Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 6, December 2021, pp. 3191~3201
ISSN: 2302-9285, DOI: 10.11591/eei.v10i6.3204

A multi-task learning based hybrid prediction algorithm for privacy preserving human activity recognition framework

Vijaya Kumar Kambala, Harikiran Jonnadula
School of CSE, VIT-AP University, Amaravathi, Andhra Pradesh, India

Article Info

Article history:

Received Aug 7, 2021
Revised Oct 3, 2021
Accepted Oct 31, 2021

Keywords:

Anonymization
Human activity recognition
Hybrid prediction algorithm
Multi-task learning
Privacy

ABSTRACT

There is ever increasing need to use computer vision devices to capture videos as part of many real-world applications. However, invading privacy of people is the cause of concern. There is need for protecting privacy of people while videos are used purposefully based on objective functions. One such use case is human activity recognition without disclosing human identity. In this paper, we proposed a multi-task learning based hybrid prediction algorithm (MTL-HPA) towards realising privacy preserving human activity recognition framework (PPHARF). It serves the purpose by recognizing human activities from videos while preserving identity of humans present in the multimedia object. Face of any person in the video is anonymized to preserve privacy while the actions of the person are exposed to get them extracted. Without losing utility of human activity recognition, anonymization is achieved. Humans and face detection methods file to reveal identity of the persons in video. Action detection is done using bidirectional long short tem memory network. We experimentally confirm with joint-annotated human motion data base (JHMDB) and daily action localization in YouTube (DALY) datasets that the framework recognises human activities and ensures non-disclosure of privacy information. Our approach is better than many traditional anonymization techniques such as noise adding, blurring, and masking.

Corresponding Author:

Harikiran Jonnadula
School of CSE, Vellore Institute of Technology
VIT-AP University, G-30, Inavolu, Beside AP Secretariat Amaravati
Andhra Pradesh 522237, India
Email: harikiran.j@vitap.ac.in

1. INTRODUCTION

In the contemporary era, there is an increased necessity for computer vision technology that supports large-scale and automated study of visual data. It has become very crucial in many applications that are useful to society. Usage of cameras associated with such computer vision applications became ubiquitous. In cities across the globe there are millions of cameras that capture visual data on 24/7 basis for leveraging different departments like policing, investigation agencies, healthcare industry, and so on. There is also need for monitoring elderly people to recognize actions like fall. Human action recognition from live videos is a popular research activity as studied in [1]-[4]. In the modern living, it is essential for different kinds of real world applications. Based on human actions, it is possible to determine the kind of activity and take some steps associated with the underlying application.

There are many advantages of using cameras in public places and select territories for monitoring. However, with respect to human action recognition, face is an important biometric used to know the identity of the person. Therefore, invasion of privacy in computer vision applications has become an increasing cause of concern, particularly video recording that is not done with prior permission. In the modern society, we

need deployment cameras to capture video and recognise events but it is essential to eliminate privacy invasion. Therefore, the need of the hour is to have mechanisms to recognise human actions from running videos but ensure that the identity of humans is not lost. Towards this end, several researchers proposed methods to ensure privacy-aware human activity recognition. For instance, [5]-[7] explored different methods for privacy preserving action recognition with zero-shot approach, automatic fall detection, position based super pixel transformation and hybrid reasoning respectively. Wu *et al*. in [8] there is GAN based approach with deep learning for privacy preserving action recognition. Similarly, in [9]-[12] GAN based approaches are discussed for computer vision applications. In different face recognition approaches are explored as the human activity recognition needs to identify face or anonymize face that prevents disclosure of identity [13]-[16].

From the literature, it is understood that most of the existing methods with adversarial learning do not consider a holistic approach with multi-task learning. This gap is filled in this paper by proposing a framework with underlying algorithm to leverage state of the art. Our approach is based on adversarial learning that involves multi-task learning such as face anonymization, face detection and face recognition where there is two-player game is characterized between generator G and discriminator D. The former tries to preserve privacy while the latter tries to break it and with continuous adversarial learning setting, they strive to perform well thus leading to better performance in privacy preserving human activity recognition. Our contributions in this paper are being as.

- We proposed a framework known as privacy preserving human activity recognition framework (PPHARF) that follows a holistic approach consisting of discriminator and generator in adversarial setting besides face anonymizer.
- An algorithm named multi-task learning based hybrid prediction algorithm (MTL-HPA) is proposed where multiple tasks are learned to ensure that the algorithm enhances privacy and ensures reliable action recognition.
- A prototype application is built using Python data science platform. It is used to explore the difference between the proposed and existing methods. The empirical results revealed that the PPHARF outperforms the existing approaches. The remainder of the paper is structured is being as. Section 2 reviews literature pertaining to human action recognition from videos with privacy preserved.

2. RELATED WORK

This section reviews literature on various aspects of the research associated with human action recognition while preserving privacy.

2.1. Generative adversarial networks

Generative adversarial network (GAN) models are found to be useful in improving performance in computer vision applications. As explored by Wu *et al*. [8], generative adversary learning has benefits to have a two-player game approach towards improving higher level of accuracy in human action recognition. GAN based research is found in different researchers such as [9]-[12]. Peng and Schmid [9] employed it for action detection where two-stream R-CNN is the deep learning technique employed. Ren and Lee [10] explored multi-task feature learning based on GAN model using synthetic imagery. Ryoo *et al*. [11] studied on privacy preserving action recognition that is based on GAN model and they considered extreme low-resolution images. Liu *et al.* [12] used GAN approach for face recognition using deep hypersphere embedding approach.

2.2. Privacy aware action recognition

Privacy refers to non-disclosure of identity of humans in this research. Kumar *et al*. [1] proposed deep learning models like convolutional neural networks (CNN) for privacy preserving human activity recognition. Wu *et al.* [8] proposed two different strategies for privacy preserving visual recognition. The strategies are known as budget model restarting and ensemble. They also employed adversarial training for better performance. Dai *et al.* [3] investigated on the trade-off between performance of action recognition while preserving privacy and number of cameras and their resolution. They found that the spatial resolution, the number of cameras used, and temporal resolution have their impact on performance from highest to lowest respectively. Similar kind of work is carried out in [17]. Lyu *et al.* [4] proposed a collaborative deep learning method that preserves privacy. They employed multiplicative perturbation for making their scheme privacy aware. Ryoo *et al*. [17] used extreme low-resolution samples for action recognition. They introduced a paradigm known as inverse super resolution (ISR) to learn image transformations optimally by creating suitable low-resolution training images.

Machot *et al*. [5] proposed a framework for human action recognition using non-visual sensors. Their framework leverages sensor data to discover unseen activities using a technique known as zero-shot learning. Zhang *et al*. [2] focused on human action recognition with respect to falling and privacy preserving. Their method involves feature extraction and representation and uses RGBD cameras. Rajput *et al*. [6] also employed RGBD cameras and deep CNN in order to achieve privacy preserving human action recognition. They reused cloud based learned models for better performance. Cippitelli *et al*. [18] used skeleton data collected from RGBD sensors for action recognition. Yet another study by Cippitelli *et al*. [19] is based on RGBD for human action recognition.

Ribono and Bettini [7] proposed an ontology-based solution based on hybrid reasoning and context-aware approaches. Zolfaghari and Keyvanpour [20] proposed smart activity recognition framework (SARF) which has different components such as sensing, pre-processing, feature extraction, feature selection, and recognition of ambient assisted living (ASL). Ciliberto *et al*. [21] explored a privacy preserving 3D model for human activity recognition. Gheid *et al*. [22] proposed a variant of KNN for privacy preserving action recognition as part of their multi-party classification protocol. Gheid and Challl [23] does similar kind of work that extends the work of [22] with optimization in security. Yonetani *et al*. [24] investigated on doubly permuted homomorphic encryption (DPHE) approach for privacy preserving visual learning. It could improve both privacy and accuracy over its predecessors.

2.3. Face recognition

Recognising face is essential in many computer vision applications. This is well explored problem found in [25]. The performance of face recognition is further enhanced with recent advanced artificial methods and available large datasets as studied in [26]-[28]. A multi-class classification problem is taken with vanilla softmax in two loss functions such as softmax loss and center loss are combined in order to jointly optimize the inter-class distance and intra-class feature distance. Wen *et al*. in [28] uses 200 million images for training and employed triplet loss function for improving performance in face recognition. The recent work in [26], [27] showed promising performance due to the usage of classification combined with metric learning. From the literature, it is understood that most of the existing methods with adversarial learning do not consider a holistic approach with multi-task learning. This gap is filled in this paper by proposing a framework with underlying algorithm to leverage state of the art.

3. PRELIMINARIES

GAN is used in many computer vision applications such image-to-image translation. This section provides an overview of GAN to understand the proposed approach which is based on adversarial learning. GAN has two important components such as generator (G) and discriminator (D). The two components are implemented using deep learning techniques. As presented in Figure 1, the G is used to generate new data denoted as (Gz) and the underlying data distribution is denoted as $p_g(z)$. The aim of GAN is to see that training sample denoted as $p_r(x)$ and $p_g(z)$ are same. On the other hand, the D takes (Gz) and real data x as input. G gets trained while D has fixed parameters. The discriminator D generates output which is denoted as D(G(z)) and between this and sample label, error rate is computed.

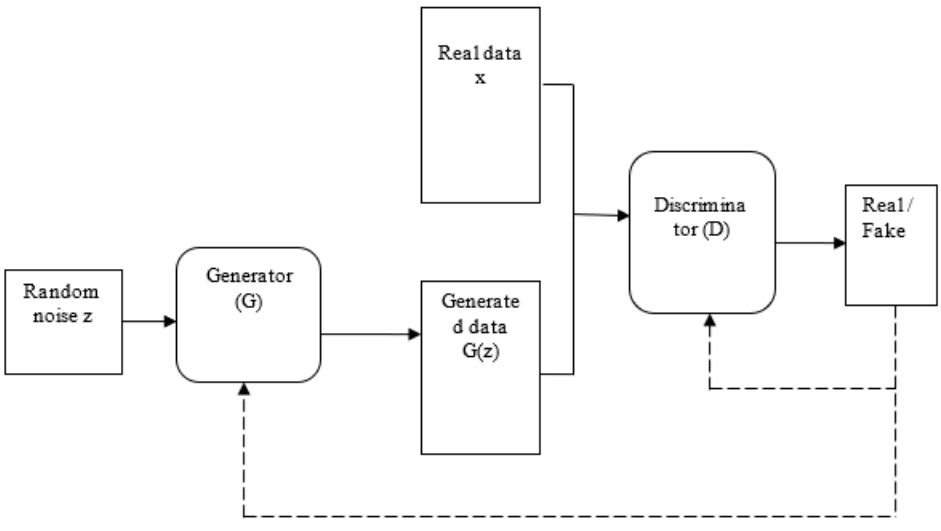

Figure 1. A typical GAN framework

Backpropagation algorithm is employed to update parameters of G. The D is aimed at finding whether input is really from given sample and provides required feedback that helps in updating G's parameters. Based on real input data is x or not, the output of D approaches to either 1 and 0 respectively. The overall process resembles min-max game played by two plyers and loss function of D is computed as in (1).

$$V(D,\theta^{(D)}) = -E_{x\sim p_r(x)}[\log D(x)] - E_{z\sim p_g(z)}[\log(1-D(g(z)))] \quad (1)$$

Similarly, the G has its loss function as in (2).

$$V(G,\theta^{(G)}) = E_{z\sim p_g(z)}[\log(1-D(g(z)))] \quad (2)$$

In the game, both players have their loss functions. In the process D maximizes $V^{(D)},(\theta^{(D)},\theta^{(G)})$while G maximizes $V^{(G)},(\theta^{(D)},\theta^{(G)})$by updating $\theta^{(D)}$ and$\theta^{(D)}$respectively. The loss functions of the players have parameter dependency. Nash equilibrium is to be achieved in order for a player to update parameters of other player and stop training.

$$\text{Min max } V(D,G) = E_{x\sim p_r(x)}[\log D(x)] + E_{z\sim p_g(z)}[\log(1-D(g(z)))] \quad (3)$$

As the GAN is denoted as min-max optimization problem, the loss function is as expressed as in (3). D makes an objective function as large as possible using read data as input. The goal of D is to ensure that output denoted as D(G(z)) is close to 1. With training the min-max game converges to Nash equilibrium. GAN models are found in many computer vision applications such as [9]-[12].

In this paper, the D tries to establish human identity while the G tries to make human face images as hard as possible to identify while ensuring human action recognition. In the proposed approach there is third component that anonymizes face region so as to see that the privacy is not lost. Reliable human action recognition while defeating disclosure of identity is the main aim of this paper. As presented in Table 1, the notations are provided to understand the symbols used in the proposed algorithm and its underlying equations.

Table 1. Notations used in the algorithm

Notation	Description
M	Face anonymizer or modifier
f or rv	input face or input face region
A	Action detector
D	The face classifier or discriminator
V	Set of frames in a video
F	Set of face images
r_v	A face region
M (r_v)	Modifying the face region given
v'	Updated frame
L_A	sum of the four losses in Faster-RCNN
M(f)	The act of anonymization of given face
L_D	Face detection loss (by discriminator)
F	Original image
λ	A weight

4. OUR APPROACH

Human action recognition from pre-recorded or live videos is an important requirement in many applications such as surveillance. However, those applications generally recognize not only human action but also identity of human. This is where the privacy of the person is lost. In this paper we aim at preventing disclosure of identity while allowing reliable action recognition.

4.1. The framework

Inspired by the generative adversarial network (GAN) explained in section 3, we considered adversarial learning phenomenon where two components compete. They include face anonymizer acting as generator (G) and face classifier (D). The former strives to ensure that human face is anonymized so as to preserve privacy while the latter tries to extract sensitive information in order to establish identity. Both G and D competes with each other with quite opposite responsibilities. This kind of adversarial learning is widely using in computer vision applications such as image to image translation [9]-[12]. Without

compromising on action detection performance, the face anonymizer strives to remove privacy sensitive information from face region of humans in given video input. Figure 2 shows the proposed framework named privacy preserving human activity recognition framework (PPHARF).

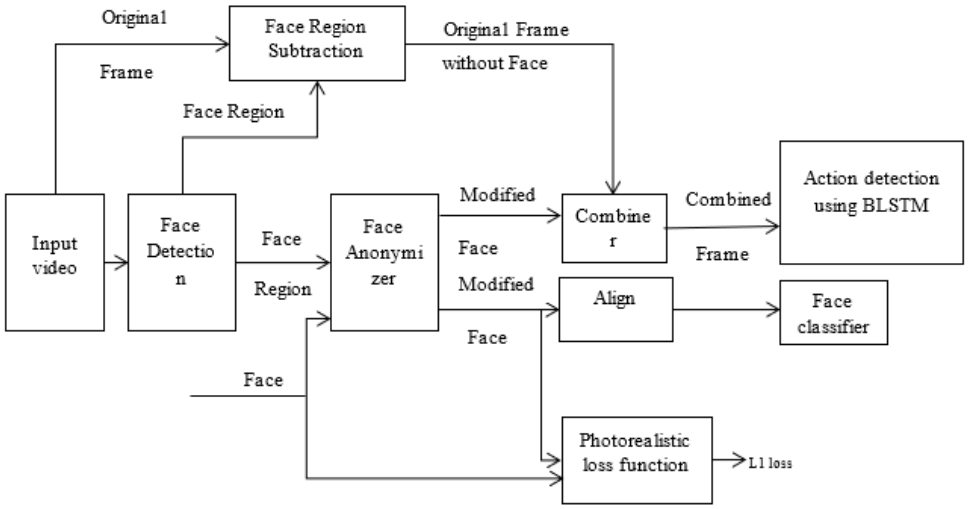

Figure 2. Overview of the proposed framework based on adversarial learning

The framework uses multi-task learning approach where the learning process is associated with different components such as face anonymizer, action detector and face classifier. The framework takes input video and finds the regions where face appears. Such face region is detected by the face detection module and that region is given to face anonymizer. Face anonymizer M takes a face (f) or face region (r_v) extracted by face detection module and modifies it by removing sensitive information resulting M(f) or M(r_v). The face region is subtracted from original video frame (v) by face region subtraction module. The combiner module takes modified face region (r_v) and combined with original frame without face region to result in a combined frame v'. The action detector module (A) tries to detect human action using v' which is devoid of sensitive information. Face classifier (the discriminator D) tries to establish identity of human (and expected to fail to do so). Its detection loss is expressed as in (4).

$$L_{det}(M, A, V) = E_{v\sim V}[L_A(v', \{b_i(v)\}, \{t_i(v)\})] \quad (4)$$

The action detector module is implemented as in [9] and [11] while the loss function is taken from [10]. The face classifier is nothing but discriminator D in adversarial learning whose aim is to identify a face. It is implemented based on the face classifier used in [12] and the adversarial loss function is modelled after the two-player game explored in [29]. The adversarial loss function is expressed in (5).

$$L_{adv}(M, D, F) = -E_{(f\sim F, i_f\sim I)}\left[L_D\left(M(f), i_f\right)\right] = -E_{(f\sim F, i_f\sim I)}\left[L_D\left(M(f, i_f\right)\right] \quad (5)$$

In order to preserve the structure such as brightness and pose of modified image, we used L1 loss which is also known as photorealistic loss. In computer vision applications this loss function is used in [30], [31]. It could force similarity between input image and modified image (some visual similarity). This loss function is expressed as in (6).

$$L_{I1} = (M, F)E_{f\sim F}\left[\lambda\left|\left|M(f) - f\right|\right|1\right] \quad (6)$$

The components in the proposed framework such as A and D are iteratively trained to perform their functionality effectively. Face detector module is the one used in [32] and another face detector used in [33] is used in order to eliminate false positives. The face anonymizer is modelled after [34] with 9 residual blocks. With respect to training Adam solver [35] is used with different parameters such as β_1=0.5 and β_2=0.999 while learning rate is set at 0.0003 for face classifier and 0.001 for face anonymizer. Total number of epochs used is 12 and the learning rate is dropped to 1/10 after 7th epoch.

4.2. Action recognition

The action recognition for face anonymized image is done using bi-directional long short term memory (BLSTM) network. long short term memory (LSTM) is an extension of recurrent neural networks. Due to its special architecture, which combats the vanishing and exploding gradient problems, it is good at handling human activity recognition up to a certain depth. The transfer process of the network can be described as follows: the input tensor is transformed, along with the tensor of the hidden layer (at the last stage), to the hidden layer by a matrix transformation. Then, the output of the hidden layer passes through an activation function to the final value of the output layer.

Recurrent neural network (RNN) is a type of artificial neural network, which is used for working with sequential data. Normal feed forward neural network has only independent data points. But for sequential data, the current data point is dependent upon the previous data point. Inorder to maintain dependencies between data points in the neural network, we use RNN which has the concept of memory which stores the information of previous data points and using this it will generate next data points in sequence. Different activation functions such as sigmoid, Tanh and Relu functions in RNN. The main advantages of RNN are ability to handle sequential data, input data having varying lengths and can store past information in memory. The main advantage of RNN is the vanishing gradient, it can remember data points information only for a short period tof time. In-order to overcome this problem, we are going to LSTM networks. When backpropagation with a deep network is used, gradients will vanish rapidly if preventative measures that permit gradients to flow deeply are not taken. Compared with the simple input concatenation and activation used in RNNs, LSTM has a particular structure for remembering information for a longer time as an input gate and a forget gate control how to overwrite the information by comparing the inner memory with the new information arriving. It enables gradients to flow through time easily. Figure 3 shows basic structure of LSTM, where vector x denotes input, vector y denotes output and vector h denotes hidden layers.

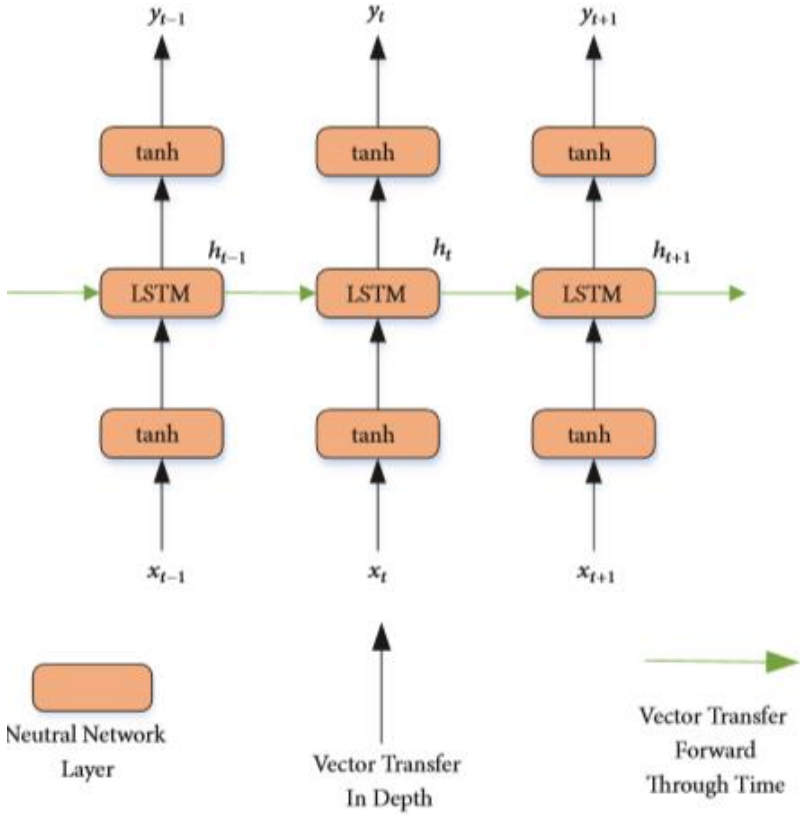

Figure 3. Basic structure of LSTM

LSTM is a special type of RNN with memory blocks connected through layers. Each memory block is smaller than a neuron with gates having capability to manage state and output. This gate uses the sigmoid activation units to add information in memory bloch and to make change of state while operating on an input

sequence. Each LSTM has three types of gates, forget gate (to remove information from the block), inouut gate (to update the memory based on input values) and output gate (values generated from the memory block based on input). These gates have weights that are leaned during the training process. In real life, human trajectories are continuous. Baseline LSTM cells predict the status based only on former information. Some important information may not be captured properly by the cell if it runs in only one direction. The improvement in bidirectional LSTM is that the current output is not only related to previous information but also related to subsequent information. For example, it is appropriate to predict a missing word based on context. Bidirectional LSTM is made up of two LSTM cells, and the output is determined by the two together. BLSTM is a kind of neural network having advantage of flowing information from in both directions means that future to past and past to future. Input flows in two directions making BLSTM different from regular LSTM. In regular LSTM, we can move information only in one direction either forward or backward. But in BLSTM, sequence information can move in bothe directions preserving future and past information in memory units. It consists of two LSTMs one taking input in a forward direction and one LSTM in a backward direction, effectively increasing the information available in the network. For providing combination of forward and backward outputs, different functions such as summation, multiplication, concatenation, and average are used. This BLSTM network is used in our work for action recognition of anonymized image. Figure 4 shows the structure of standard BLSTM.

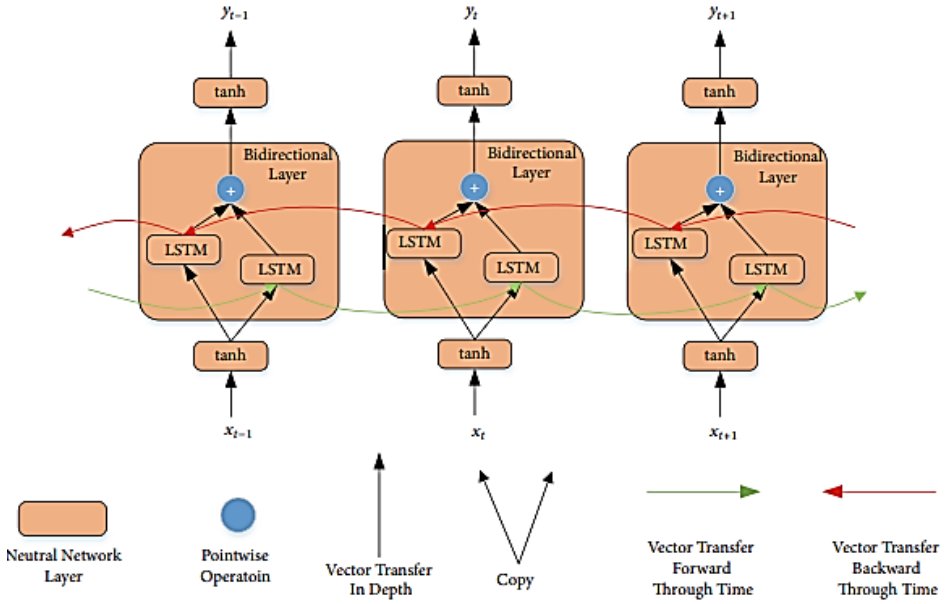

Figure 4: Standard BLSTM structure

4.3. Algorithm design

An algorithm named multi-task learning based hybrid prediction algorithm is proposed to realize the functionality of our approach. As presented in Algorithm 1, it takes set of video frames V, a discriminator D, set of face images F and set of identity labels are used as input. There is an iterative process that works for each video frame. From the video frame, the face region is identified and that is anonymized. On the other hand, the subtraction module removes face region from the video frame and the result is assigned to s. After anonymizing face image, it is combined with the subtracted frame in order to have better anonymized face image with visual appearance. Such anonymized face image is taken by action detector that identifies human action associated with face image. At the same time, the discriminator D (face classifier) tries to establish identity of the face as part of its adversarial setting. Every time, the fame modifier (M) and the action detector (A) are updated with improved knowledge. This process continues for all the frames in the given video. If there are multiple images in a single frame, there will be a sub process to repeat steps for each fame image in the video frame v.

Algorithm: Multi-Task Learning based Hybrid Prediction Algorithm

Input: D, V, F, identity labels
Output: M, A

1. For each *v* in V
2. *f*←DetectFace(*v*)
3. s ← Subtract(*f*)
4. *f* ' ← M(*f*)
5. Compute L1 loss

$$L_{l1} = (M,F)E_{f\sim F}\left[\lambda\left|\left|M(f) - f\right|\right|1\right]$$

6. *v*'←Combiner (*f*', *s*)
7. a ← A(*v*')
8. Compute detection loss

$$L_{det}(M,A,V) = E_{v\sim V}[L_A(v',\{b_i(v)\},\{t_i(v)\})]$$

9. det=D (*f* ')
10. Compute adversarial loss

$$L_{adv}(M,D,F) = -E_{(f\sim F,i_f\sim I)}\left[L_D\left(M(f),i_f\right)\right] = -E_{(f\sim F,i_f\sim I)}\left[L_D\left(M(f,i_f\right)\right]$$

11. Update M
12. Update A

5. RESULTS AND DISCUSSION

5.1. Datasets

We used two datasets namely DALY and JHMDB in this dataset. Daily action localization in YouTube (DALY) is the dataset introduced in [36] and the dataset is obtained from [37]. The dataset has more than 30 hours of YouTube videos that are annotated in spatial and temporal domains. It consists of 10 human actions that are witnessed every day with 3600 total number of instances. Action classes are considered with clearly set temporal boundaries. This is essential to overcome ambiguities associated with noise. Some of the action classes include brushing teeth, phoning, taking photos, drinking, applying makeup on lips and playing harmonica as presented in Figure 5.

Figure 5. Representative pictures with human action annotated for10 classes of DALY dataset

JHMDB is another dataset introduced in [25] and it is collected from [38]. It is the dataset which is benchmark for human detection, human actions and pose estimation. JHMDB is derived from the HMDB51 dataset [39] where 5,100 videos consisting of 51 human actions. JHMDB is a subset of HMDB51 with 21 categories that involve a single human with certain actions as presented in Figure 6. Some of the actions include hug, kick, jump, run, and shoot.

5.2. Results

The experiments are made with DALY and JHMDB datasets. The experimental results are observed in terms of face verification error and mean average precision. Our anonymization approach is compared with many states of the art or baseline approaches. They include Blur x 3, masked, noise x 3 and edge.

The qualitative results shown provide the visual difference before and after the face modification. This is made due to the fact that our algorithm anonymizes prior to performing action recognition. The user study conducted with famous personalities convinced that our anonymization approach has acceptable performance.

Figure 6. Some of the images from JHMDB dataset with annotated human actions

As presented in Table 2, the results revealed that for many actions, the proposed approach showed better performance. The aim of the research is to detect human actions reliably while preventing disclosure of human identity. This has been fulfilled as observations are made in the experimental study. The results also reveal that the performance of the proposed method is better over baselines in terms of anonymization as well. It is evidenced in both empirical study and user study.

Table 2. Shows accuracy in action detection using DALY dataset

MTL-HPA	Edge	Masked	Noise(σ 2=0.6)	Noise (σ 2=0.4)	Noise (σ 2=0.2)	Blur(24x24)	Blur(16x16)	Un-anonymized	Action
90.2892	80.3803	67.1271	83.7136	87.7276	87.7977	92.1721	82.1521	92.3923	Lip
35.1131	29.4895	15.2052	21.4715	25.005	31.4014	39.8798	32.4324	51.3113	Brush
79.1971	78.098	78.9389	81.4013	78.6686	78.4884	80.0099	79.6095	76.8067	Floor
34.5926	30.5405	26.6066	29.6196	32.7127	31.9019	31.8018	32.1721	27.8979	Window
24.9529	10.3203	6.59659	7.43743	8.34834	12.4224	15.2452	10.7507	31.2612	Drink
34.8939	32.6726	25.5555	29.1091	35.3653	34.8348	34.995	31.3814	32.7027	Fold
79.1471	79.2292	73.023	78.048	75.035	76.5765	74.7247	74.9849	75.3753	Iron
48.5665	35.1852	27.8178	33.9639	40.1601	42.4124	46.3763	37.007	51.5515	Phone
57.3753	54.7447	21.3413	27.3774	36.6466	50.1901	51.8318	48.4184	73.9839	Harmonica
57.5955	49.6696	46.8068	46.3964	45.4654	53.5335	53.1431	52.5625	55.7957	Photo
54.37232	48.03299	38.90186	43.85381	46.51347	49.95591	52.01797	48.1471	56.90785	mAP

6. CONCLUSION

We proposed a framework named privacy preserving human activity recognition framework (PPHARF) with an underlying algorithm known as multi-task learning based hybrid prediction algorithm (MTL-HPA). The framework is aimed at supporting adversarial learning based multi-task formulation that accomplishes human action recognition, anonymization of human face and face detection (to evaluate anonymization). Face anonymizer is part of the framework that is designed to ensure preserving of privacy. In other words, non-disclosure of human identity and at the same time recognition of action are important considerations. The face anonymizer is designed to confuse humans and applications in recognition of human based on face while supporting action recognition accurately. The proposed framework is evaluated with a prototype using JHMDB and DALY datasets. The proposed approach outperformed traditional face modification approaches and it showed significantly better performance over the state of the art. In future, we intend to explore advanced GAN approaches such as SingleGAN with possible improvements.

REFERENCES

[1] K. V. Kumar, J. Harikiran, M.A.R. Prasad, and U. Sirisha, ''Privacy-Preserving Human Activity Recognition and Resolution Image using Deep Learning Algorithms Spatial relationship and increasing the attribute value in OpenCV,'' *International Journal of Advanced Science and Technology*, vol. 29, no. 7, pp. 514-523, 2020.

[2] C. Zhang, Y. Tian, and E. Capezuti, ‘‘Privacy Preserving Automatic Fall Detection for Elderly Using RGBD Cameras,’’*International Conference on Computers for Handicapped Persons*, 2012, pp. 625-633, doi: 10.1007/978-3-642-31522-0_95.

[3] J. Dai, J. Wu, B. Saghafi, J. Konrad, and P. Ishwar, ‘‘Towards Privacy-Preserving Activity Recognition Using Extremely Low Temporal and Spatial Resolution Cameras,’’ *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 68-76, 2015.

[4] L. Lyu, X. He, Y. W. Law, and M. Palaniswami, ‘‘Privacy-Preserving Collaborative Deep Learning with Application to Human Activity Recognition,’’ *ACM*, pp. 1219-1228, Nov. 2017, doi: 10.1145/3132847.3132990.

[5] F. Al Machot, M. R. Elkobaisi, and K. Kyamakya, ‘‘Zero-Shot Human Activity Recognition Using Non-Visual Sensors,’’ *Sensors*, vol. 20, no. 3, p. 825, doi: 10.3390/s20030825.

[6] A. S. Rajput, B. Raman, J. Imran, ‘‘Privacy-preserving human action recognition as a remote cloud service using RGB-D sensors and deep CNN,’’ *Expert Systems with Applications*, vol. 152, p. 113349, August 2020, doi: 10.1016/j.eswa.2020.113349.

[7] D. Riboni and C. Bettini, ‘‘COSAR: hybrid reasoning for context-aware activity recognition,’’ *Personal and Ubiquitous Computing*, vol. 15, no. 3, pp. 271-289, 2011, doi: 10.1007/s00779-010-0331-7.

[8] Z. Wu, Z. Wang, Z. Wang, and H. Jin, ‘‘Towards Privacy-Preserving Visual Recognition via Adversarial Training: A Pilot Study,’’ *Springer*, pp. 1-19, 2018.

[9] X. Peng and C. Schmid, ‘‘Multi-region two-stream r-cnn for action detection,’’in: *European conference on computer vision (ECCV)*, October 2016, pp. 744-759, doi: 10.1007/978-3-319-46493-0_45.

[10] Z. Ren and Y. J. Lee, ‘‘Cross-domain self-supervised multi-task feature learning using synthetic imagery,’’in: *CVPR*, 2018, pp. 762-771.

[11] M. S. Ryoo, B. Rothrock, C. Fleming, ‘‘Privacy-preserving egocentric activity recognition from extreme low resolution,’’*Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017, pp. 4255-4262.

[12] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, ‘‘Sphereface: Deep hypersphere embedding for face recognition,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 212-220.

[13] Y. Sun, X. Wang, and X. Tang, ‘‘Deep learning face representation from predicting 10,000 classes,’’*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1891-1898

[14] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, ‘‘Deepface: Closing the gap to human-level performance in face verification,’’*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1701-1708.

[15] Y. Mao, Shanhe Yi et.al., ‘‘A Privacy Preserving Deep Learning Approach for Face Recognition with Edge Computing’’*Conference proceedings, usenix hotedge18-papers*, pp.1-6, 2016.

[16] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, ‘‘Face recognition: A literature survey,’’in: *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399-458, December 2003, doi: 10.1145/954339.954342.

[17] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, ‘‘Privacy-Preserving Human Activity Recognition from Extreme Low Resolution,” *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017, pp. 4255-4262.

[18] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, ‘‘A Human Activity Recognition System Using Skeleton Data from RGBD Sensors,’’*Computational Intelligence and Neuroscience*, vol. 2016, pp. 1-14, 2016, doi: 10.1155/2016/4351435.

[19] E. Cippitelli, E. Gambi, and S. Spinsante, ‘‘Human Action Recognition with RGB-D Sensors,’’ *Motion Tracking and Gesture Recognition*, pp. 100-115, 2017, doi: 10.5772/68121.

[20] S. Zolfaghari and M. R. Keyvanpour, "SARF: Smart activity recognition framework in Ambient Assisted Living," *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2016, pp. 1435-1443.

[21] M. Ciliberto, D. Roggen, and F. J. O. Morales, ‘‘Exploring human activity annotation using a privacy preserving 3D model,’’*Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct-UbiComp ’16*, September 2016, pp. 803-812, doi: 10.1145/2968219.2968290.

[22] Z. Gheid, Y. Challal, X. Yi, and A. Derhab, ‘‘Efficient and privacy-aware multi-party classification protocol for human activity recognition,” *Journal of Network and Computer Applications*, vol. 98, pp. 84-96, November 2017, doi: 10.1016/j.jnca.2017.09.005.

[23] Z. Gheid and Y. Challal, "Novel Efficient and Privacy-Preserving Protocols for Sensor-Based Human Activity Recognition," *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, 2016, pp. 301-308, doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0062.

[24] R. Yonetani, V. N. Boddeti, K. M. Kitani, and Y. Sato, ‘‘Privacy-Preserving Visual Learning Using Doubly Permuted Homomorphic Encryption,” *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2040-2050.

[25] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, ‘‘Towards understanding action recognition,’’in: *International Conf. on Computer Vision (ICCV)*, December 2013, pp. 3192-3199.

[26] X. Yu, X. Cai, Z. Ying, T. Li, and G. Li, ‘‘SingleGAN: Image-to-Image Translation by a Single-Generator Network using Multiple Generative Adversarial Learning,” *Asian Conference on Computer Vision. Springer, Cham*, December 2018, pp. 341-356, doi: 10.1007/978-3-030-20873-8_22.

[27] X. Peng and C. Schmid, ‘‘Multi-region two-stream r-cnn for action detection,’’in: *European conference on computer vision (ECCV)*, October 2016, pp. 744-759, doi: 10.1007/978-3-319-46493-0_45.
[28] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, ‘‘A discriminative feature learning approach for Deep face recognition,’’*European conference on computer vision*, 2016.
[29] I. Goodfellow *et al*., ‘‘Generative adversarial nets,’’*Advances in neural information processing systems*, vol. 27, pp. 1-9, 2014.
[30] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, ‘‘Image-to-image translation with conditional adversarial networks,’’ *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125-1134.
[31] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, ‘‘Unpaired image-to-image translation using cycle-consistent adversarial networks,’’*Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223-2232.
[32] M. Najibi, P. Samangouei, R. Chellappa, and L. Davis, ‘‘SSH: Single stage headless face detector,’’*Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4875-4884.
[33] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," in *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.
[34] J. Johnson, A. Alahi, and L. Fei-Fei, ‘‘Perceptual losses for real-time style transfer and super-resolution,’’*European conference on computer vision. Springer, Cham*, October 2016, pp. 694-711, doi: 10.1007/978-3-319-46475-6_43.
[35] D. P. Kingma and J. Ba, ‘‘Adam: A method for stochastic optimization,’’*arXiv preprint arXiv:1412.6980*, 2014.
[36] P. Weinzaepfel, X. Martin, and C. Schmid, ‘‘Human action localization with sparse spatial supervision,’’*arXiv preprint arXiv:1605.05197*, 2016.
[37] Thoth Inria, ‘‘Daily Action Localization in Youtube videos,’’[Online]. Available: http://thoth.inrialpes.fr/daly/.
[38] JHMDB Dataset, ‘‘A fully annotated data set for human actions and human poses,’’[Online]. Available: http://jhmdb.is.tue.mpg.de/.
[39] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," *2011 International Conference on Computer Vision*, 2011, pp. 2556-2563, doi: 10.1109/ICCV.2011.6126543.

BIOGRAPHIES OF AUTHORS

Vijaya Kumar Kambala working as Part Time Research Scholar in School of CSE VIT-AP. he is working as Assistant Professor, in PVP Siddhartha Institute of Technology, Vijayawada. His research interest includes Image Processing, Video Analysis, human activity recognition, privacy preserving human activity recognition.

Harikiran Jonnadula working as Assistant professor, School of CSE, Vellore Institute of Technology, VIT-AP. His reserach interest include microarray image analysis , Hyperspectral Imaging and Video Analytics.