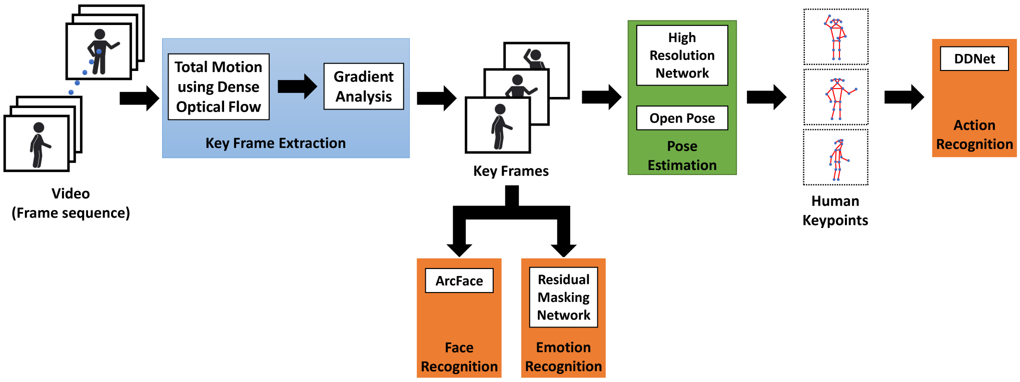


Figure 1. The proposed framework for the real-time vision system of a social robot that incorporates action recognition with face and emotion recognition.

3.1. Key Frame Extraction

Instead of frame-by-frame analysis of the entire video sequence for the extraction of human skeleton features, which could then be used for action recognition, we propose the extraction of key frames from the input video, which are subsequently analyzed for the extraction of human skeleton features. These human skeleton features, also known as human keypoints, are then used to distinguish between different actions using a classifier. Key frames dramatically reduce the amount of data that must be processed to extract the features that we propose to use for action recognition and capture only those parts of the video which undergo a significant transition. Key frames are identified by utilizing the gradient of the optical flow displacement field between every pair of consecutive video frames. Specifically, a two-step procedure is followed to determine the key frames.

First, dense optical flow is used to compute the changes in intensity $I(x, y, t)$ or velocity that each pixel undergoes in two successive frames [40] along both the horizontal and vertical directions. The velocity's horizontal and vertical components are summed over all the pixels for each frame to compute an estimate of the total velocity, which is used as a motion metric $M(t)$ for every frame t. The relation for computing the motion metric $M(t)$ is given in Equation (1):

$$M(t) = \sum_{1}^{m}\sum_{1}^{n} V_x(x, y, t) + V_y(x, y, t), \tag{1}$$

where m represents the width of the image in pixels, n represents the height of the image in pixels, $V_x(x, y, t)$ is the horizontal or x component of the velocity or optical flow that estimates the gradient of intensity along the x-axis for the pixel at coordinates (x, y) between two successive frames t and $t + 1$, which is given by

$$V_x(x, y, t) = \frac{\partial I(x, y, t)}{\partial x}, \tag{2}$$

and $V_y(x, y, t)$ is the vertical or y component of the velocity or optical flow that estimates the gradient of intensity along the y-axis for the pixel at coordinates (x, y) between two successive frames t and $t + 1$, which is given by

$$V_y(x, y, t) = \frac{\partial I(x, y, t)}{\partial y}. \tag{3}$$

As dense optical flow tracks the intensity variations across all the pixels of a frame, therefore, the motion metric in Equation (1) provides an estimate of the amount of motion along consecutive frames.

Secondly, the proposed motion metric $M(t)$ is analyzed as a function of time by looking at its variation across different frames. The gradient of the proposed motion metric $\frac{dM}{dt}$ can be used to identify the frames between which there is significant activity. This can be easily performed by looking at the local extrema of $\frac{dM}{dt}$ to identify the key frames (i.e., frames where there is a significant transition or change taking place). Thus, by quantifying motion or the dense optical flow across consecutive video frames into a single metric, and then by analyzing the gradient of this motion metric, we identify the key frames, which are then used for the extraction of skeleton features using pose estimation models as described in Section 3.2.

3.2. Extraction of Skeleton Features

We propose that once the key frames have been extracted, human pose estimation algorithms should be applied to those key frames to extract the skeleton features, which can then be used for the classification of different actions. The skeleton features are nothing but the pixel coordinates of the human skeletal joints in the key frame. They are also referred to as human keypoints. Human pose estimation schemes try to identify the relative placement or layout of different human joints and body parts in an image, which can then be used for action recognition, as we demonstrate in this study.

There are two popular approaches to human pose estimation (i.e., the *top-down* and *bottom-up* approaches). In the *top-down* approach, first a human body is detected in an image and bounded by a bounding box. Then, the skeletal joints and body parts are discovered within each bounded box (i.e., for each human body in the image). The coordinates of these joints and body parts are called the keypoints, which can be used to identify a human pose. Meanwhile, in the *bottom-up* approach, first the skeletal joints and body parts are discovered in an image, and then they are joined together to constitute different poses. The *top-down* approaches can be computationally expensive, as they first involve the identification of human objects in an image and then pose estimation for each human object in that image. This also entails that separate keypoint configurations be maintained for every human object or bounding box in the image. In contrast, *bottom-up* approaches are generally faster as they directly identify all the keypoints in an image without first performing any human object detection.

In this study, we use two different pose estimation models (i.e., the high-resolution network (HRNet) [22] and OpenPose [12]) for the extraction of skeleton features from key frames. Both of these models delivered the best performance on the *COCO* keypoint dataset [41], which includes around 200,000 images with about 250,000 instances of persons labeled with 17 joints. OpenPose [12] uses the bottom-up approach for the extraction of keypoints, whereas HRNet [22] works in a top-down fashion. OpenPose can be used for the detection of multiple persons in a single image in real time and can extract up to 135 keypoints. The running time complexity of OpenPose is not affected by the number of persons in the image, as it first recognizes the body parts in an image and then estimates the pose from those body parts. In either case, the extracted features are used to construct training vectors for the classifier, which is discussed in more detail in Section 3.3. The trained classifier can then be used for the recognition of the action represented by a vector of skeleton features extracted from the key frames alone.

3.3. Action Recognition Using Ddnet

Once the human skeleton features or keypoints are extracted, then they must be used to recognize the action that they represent. We propose using the double-feature double-motion network (DDNet) [30] as a classifier for the recognition of actions represented by the feature vectors of keypoints extracted using pose estimation models. DDNet is fast because it uses a very lightweight network architecture with only 0.15 million parameters [30], which allows it to operate in real time. Moreover, the relatively fewer parameters mean that it uses a relatively simpler model, which enables it to achieve better generalization performance and detection accuracy compared with the more complex models, as demonstrated in Section 5, where we test it on the University of Central Florida (UCF) sports [18] and on the self-collected HNH datasets.

3.4. Face Recognition and Emotion Recognition

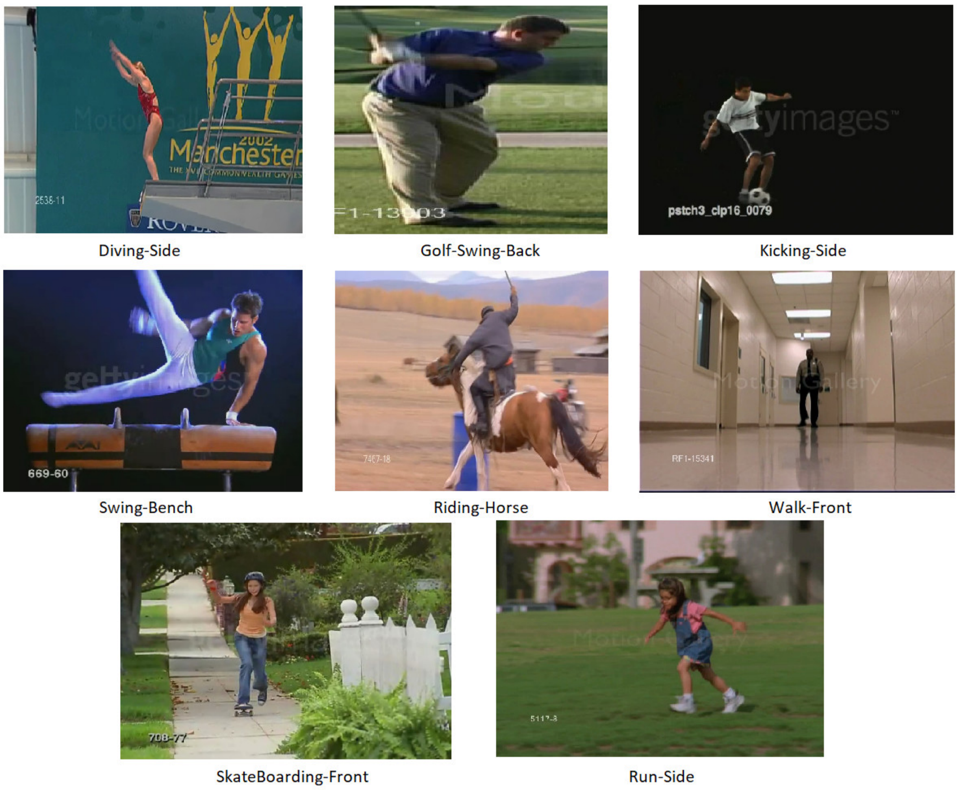
In addition to skeleton feature-based action recognition, we also incorporate face and emotion recognition into our proposed framework to make human interaction with the robot more personal. Face recognition involves the identification of a face from a given face dataset. There are several techniques that can be used to accomplish face recognition, such as SphereFace [42], which uses an angular margin as the loss function to learn those features for face recognition which are angularly discriminative, CosFace [43], an improvement over SphereFace that uses a cosine margin penalty as the loss function, and ArcFace [44], which uses the additive angular margin loss function for deep face recognition with better results than both SphereFace and CosFace. Hence, we propose to use ArcFace [44] for face recognition in our proposed framework for robot vision. Moreover, for emotion recognition, we employ a residual masking network [45] for the extraction of facial emotion features and their subsequent classification. The residual masking network that we utilize in our study comprises four layers, where each layer consists of a residual layer and a masking block. We test our proposed residual masking network on the FER2013 dataset [20] for facial emotion recognition, which comprises more than 32,000 images 48 × 48 pixels in size each and split across 7 emotion categories. Furthermore, we use the Dlib toolbox (dlib.net, accessed on 8 December 2021) for face detection and alignment.

4. Data and Experimental Set-Up

The proposed methodology for action recognition was tested on two datasets: the University of Central Florida (UCF) Sports dataset [18] and the self-collected HNH human action dataset. The UCF Sports dataset is a moving camera dataset that includes realistic sports videos from broadcast television networks like the BBC and ESPN. It includes around 150 sports videos containing 10 action classes: running, lifting, diving, skateboarding, kicking, swing-golf, walking, swing-side, horse-riding and swing-bench. The frame rate of each sports video is 10 frames per second (fps), whereas the resolution of these videos is 720 × 480 pixels, which is comparatively higher than other datasets, thereby making it relatively more challenging. Moreover, the number and duration of clips for each action class is different (e.g., actions like kicking are shorter and less periodic compared with actions such as walking and running).

Figure 2 shows snapshots from different videos in the UCF Sports dataset to illustrate the different types of action classes used for testing the proposed action recognition methodology. In addition to the UCF dataset, the proposed methodology was also tested on the HNH moving camera dataset, which is a self-collected dataset comprising 180 video clips recorded at different resolutions in both indoor and outdoor environments. Each video clip in this dataset is labeled as one of six action classes: walking, greeting, shaking hands, dancing, sitting, and standing. The video clips labeled as videos containing the *shaking hands* action contain two persons, whereas the rest of the video clips contain only one person. Similar to the UCF Sports dataset, the number of clips is not the same for each action class. The dataset is divided into a training and validation subset of 140 video clips and a test set of 40 video clips.

The face recognition component of the proposed methodology was tested using the standard Labeled Faces in the Wild (LFW) dataset [19,46] and the self-collected HMTA dataset for face recognition tasks, comprising 100 identities. Figure 3 shows some sample face images from the HMTA dataset. The LFW is perhaps the most widely used benchmark dataset for unconstrained face recognition in videos and images. It comprises 13,233 images with 5749 people [19].



Diving-Side

Golf-Swing-Back

Kicking-Side

Swing-Bench

Riding-Horse

Walk-Front

SkateBoarding-Front

Run-Side

Figure 2. Snapshots of different videos from the UCF Sports dataset depicting the different types of action classes.

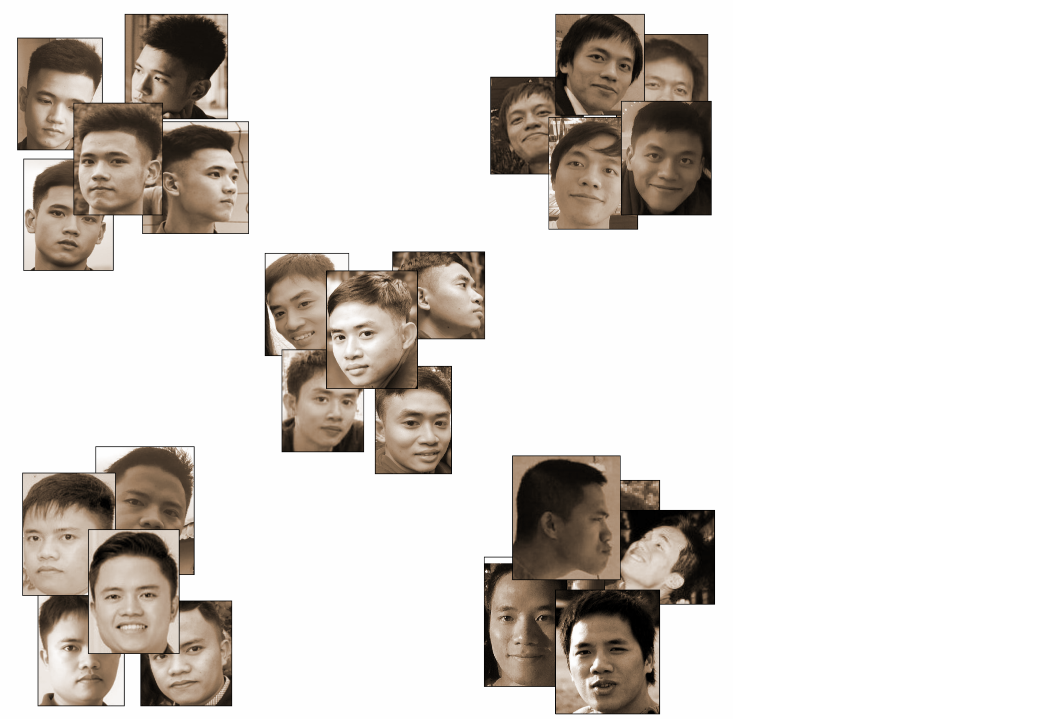


Figure 3. Sample face images from the self-collected HMTA dataset for face recognition.

Moreover, the emotion recognition component of the proposed framework was tested on the FER2013 dataset [20] for facial emotion recognition, which comprises more than 32,000 images 48 × 48 pixels in size each that are split across 7 different types of emotions.

The proposed methodology was tested and compared with other techniques by implementing it on a self-developed, intelligent, humanoid robot named *BonBon*, shown in Figure 4. (The word *BonBon* is used as a common nickname for kids in Vietnam and is used in the meaning of *cute*. The word *Bon* has its roots in French and means *good*). *BonBon* has been developed to support the instruction of the English language to non-native speakers in elementary schools. It has the appearance of a boy, as shown in Figure 4, is 1.2 meters tall and weighs 35 kilograms. The robot consists of an upper body with 21 degrees of freedom, a head, two arms, two hands, a ribcage and a mobile platform with three omnidirectional wheels. The robot's body houses two computers that control the entire operation of the robot, including motion planning, voice recognition and synchronization, face, emotion and action recognition, receiving and processing commands from an external control and monitoring station, receiving and processing input from a microphone array, cameras, and other sensors and receiving and sending signals to the mobile module and the upper body controllers. The microphones, speakers and cameras are located on the head and chest of the robot to perform voice communication and image acquisition functions. A touch screen is provided on the robot's chest to allow people to interact with and display necessary information. *BonBon* can communicate with people through a voice and can perform actions such as greetings, dancing, singing and expressing emotions verbally. The proposed methodology presented in Section 3 and the techniques to which it was compared in Section 5 were implemented in *BonBon* for comparison and real-time performance evaluation.

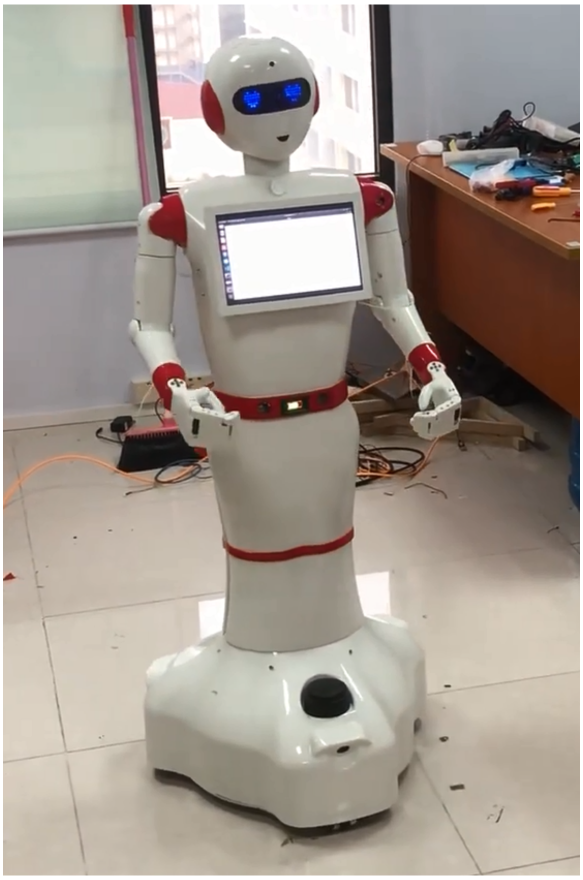


Figure 4. The self-developed intelligent humanoid robot *BonBon* that was used for testing the proposed framework.

BonBon houses two computers: an embedded computer with an Intel Core i7-8559U CPU to manage all the motion control tasks and safety functions of the robot and a central control computer (i.e., an NVIDIA Jetson AGX Xavier with a 512-core GPU and an 8-core, 64-bit ARM CPU with 16Gb RAM) for managing the vision and speech processing systems as well as the task planning function of the robot. The robot control software was developed using the robotic operating system (ROS) [47]. Dlib (dlib.net, accessed on 8 December 2021), which is a modern C++ toolkit, is used for face detection and alignment.

5. Results and Discussion

As discussed in Section 3.1, the proposed methodology began by first extracting key frames from the video clips in the datasets used in this study as described in Section 4. The key frames were extracted based upon the gradient of the total motion metric given in Equation (1) across consecutive frames in the video. The total motion metric was computed using the Gunnar Farneback algorithm and dense optical flow [40]. Before computing the changes in the intensity $I(x, y, t)$ of pixels across successive frames, they were first converted into the hue, saturation and value (HSV) format, and then the pixels undergoing intensity changes in successive frames were highlighted for better visibility. From a set of flow vectors, the algorithm calculated the amount and direction of optical flow. The total motion metric $M(t)$, as given in Equation (1), was then computed to quantify the motion over both the horizontal and vertical directions for a frame t. The gradient of the optical flow function $M(t)$ was then computed to select the key frames based upon the local extrema of the total motion in the video. Figure 5 shows the gradient of the total motion metric for a walking video from the UCF Sports dataset [48].

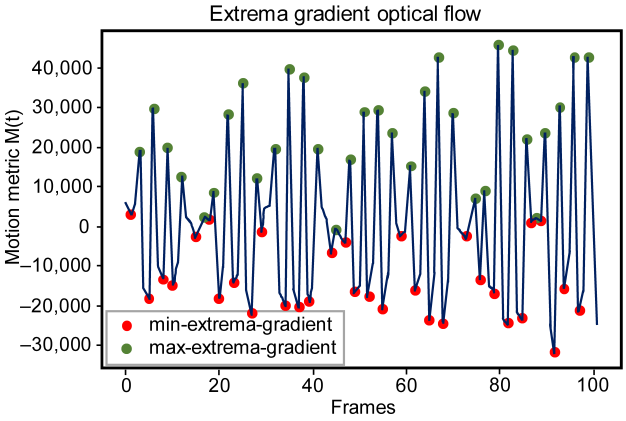


Figure 5. The gradient of the total motion metric for a walking video from the UCF Sports dataset. The red dots are the local minima, whereas the green dots are the local maxima of the gradient function.

The number of local extrema depends upon the contents of the video (i.e., videos that record relatively more complex and abrupt movements or activities would have more local extrema, whereas videos that record relatively simpler and smoother activities and movements would have fewer local extrema). Therefore, to capture the most significant changes in motion that were important for the identification of an action, we selected k frames corresponding to the highest k local extrema. Figure 6 shows the selected key frames, which were k in number, for a video from the UCF Sports dataset, labeled as a *walking video.*

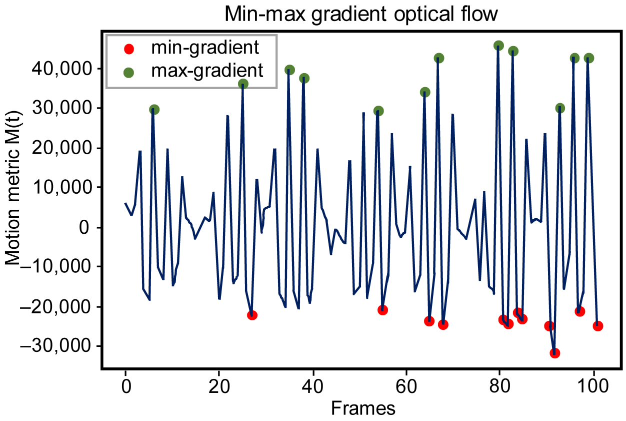


Figure 6. The selected *k* frames for a video from the UCF Sports dataset, labeled as a *walking video* based upon the highest *k* values of the gradient of the total motion metric. The red dots show the local minima, whereas the green dots show the local maxima.

The gradient of the total motion metric allowed the proposed algorithm to sum up the motion in successive frames of a video and to capture the motion information that better represented an action. Moreover, it also rendered a proposed method that was simpler and more efficient as motion information from the entire video sequence was squeezed into a few key frames (i.e., *k* number of key frames, which for the purpose of this study was taken to be 24). Twenty-four key frames amount to roughly 10–20% of the total frames in each video in the UCF Sports dataset. This is also helpful in reducing the dimensionality of the feature vectors comprising the human keypoints extracted from these key frames using pose estimation models, as we discarded 80–90% of the frames in each video. That translated into relatively faster training and more accurate classification of different actions by the DDNet in the latter stages.

Once the key frames were extracted from a video, pose estimation models were applied to them to extract the human skeleton features, also known as human keypoints. These keypoints represent the coordinates of the different joints in a given frame that constitute a pose. In this study, two pose estimation models (i.e., HRNet-W48 [22] and OpenPose [12]) were used to extract the human keypoints from the extracted key frames. The HRNet with 48 channels extracted 17 keypoints from the input images, which were resized to 384 × 288 pixels, whereas OpenPose extracted 18 keypoints from the input images, which were resized to 368 × 368 pixels. Certain hyperparameters of the pose estimation models (i.e., the number of keypoints in the action capture model) were tuned to find out the values that would yield the best results. Different numbers of keypoints (i.e., 11, 13, 15 and 17) were extracted from the key frames using the pose estimation models, and the best results were obtained with 17 keypoints. Furthermore, while applying the HRNet model for pose estimation, we proposed using Yolov5 Tiny instead of YoloV3 for detection of human objects in a frame to improve both speed and accuracy.

Figure 7 shows an example of the human keypoints extracted from a key frame of one of the videos in the self-collected HNH dataset. Multiple human objects can be seen in Figure 7 along with other objects, and it can also be observed that the pose estimation model correctly identified the human objects and the relative positions of their joints for subsequent action recognition. In addition to action recognition, the proposed framework simultaneously performed face and emotion recognition on the selected key frame in real time, the results of which are depicted in the text labels of the bounding boxes in Figure 7. Moreover, it is pertinent to mention here that the video clip to which the frame in Figure 7 belongs was not used to train the DDNet model for action recognition. Instead, the DDNet model was pretrained using a subset of our self-collected HNH dataset for action recognition. Similarly, the face recognition model ArcFace [44] was pretrained using the self-collected HMTA dataset for face recognition. The training data for face recognition did not include the video clip to which the frame in Figure 7 belongs. Likewise, the residual masking network [45] that we employed for emotion recognition was pretrained using a subset of the self-collected HMTA dataset and tested both on the FER2013 dataset [20] and the HMTA dataset. A description of these datasets is available in Section 4.

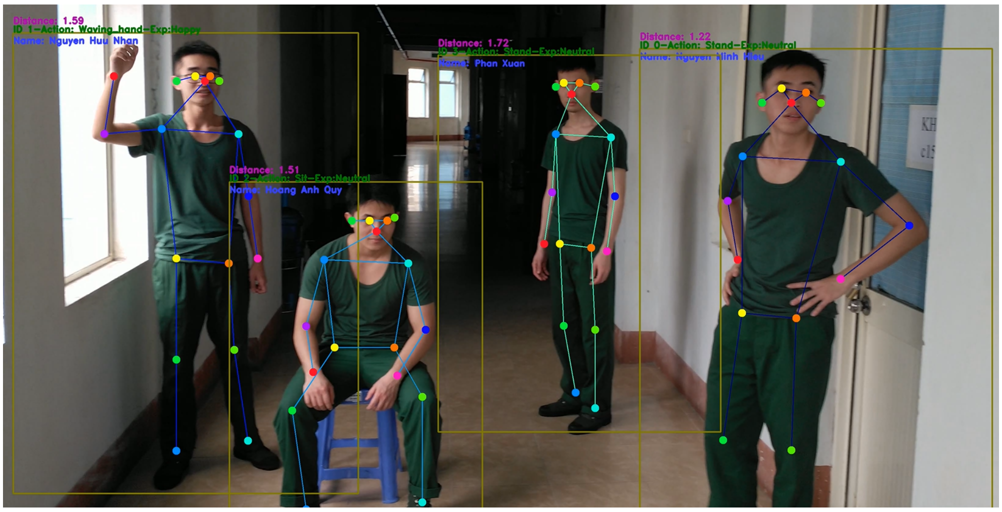


Figure 7. An example of the human keypoints extracted by the proposed method from a key frame of one the videos in the self-collected datasets. The proposed framework simultaneously performs action, face and emotion recognition on the selected key frame in real time. The human objects in this image have been labeled with their estimated distances, identified actions, emotions and identities through face recognition. For instance, from left to right, the identified actions and emotions for each human object are (waving hand, happy), (sitting, neutral), (standing, neutral), and (standing, neutral).

The keypoints extracted using both HRNet and OpenPose were used to train the DDNet [30] for action recognition. The human skeletal keypoints were used to train the DDNet directly without any preprocessing. The DDNet model was trained using Adam [49] as the optimizer with the parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Different values were tried for the learning rate, including 1×10^{-2}, 1×10^{-3} and 1×10^{-4}. The training was performed for a different number of epochs ranging from 50 to 600 with a batch size of 64. During training, the value of the loss function showed a steady decrease. On the validation set, the best accuracy was achieved at a learning rate of 1×10^{-4} and at the 350th epoch. Hence, we used a learning rate of 1×10^{-4} and continued the training process for 350 epochs. Moreover, the human skeleton features extracted using HRNet yielded better results in terms of action recognition on several datasets; however, it was not as fast as OpenPose, especially when the number of human objects in a video frame increased.

5.1. Action Recognition Performance

The performance of the proposed methodology for action recognition on the UCF Sports dataset was evaluated using the leave-one-out (LOO) cross validation scheme (i.e., for each action class), and the DDNet was trained on all the videos for that action class except one, which was left out for testing. This process was repeated in a cyclical manner for every video of each action class in the dataset (i.e., for a particular action class, each video for that action class was used one by one for testing, whereas the remaining videos were used for training the DDNet). For each cycle, the weights of the DDNet were initialized (i.e., the weights from the previous cycle were not considered). The overall recognition accuracy for a particular action class was determined by calculating the average accuracy for all the videos. The average accuracy over all the iterations of this process for each action class was then used to measure and compare the action recognition performance of different methods. The confusion matrix of the proposed methodology for the UCF Sports dataset is given in Figure 8, which shows the performance of the proposed methodology for the *10* different action categories in the UCF Sports dataset. In Figure 8, these action categories are represented by labels from 0 to 9. Table 1 provides a description or specification of these labels (i.e., the action category that each of these labels corresponded to along with the number of video clips in the dataset for each action category). The blue color bar to the right of the confusion matrix in Figure 8 represents the number of video clips corresponding to each action category, which is then highlighted in the confusion matrix using the appropriate color from this blue color bar corresponding to the number of video clips for each action category. We can observe in Figure 8 that for 7 out of 10 action classes, the proposed methodology achieved 100% classification accuracy. However, for three action categories—*horse-riding*, *swing-side* and *swing-golf*—its classification performance deteriorated. This was perhaps the result of similarity between the spatiotemporal changes for these action categories.

Table 1. Description of the different labels used in Figure 8.

Label	Specification	No. of Video Clips
0	Diving	14
1	Kicking	20
2	Lifting	6
3	Horse-riding	12
4	Running	13
5	Swing-side	13
6	Skateboarding	20
7	Swing-bench	12
8	Walking	22
9	Swing-golf	18

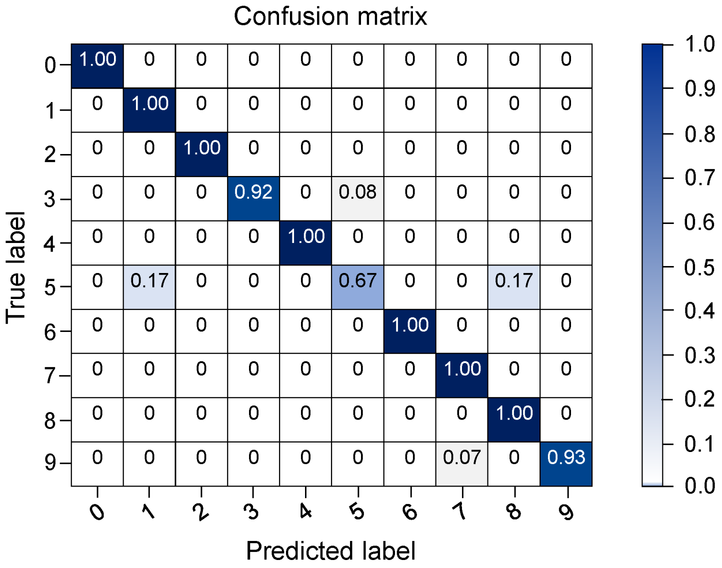


Figure 8. The confusion matrix of the proposed methodology for the UCF Sports dataset.

In Table 2, we provide a comparison of the action recognition performance of the proposed method and several state-of-the-art methods [33–38] on the UCF Sports dataset. The proposed method achieved an accuracy of 93.1%, which was better than all the state-of-the-art methods, except for the one proposed in [38] that uses 3D convolutional neural network (CNN)-based fused feature maps and a long short-term memory (LSTM) recurrent neural network (RNN). The method proposed in [38] achieved only a marginal performance gain of 0.8% but at a much higher computational cost. The proposed key frame-based skeleton feature estimation and action recognition network (KFSENet) had comparatively lower computational complexity but achieved action recognition performance that was better than most of the existing methods. Furthermore, our results also indicate the advantages of key frame extraction, as key frame extraction not only reduced the computational cost of extracting skeleton features, since only 10–20% of the frames in a video have to be processed, but it also improved the action recognition performance by almost 4%. Moreover, it was also helpful to increase the frame rate from 18 to 24 frames per second (FPS) (i.e., the proposed framework can process video input at a higher frame rate of 24 FPS for action, emotion and face recognition in real-time). Extraction of skeleton features from the key frames alone reduced the size of the training vectors by leaving out redundant and potentially irrelevant features, which resulted in better generalization performance and reduced training time for the DDNet.

Table 2. The performance evaluation of different methods for action recognition on the UCF Sports dataset.

Method	Accuracy (%)
Harris3D detector + HOG/HOF descriptors [33]	78.1
ST-SIFT detector + HOG3D descriptors [36]	80.5
Dense sampling + HOG/HOF descriptors [33]	81.6
MBH + dense trajectories [37]	84.2
Stacked convolutional independent subspace analysis [35]	86.5
Kovashka et al. [34]	87.3
3DCNN-based fused feature maps + LSTM [38]	93.9
The proposed method without key frame extraction	89.3
The proposed method with key frame extraction	93.1

Table 3 presents the comparison of different methods for action recognition on the self-collected HNH dataset. The proposed key frame-based skeleton feature estimation and action recognition network (KFSENet) achieved an accuracy of 98.9%. The other two methods were different from the proposed method in only one respect: the human skeleton features for those two methods were extracted from all the frames in the video, whereas for the proposed KFSENet, the human skeleton features were extracted only from the key frames. Thus, key frame extraction had a significant contribution toward the comparatively better accuracy of the proposed method. Moreover, the results in Table 3 also demonstrate that HRNet yielded higher accuracy than OpenPose; however, Openpose could achieve a relatively higher frame rate compared with HRNet.

Table 3. Performance evaluation of different methods for action recognition on the self-collected HNH dataset.

Method	Accuracy (%)
HRNet + DDNet	94.4
OpenPose + DDNet	92.8
The proposed method with key frame extraction	98.9

5.2. Face and Emotion Recognition Performance

The face recognition component of the proposed methodology was evaluated using the LFW dataset [19,46], which is a benchmark dataset that was described in Section 4 and is the most widely used for unconstrained face recognition in images and videos. The performance results for face recognition were reported while following the unrestricted images with labeled outside data protocol. We proposed to use ArcFace for face recognition in our integrated approach to robot vision, and it achieved an accuracy of 99% [44] on the LFW dataset [19]. When evaluated using the self-collected HMTA dataset with facial images for 100 people, ArcFace achieved an accuracy of 99.2%. Moreover, for emotion recognition, we employed a residual masking network [45] for the extraction of facial emotion features and their subsequent classification. The residual masking network that we utilized in our study comprises four layers, where each layer consists of a residual layer and a masking block. When tested on the FER2013 dataset [20] for facial emotion recognition, the residual masking network resulted in an accuracy of 74.14%. We compared the emotional recognition performance of the residual masking network with a pretrained IR50 backbone network [44,50] combined with a softmax classifier. The model was trained using stochastic gradient descent (SGD) as the optimizer to optimize a logarithmic loss function at a learning rate of 1×10^{-3}.The training was performed for 200 epochs with a batch size of 64. The IR50 backbone network was used to extract features from the images for facial emotion recognition. These features were then subsequently used to train the softmax classifier. The IR50 backbone and softmax classifier delivered an accuracy of 73.04% on the FER2013 dataset [20] for facial emotion recognition, as given in Table 4.

Table 4. Performance evaluation of different methods for facial emotion recognition on the FER2013 dataset.

Method	Accuracy (%)
IR50 Backbone Network [50] + Softmax Classifier	73.04
Residual Masking Network	74.14

Face and emotion recognition make the proposed framework very well-suited for social robots, which are expected to have a more personal interaction with humans.

6. Conclusions

In this paper, we proposed an integrated approach to robot vision—the key frame-based skeleton feature estimation and action recognition network (KFSENet)—which incorporates action recognition with face and emotion recognition to afford social robots a more personal interaction with humans. We propose a key frame-based approach to the extraction of human skeleton features or keypoints using pose estimation models. The key frames are selected based upon the gradient of a proposed total motion metric that is computed using dense optical flow. We use the human skeleton features extracted from the key frames to train a deep neural network (i.e., the double-feature double-motion network (DDNet)) for action recognition. The proposed KFSENet utilizes a simpler model to learn about different action classes, is computationally simpler and more efficient and yields better action recognition performance compared with existing methods. The use of key frames allows the proposed method to eliminate unnecessary and redundant frames, thereby improving classification accuracy and decreasing the computational cost. The proposed method was tested and compared to existing methods using both standard benchmark datasets and self-collected datasets. Furthermore, we integrated face and emotion recognition into the proposed framework to enable social robots to engage in more personal interaction with humans.

Author Contributions: Conceptualization, H.H.H.; Data curation, V.-A.T.; Formal analysis, T.-H.N.; Funding acquisition, D.-S.L. and T.-H.N.; Investigation, T.-H.N.; Methodology, D.-S.L.; Resources, H.H.H.; Software, H.-H.P. and V.-A.T.; Validation, H.-H.H. and D.-Q.N.; Writing—original draft, D.-S.L., H.-H.P. and D.-Q.N.; Writing—review and editing, D.-Q.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Vietnam national project No. KC-4.0-04/19-25.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pachidis, T.; Vrochidou, E.; Kaburlasos, V.; Kostova, S.; Bonković, M.; Papić, V. Social robotics in education: State-of-the-art and directions. In Proceedings of the International Conference on Robotics in Alpe-Adria Danube Region, Patras, Greece, 6–8 June 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 689–700.
2. Akalin, N.; Loutfi, A. Reinforcement learning approaches in social robotics. *Sensors* **2021**, *21*, 1292. [CrossRef] [PubMed]
3. Sheridan, T.B. A review of recent research in social robotics. *Curr. Opin. Psychol.* **2020**, *36*, 7–12. [CrossRef] [PubMed]
4. Share, P.; Pender, J. Preparing for a robot future? Social professions, social robotics and the challenges ahead. *Ir. J. Appl. Soc. Stud.* **2018**, *18*, 4.
5. Lei, Q.; Du, J.X.; Zhang, H.B.; Ye, S.; Chen, D.S. A survey of vision-based human action evaluation methods. *Sensors* **2019**, *19*, 4129. [CrossRef] [PubMed]
6. Ren, B.; Liu, M.; Ding, R.; Liu, H. A survey on 3d skeleton-based action recognition using learning method. *arXiv* **2020**, arXiv:2002.05907.
7. Yasin, H.; Hussain, M.; Weber, A. Keys for action: An efficient keyframe-based approach for 3D action recognition using a deep neural network. *Sensors* **2020**, *20*, 2226. [CrossRef] [PubMed]
8. Yan, X.; Gilani, S.Z.; Feng, M.; Zhang, L.; Qin, H.; Mian, A. Self-supervised learning to detect key frames in videos. *Sensors* **2020**, *20*, 6941. [CrossRef] [PubMed]
9. Yan, X.; Gilani, S.Z.; Qin, H.; Feng, M.; Zhang, L.; Mian, A. Deep keyframe detection in human action videos. *arXiv* **2018**, arXiv:1804.10021.
10. Phan, H.H.; Vu, N.S.; Nguyen, V.L.; Quoy, M. Action recognition based on motion of oriented magnitude patterns and feature selection. *IET Comput. Vis.* **2018**, *12*, 735–743. [CrossRef]
11. Gong, D.; Medioni, G.; Zhao, X. Structured time series analysis for human action segmentation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1414–1427. [CrossRef] [PubMed]
12. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
13. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 466–481.
14. Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.P.; Xu, W.; Casas, D.; Theobalt, C. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph. (TOG)* **2017**, *36*, 1–14. [CrossRef]
15. De Smedt, Q.; Wannous, H.; Vandeborre, J.P.; Guerry, J.; Saux, B.L.; Filliat, D. 3d hand gesture recognition using a depth and skeletal dataset: Shrec'17 track. In Proceedings of the Workshop on 3D Object Retrieval, Lyon, France, 23–24 April 2017; pp. 33–38.
16. Hou, J.; Wang, G.; Chen, X.; Xue, J.H.; Zhu, R.; Yang, H. Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
17. Zolfaghari, M.; Oliveira, G.L.; Sedaghat, N.; Brox, T. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2904–2913.
18. Soomro, K.; Zamir, A.R. Action recognition in realistic sports videos. In *Computer Vision in Sports*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 181–208.
19. Huang, G.B.; Mattar, M.; Berg, T.; Learned-Miller, E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In Proceedings of the Workshop on faces in'Real-Life'Images: Detection, Alignment, and Recognition, Marseille, France, 12–18 October 2008.
20. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the International Conference on Neural Information Processing, Daegu, Korea, 3–7 November 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 117–124.
21. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
22. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
23. De Smedt, Q.; Wannous, H.; Vandeborre, J.P. Skeleton-based dynamic hand gesture recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1–9.
24. Choutas, V.; Weinzaepfel, P.; Revaud, J.; Schmid, C. Potion: Pose motion representation for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7024–7033.
25. Sevilla-Lara, L.; Liao, Y.; Güney, F.; Jampani, V.; Geiger, A.; Black, M.J. On the integration of optical flow and action recognition. In *Proceedings of the German Conference on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 281–297.
26. Gowda, S.N.; Rohrbach, M.; Sevilla-Lara, L. Smart frame selection for action recognition. *arXiv* **2020**, arXiv:2012.10671.
27. Sawant, C. Human activity recognition with openpose and Long Short-Term Memory on real time images. *EasyChair* **2020**, Preprint.
28. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.

29. Lee, I.; Kim, D.; Kang, S.; Lee, S. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1012–1020.
30. Yang, F.; Wu, Y.; Sakti, S.; Nakamura, S. Make skeleton-based action recognition model smaller, faster and better. In Proceedings of the ACM Multimedia Asia, Beijing, China, 15–18 December 2019; pp. 1–6.
31. Müller, M.; Röder, T.; Clausen, M.; Eberhardt, B.; Krüger, B.; Weber, A. Documentation Mocap Database hdm05. 2007. Available online: https://resources.mpi-inf.mpg.de/HDM05/ (accessed on 12 April 2022)
32. CMU Graphics Lab Motion Capture Database. Available online: http://mocap.cs.cmu.edu/ (accessed on 12 April 2022).
33. Wang, H.; Ullah, M.M.; Klaser, A.; Laptev, I.; Schmid, C. Evaluation of local spatio-temporal features for action recognition. In Proceedings of the Bmvc 2009-British Machine Vision Conference, London, UK, 7–10 September 2009; BMVA Press: Swansea, UK, 2009; pp. 1–124.
34. Kovashka, A.; Grauman, K. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 2046–2053.
35. Le, Q.V.; Zou, W.Y.; Yeung, S.Y.; Ng, A.Y. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3361–3368.
36. Al Ghamdi, M.; Zhang, L.; Gotoh, Y. Spatio-temporal SIFT and its application to human action classification. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 301–310.
37. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.L. Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vis.* **2013**, *103*, 60–79. [CrossRef]
38. Arif, S.; Wang, J.; Ul Hassan, T.; Fei, Z. 3D-CNN-based fused feature maps with LSTM applied to action recognition. *Future Internet* **2019**, *11*, 42. [CrossRef]
39. Kim, J.; Lee, D. Activity Recognition with Combination of Deeply Learned Visual Attention and Pose Estimation. *Appl. Sci.* **2021**, *11*, 4153. [CrossRef]
40. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In Proceedings of the Scandinavian Conference on Image Analysis, Halmstad, Sweden, 29 June–2 July 2003; Springer: Berlin/Heidelberg, Germany, 2003; pp. 363–370.
41. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
42. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L. Sphereface: Deep hypersphere embedding for face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 212–220.
43. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. Cosface: Large margin cosine loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5265–5274.
44. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4690–4699.
45. Pham, L.; Vu, T.H.; Tran, T.A. Facial expression recognition using residual masking network. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 4513–4519.
46. Huang, G.; Mattar, M.; Lee, H.; Learned-Miller, E. Learning to align from scratch. *Adv. Neural Inf. Process. Syst.* **2012**, *25* .
47. Laboratory, S.A.I. Robotic Operating System. 2018. Available online: https://www.ros.org (accessed on 8 December 2021).
48. Rodriguez, M.D.; Ahmed, J.; Shah, M. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, 23–28 June 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–8.
49. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
50. Wang, Q.; Zhang, P.; Xiong, H.; Zhao, J. Face. evolve: A high-performance face recognition library. *arXiv* **2021**, arXiv:2107.08621.