Advances and Applications in Mathematical Sciences
Volume 20, Issue 12, 2021, Pages 3333-3350

EFFICIENT HUMAN ACTIVITY RECOGNITION WITH PRIVACY SOLVING THE PUBLIC AND PRIVATE ACTIVITIES OF VIDEOS AND IMAGES VIA DEEP ENSEMBLE LEARNING

VIJAYA KUMAR KAMBALA, J. HARIKIRAN and B. SAI CHANDANA

Assistant Professor
Department of CSE
Prasad V Potluri Siddhartha
Institute of Technology
and Research Scholar
School of CSE
Vellore Institute of Technology-AP
VIT-AP, Amaravathi, India
E-mail: kvkumar@pvpsiddhartha.ac.in
vijaya.18phd7017@vitap.ac.in

School of Computer
Science and Engineering
VIT-AP University
Vijayawada, India
E-mail: harikiran.j@vitap.ac.in

Department of CSE
GEC, Gudlavalleru Engineering College
Gudlavalleru, India
E-mail: bschandana@gmail.com

Abstract

Now a day's everybody thinks about a secure and serene life with no problems at all, for which he/she needs privacy protection of their data. We can see that at every place the

2010 Mathematics Subject Classification: 68.
Keywords: Human Activity Recognition, Inverse super resolution (ISR), Image Classification, Low Resolution Images.
Received November 20, 2020; Accepted December 17, 2020

movement and activities of people are captured and recorded, causing greater worry to the privacy of individuals because of surreptitious video recordings. To some extent, these images and videos assist the authorities and individuals in controlling crime, but one cannot ensure privacy of a recorded video/image. The privacy of individuals is at stake. The research paper presents fundamental concepts to provide security to the videos and images of individuals by recognizing the various actions of humans with extreme low-resolution (e.g., 16×12). This can be achieved by introducing the paradigm of Inverse Super Resolution. The concept of learning the optimal set of video transformations is needed to generate multiple low-resolution (LR) training videos or images from a single video or image. The inverse super resolution tells various types of different sub pixel transformations optimized for the human activity classification and reorganization, taking the classifier to take advantage of existing high-resolution images and videos (e.g., any recorded or YouTube videos) and by taking the multiple low resolution training anonym zed videos, tailored for different problems. Some real videos are collected from capturing cameras installed at various public and private places at different locations. The proposed method recognizes human activities in videos with the concept of privacy.

1. Introduction

A more fundamental solution toward the construction of a privacy-preserving vision system is the use of anonymized videos. The idea is that, if we are able to develop reliable computer vision approaches that only utilize such anonymized videos, we will be able to do the recognition while preserving privacy. The videos are becoming increasingly pervasive. Wherever we go, many surveillance cameras are found in public and private places. People are using different types of devices and hidden cameras to record the activities of others and later upload them on social websites. This kind of anonymized video, recording at every place has drastically increased, especially in shopping malls, educational institutions, residential areas. [1]. Different videos of the human actions will be recognized and traced according to their work and their need.

All the recoded videos of the human activity will be stored and watched. But in real time applications the data taken from different locations will be violated [2]. Therefore, the data of human activity will be taken and using it for a different purpose and posting in various social media. So, we want to take step to finalize the videos analyze them based on our requirements. Because the data, which is finalized, may contain private and public information, hence, should not be shared or kept in public platforms because of privacy issues. For example, if the data of any hospitals is taken, it may be video, there may be contradiction that it should not be shared [1] [2]. The recorded videos are converted in to low resolution to high resolution sing the inverse super resolution (ISR). The images that are taken in to high resolution are classified by polling techniques. The converted data will be compared with other images by different layers of the neural network.

Figure 1. Throwing a ball the image of LR from a 1st-person.

Figure 2. LR image with a different camera transformation.

The above two figures clearly show images of human activity in different directions. Figure 1 shows throwing of a ball in different style and by different persons and moreover the images are captured in different directions. However, the two persons are throwing the ball comparing the two messages and recognized and identified based comparative measures [2] [3]. In the figure 2 people are going on different bikes with different directions and their images are captured and compared. The figures 1 and 2 are compared separately with nearest clustered pixels and trained accordingly.

Human will always need to free in every place [4]. Human activity recognition is a very complex and challenging task. Idea behind of Human activity reorganization is to collect the data with video and images of the all the collected and stored data should provide the privacy and protect the data. The objective is not to watch the complete video but should be converted to low resolution for privacy. Inverse super resolution is a novel concept to solve the image resolution conversions.

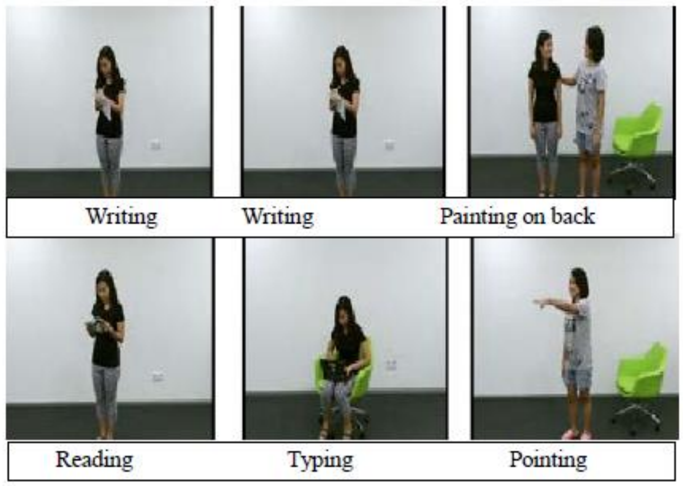

Figure 3. Human-activities compared with different images.

The above figures show the different types of human activities are shown. The images are in writing with different activity, painting back of the other human, standing in a position with reading activity, typing, and last pointing at standing position. Taking the videos of human with high resolution and making it to low resolutions and the images cannot be finalized and taking this type of data will not be given to required persons. The generating and combining of informative low-resolution images from one high resolution image. Therefore, the low resolution of privacy protection, the images of an access to a rich set of videos publicly available and taken from the real time videos and will be taken to convert (e.g., YouTube). The concept that will be taken from the low resolution training of inverse resolution is that, will give amount of data to be taken to compare [7] [8]. The approach decides the ideal transformations of the images and there is possibility for generating the resolutions, uses the generated LR videos to obtain various decision boundaries. The image based on the stored data will be finalized.

2. Related Works

Inverse super resolution (ISR) is the concept of generating a set of low-resolution training images from a single high-resolution image, by 'learning' different image transforms optimized for the recognition task. Such transforms may include sub-pixel translation, scaling, rotation, and other affine transforms emulating possible camera motion. The various activities of human activity is a computer vision and is challenging issues from real time day to day life (Aggarwal and Ryoo [17]). The information getting from various places from public and private. The videos are shown in social media and videos from you tube and cc cameras from residential apartments. Figure 2 also the recording the videos and the usage of the recorded videos are taken for different purposes and posting them in various unwanted purposes. Problems are recognized and focused with reliable resolution and they dint not concentrated on human privacy on images and videos [9].

We focused on human privacy of images and videos as explained in introduction part, there are many research works with different problems with solutions, specifically focused on privacy concerns capturing the images and recording the videos. The different techniques has been used to automatically verify and detect various locations. Studied human activity recognition from extreme low-resolution videos, different from conventional activity recognition literature that was focusing mostly on methods for images and videos with sufficient resolutions. Although their work only focused on recognition from 3rd person videos captured with static cameras, they showed the potential that computer vision features, and classifiers can also work with very low resolution videos. However, they followed the 'conventional paradigm' described. The images or videos taken and converting low resolution to high resolution [10]. Resizing the images and taking the final data for consideration.

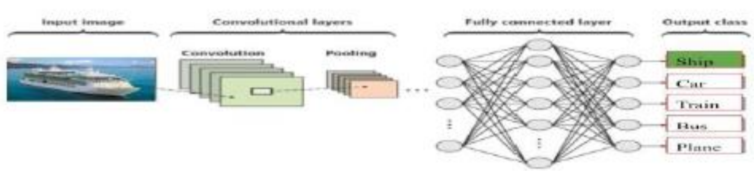

Figure 4. CNN architecture.

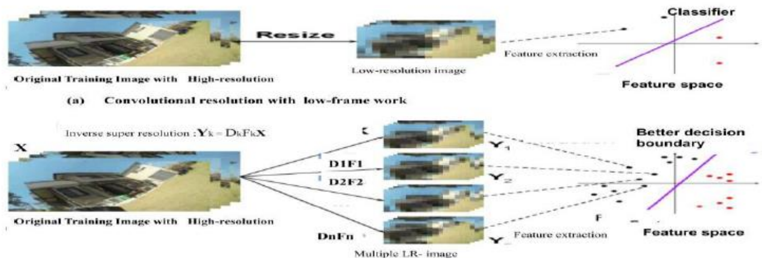

Figure 5. Our proposed learning using inverse-super-resolution.

3. Super Resolution of Image with Inverse

Explains the images or videos taken from the various data sets or from the stored data recording or live data recordings with resolutions with high and this will generated to a set of low-resolution image, taking the different images from various public are private work places. The data is trained by taking the algorithm of convolutional neural network and by applying pooling each pixel will be taken distance is calculated from each image samples [10] [9]. Transforms may include sub-pixel translation, scaling, rotation images will be different shapes and may be applied to different ways of representation.

The inverse resolution technique does things more real time data with realistic scenario where the system is prohibited from obtaining high-resolution videos. The data taken from the various systems and will be finalized and possibly this can tested in software levels where the requirements gathered and analyzed things and therefore it will be tested. After taking the tested data the final low-resolution data will be taken and pooled by trained in neural network [13]. The resolution taken from different images perspective makes various ways of using the super resolution formulation, the resolution of various images and videos will be comparable amount of data taken with different resolutions [14].

3.1.1. Formulation of Resolution with Inverse

Ultimate thing is that or we can also say that goal of this inverse of image is to take a series of low resolution images *YK* and generate a high resolution output image *X*. This formation will be continued by taking the sequence of images *YK* to be different views of the high-resolution image *X*, subject to camera motion, lens blurring, down-sampling, and noise. Each of these effects is modified according to the sequential machine minder, this can take in sequence of images [11] [12].

$$Y_k = DHFX_k + V_k^1, K = 1_k, \ldots, n. \tag{1}$$

Where F_x is the transformation, H_k models the blur ring effects, Y_k is the down-sampling operator, and V_k is the noise term. The images flattened into 1D vectors are X, Y_k. The operators that are taken for the solving the image resolution with reverse will know with exact pixel size, finally it will be in accurate levels. So the estimating the image resolution research and can be focuses on estimating these values of X by adding additional requirements.

The generation of multiple of n, the images Y_k resolution for every pixel calculation and high resolution image X, to solve the optimal set of image transformations F_k, H_k and D_k. By calculation the formulation by deleting the noise term V_k and the lens blur term H_k because of these reasons to know. The resolution with low and images that are to be deviated [15] [16].

$$Y_k = D_k F_k X, k = 1, \ldots, n. \tag{2}$$

By considering the combinations of shifting, scaling, and rotation transforms as our F_k. Using the standard average down sampling as our D_k. Prominent solution that can be best by converting the images to low to high from specified tracks. Calculating the image comaprasoion techniques of various levels and finding the low resolution can possible by calculating the Y_k in verified and tested pixels of the image or video samples taken and convoluted and polled[14] [15].

4. Transformation Learning

The advantages of the frame work that is taken with low resolution of images and videos can be made explained clearly. By taking the set of images and trained accordingly with different motions and video blurred [16]. This type of motions can be taken in the various solutions taken in various research works in process. $S = \{F_k\}^n$. The images and videos data set that taken and polled from the convolution neural network (CNN). The S is taken from the images and videos of the particular data from and taken as final trained with various transformation techniques applied to get the variable results. The Low-resolution videos and images from various public and private of a educational, hospital rooms are captured and enhanced with a privacy preserving model, before the frames are fed to an action recognition model to perform tasks such as hand-hygiene monitoring or activity logging. We used the original 224, 224 depth images by different scales. Methods are to distort images include Gaussian blurring or super pixel clustering, the images that are for suitable for depth images and videos that do not have many sharp edges for any images. By collecting the high-fidelity images altogether by using LR camera hardware, settings to reduce memory for data storage and can be analysed. The resolution sizes of different images will be 16_16 and will be analyzed. There is not enough visual information to discern facial features and can be shown in showing with different ways that general privacy guideline for the levels [10] [11].

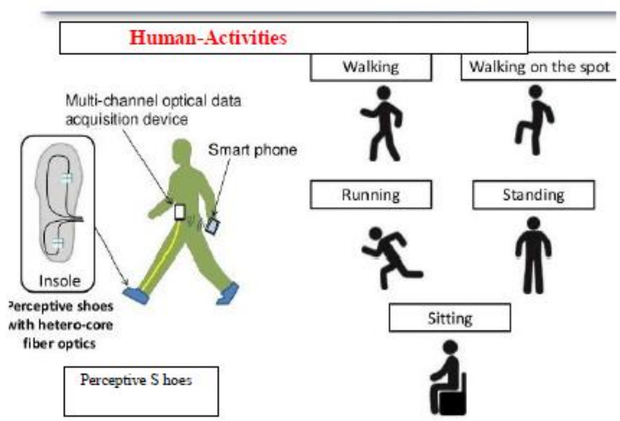

Figure 6. Human-Activities recognized and control.

The different activities of human are shown in above figure-6.Every activity is control by system and recognized how the positions of the human is in different style. Downsizing the images starting from 224_224, we can downsize our images by 16 to 14_14 images. The dataset consists of full-body viewpoints where the face takes up at most 56_56 of an image. We also perform experiments down sampling by 6x to 36_36 that provides somewhat less privacy assurances are shown. Private resolution is used to perform super resolution. The Deep ensemble is a CNN consisting of a feature network with extraction with reconstruction network, and is trained by feeding in pairs of low resolution and high resolution images to find the optimal super-resolution weights. The models on images trained and sampled with resolutions of 36_36 and 14_14 on the open source action recognition dataset [10] [14]. Taking the models and training on public datasets disjoint from our dataset, considering the data with privacy can be trained again with various multiple ways of sampling by taking the down sampling of images can be placed, so that this can be taken in to account by providing the sample inputs from various data sets.

5. Experiments

Performing the experiments by taking the sampling of video recording. By training an Imagenet and pretrained [7] model. We include examples of down sampled images and respective DCSCN outputs. Due to many classes, there will be no proper imbalance present in our dataset, we applied and performed the data augmentation on our dataset and sample images which we have taken for training. Taking the different image and video recordings of the human and applied random transformations and giving the input with sequentially in equal numbers of positive or negative disposer of usage frames during the training of different phase. Scaling down the hand-hygiene dataset does not because significant drop in utility of the usages in real time. Our experiments show that basic down sampled images already preserve a practical amount of utility for each task taken from different images. The original images that are taken from different data sets the resolutions of low can be performed in the outputs taken and the images cannot be reducing the resolution. Applying the polling in number times and will be trained on machine with number of networks. We use a Resnet-18 to process the ICU actions, using data augmentation to balance our classes. Improve the accuracy for 14 _ 14 images and only slightly improves the accuracy for 56 _ 56 images, it improves the AUC for several classes such as 2 and 4 for 14 _ 14 images and class 1 and 4 for 16 _ 16 images. In addition, the super-resolution enhanced images are more visually interpretable and easier to annotate. Most importantly, we can see how it is visually impossible to discern any personally identifying information from a 14 _ 14 frames.

The human actions will be identified by using the CNN algorithms. Taking the different samples problem setting Interaction Dataset. If two people interacted and the images can be set in dataset for video-based actions are recognized, taking samples of different images and will be analyzed based on the different behaviors and captured with different actions. We note that the target trade-off is highly. The supple of the images can be taken, the actor pair recognition task can be achieved over 96% accuracy on the original dataset and stands robust even when the frames are down sampled 35 times, the human action recognized with good performance [16]. The below methods. 1. Method-I. Naive down sampling using raw RGB frames under different down-sampling rates this can be in any formations and will be stated that the possibilities of the different native methods will be considered. Method-II proposed restarting and taking the trained images and applied and the images and videos of different pixel representations cannot be considered. So various features can be shown. Method-III proposed applying the predicted data can be trained and final RGB frames, using budget model ensemble with restarting. Method=IV. Detecting and cropping out faces from RGB frames. Method -V detecting and cropping the images of everything from RGB frames. Method-I follows, while Methods IV and V are inspired. Implementation details combined into video sequencing of the different frames, and use those frame groups as our default input data *X*. We use the C3D net [60] as the default action recognition model, i.e. *fT*. For the *fb* identity recognition model, we choose Mobile Net to identify actor pair in each frame and use average pooling to aggregate the frame-wise predictions. We then freeze them both, and start initializing *fb* (MobileNet) for the actor pair identification task, by adapting it to the output of the currently trained *fd*. The values will be taken and finalized according to input taken from the images and videos from the recordings of taken as in stored data [13] [14].

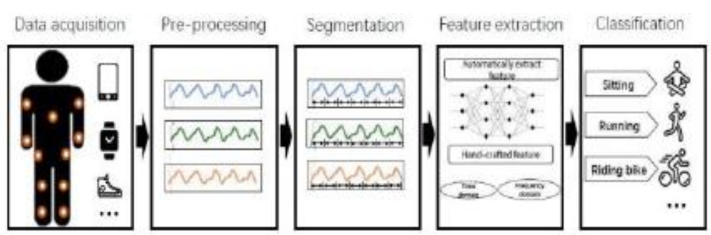

Figure 7. Image processing with different classes.

The figure 7 shows clearly the image processing with different levels and the image data acquisition. The processing can be noted from the above figure with images. The processing can be done and images can be traced with different samples. Segmentation of the processed data can be followed in various data levels [15] [16]. The feature extraction can be placed on the image levels and mainly the extraction can be done in various levels and will be placed in segmentation levels. Based on the image segmentation and extraction levels this can be replaced accordingly. In above figure shows the images with different levels can be sitting, running, riding bike.

6. Pre-Processing of Data

6.1.1. Processing of Data instance classifier:

The data taken from the different images and will be processed the processed data can be processed with known images. For each sample data, time domain features including mean, variance, root mean square, and the maximum and mini- mum values.

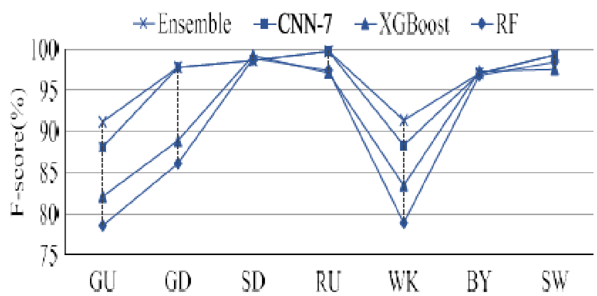

Figure 8. Graph shows the image ensemble with CNN.

Table 1. Distribution of data.

Activity	Distribution
Throwing a ball	43172(18.25%)
Distance of thrown	25364(15.56%)
Ball in the air to ground	45152(22.15%)
Riding the Bike	45875(17.55%)

Distance of two persons	64252(16.45%)
Movement of a person	24451(14.53%)

Algorithm: Ensemble of CNN with Framework

Input: Image conversion with high resolution with low resolution.

Output:

Human activity reorganization with privacy.

1. Step 1: Taking samples of human activity of images.
2. For i to n samples repeat.
3. CNN with 9 networks shows the predicated activity.
4. Checking the activity, and converted it in to High to low resolution.
5. If images show the nearest activity then repeat.
6. Probility of getting the same image with GU and k with $p^1_k p^2_k p_w^2$.
7. CNN-2 network predicts the results with $p^2_k p^2_k p^2_k$.
8. $\alpha = P^1_\mu$, $\beta = p^1_w$.
9. else
10. $P_k = a + P^1 + (1-a) \times P^2$.
11. end if
12. return Predicated activity.
13. else
14. return predicated activity from CNN-6
15. end STOP.

The table 1 show the data distribution of the human activity of different locations showed in taken from different locations. The images are compared with distance. The riding the persons with different persons and their movements and shown with different angles. The table showed clearly the various levels of distribution the data with values. The percentages of values will be taken from the each location of the images [12] [16]. The various images locations can be with captured with different styles in public places. This is what happing in public places where the privacy is lagging and this will be the same and different situations and can be analyzed and why the particular video or images are taken.

7. Data Pre-Processing of Deep Classifier

In the end-to-end architecture the images can be shown clearly, the additional features can be taken for processing the image of human activities.

The out of the data can be taken according the images captured. The terms GU(going under)shows the various levels that can be placed. GD (going down) this will be showing the ball is taking from the ground. SD (Standing) this shows the best way of image classification. WK (walking image shows the various places in the human will be public places. To meet the format requirement of the proposed CNN model, a sliding window segmentation approach with fixed step size of fifty seconds is applied to each sensor data. In our work, the raw sensor data stream is cropped into the same size with an overlap of 15%. [15].

8. Architecture of Deep Learning

Deep learning models are a class of machine learning models that can learn features from raw data. Deep learning techniques have been applied to several domains such as object recognition [6], human tracking [7], natural language processing [8], human action recognition [6]. We have given the learning tool used by the techniques presented. Supervised and Unsupervised learning algorithms can be supervised or unsupervised. Both are different approaches. Supervised techniques need training data while unsupervised techniques do not need any training data. The proposed CNN-based model has 3 kinds of layers: 1. Is an input layer, as explained in the above 2. Convolutional layers is specifies the image of human activity from input data; 3. Max-pooling layers shows clearly that the images from human can be reduce the size of extracted features and enhance the strongest of the known images of the human activity.4.Fully connected layers integrate all features extracted.5.An output layer of the softmax function represents a categorical distribution over seven different activities[16].

9. Convolutional Layers

Convolutional Neural Networks (CNNs) are a type of deep models that uses multiple layers, shared weights, local connections in the feature maps (units) of the previous layer through filter bank (set of weights) for learning complex features and a pooling layer to alternating combine semantically similar features into one from raw input [9]. CNN is a special network of networks when comparing with other layers of network in terms of sparse connectivity can be done with sparse. The layers of the many different layers and values sharing the pooled of the networks. The layers can be viewed in ways of layer are only connected to a local subset of units in the input layer. The network uses the different filters. The convolution neural network can be explained in terms of the images can be polled in and classified in terms of the cropping can be done. The ensemble models can be shown in different tables, and the test values of different images taken from human activity with various values in percentages.

Table 1. Ensemble model (%).

	GU	GD	SD	RU
Test-1	98.56	99.65	87.85	99.45
Test-2	89.78	98.68	99.78	96.25
Test-3	98.59	97.89	87.85	87.85
Test-4	99.65	87.85	99.45	99.78
Test-5	98.59	97.89	98.59	97.89
Test-6	87.85	87.85	87.85	99.65
Test-7	99.78	87.85	87.85	97.89
Test-8	99.65	97.28	98.59	97.89
Test-9	98.68	97.89	98.59	97.89
Test-10	98.68	99.78	96.25	97.28

The table-1 shows with values of various human activity images are captures in public places. The values shows nearest to the compared images in the different know values of the images. The TEST-1 to TEST-10, the known can be clearly explains the image cropping or converting to high resolution to the low level resolution in order to achieve the privacy.

Table 2. Ensemble model (%).

	GU	GD	SD	RU
Test-1	98.56	99.65	87.85	99.45

Test-2	89.78	98.68	99.78	96.25
Test-3	98.59	97.89	87.85	87.85
Test-4	99.65	87.85	99.45	99.78
Test-5	98.59	97.89	98.59	97.89
Test-6	87.85	87.85	87.85	99.65
Test-7	99.78	87.85	87.85	97.89
Test-8	99.65	97.28	98.59	97.89
Test-9	98.68	97.89	98.59	97.89
Test-10	98.68	99.78	96.25	97.28

Table-3. Ensemble model (%).

	GU	GD	SD	RU
Test-1	98.56	99.65	87.85	99.45
Test-2	89.78	98.68	99.78	96.25
Test-3	98.59	97.89	87.85	87.85
Test-4	99.65	87.85	99.45	99.78
Test-5	98.59	97.89	98.59	97.89
Test-6	87.85	87.85	87.85	99.65
Test-7	99.78	87.85	87.85	97.89
Test-8	99.65	97.28	98.59	97.89
Test-9	98.68	97.89	98.59	97.89
Test-10	98.68	99.78	96.25	97.28

10. Summary of Outputs

The output values are given in table. The data is provided with comparing the previous data and standard values, camera low resolutions.

Observations	Camera values	Standards Values	Test-Values(Table-1&2&3)
Average-Values	20×20, 40 Hz, 4	99.85%	2.74%

Frame-value	10×10, 2 Hz, 5	96.89%	2.68
All-Cameras	20×20, 40 Hz, 1	89.69%	2.15
less Spatial values	2×2, 40 Hz, 5	88.85%	2.18
Below average	1.5×1.5, 3 Hz,	55.55%	2.99

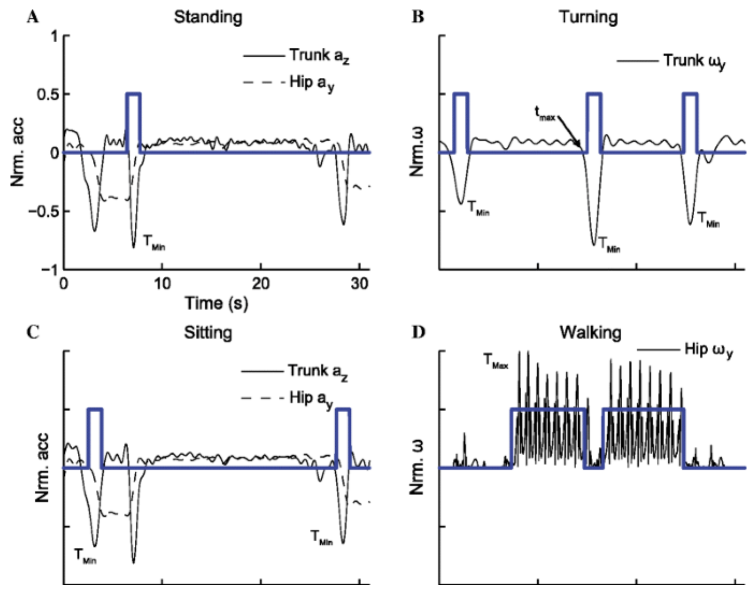

Figure 9. Enhanced human activity recognition.

11. Conclusions

The Human Activities Reorganization will be finding in different public and private areas and can be recorded and captures without any privacy. We are converting these captured images to high resolution to the low resolution so, that we provide the privacy for recorded videos and images. Based on experimentally confirm its effectiveness of different image datasets. Final outputs can be recognized for a quality performance with various datasets and our approach better than previous techniques, in future work to Enhanced Advanced Deep Learning Techniques. We trained a model using depth wise convolution neural network for human activity recognition. Result shows that our method provides privacy for human activities.

12. Acknowledgements

Vijaya Kumar Kambala, J. Harikiran and B. Sai Chandana have done the research work with experimental results Efficient Human Activity Recognition with privacy solving the public and private Activities of videos and images via Deep Ensemble Learning.

References

[1] C. V. K. Vijaya and Kumar J. Harikiran, Privacy-Preserving Human Activity Recognition and Resolution Image using Deep Learning Algorithms Spatial relationship and increasing the attribute value in Open International Journal of Advanced Science and Technology 29(7) (2020), 514-523.

[2] S. Ryoo and Brandon Rothrock, Privacy-Preserving Human Activity Recognition from Extreme Low Resolution, Michael Dec. 2016.

[3] Zhongzheng Ren and Yong Jae Lee, Learning to Anonymize Faces for Privacy Preserving Action Detection, 30, March, 2018.

[4] Yucai Bai, Qiang Dai and Long Chen, Extreme Low Resolution Activity Recognition with Spatial-Temporal Attention Transfer, 9 Sep. 2019.

[5] Partho Ghosh, Md. Abrar Istiak and Nayeeb Rashid, privacy-aware activity classification from first person office videos. 11 Jan., 2020.

[6] Qingchang Zhu, A Novel Semi supervised Deep Learning Method for Human Activity Recognition, IEEE Transactions on Industrial Informatics 15(7) July(2019).

[7] Deep Learning Class Discrimination Based on Prior Probability for Human Activity Recognition. Received December 1, 2018, accepted December 19, 2018, date of publication, January 11, 2019, date of current version February 8, 2019.

[8] Improving Cross-Subject Activity Recognition via Adversarial Learning. Received March 19, 2020, accepted May 5, 2020, date of publication, May 11, 2020, date of current version, May 27, 2020.

[9] A Hybrid Network Based on Dense Connection and Weighted Feature Aggregation for Human Activity Recognition Received March 23, 2020, accepted April 3, 2020, date of publication April 7, 2020, date of current version April 22, 2020.

[10] Efficient Human Activity Recognition Solving the Confusing Activities Via Deep Ensemble Learning. Received May 7, 2019, accepted May 21, 2019, date of publication June 10, 2019, date of current version June 21, 2019.

[11] Improving Cross-Subject Activity Recognition via Adversarial Learning. Received March 19, 2020, accepted May 5, 2020, date of publication May 11, 2020, date of current version May 27, 2020.

[12] Vijaya Kumar Kambala and J. Harikiran, Privacy-Preserving Human Activity

Recognition and Resolution of the image using Deep Learning Algorithms of Convolutional neural network and recurrent neural network, Journal of Shanghai Jiaotong University, Volume 16, Issue 7, July 2020.

[13] Deep Learning for Multiple-Image Super-Resolution, IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 6, June 2020.

[14] Bit-Scalable Deep Hashing With Regularized Similarity Learning for Image Retrieval and Person Re-Identification, IEEE Transactions on Image Processing 24(12) December 2015.

[15] A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction. IEEE Transactions on Medical Imaging 37(2), February 2018.

[16] Multi-Scale Deep Residual Learning-Based Single Image Haze Removal via Image Decomposition, IEEE Transactions on Image Processing 29, 2020.

[17] Aggarwal, J. and Ryoo, M. Human Activity Analysis: A Review. ACM Computing Surveys, 43, 16:1-16:43 (2011).