ENHANCING HUMAN IDENTIFICATION THROUGH COMPUTER VISION TECHNIQUE.

Supported by tetfund

J.O. Okelola[1,*], V.I. Yemi-Peters[2], S.E. Adewumi[3.]

[1,2,3] Department of Computer Science, Federal University, Lokoja, Nigeria,.
* Corresponding author: john.okelola-pg@fulokoja.edu.ng.

Received: September 14, 2024 Accepted: November 28, 2024

Abstract: A sophisticated system for item detection and recognition is in high demand globally. The decision to focus on this research article was influenced, in part, by the increasing security concerns in Nigeria. Despite the implementation of various traditional methods to combat these issues, the threat continues to persist. Therefore, it is imperative to transition from outdated identification techniques to more contemporary ones. This study introduces a computer vision system using the Convolutional Neural Network (CNN) methodology with yoloV4 as the algorithm to offer a more efficient and effective approach to detecting human presence and verifying their identity. YOLO V4 is used to detect the human while CNN extracts the features of the human images from the recorded video. To assess performance, the system is compared against various models such as YOLO and YOLO V2. The proposed system demonstrated an identification accuracy of 72% on the MSMT17 datasets and a detection accuracy of 80.1% on the MS-COCO datasets, respectively.

Keywords: Human Identification, Object Detection, Person Re-identification, Computer Vision, Pattern Recognition.

Introduction

Considering the escalating global threat posed by terrorism and violence, there has been a notable surge in interest surrounding video surveillance as an identification technique (Kim et al., 2012). Person identification has been a thriving research field for the past twenty years, owing to its pivotal role in verifying individuals in sophisticated applications like surveillance, forensics, and access control (Khan et al., 2021). Smart Human Identification (SHI) pertains to video-level processing mechanisms utilized for the recognition of humans from live footage.

The field of computer vision has extensively explored image classification, resulting in outstanding outcomes in global competitions like ILSVRC, PASCAL VOC, and Microsoft COCO through the implementation of deep learning techniques (Krizhevsky et al., 2017). The identification of human beings holds great importance in various intelligent applications, including but not limited to smart homes, traffic control systems, and human-computer interfaces. Due to the current high-security challenges facing nations worldwide, computer vision has become increasingly valuable for identifying individuals involved in criminal activities and aiding in human identification and tracking.

The escalating insecurity issues plaguing Nigeria and other similar cases of identity theft have spurred the focus of this research paper. Although several conventional models have been employed towards resolving these problems, they seem inadequate against their persistent nature. Therefore, there is a pressing need to transition from traditional methods of personal identification towards more advanced technological approaches that incorporate smart computer vision systems capable of accurately recognizing individuals.

Related Work

Object Detection

Object detection is the process of identifying the specific category to which an object belongs and estimating its spatial location by generating a bounding box that surrounds it (Pathak et al., 2018). Extensive employment of deep convolutional neural networks (CNNs) has been witnessed in the field of object detection. A feed-forward neural network, CNN operates on the principle of weight sharing. Before the advent of deep learning, the deformable part model technique (Felzenszwalb et al., 2010) was predominantly employed for object detection. This technique facilitates multi-scale-based object detection and localization.

Person Re-identification (Re-ID)

When provided with a query individual of interest, the objective of Re-identification (Re-ID) is to ascertain if this individual has been recorded in a separate location at a different time by a separate camera, or even the same camera at a different time frame (Ming et al., 2022). Person re-identification falls in between image classification and instance retrieval regarding the correlation between the classes used in training and testing (Gheissari et al., 2006).

A hand gesture recognition model that works well in real-time applications was presented by (Mopidevi et al., 2023). Tensorflow in OpenCV and Python programming languages, the Google media pipe framework, and feed-forward neural networks using Keras models for categorization are used in the development of the model. With a 95.7% accuracy rate, the suggested model can identify 10 different hand gestures: the thumbs up, thumbs down, peace sign, smile, rock on, okay, fist bump, live long and prosper, call-me, and stop signs.

A complex method called Gate-ID was presented by (Zhang et al., 2021) that can reliably identify people's identities regardless of which way they are walking. The study illustrates how antenna array orientations and walking directions contribute to the mirror-like patterns seen in WiFi signals using both theoretical communication models and real-world data. An innovative heuristic technique is presented that can accurately determine a person's direction of walking. The findings demonstrate that Gate-ID achieves remarkable accuracy rates for human identification, ranging from 90.7% for six-person groups to 75.7% for twenty-person groups, respectively.

Benkaddour et al. (2021) devised a convolutional neural network-driven gender prediction and age estimation system that operates on both face images and real-time videos. The findings of this study indicate that the utilization of CNN networks significantly enhances the performance and recognition accuracy of the system.

Koo et al. (2023) suggested a design for a cascaded model that could categorize 14 upper-body workouts by employing z-axis acceleration data from an IMU sensor. This model utilized a decision tree as the first stage and a one-dimensional

convolutional neural network as the subsequent phase. The outcomes demonstrated an overall enhancement in exercise classification precision using this method, with accuracy ratings reaching approximately 92%, surpassing the results of the 1D-CNN model which was recorded at only 82.4%.

Research Methodology

Research Design

This study aims to design an efficient system framework for human identification using the YOLOv4 algorithm and evaluate its performance. This study employs a quantitative approach, leveraging computer vision and deep learning techniques to achieve accurate human identification.

Data Collection

For the development of this system, two different datasets have been chosen for use: the MS-COCO datasets and the MSMT17 datasets with a few custom datasets. The MSMT17 dataset was specifically selected for its relevance to the re-identification task at hand. MS-COCO provides a wide range of images for object detection, while the MSMT17 dataset focuses more on person re-identification, making it a crucial component of the system's development. By incorporating these two datasets, the system will have a solid foundation for accurate and efficient performance in the detection and re-identification task.

Data Annotation

Data annotation acts as the bridge between unstructured data and structured data, transforming it into a format compatible with machine learning applications (Kokhan, 2024). Data annotation involves the systematic categorization and labeling of data to facilitate the effective implementation of artificial intelligence applications. The process of annotating the datasets for this research project follows a simple approach using an annotation tool called labelling.

Data Augmentation

Data augmentation is a term used to describe the utilization of unobserved data or latent variables in order to create iterative optimization or sampling algorithms (Dyk & Meng, 2001). To ensure the model learns all the essential features in the provided datasets, a yolov4 freebie was conducted. This Computer Vision freebie also acts as a regularizer that helps reduce the overfitting of the model. The pixel values are then normalized to ensure they are within the range of [0, 1], which aids in optimizing training and achieving convergence.

Model Setup

CUDA Toolkit 10.0 is downloaded via this link: https://developer.nvidia.com/cuda-toolkit-archive and was installed on a Ubuntu operating system. CUDNN was installed and YOLO V4 (clone) GitHub repo was downloaded via https://github.com/AlexeyAB/darknet. The algorithm is built using the CMake command in the darknet folder of the cloned GitHub. The implementation of this thesis is carried out in three phases with different steps. These phases include the training phase the testing phase, and the integration phase.

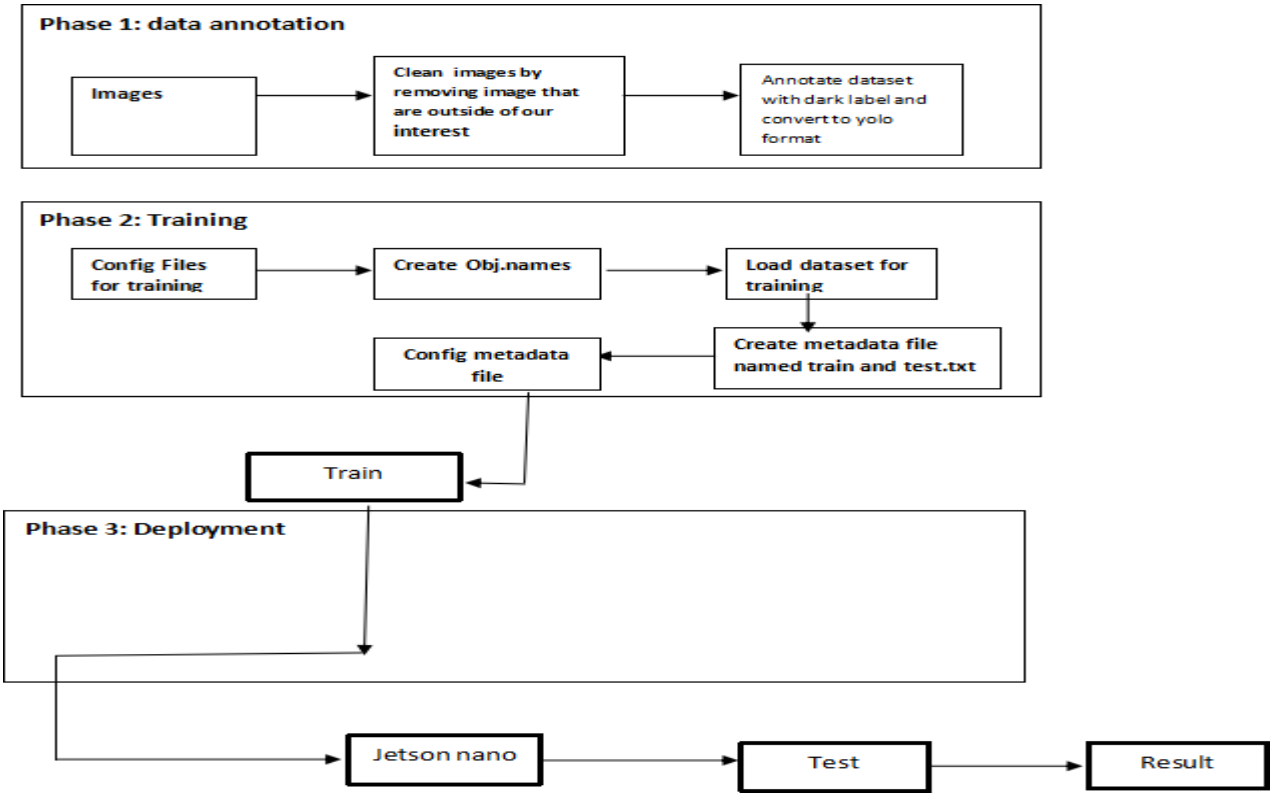

Figure 1: Development Process for the Proposed System

Experimental Results and Discussion

Dataset

MSMT17 Dataset

The MSMT17 dataset is considered one of the largest collections for person re-identification, containing over 126,000 images that depict 4,101 unique individuals. This dataset is particularly notable for its vast number of identities and images, as highlighted by (Sovrasov & Sidnev, 2021). The datasets were gathered in a real-world environment, capturing various perspectives, occlusions, and lighting scenarios. This diverse range of conditions makes it a valuable resource for this project.

COCO Datasets

There are 91 common object categories in the Microsoft Common Objects in Context (MS COCO) datasets, and 82 of them have more than 5,000 annotated instances. The dataset is a popular resource for posture estimation models because it includes 2,500,000 labeled instances among 328,000 images (Lin et al., 2014). The COCO datasets contain more instances per category than the popular ImageNet but have fewer categories overall (Deng et al., 2009). A wide range of annotated data is included in these datasets for tasks like object detection, semantic segmentation, instance segmentation, key-point detection, and picture captioning (Hasib et al., 2021). A JSON file contains the annotated picture storage.

Algorithms and Model

YOLO-V4

YOLO-v4 is an advanced and efficient One-Stage object detection algorithm developed in 2020, incorporating features from previous versions such as YOLO-v1, YOLO-v2, and YOLO-v3 (Yu & Zhang, 2021). It has achieved the optimal balance between detection speed and accuracy trade-offs, setting a new standard in the field. YOLOv4 is renowned for its significant improvements in Average Precision (AP) and Frames Per Second (FPS).

There are two classes of models in object detection: single-stage and two-stage detectors. Two-stage detectors operate in separate stages, first identifying important regions and then classifying those regions to determine if the object is present. YOLOv4, as a single-stage object detector, offers greater accuracy and speed compared to two-stage detectors such as R-CNN and Fast R-CNN. Yolo algorithms differ from most neural models as they employ a single convolutional network to predict bounding boxes and their associated probabilities. These bounding boxes are assigned weights based on the probabilities, influencing the model's detection process. This allows for the direct maximization of the model's end-to-end output, resulting in rapid image production and processing. Each bounding box can be described using four descriptors:

i. Class Number
ii. Object center coordinates in x
iii. Object center coordinates in y
iv. Object width and height

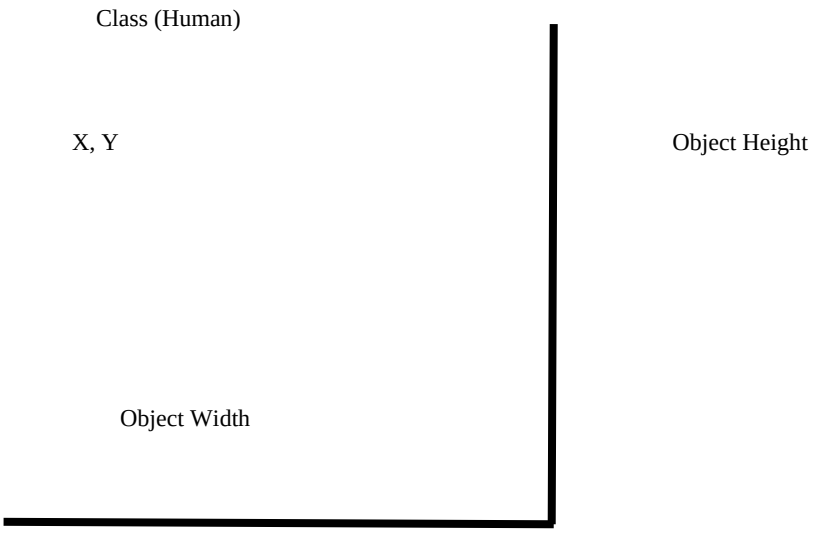

Figure 2: Graph showing the bounding boxes.

When there is an object detected, the class of the object is identified and a rectangle bounding box is put around the detected object.

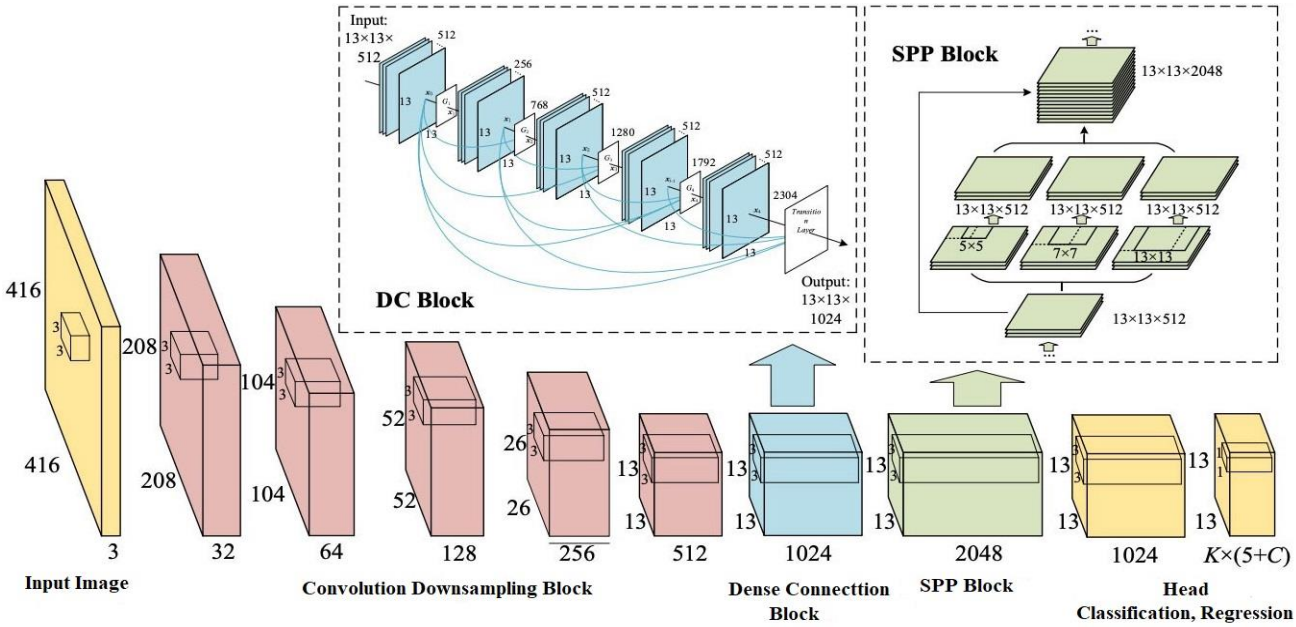

Figure 3: The architecture of the YOLO-v4 (Source:(Jeny et al., 2022))

In modern object detection systems, there is a common practice of introducing intermediate layers between the backbone and the head. These layers serve the purpose of gathering feature maps from various stages of the network (Bochkovskiy et al., 2020). The object detector comprises the following features:

A. Backbones

YOLOv4 though basically uses one of the three models as its backbone, however, for this study the backbone choice was darknet which is mostly used. Other feature extractor models include: CSPResNext50 and EfficientNet-B3

B. Neck

The neck region, situated between the backbone and the output, primarily focuses on feature aggregation to enhance the precision of detecting small objects (Zhou et al., 2022). The features formed at the backbone stage are gathered at the stage and later, which are later fed to the head for detection.

C. Head

In YOLOv4, the head's primary goal is to carry out prediction, which involves bounding box regression and categorization. The CIOU_loss loss function and DIOU_nms non-maximum suppression (NMS) technique are the main improvements of YOLO v4 in the Head module (LeCun & Bengio, 1998).

Convolutional Neural Network (CNN) Model

LeCun was a pioneer in the CNN Model's creation in 1996. The convolution layer, pooling layer, and full connection layer are the three layers that make up this neural network (Guyon et al., 1989). Convolutional networks use three architectural ideas—local receptive fields, shared weights or weight replication, and occasionally spatial or temporal subsampling—to provide a certain level of shift and distortion invariance. A convolutional layer typically consists of several feature map vectors with varying weights so that numerous features can be extracted at each location (Deshmukh & Jagade, 2022). The convolutional neural network can autonomously acquire successive stages of invariant features specific to a given task, unlike traditional methods that rely on pre-designed characteristics (Chatfield et al., 2014).

Convolutional Neural Networks (CNNs) are a class of sophisticated deep learning models that employ trainable filters and local neighborhood pooling operations in a layered fashion on the initial input images (Chatfield et al., 2014). This process leads to the development of a hierarchical structure of progressively intricate features. Research has demonstrated that CNNs when appropriately trained with regularization techniques(Chaudhary & Murala, 2019), can exhibit exceptional performance in tasks related to visual object recognition.

DetNet

The proposed backbone architecture for object detection in this study is DetNet 59, which features an additional stage compared to the conventional classification network ResNet-50. There are two obstacles to creating a robust and successful framework for object detection (Li et al., 2018). Firstly, retaining the spatial resolution required by deep neural networks demands excessive time and memory resources. Secondly, decreasing the down-sampling factor leads to a reduction in the valid receptive field, which can detrimentally impact various vision tasks including image classification and semantic segmentation.

DetNet functions as a backbone structure for a convolutional neural network that was primarily developed for object detection. It also aids in preserving the spatial resolution of features, even with the addition of supplementary stages.To maintain efficiency, DetNet employs a dilated bottleneck structure with low complexity. This is in contrast to conventional pre-trained models designed just for ImageNet classification and emphasizes the necessity for specialized networks that are tailored to object recognition tasks.

The datasets undergo processing by the yolov4 object detection module. Within the yolov4 module, Yolo initially processes an input image, segmenting it into a grid structure, for example, 3 x 3. Subsequently, image classification and localization are executed on each grid. YOLO then generates predictions for bounding boxes and associated class probabilities of objects. This enables the identification of human presence among the

input datasets, which consist of over 80 classes of objects in yolov4.

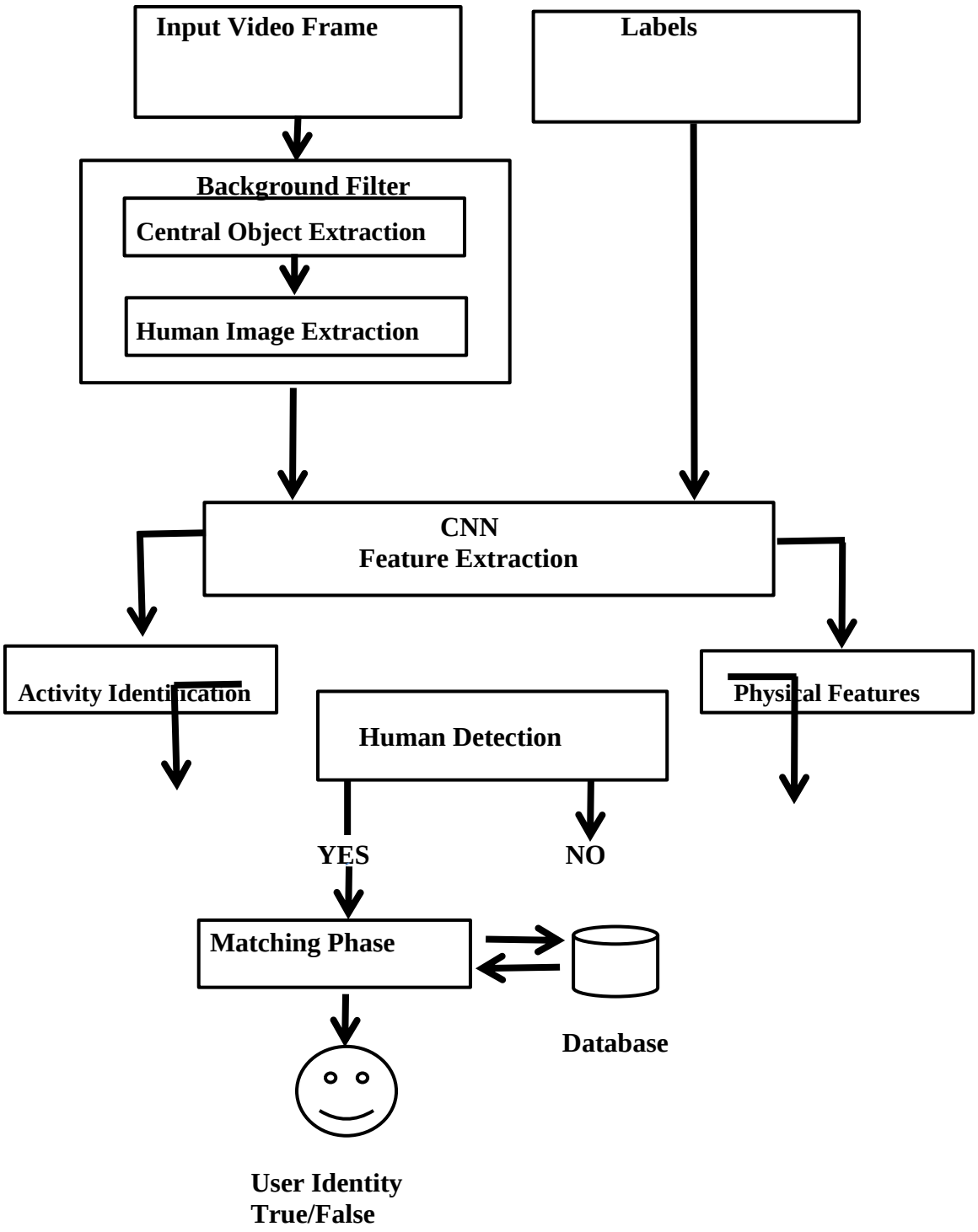

Figure 4: Framework of the Proposed System for object detection and identification

Performance Evaluation Metrics

The accuracy of classification is assessed using the Matthews Correlation Coefficient (MCC). Initially introduced by B.W. Matthews in 1975 for the assessment of chemical structures, MCC was later reintroduced by (Baldi et al., 2000) as a universal performance measure for machine learning, with a seamless expansion to accommodate multi-class scenarios. Matthews Correlation Coefficient considers both true and false positives and negatives, making it a well-balanced measure that remains effective even when dealing with classes of varying sizes.

MCC =

$$\frac{(\mathrm{TP} \times \mathrm{PN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})}}$$

(worst value: –1; best value: +1)

PRECISION = $\frac{TP}{TP+FP}$

RECALL= $\frac{TP}{TP+FN}$

F1-Score= $2\frac{Precision \times Recall}{Precision \times Recall}$

MEAN AVERAGE PRECISION (mAP)= $\frac{1}{N}\sum_{i=1}^{N} APi$

True positive (TP) = the amount of positive classes the model accurately identify

False Negative (FN) = the amount of negative classes the model wrongly identify

True negative (TN) = the amount of negative classes the model accurately identify

False Positive (FP) = the amount of positive classes the model wrongly identify

Findings and Discussion

Table 1: Performance Comparison between the backbones on MS-COCO dataset

Method	datasets	mAP	Detection Accuracy
Fast R-CNN	MS-COCO	63.4	76.3
Yolo	MS-COCO	66.2	74.2
YoloV$_2$ 544	MS-COCO	71.1	77.5
YoloV4 (Darknet-53)	MS-COCO	74.9	80.1

The YOLO V4 (DarkNet-53) has demonstrated a better performance compared to other models when trained on identical datasets, as evidenced by the results presented in Table 4.1. The model achieved an impressive overall detection accuracy of 80.1%, showcasing its effectiveness in identifying objects within images on MSMT17 datasets. This remarkable accuracy indicates the potential of YOLO V4 (DarkNet-53) to be a valuable tool in various applications that require precise object detection capabilities.

Figure 5: Input datasets for object detection

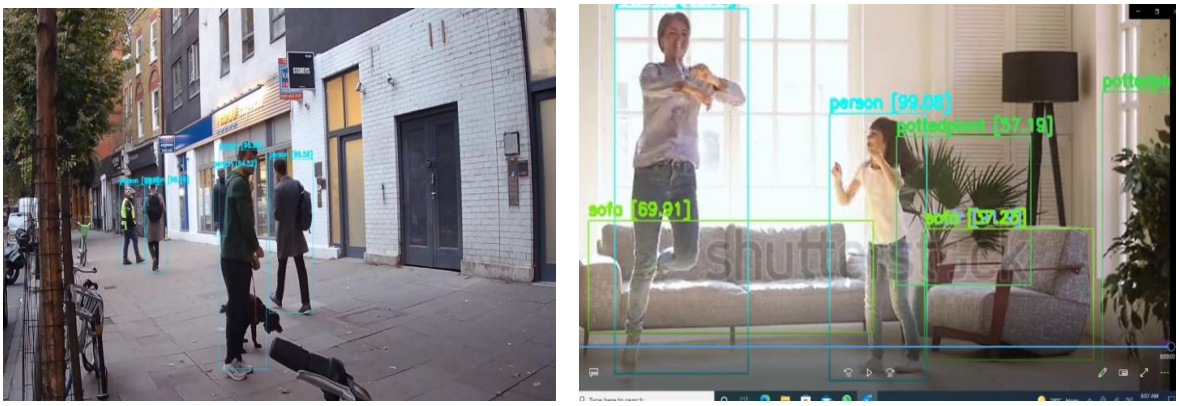

Figure 6: Results of detection from input datasets

Table 2: Performance Comparison between the backbones on MSMT17 datasets

Method	datasets	mAP	Re-identification Accuracy
Fast R-CNN	MSMT17	68.4	70
Yolo	MSMT17	57.9	63.5
YoloV$_2$ 544	MSMT17	73.4	68.7
YoloV4 (Darknet-53)	MSMT17	77.9	72

The re-identification results can be found in Table 4.2, which includes a comparison with various models using the MSMT17 datasets. Among these models, YOLO V4 (Darknet-53) stood out for its exceptional performance, achieving an impressive identification accuracy rate of 72%. This indicates that YOLO V4 (Darknet-53) outperformed the other models in the datasets when it comes to re-identification tasks.

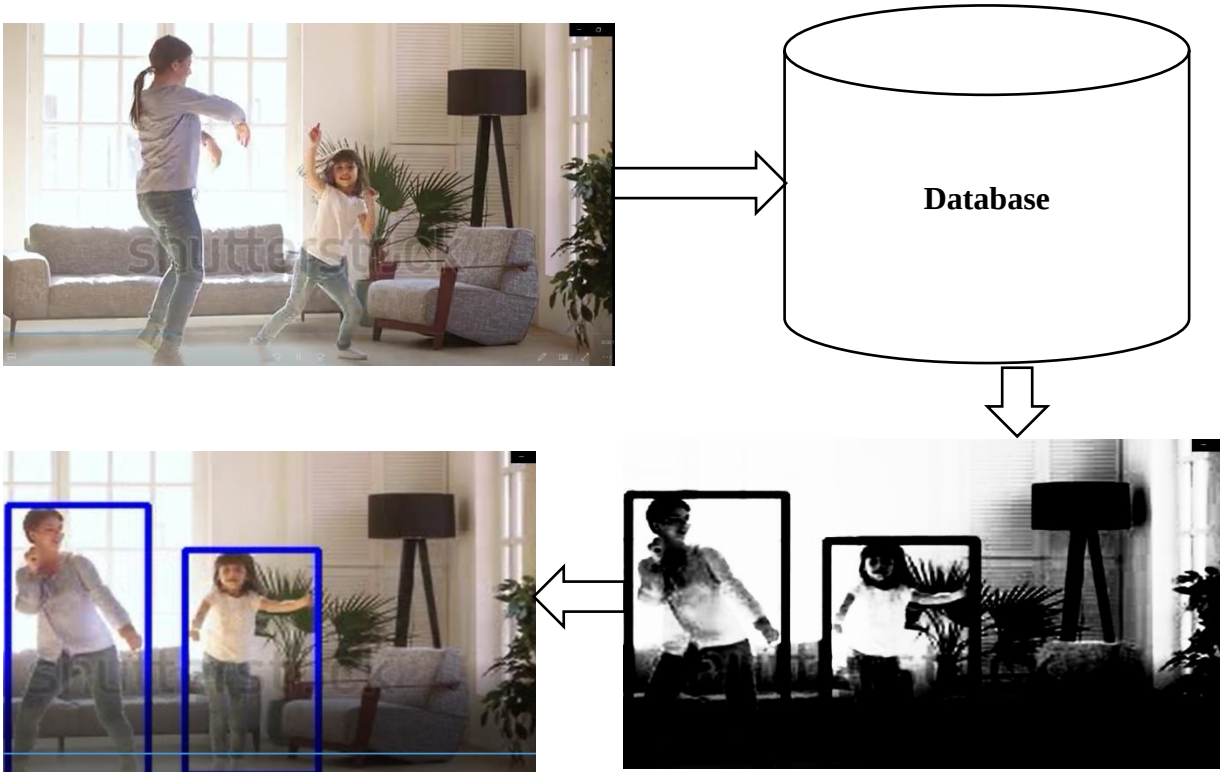

Figure 7: Results of the re-identification stages from input datasets

Conclusion

In this research work, an in-depth analysis of the development of a Smart Human Identification System using a Computer Vision Technique was presented. The proposed system was evaluated using various datasets and real-world scenarios to facilitate human detection and re-identification using Yolo V4 (DarkNet 53) architecture.

Moreover, the model was compared with pre-existing models such as Fast R-CNN, Yolo, and Yolo V2 544. The model exhibited superior performance, achieving a re-identification accuracy of 72% on the MSMT17 datasets and a detection accuracy of 80.1% on the MS-COCO datasets. In future work, YOLO V4 can be combined with algorithms such as Sort or DeepSort for tracking purposes.

Reference

Baldi, P. *et al.* (2000) 'Assessing the accuracy of prediction algorithms for classification: an overview,' *Bioinformatics*, 16(5), pp. 412–424. https://doi.org/10.1093/bioinformatics/16.5.412.

Benkaddour, M.K., Lahlali, S. and Trabelsi, M. (2021) 'Human Age And Gender Classification using Convolutional Neural Network,' *2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, pp. 215–220. https://doi.org/10.1109/ihsh51661.2021.9378708.

Bochkovskiy, A., Wang, C.-Y. and Liao, H.-Y.M. (2020) *YOLOV4: Optimal speed and accuracy of object detection*. http://arxiv.org/abs/2004.10934.

Chatfield, K. *et al.* (2014) *Return of the Devil in the Details: Delving Deep into Convolutional Nets*. http://arxiv.org/abs/1405.3531.

Chaudhary, S. and Murala, S. (2019) 'Deep network for human action recognition using Weber motion,' *Neurocomputing*, 367, pp. 207–216. https://doi.org/10.1016/j.neucom.2019.08.031.

Deng, J. *et al.* (2009) 'ImageNet: A large-scale hierarchical image database,' *2009 IEEE Conference on Computer Vision and Pattern Recognition* [Preprint]. https://doi.org/10.1109/cvpr.2009.5206848.

Deshmukh, S. , & Jagade, S. (2022) 'An effective approach for detecting and identifying human hand gestures using convolutional neural network', *NeuroQuantology*, *20*(13), pp. 1006–1013. https://doi.org/10.14704/nq.2022.20.13.NQ88128

Van Dyk, D.A. and Meng, X.-L. (2001) 'The art of data augmentation,' *Journal of Computational and Graphical Statistics*, 10(1), pp. 1–50. https://doi.org/10.1198/10618600152418584.

Felzenszwalb, P.F. *et al.* (2009) 'Object Detection with Discriminatively Trained Part-Based Models,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), pp. 1627–1645. https://doi.org/10.1109/tpami.2009.167.

Gheissari, N., Sebastian, T.B. and Hartley, R. (2006) 'Person Reidentification Using Spatiotemporal Appearance,' *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 1528–1535. https://doi.org/10.1109/cvpr.2006.223.

Guyon, N. *et al.* (1989) 'Comparing different neural network architectures for classifying handwritten digits,' *International 1989 Joint Conference on Neural Networks*, pp. 127–132 vol.2. https://doi.org/10.1109/ijcnn.1989.118570.

Hasib, R. *et al.* (2021) 'Vision-based Human Posture Classification and Fall Detection using Convolutional Neural Network,' *2021 International Conference on Artificial Intelligence (ICAI)*, pp. 74–79. https://doi.org/10.1109/icai52203.2021.9445263.

Jeny, A.A., Junayed, M.S. and Islam, M.B. (2021) 'PoseTED: a novel Regression-Based technique for recognizing multiple pose instances,' in *Lecture notes in computer science*, pp. 573–585. https://doi.org/10.1007/978-3-030-90439-5_45.

Khan, M.H., Farid, M.S. and Grzegorzek, M. (2021) 'Vision-based approaches towards person identification using gait,' *Computer Science Review*, 42, p. 100432. https://doi.org/10.1016/j.cosrev.2021.100432.

KimKim, M.-G. *et al.* (2012) 'A survey and proposed framework on the soft Biometrics technique for human identification in intelligent video surveillance system,' *Journal of Biomedicine and Biotechnology*, 2012, pp. 1–7. https://doi.org/10.1155/2012/614146.

Kokhan, O. (2024) *Understanding the Power of Data: Data Labeling vs. Annotation*. https://www.linkedin.com/pulse/understanding-power-data-labeling-vs-annotation-olga-kokhan-dqhje/.

Koo, B., Nguyen, N.T. and Kim, J. (2023) 'Identification and classification of human body exercises on smart textile bands by combining decision tree and convolutional neural networks,' *Sensors*, 23(13), p. 6223. https://doi.org/10.3390/s23136223.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) 'ImageNet classification with deep convolutional neural networks,' *Communications of the ACM*, 60(6), pp. 84–90. https://doi.org/10.1145/3065386.

LeCun, Y., & Bengio, Y. (1998) 'Convolutional networks for images, speech, and time series | The handbook of brain theory and neural networks' (no date) *Guide Books* [Preprint]. https://doi.org/10.5555/303568.303704.

Li, Z. *et al.* (2018) *DetNet: A Backbone network for Object Detection*. http://arxiv.org/abs/1804.06215.

Lin, T.-Y. *et al.* (2014) 'Microsoft COCO: Common Objects in context,' in *Lecture notes in computer science*, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

Ming, Z. *et al.* (2022) 'Deep learning-based person re-identification methods: A survey and outlook of recent works,' *Image and Vision Computing*, 119, p. 104394. https://doi.org/10.1016/j.imavis.2022.104394.

Mopidevi, S. *et al.* (2023) 'Hand gesture recognition and voice conversion for deaf and Dumb,' *E3S Web of Conferences*, 391, p. 01060. https://doi.org/10.1051/e3sconf/202339101060.

Pathak, A.R., Pandey, M. and Rautaray, S. (2018) 'Application of deep learning for object detection,' *Procedia Computer Science*, 132, pp. 1706–1717. https://doi.org/10.1016/j.procs.2018.05.144.

Sovrasov, V. and Sidnev, D. (2021) 'Building Computationally Efficient and Well-Generalizing Person Re-Identification Models with Metric Learning,' *2022 26th International Conference on Pattern Recognition (ICPR)*, abs 1905 3422, pp. 639–646. https://doi.org/10.1109/icpr48806.2021.9412598.

Yu, J. and Zhang, W. (2021) 'Face Mask Wearing Detection algorithm based on improved YOLO-V4,' *Sensors*, 21(9), p. 3263. https://doi.org/10.3390/s21093263.

Zhang, J. *et al.* (2020) 'Gate-ID: WiFi-Based Human identification irrespective of walking directions in smart home,' *IEEE Internet of Things Journal*, 8(9), pp. 7610–7624. https://doi.org/10.1109/jiot.2020.3040782.

Zhou, X. *et al.* (2022) 'Human Detection Algorithm Based on Improved YOLO v4,' *Information Technology and Control*, 51(3), pp. 485–498. https://doi.org/10.5755/j01.itc.51.3.30540.