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

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. 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.

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

This is an open access article under the CC BY-SA license.

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


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](image)

*A multi-task learning based hybrid prediction algorithm for privacy preserving ... (Vijaya Kumar Kambala)*
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 players and loss function of $D$ is computed as in (1).

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

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

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

In the game, both players have their loss functions. In the process $D$ maximizes $V(D)$ while $G$ maximizes $V(G)$ by updating $\theta^{(D)}$ and $\theta^{(G)}$ 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.

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

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. As presented in Figure 2, the original image (fame) from a video showed that the child is brushing teeth. 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.

![Figure 2. Original image (left) vs. anonymized image (right)](image)

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>Face anonymizer or modifier</td>
</tr>
<tr>
<td>$f$ or $rv$</td>
<td>input face or input face region</td>
</tr>
<tr>
<td>$A$</td>
<td>Action detector</td>
</tr>
<tr>
<td>$D$</td>
<td>The face classifier or discriminator</td>
</tr>
<tr>
<td>$V$</td>
<td>Set of frames in a video</td>
</tr>
<tr>
<td>$F$</td>
<td>Set of face images</td>
</tr>
<tr>
<td>$r_a$</td>
<td>A face region</td>
</tr>
<tr>
<td>$M(r_a)$</td>
<td>Modifying the face region given</td>
</tr>
<tr>
<td>$\nu'$</td>
<td>Updated frame</td>
</tr>
<tr>
<td>$L_f$</td>
<td>sum of the four losses in Faster-RCNN</td>
</tr>
<tr>
<td>$M(f)$</td>
<td>The act of anonymization of given face</td>
</tr>
<tr>
<td>$L_a$</td>
<td>Face detection loss (by discriminator)</td>
</tr>
<tr>
<td>$F$</td>
<td>Original image</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>A weight</td>
</tr>
</tbody>
</table>
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 3 shows the proposed framework named privacy preserving human activity recognition framework (PPHARF).

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) extracted by face detection module and modifies it by removing sensitive information resulting M(f) or M(r). The face region is subtracted from original video frame (v) by face region subtraction module. The combiner module takes modified face region (r) 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).

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

\[ L_{adv}(M,D,F) = -E_{(f \sim F,i \sim I)}[L_d(M(f), i)] = -E_{(f \sim F,i \sim I)}[L_d(M(f), i)] \]  (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).

\[ L_{11} = (M, F)E_{f \rightarrow f}[\lambda ||M(f) - f||1] \]  

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 $\beta_1$=0.5 and $\beta_2$=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 7\textsuperscript{th} epoch.

4.2. 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.

<table>
<thead>
<tr>
<th>Algorithm: Multi-Task Learning based Hybrid Prediction Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> D, V, F, identity labels</td>
</tr>
<tr>
<td><strong>Output:</strong> M, A</td>
</tr>
<tr>
<td>1. For each v in V</td>
</tr>
<tr>
<td>2. f ≜ DetectFace(v)</td>
</tr>
<tr>
<td>3. s ≜ Subtract(f)</td>
</tr>
<tr>
<td>4. f' ≜ M(f)</td>
</tr>
<tr>
<td>5. Compute L1 loss</td>
</tr>
<tr>
<td>6. v' ≜ Combiner (f', s)</td>
</tr>
<tr>
<td>7. a ≜ A(v')</td>
</tr>
<tr>
<td>8. Compute detection loss</td>
</tr>
<tr>
<td>9. det+D (f')</td>
</tr>
<tr>
<td>10. Compute adversarial loss</td>
</tr>
<tr>
<td>11. Update M</td>
</tr>
<tr>
<td>12. Update A</td>
</tr>
</tbody>
</table>

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 4.

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 5. Some of the actions include hug, kick, jump, run, and shoot.

Figure 4. Representative pictures with human action annotated for 10 classes of DALY dataset

Figure 5. Some of the images from JHMDB dataset with annotated human actions

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, super-pixel and edge. The results are also observed in the form of modified images as shown in Figure 6.

Figure 6. Input images from outside the datasets before and after anonymization
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. The anonymized samples utilized for user study are shown in Figure 7. Experiments made with DALY and JHMDB with respect to face verification by discriminator in the adversarial learning led to the observations made in Figure 8.

As presented in Table 2, the results revealed that for many actions, the proposed approach showed better performance. As presented in Figure 8, the face recognition performance is compared against different anonymization methods. The results are observed for both JHMDB and DALY datasets. Since the face recognition is to be prevented as per the adversarial setting, the high error rate in recognition indicates higher performance. Accordingly, the proposed approach has shown better performance over the state of the art. From the empirical study it is understood that there is improvement in the privacy preservation and also action recognition. 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.

![Figure 7. Anonymized images of the proposed system for user study](image)

![Figure 8. Face recognition performance using. (a) JHMDB (left), (b) DALY (right) (Note: high error rate means high performance)](image)
REFERENCES


A multi-task learning based hybrid prediction algorithm for privacy preserving ... (Vijaya Kumar Kambala)


A multi-task learning based hybrid prediction algorithm for privacy preserving … (Vijaya Kumar Kambala)