IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 4, 2020, accepted February 26, 2020, date of publication March 4, 2020, date of current version March 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2978249

Vision-Based Fall Detection With Multi-Task Hourglass Convolutional Auto-Encoder

XI CAI, SUYUAN LI, XINYUE LIU, AND GUANG HAN
School of Computer and Communication Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China

Corresponding author: Guang Han (coldlight919@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61601108 and Grant 61701098, and in part by the Natural Science Foundation of Hebei Province under Grant F2018501084.

ABSTRACT Fall detection is a hot research issue in intelligent video surveillance. Falls can generate physical and psychological damage, especially for the elderly. Different from most conventional vision-based fall detection methods typically relying on hand-crafted features, fall detection methods based on deep learning techniques can automatically learn features and hence have got widespread concern recently. However, as deep networks are increasingly applied to fall detection, the problem of information loss in the deep networks can not be ignored, because this will ultimately affect the performance of fall detection. To solve the above problem, we propose a vision-based fall detection method using multi-task hourglass convolutional auto-encoder (HCAE). In this method, hourglass residual units (HRUs) are introduced into the encoder of the HCAE to extract multiscale features by expanding receptive fields of neurons. A multi-task mechanism is presented to enhance the feature representativeness of the network by completing an auxiliary task of frame reconstruction while realizing the main task of fall detection. Experimental results demonstrate that, the proposed method can effectively achieve accurate fall detection with the shallow-layer network, and outperforms several state-of-the-art methods.

INDEX TERMS Fall detection, deep learning, hourglass convolutional auto-encoder (HCAE), hourglass residual unit, multi-task mechanism.

I. INTRODUCTION

Fall is a sudden, involuntary, unintentional postural change which may endanger people's lives, especially for elderly people. Due to degradation of body functions, the elderly have a greater possibility of falls in their daily lives. Statistics have shown that one-third of the elderly over the age of 65 fall every year [1]. Nowadays, fall is becoming one of the most terrible accidents threatening the health of the elderly. Specifically, falls may lead to injuries such as sprains, bruises and lacerations, and more seriously, may even result in disabilities or deaths. It has been reported that, falls cause over 37.3 million severe injuries and 646,000 deaths yearly, and hence have become a global public health issue [2]. It has become very important to develop intelligent surveillance systems, especially vision-based systems, which can automatically monitor and detect falls [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

Plenty of research has been done to develop systems and methods for highly-accurate automatic fall detection. The intelligent systems and methods that can be widely promoted should have good user-friendliness [4], such as easy to use, non-invasive, minimally restricting the user's normal activities, avoiding high-frequency radiation especially for people who are wearing pacemakers and sensitive to electromagnetic interference, *etc.*. Therefore, in comparison to wearable-sensor-based and ambient-sensor-based technologies, vision-based fall detection technology will have good universality and feasibility in the future promotion [3], because of its advantages in rich monitoring information, non-contact monitoring manner, and zero electromagnetic interference monitoring environment.

Different from traditional vision-based fall detection methods relying on hand-crafted features, vision-based fall detection methods using deep learning networks can automatically extract features for detection after learning and analyzing a mass of data, and hence have recently received widespread attention [5], [6]. Deep networks have been increasingly

applied to fall detection [7]–[11]. However, information may be lost after multiple layers of a deep network, which will lead to representativeness reduction of the feature in the network and further affect the accuracy of fall detection.

To solve the above problem, in this paper, we propose a vision-based fall detection method using multi-task hourglass convolutional auto-encoder (HCAE). In the proposed HCAE, hourglass residual units (HRUs) are introduced into the encoder to capture multiscale features by neurons with expanding receptive fields. Addtionally, a multi-task mechanism, including a main task of fall judgment and an auxiliary task of frame reconstruction, is proposed in which the auxiliary task is used to enhance the representativeness of the feature in the network and further help complete the main task of fall detection. Experimental results prove that the proposed method can produce accurate detection results with the shallow-layer network and outperforms several state-of-the-art methods.

The rest of this paper is organized as follows. In Section II, we briefly review the related work. In Section III, we present our fall detection method using the multi-task HCAE. Section IV describes our experimental results and verifies the validity of the proposed method through comparing its performance with other five fall detection methods. Finally, the conclusion is drawn in Section V.

II. RELATED WORK

Fall detection is one of the hot issues in the field of public healthcare. According to devices involved, the existing methods can be roughly categorized into three classes [3], [12]: methods based on wearable sensors, methods based on ambient sensors, and methods based on video cameras.

Wearable sensors mainly include tilt switch, accelerometer, gyroscope and barometric pressure sensor [13]. These sensors are usually attached to the chest, waist or wrist to obtain acceleration or other motion information. Lai *et al.* placed triaxial acceleration sensors in several key body parts to analyze whether horizontal acceleration and normal acceleration of human body exceed regular ranges [14]. Tmaura *et al.* used a triggering airbag to build a fall detection system with accelerometers and gyroscopes [15]. The wearable sensors are inexpensive and able to measure activities directly. However, the wearable sensors must be firmly fixed in specific areas of a body, and may cause inconvenience to the elderly, especially in daily home monitoring.

Ambient sensors are used to collect information from monitoring environments when a fall event occurs. Piezoelectric sensors, acoustic sensors and infrared sensors are commonly used ambient sensors. Since falls generally cause vibrations and sounds in the duration of hitting the ground, vibrations and sounds are usually used to detect fall events. Li *et al.* used a circular microphone array to collect sound signals of the environment to detect falls [16]. In [17], a fall detection method using piezoelectric sensors to measure floor vibration was introduced. Ambient sensors can be installed optionally into environments without interfering with elders' daily lives. However, data collected by these sensors is limited and susceptible to external noises, which may lead to low detection accuracy of fall detection methods based on ambient sensors.

Video cameras have the advantages of rich information, non-contact acquisition, no electromagnetic interference, good user experience, low cost and multi-tasking parallelism, and have been widely employed for fall detection in this decade. Traditionally, vision-based fall detection methods usually segment human silhouettes from videos captured by RGB cameras or Kinects, and extract features to discriminate falls. Commonly used hand-crafted features mainly include shape related features, head trajectories and human motion features [3], [18].

Shape related feature is often used as an important basis for detecting a fall. Wu *et al.* detected falls according to the ratio of the height to bottom of a triangle formed by the head and two feet [19]. In [20], dynamic shapes were represented as points moving on a unified Riemannian manifold and employed for fall detection. A shape descriptor named silhouette orientation volume was used to represent actions and classify falls in [21]. In [22], ratios of five partial occupancy areas of body were used as the feature and input to four different machine learning algorithms to compare the performance of fall discrimination. In [23], fall detection was first achieved by using a MEWMA strategy according to shape feature of five partial occupancy area extracted from each frame, and then a SVM classifier was used to further distinguish falls and fall-like behaviors.

Head trajectory is also a feature of interest in fall detection, because the head is generally not easily occluded. Bian *et.al* proposed a fall detection method based on 3D trajectory of head joint extracted by a pose-invariant randomized decision tree [24]. In [25], particle filter was employed to track head, and 3D horizontal and vertical velocities of the head were utilized to detect the occurrence of a fall. In [26], 3D trajectory of head was tracked by a single calibrated camera, and used to compute velocity characteristics for fall detection.

Human motion feature is often utilized as the clue for fall detection. In [27], a fall detection method was proposed based on combination of integrated time motion image and eigenspace technique. Su *et al.* extracted motion features from spatio-temporal interest points to indicate degree of impact shock and body vibration and then achieved fall discrimination [28]. In [29], speed and direction of body motion were computed according to optical flows of interest points, and then used to discriminate fall events.

With the development of deep learning theory, deep learning technology has been applied to fall detection in recent years. In [5], convolutional neural network (CNN) was applied to each frame of a video to learn human shape deformation features that describe different postures of the human and determine if a fall occurs. In [6], 3D-CNN combined with a long short-term memory (LSTM) scheme was developed to extract motion information within a region of interest to implement fall detection.

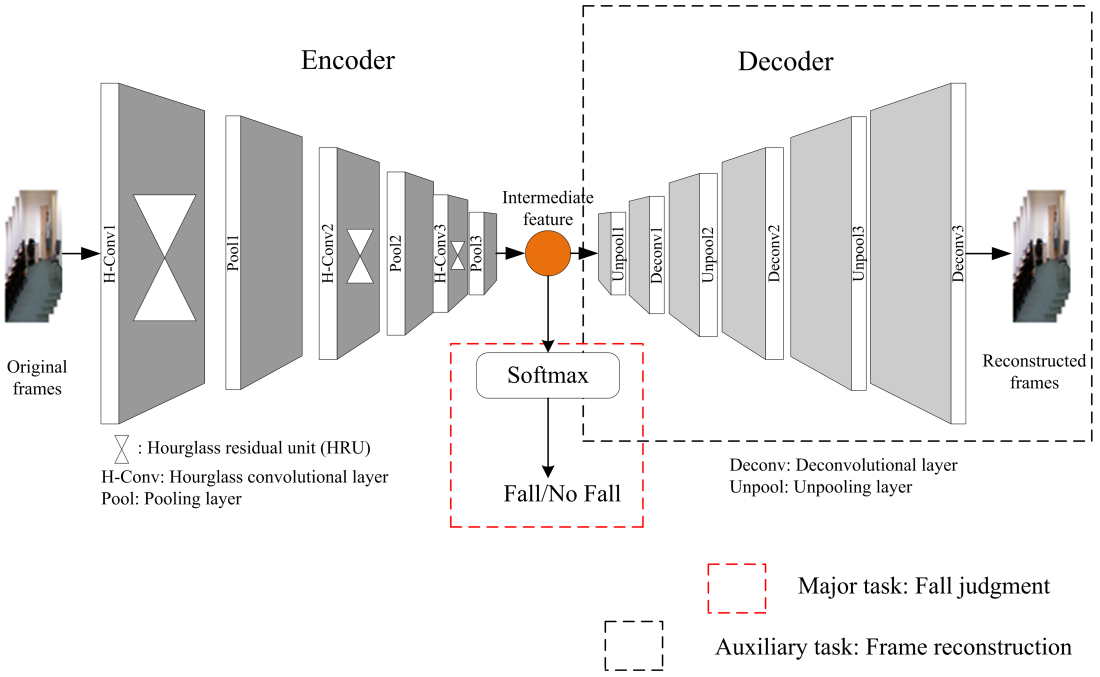

FIGURE 1. Framework of the proposed HCAE-FD method.

Most existing methods based on deep learning tend to utilize deeper neural networks to analyze and learn features from the mass data. In [7], a pre-trained deep AlexNet combined with transfer learning was used to detect fall events. Co-saliency-enhanced deep recurrent convolutional networks were utilized for fall detection [8]. VGG-16 net was employed to receive optical flows and to complete fall discrimination in [9]. In [10], VGG-16 net combined with an attention-guided LSTM was adopted to capture spatio-temporal features for fall detection. In [11], an extremely deep residual network and a recurrent neural network with LSTM were utilized for fall detection.

Comparing with conventional video-based fall detection methods, the methods based on deep learning can automatically learn proper features for detection after analyzing a mass of data, with no need to extract hand-crafted features in advance. However, information may be lost after going through multiple layers of a deep network, which will decrease representativeness of the feature in the network and further affect the accuracy of fall detection. Accordingly, we propose a fall detection algorithm based on hourglass convolutional auto-encoder (HCAE-FD) in this paper. To capture more abundant information, we adopt HRUs in the encoder of the HCAE to extract multiscale features by expanding receptive fields of neurons. Besides, a multi-task mechanism is designed to improve the feature representativeness of the network by assigning a secondary task of frame reconstruction along with a main task of fall detection.

III. PROPOSED METHOD

In this work, a novel method based on the HCAE with HRUs and a multi-task mechanism is proposed for fall detection. Framework of the proposed HCAE-FD method is illustrated in Fig. 1. In the first phase, original video frames are input into the HCAE, and the encoder of HCAE with the HRU is used to obtain an intermediate feature containing more abundant behavior information. In the second phase, the intermediate feature is utilized in a multi-task mechanism, in which fall detection is completed as the main task by classifiying the intermediate feature using a classifier and frame reconstruction is realized as the secondary task by the decoder to enhance the representativeness of the intermediate feature. Optimal HCAE model can be gained after optimizing a multi-task loss function in a training process, and then can be utilized to detect fall events in a testing process.

A. HOURGLASS CONVOLUTIONAL AUTO-ENCODER

The HCAE consists of an encoder and a decoder. In order to preserve more abundant information at multiple scales, we introduce HRUs into the encoder of the HCAE, and improve the convolutional layers into hourglass convolutional layers. Compared with a conventional convolutional unit, the HRU can expand receptive field and capture multiscale features. In this way, the proposed HCAE can avoid loss of features within a shallow-layer network (with three hourglass convolutional layers), and hence can further help improve the accuracy of the HCAE-FD method.

1) HOURGLASS RESIDUAL UNIT

The HRU proposed in the HCAE is adopted to capture image information at multiple scales. When identifying features like faces and hands, local evidence is imperative; nonetheless, for a judgment of a behavior, a coherent understanding of the whole behavior is required. As a consequence,

we utilize HRUs, expanding receptive field (i.e. contextual regions of neurons), to obtain multiscale features. In this work, a single pipeline with skip layers is used to preserve spatial information at each resolution, and a HRU consists of three branches, i.e. an identity mapping branch, an hourglass mapping branch, and a residual mapping branch (as illustrated in Fig. 2).

The identity mapping branch is used to retain information in original resolution; the hourglass mapping branch is adopted to extract global features in low-resolution images; the residual mapping branch is employed to capture local features in high-resolution images (in comparison to the hourglass mapping branch). Synthetically, three branches are integrated to achieve a comprehensive understanding of the overall human body posture and motion.

To be specific, the HRU used in the HCAE can be expressed as:

$$\begin{aligned} z &= z_A + z_B + z_C \\ &= \sigma(\boldsymbol{W}^I * \boldsymbol{x}) + u\{\sigma\{\boldsymbol{W}^{H_2} * \sigma[\boldsymbol{W}^{H_1} * p(\boldsymbol{x})]\}\} \\ &\quad + \sigma\{\boldsymbol{W}^{R_3} * \sigma[\boldsymbol{W}^{R_2} * \sigma(\boldsymbol{W}^{R_1} * \boldsymbol{x})]\} \end{aligned} \tag{1}$$

in which $\boldsymbol{x}$ and z indicate the input and the output of the HRU respectively, σ denotes the Relu nonlinear activation function, and the symbol $*$ represents the convolution.

z_A denotes the identity mapping branch, in which images in original resolution are scanned with a convolutional kernel $\boldsymbol{W}^I$ of size 1×1, and then processed by a Relu nonlinear activation function. The identity mapping branch is designed to avoid loss of image information by using the skip connections.

z_B denotes the hourglass mapping branch in which the features in low-resolution images are captured. First of all, in order to obtain low-resolution images, pooling (i.e. the function p in (1)) is used on original resolution images. Then, small convolutional kernels $\boldsymbol{W}^{H_1}$ and $\boldsymbol{W}^{H_2}$ are consecutively adopted to capture features, followed by a Relu nonlinear activation function for each. Finally, an upsampling operation (i.e. the function u in (1)) is applied to recover images of the original resolution for easy combination with other branches. Consequently, the hourglass mapping branch is utilized to macroscopically analyze features of human behavior.

z_C denotes the residual mapping branch. Multiple convolutional kernels $\boldsymbol{W}^{R_1}$, $\boldsymbol{W}^{R_2}$, and $\boldsymbol{W}^{R_3}$ are stacked to reduce parameters and to increase more nonlinear functions, and Relu nonlinear activation function is utilized after each convolution. The residual mapping branch can extract features from high-resolution images, such as local information of human posture and motion.

In conclusion, as shown in Fig. 2, to obtain output of the HRU, the features of three branches are effectively processed and integrated through the network. Compared with the traditional convolutional unit, the HRU can help capture multiscale features, expanding the receptive field to obtain richer information.

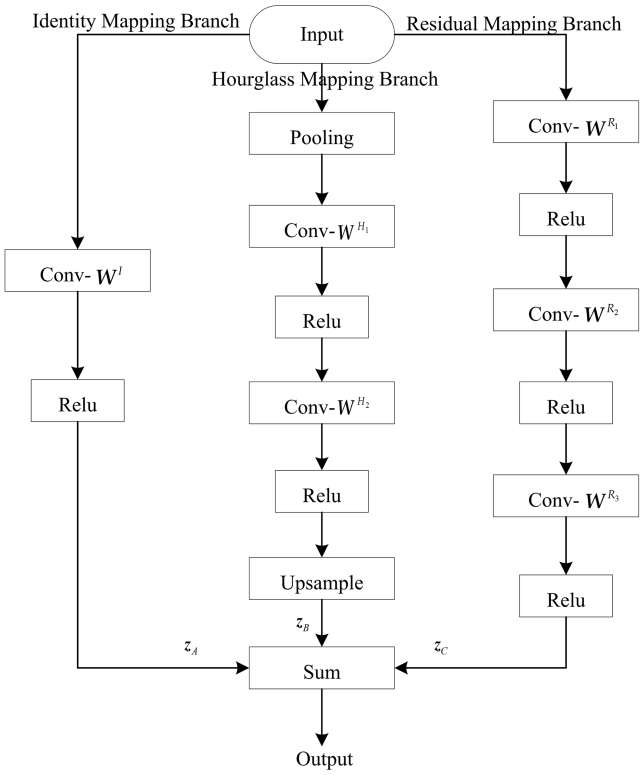

FIGURE 2. An illustration of the hourglass residual unit.

2) ENCODER AND DECODER IN HCAE

The proposed HCAE is presented to automatically obtain optimal intermediate feature from original frames which is propitious to detect falls. The architecture of the HCAE is composed of two parts: an encoder and a decoder (as shown in Fig. 1). First, the HCAE compresses the input into an intermediate feature by using the encoder; then, the decoder recovers approximate original frames from the intermediate feature; finally, the HCAE compares the error value between the encoded-decoded frames and the original frames, applying back-propagation for self-tuning.

In the proposed model, to capture multiresolution features from the frames, the HRU is used in the hourglass convolutional layers in the encoder. The kth ($k = 1, 2 \ldots, K$) hidden feature $\boldsymbol{h}_k$ of the hourglass convolutional layer can be expressed as:

$$\boldsymbol{h}_k = I(\boldsymbol{x}, \boldsymbol{W}_k^I) + H(\boldsymbol{x}, \{\boldsymbol{W}_k^{H_j}\}) + R(\boldsymbol{x}, \{\boldsymbol{W}_k^{R_i}\}) \tag{2}$$

in which $\boldsymbol{x}$ denotes the input of the hourglass convolutional layer; I, H, and R, respectively, denote the functions corresponding to the identity mapping branch, the hourglass mapping branch and the residual mapping branch of the hourglass convolutional layer; i and j respectively indicate the index of convolutional layer in the hourglass mapping branch and residual mapping branch; $\boldsymbol{W}_k^I$ denotes the weight matrix which should be learnt in the identity mapping branch to retain the information of original frames; $\{\boldsymbol{W}_k^{H_j}\}$ indicates a set of weight matrixes which

should be learnt in the hourglass mapping branch to capture the global information of a behavior; $\{\boldsymbol{W}_k^{R_i}\}$ signifies a set of weight matrixes which should be learnt in the residual mapping branch to extract local information of the frames.

Pooling layer down-samples the feature cube which consists of the hidden features obtained through the hourglass convolutional layer. Pooling operation, such as choosing the maximum value over non-overlapping rectangular subregions, reduces the spatial size of the representation and the amount of parameters. Therefore, the feature size and the parameters in the fully connected layer can be reduced effectively, and the computation speed can be accelerated.

In the proposed HCAE, after three times of hourglass convolution and pooling, the encoder can obtain the intermediate feature, the most critical part in HCAE. Then, the intermediate feature will be fed not only into a classifier to judge falls, but also into the decoder to reconstruct the original frames.

The main function of the decoder in the proposed HCAE is to recover frames from the intermediate feature. In general, unpooling and deconvolution are adopted to realize the decoder. Unpooling layer performs a reverse operation of pooling, and reconstructs the original size of each rectangular subregion. Then, deconvolution, an inverse operation of convolution, is applied to unpooled feature cube. The main steps of the deconvolutional layer can be expressed as:

$$\boldsymbol{y} = \sigma(\sum_{k=1}^{K} \boldsymbol{h}_k * \tilde{\boldsymbol{W}}_k) \tag{3}$$

where $\boldsymbol{y}$ denotes the output, and $\tilde{\boldsymbol{W}}_k$ is the weight matrix to be learnt in the deconvolutional layer.

After three times of deconvolution and upsampling, recovered frames can be output from the decoder. Finally, by minimizing the error between the original frames and the reconstructed frames, we can obtain the optimal intermediate feature, effectively representing the original frames, by using the HCAE.

The decoder in the proposed HCAE is used as a weak supervisor to reconstruct the original frames, and hence has a better correction effect on the intermediate feature extracted by the encoder.

B. MULTI-TASK MECHANISM

In order to acquire more representative features by the neural network, a multi-task mechanism with a multi-task loss function is proposed in the HCAE-FD method. Specifically, we assign two tasks to the HCAE, including a main task of fall judgment and an auxiliary task of frame reconstruction. Correspondingly, we also design a two-task loss function for the HCAE.

The goal of the main task (i.e. fall judgment) is to classify the intermediate feature of the HCAE using a classifier to obtain a fall or non-fall judgment; whereas the purpose of the auxiliary task (i.e. frame reconstruction) is to make the intermediate feature have the ability to accurately reconstruct the original frames. Accordingly, a cross-entropy function is set as the loss function for the main task, and a mean square error loss function is assigned to the auxiliary task. To achieve the multi-task goal, the loss functions of the main task and the auxiliary task are weighted summed into a two-task loss function.

When optimizing the two-task loss function, objectives of both tasks can be accomplished simultaneously, and an optimal HCAE model can be obtained. That is, for the main task, the HCAE model is gradually improved to capture more suitable intermediate feature for accurate judgments of falls, and for the auxiliary task, the HCAE model is continuously enhanced to gain the intermediate feature better representing the original frames.

The two tasks are closely related, for they share the intermediate feature and the network parameters to jointly realize the ultimate fall classification. The completion of the auxiliary task can help make the intermediate feature better represent the original frames, which can enhance the ability of feature expression of the HCAE model and hence can further improve the accuracy of fall detection in the main task.

The cross-entropy loss function for the main task is used to measure the error between the predicted classification results and the actual labels, and can be expressed as:

$$J_1 = -\frac{1}{n}\sum_{i=1}^{n}[t_i \cdot \log p_i + (1 - t_i) \cdot \log(1 - p_i)] \tag{4}$$

in which t_i denotes the true classification label of the ith sample, $t_i = 0$ denotes a fall in the ith sample, and $t_i = 1$ denotes a non-fall in the ith sample; p_i indicates the classification result predicted by the HCAE-FD method; n is the total number of the samples.

The loss function for the auxiliary task can be expressed as

$$J_2 = \min \sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \boldsymbol{y}_i \right\|_2^2 \tag{5}$$

in which $\boldsymbol{x}_i$ indicates the frames of the ith sample, $\boldsymbol{y}_i$ denotes the reconstructed frames corresponding to the ith sample, and n is the total number of the samples.

Finally, two loss functions are proportionally integrated into a multi-task loss function as:

$$J = \alpha \cdot J_1 + \beta \cdot J_2 \tag{6}$$

in which α and β are respectively the weights of the main and the auxiliary tasks, controlling the influence of the loss of both tasks in the multi-task learning. After multiple tests, we choose the optimal values from the candidate range [0,1] for these parameters, that is, $\alpha = 0.7$ and $\beta = 0.3$.

C. PROCEDURE OF THE METHOD

The framework of the HCAE-FD algorithm is given in Fig. 1. The main purpose of HCAE-FD algorithm is to detect fall behaviors from videos, and the algorithm can be mainly divided into training phase and testing phase.

In both phases, 10 consecutive frames are input as a stack into the encoder of the HCAE at a time, so as to take

advantage of the temporal information. After the frames are input into the network as 10-channel data, the processing of each convolution kernel is to sum the results of the separate convolutions of the 10 channels. That is, every feature in the HCAE contains spatio-temporal information extracted from all the input frames.

1) TRAINING

The training phase is a process of searching for optimal parameters of the HCAE model by minimizing the value of the multi-task loss function. In the proposed method, supervised learning is adopted to train the network, and the training data is used to learn the correlation relationship between the input and the output of the HCAE, which can be realized by constantly updating the parameters of the network. The flow chart of the training process is shown in Fig. 3.

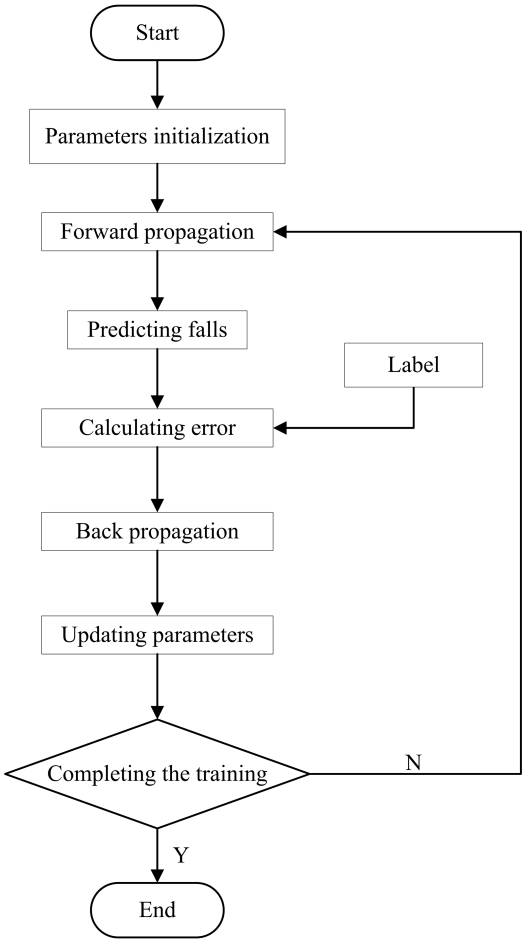

FIGURE 3. The flow chart of the training process.

The training process is divided into two stages: forward propagation and back propagation. In the forward propagation, the encoder of the HCAE is adopted to capture the intermediate feature from the training data. The intermediate feature is then used to discriminate falls by using a softmax classifier and meanwhile used to reconstruct the frames by using the decoder of the HCAE. In the back propagation, error between the expected output and the actual output of the proposed method can be calculated according to the multi-task loss function, and a back propagation algorithm is adopted to calculate the gradient, which can be used to update the weights and the bias values of each layer in the encoder, the decoder, and the classifier in the HCAE-FD method. When the loss function converges, the optimal HCAE model is obtained.

2) TESTING

When testing, frames in the test data are input into the HCAE sequentially. Then, by using the trained hourglass convolutional encoder and the trained softmax classifier, we can finally obtain a predicted probability p. $\leq p < 0.5$ indicates that the predicted value p is closer to the label $t = 0$ (a label for a fall event), that is, a fall is detected; whereas $0.5 \leq p \leq 1$ means that the predicted value p is closer to the label $t = 1$ (i.e. no falls detected).

IV. VALIDATION EXPERIMENTS

All experiments ran on a dedicated GPU server with an Intel i5-7400 running at 3.3 GHz, 4GB of memory, and four Nvidia Tesla K80 GPU accelerators. The proposed method was mainly implemented in python based on the tensorflow framework, and the performance was evaluated on the UR fall dataset [30] (URFD).

A. DATASET

In the experiments, we used the UR fall dataset which was proposed by researchers of Computational Modelling University of Rzeszow in 2014 [30]. The dataset includes RGB videos and depth videos recorded by two Microsoft Kinect cameras with a frame rate of 32 fps and a resolution of 640×480, and also acceleration data recorded by accelerometers.

RGB videos of camera 0 in the UR fall dataset, including 30 sequences of falls and 40 sequences of daily activities, were adopted in our experiments, and were divided into a training set and a testing set according to a ratio of 4:1. There are in total 900 fall frames and 11036 non-fall frames used in our experiments.

Non-fall frames mainly include common daily activities, such as walking, squatting, bending, etc. Fall frames mainly include fall behaviors which are performed by the participants, such as falling while walking and falling off a chair.

B. EVALUATING METRICS

The fall detection is a binary classification problem which determines whether or not there is a fall event in a particular sequence of a video. The most common metrics to evaluate the performance of such a classification are sensitivity/recall, precision, specificity, accuracy, and F-score. These metrics are not affected by distribution of unbalanced categories, which makes them more suitable for fall detection datasets, because the fall samples are usually much fewer than the non-fall samples in most datasets.

Sensitivity/recall is the proportion of correctly detected falls in all the falls, that is,

$$Sensitivity/\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

Precision is the proportion of correctly detected falls in all the detected falls, i.e.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

Specificity is the proportion of correctly detected non-fall behaviors in all the non-fall events, i.e.

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

Accuracy is the proportion of correctly detected falls and non-fall behaviors, i.e.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

F-score is a harmonic mean of recall and precision, and has been proven to be the most relevant evaluation index for the overall performance of detection algorithms [31], i.e.

$$\text{F-score} = 2\frac{Recall \times Precision}{Recall + Precision} \tag{11}$$

in which *TP* refers to the number of true positives, that is, the number of fall clips correctly classified as ''fall''; *TN* denotes the number of true negatives, i.e. the number of non-fall clips correctly classified as ''no fall''; *FP* indicates the number of false positives, i.e. the number of non-fall clips wrongly classified as ''fall''; and *FN* is the number of false negatives, i.e. the number of fall clips wrongly classified as ''no fall''.

C. EXPERIMENTAL RESULTS

1) PARAMETER SETTING FOR THE MULTI-TASK LOSS FUNCTION

In the proposed method, a multi-task loss function is designed by weighting the loss function of fall detection and the loss function of frame reconstruction. The weight parameters α and β, respectively, control the strengths of the main task (i.e. fall detection) and the auxiliary task (i.e. frame reconstruction) in the multi-task learning.

To investigate the impact of these parameters on the performance of our method, we tuned the parameters in the candidate range [0,1] by using 0.1 as the step size, and performed multiple tests so as to choose the optimal values. Table I provides the values of fall detection accuracy of our method when using different weight parameters.

When $\alpha = 0$ and $\beta = 1.0$, the loss function of the HCAE-FD method only contains the mean square error loss function of the auxiliary task of frame reconstruction, without the loss function of the main task of fall detection. In this case, the proposed method can only complete the frame reconstruction, without fall detection, when $\alpha = 0$ and $\beta = 1.0$.

As α increases, the fall detection accuracy gradually improves first, and then starts to drop. The best performance of our method is achieved when $\alpha = 0.7$ and $\beta = 0.3$.

TABLE 1. Fall detection accuracy with different weights.

Parameters	Accuracy
$\alpha = 0$ $\beta = 1.0$	0
$\alpha = 0.1$ $\beta = 0.9$	0.944
$\alpha = 0.2$ $\beta = 0.8$	0.946
$\alpha = 0.3$ $\beta = 0.7$	0.950
$\alpha = 0.4$ $\beta = 0.6$	0.954
$\alpha = 0.5$ $\beta = 0.5$	0.955
$\alpha = 0.6$ $\beta = 0.4$	0.960
$\alpha = 0.7$ $\beta = 0.3$	0.962
$\alpha = 0.8$ $\beta = 0.2$	0.957
$\alpha = 0.9$ $\beta = 0.1$	0.954
$\alpha = 1.0$ $\beta = 0$	0.945

Therefore, we set the weight parameters with the optimal values $\alpha = 0.7$ and $\beta = 0.3$ in the following experiments. With these optimal weight parameters, the task of fall detection dominates the HCAE-FD method, and the task of frame reconstruction just acts as an secondary task to increase the representativeness of features of the network.

It is worth noting that, when $\alpha = 1.0$ and $\beta = 0$, the HCAE-FD method only completes the task of fall detection without the task of frame reconstruction, and yields a worse result than the optimal situation considering both the main and the auxiliary tasks. This has further verified that, the auxiliary task plays a good role in helping better complete the fall detection (i.e. the main task).

2) ADVANTAGE OF THE HRUS

To further demonstrate the advantage of the HRUs used in the proposed method, we compared the HCAE-FD method with a fall detection method (called CAE-FD for short) which is the same as the proposed method except that it utilizes traditional convolutional layers without HRUs. For purpose of a fair comparison, the CAE-FD method employs the same network structure as the HCAE-FD method does. Considering that training frames are randomly selected from the training set, we compute average values of the accuracy after five times of experiments. Table 2 provides the average values of accuracy with different iterations for the CAE-FD method and the HCAE-FD method.

As shown in Table 2, the proposed method obtains a great improvement in terms of fall detection accuracy in the 10 iterations by using the HRUs. These results thus confirm that, the HRUs extracting multiscale features can capture more abundant information with fewer convolutional layers, and can further help ensure the high accuracy of fall detection with the shallow-layer network of the proposed method.

TABLE 2. Comparison of average accuracy of CAE-FD and HCAE-FD.

Iterations	CAE-FD	HCAE-FD
1	0.661	0.828
2	0.773	0.921
3	0.829	0.932
4	0.865	0.948
5	0.889	0.949
6	0.905	0.975
7	0.920	0.975
8	0.934	0.970
9	0.944	0.973
10	0.942	0.972

3) ADVANTAGE OF THE MULTI-TASK MECHANISM

To verify the effectiveness of designing the multi-task mechanism in the proposed method, we compared the HCAE-FD method with a fall detection method (called HC-FD for short). The HC-FD method adopts the same encoder (with hourglass convolutional layers) and the same classifier as the proposed method does; however, unlike the multi-tasking goal of the proposed method, the HC-FD method has only one task to detect falls using the classifier. Since the selection of training data is random, we also calculate average values of fall detection accuracy for these two methods after five times of experiments. Table 3 provides the average accuracy of the HC-FD method and the HCAE-FD method with different iterations.

TABLE 3. Comparison of average accuracy of HC-FD and HCAE-FD.

Iterations	HC-FD	HCAE-FD
1	0.806	0.828
2	0.906	0.921
3	0.916	0.932
4	0.922	0.948
5	0.934	0.949
6	0.951	0.975
7	0.955	0.975
8	0.971	0.970
9	0.973	0.973
10	0.970	0.972

As can be seen from Table 3, for iterations no more than 7, the accuracy of the HCAE-FD method shows a significant improvement in contrast with that of the HC-FD method; for the latter 3 iterations, the results of these two methods are relatively close, due to the single scene of the UR fall dataset. Through the comparison of these two methods, the HCAE-FD method can extract more representative and effective intermediate feature for fall detection, which is benefited from the constraints of the auxiliary task (i.e. frame reconstruction) in the multi-task mechanism on the intermediate feature.

4) COMPARISON WITH THE STATE-OF-THE-ART METHODS

To objectively evaluate the HCAE-FD method, we compared the performance of our method with five state-of-the-art methods (including three methods based on hand-crafted features and two methods based on deep learning) using the public UR fall dataset. The five methods for comparison are listed as follows.

1. A fall detection method based on points of interest (called Shi-Tomasi-FD method for short) [29]. This method first utilizes Shi-Tomasi algorithm to find interest points, and then tracks these points and computes their maximum displacement to obtain speed and direction of body motion to detect a fall.

2. A fall detection method based on area ratios of human body (called Area-FD method for short) [22]. In this method, ratios of five partial occupancy areas of the body are used as the feature, and input into machine learning algorithm to detect and classify falls.

3. A fall detection method using a MEWMA strategy based on area ratios (called MEWMA-FD method for short) [23]. In this method, fall detection is first achieved by using the MEWMA monitoring scheme according to area ratios of five partial areas constituting the human body extracted from each frame, and then a classification stage based on SVM is applied on the detected frames to further distinguish falls and fall-like behaviors.

4. A fall detection method based on convolutional neural network (CNN-FD method for short) [9]. This method employs VGG-16 net (including thirteen convolutional layers and three full connected layers) to receive optical flow images as input and to decide if a sequence of frames contains a fall event.

5. Fall detection method based on CNN and attention-guided LSTM (CNN-LSTM-FD method for short) [10]. In the method, VGG-16 net and an attention guided LSTM model are adopted to detect falls.

For purpose of a fair comparison, we employ the 5-fold cross-validation training as the CNN-FD algorithm did. Table 4 provides the evaluating metrics of the proposed method and the other five methods.

As can be seen from Table 4, the proposed method yields detection results with the best F-score and the second best accuracy when compared with the state-of-the-art methods. By contrast with the traditional vision-based fall

TABLE 4. Comparison with the state-of-the-art methods.

Method	Sensitivity /Recall	Specificity	Accuracy	Precision	F-score
Shi-Tomasi -FD [29]	0.967	-	0.957	0.935	-
Area-FD [22]	0.980	0.894	0.940	0.830	0.900
MEWMA-FD [23]	1	0.949	0.967	0.936	0.952
CNN-FD [9]	1	0.920	0.950	-	-
CNN-LSTM -FD [10]	0.914	-	-	0.948	0.931
Proposed method	1	0.930	0.962	0.923	0.960

detection methods (such as the Shi-Tomasi-FD, Area-FD, and MEWMA-FD methods), the HCAE-FD method does not require hand-crafted features, and has powerful data analysis ability with the help of HCAE, which significantly improves the performance in fall detecion. Compared with the CNN-FD and CNN-LSTM-FD methods based on the deep network, the proposed method also performs better with a shallow-layer network. This mainly results from the HRUs helping capture multiscale features and avoid information loss, and also from the multi-task mechanism which helps enhance the feature representativeness by completing the auxiliary task and further improves the performanc of the main task (i.e. fall detection).

V. CONCLUSION

In this paper, we propose a novel fall detection method based on the multi-task HCAE. The HRUs are adopted in the encoder of the HCAE to improve the convolutional layers into the hourglass convolutional layers and to extract multi-scale features from original frames. Furthermore, the multi-task mechanism is utilized to make the intermediate feature abundant for behavior information and further appropriate for classification, which is benefit to improve the accuracy of fall detection. The experimental results have shown that, the HCAE-FD method can effectively achieve accurate fall detection with the shallow-layer network, and outperforms several state-of-the-art methods. In the future, we will apply the method to complicated environments to further ensure the lives of the elderly.

REFERENCES

[1] S. R. Lord and J. Dayhew, "Visual risk factors for falls in older people," *J. Amer. Geriatrics Soc.*, vol. 49, no. 5, pp. 508–515, Dec. 2001.

[2] G. Santos, P. Endo, K. Monteiro, E. Rocha, I. Silva, and T. Lynn, "Accelerometer-based human fall detection using convolutional neural networks," *Sensors*, vol. 19, no. 7, p. 1644, Apr. 2019.

[3] M. Mubashir, L. Shao, and L. Seed, "A survey on fall detection: Principles and approaches," *Neurocomputing*, vol. 100, pp. 144–152, Jan. 2013.

[4] D. H. Stefanov, Z. Bien, and W.-C. Bang, "The smart house for older persons and persons with physical disabilities: Structure, technology arrangements, and perspectives," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 12, no. 2, pp. 228–250, Jun. 2004.

[5] X. Li, T. Pang, W. Liu, and T. Wang, "Fall detection for elderly person care using convolutional neural networks," in *Proc. 10th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Shanghai, China, Oct. 2017, pp. 1–6.

[6] N. Lu, Y. Wu, L. Feng, and J. Song, "Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 1, pp. 314–323, Jan. 2019.

[7] H. Yhdego, "Towards musculoskeletal simulation-aware fall injury mitigation: Transfer learning with deep CNN for fall detection," in *Proc. Spring Simulation Conf. (SpringSim)*, Tucson, AZ, USA, 2019, pp. 1–12.

[8] C. Ge, I. Y.-H. Gu, and J. Yang, "Co-Saliency-Enhanced deep recurrent convolutional networks for human fall detection in E-Healthcare," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 1572–1575.

[9] A. Marcos, G. Azkune, and I. Arganda-Carreras, "Vision-based fall detection with convolutional neural networks," *Wireless Commun. Mobile Comput.*, vol. 2017, pp. 1–16, Dec. 2017.

[10] Q. Feng, C. Gao, L. Wang, Y. Zhao, T. Song, and Q. Li, "Spatio-temporal fall event detection in complex scenes using attention guided LSTM," *Pattern Recognit. Lett.*, vol. 130, pp. 242–249, Feb. 2020, doi: 10.1016/j.patrec.2018.08.031.

[11] W. Lie, A. T. Le, and G. Lin, "Human fall-down event detection based on 2D skeletons and deep learning approach," in *Proc. Int. Workshop Adv. Image Technol. (IWAIT)*, Chiang Mai, Thailand, Jan. 2018, pp. 1–4.

[12] X. Yu, "Approaches and principles of fall detection for elderly and patient," in *Proc. 10th Int. Conf. e-Health Netw., Appl. Services (HealthCom)*, Singapore, Jul. 2008, pp. 42–47.

[13] N. Pannurat, S. Thiemjarus, and E. Nantajeewarawat, "Automatic fall monitoring: A review," *Sensors*, vol. 14, no. 7, pp. 12900–12936, Jul. 2014.

[14] C.-F. Lai, S.-Y. Chang, H.-C. Chao, and Y.-M. Huang, "Detection of cognitive injured body region using multiple triaxial accelerometers for elderly falling," *IEEE Sensors J.*, vol. 11, no. 3, pp. 763–770, Mar. 2011.

[15] T. Tamura, T. Yoshimura, M. Sekine, M. Uchida, and O. Tanaka, "A wearable airbag to prevent fall injuries," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 6, pp. 910–914, Nov. 2009.

[16] Y. Li, K. C. Ho, and M. Popescu, "A microphone array system for automatic fall detection," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1291–1301, May 2012.

[17] M. Popescu and A. Mahnot, "Acoustic fall detection using one-class classifiers," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Minneapolis, MN, USA, Sep. 2009, pp. 3505–3508.

[18] L. Ren and Y. Peng, "Research of fall detection and fall prevention technologies: A systematic review," *IEEE Access*, vol. 7, pp. 77702–77722, 2019.

[19] L.-H. Juang and M.-N. Wu, "Fall down detection under smart home system," *J. Med. Syst.*, vol. 39, no. 10, pp. 1–12, Aug. 2015.

[20] Y. Yun and I. Y.-H. Gu, "Human fall detection via shape analysis on Riemannian manifolds with applications to elderly care," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, USA, Sep. 2015, pp. 3280–3284.

[21] E. Akagündüz, M. Aslan, A. Şengür, H. Wang, and M. C. İnce, "Silhouette orientation volumes for efficient fall detection in depth videos," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 3, pp. 756–763, May 2017.

[22] N. Zerrouki, F. Harrou, A. Houacine, and Y. Sun, "Fall detection using supervised machine learning algorithms: A comparative study," in *Proc. 8th Int. Conf. Model., Identificat. Control (ICMIC)*, Algiers, Algeria, Nov. 2016, pp. 665–670.

[23] F. Harrou, N. Zerrouki, Y. Sun, and A. Houacine, "Vision-based fall detection system for improving safety of elderly people," *IEEE Instrum. Meas. Mag.*, vol. 20, no. 6, pp. 49–55, Dec. 2017.

[24] Z.-P. Bian, J. Hou, L.-P. Chau, and N. Magnenat-Thalmann, "Fall detection based on body part tracking using a depth camera," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 2, pp. 430–439, Mar. 2015.

[25] M. Yu, S. M. Naqvi, and J. Chambers, "Fall detection in the elderly by head tracking," in *Proc. IEEE/SP 15th Workshop Stat. Signal Process.*, Cardiff, Wales, Aug. 2009, pp. 357–360.

[26] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Monocular 3D head tracking to detect falls of elderly people," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, New York, NY, USA, Aug. 2006, pp. 6384–6387.

[27] H. Foroughi, A. Naseri, A. Saberi, and H. Sadoghi Yazdi, "An eigenspace-based approach for human fall detection using integrated time motion image and neural network," in *Proc. 9th Int. Conf. Signal Process.*, Oct. 2008, pp. 1499–1503.
[28] S. Su, S.-S. Wu, S.-Y. Chen, D.-J. Duh, and S. Li, "Multi-view fall detection based on spatio-temporal interest points," *Multimedia Tools Appl.*, vol. 75, no. 14, pp. 8469–8492, Jul. 2015.
[29] S. Bhandari, N. Babar, P. Gupta, N. Shah, and S. Pujari, "A novel approach for fall detection in home environment," in *Proc. IEEE 6th Global Conf. Consum. Electron. (GCCE)*, Nagoya, Japan, Oct. 2017, pp. 1–5.
[30] *UR Fall Dataset*. Accessed: Jun. 1, 2018. [Online]. Available: http://fenix.univ.rzeszow.pl/~mkepski/ds/uf.html
[31] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.Net: A new change detection benchmark dataset," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 1–8.

XI CAI received the B.Eng. and Ph.D. degrees from the School of Electronic and Information Engineering, Beihang University, Beijing, China, in 2005 and 2011, respectively. She is currently an Associate Professor with Northeastern University at Qinhuangdao, China. Her research interests include digital image processing and video analysis.

SUYUAN LI received the B.S. degree from Liaoning Normal University, Dalian, in 2017. He is currently pursuing the Ph.D. degree with Northeastern University at Qinhuangdao, China. His research interests include fall detection based on deep learning and image processing.

XINYUE LIU received the B.S. degree from Shenyang Ligong University, Shenyang, in 2018. She is currently pursuing the Ph.D. degree with Northeastern University at Qinhuangdao, China. Her research interests include fall detection based on deep learning and image processing.

GUANG HAN received the B.Eng. and M.Eng. degrees from the School of Electronic and Information Engineering, Beihang University, Beijing, China, in 2005 and 2008, respectively, and the Ph.D. degree from the School of Computer Science and Engineering, Northeastern University, Shenyang, China, in 2017. He is currently an Associate Professor with Northeastern University at Qinhuangdao, China. His research interests include digital image processing and video analysis.