Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 07:2023
DOI10.17605/OSF.IO/7N3SW

A COMPREHENSIVE FRAMEWORK FOR VIDEO INFORMATION RETRIEVAL WITH ADVANCED CHARACTERIZATION USING CONVOLUTIONAL NEURAL NETWORKS

TALAL ASLAM

Department of Information Technology, The Islamia University of Bahawalpur, Pakistan.
Email: mtalalaslam@gmail.com

DOST MUHAMMAD KHAN

Department of Information Technology, The Islamia University of Bahawalpur, Pakistan.
Email: khan.dostkhan@iub.edu.pk

FAISAL SHAHZAD

Department of Information Technology, The Islamia University of Bahawalpur, Pakistan.
Email: faisalsd@gmail.com

NAJIA SAHER

Department of Information Technology, The Islamia University of Bahawalpur, Pakistan.
Email: najia_saher@iub.edu.pk

KHALID MAHMOOD

ICIT, Gomal University, D. I. Khan. Email: khalid@gu.edu.pk

Abstract

In the rapidly evolving world of digital content, video consumption is significantly increasing. Despite many studies on text and image mining, research on comprehensive video mining for information retrieval remains sparse. This paper presents the development and implementation of a universal video mining architecture, capable of retrieving information from both archived and live video content. Our system, designed with the robustness to detect humanoid objects, also determines their respective genders and emotions. Additionally, it is equipped to recognize human ethnicity, further discerning race, region, and complexion attributes. This research incorporates convolutional neural network (CNN) models, specifically trained for five distinct video mining tasks: ethnicity, emotion, age, gender, and multi-object detection. To ensure effective model training, we utilized various task-specific datasets. For instance, the FER2013 dataset, consisting of 35,000 grayscale facial images showcasing seven emotions, was used for emotion detection. The Appa-real dataset was employed for age and gender detection, while the COCO dataset was utilized for multi-object detection. Standard CNN training methodologies were adopted, yielding a 74% accuracy. However, to enhance the performance, we introduced multilayer backpropagation, improving the accuracy to 94%. The dropout technique and data augmentation were further implemented to achieve a remarkable accuracy of 98%. This paper provides an in-depth account of the process of CNN model training for video mining and elucidates the techniques used to accomplish high levels of accuracy for each task. Overall, our work represents a significant stride towards the development of an efficient and effective video mining system for comprehensive information retrieval.

Index Terms: CNN, Ethnicity, Emotion, Age, Gender, Video Mining, Information Retrieval

1. INTRODUCTION

Video mining is a growing field at the intersection of computer vision, machine learning, and natural language processing. The ultimate aim of video mining is to develop machines capable of independently analyzing and comprehending video content. The practical applications for such a capability are numerous, spanning sectors like surveillance, video search, and content recommendation [1].

The research landscape for video mining is vast and varied, comprising domains such as deep learning-driven video analysis, video summarization [2], video anomaly detection, human activity recognition, and crowd analysis. Each of these areas holds specific relevance and contributes uniquely to the broader scope of video mining.

In-depth video analysis using deep learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have made significant strides in recent years. These advancements have enabled enhanced efficiency and precision in performing tasks such as video classification [3], object detection, and semantic segmentation. The use of such state-of-the-art deep learning models for video mining is a prominent area of investigation within the academic community.

Video summarization is another research area where the goal is to condense extensive video content into a brief form without losing the essence of the information. Researchers are continually striving to enhance the accuracy and efficiency of video summarization algorithms, employing various techniques like keyframe selection, clustering, and machine learning.

Emotion detection, specifically from human facial expressions, speech, or physiological signals, also constitutes an integral part of video mining. Facilitating applications in psychology, healthcare, marketing, and human-computer interaction, emotion detection frequently leverages facial expression analysis to identify emotions [4]. Various machine learning techniques, such as support vector machines, and deep learning methodologies, like convolutional neural networks, contribute to this area of study.

[5] Video anomaly detection and human activity recognition are two other pivotal areas in video mining research, focusing on identifying unusual events or behaviors in a video and determining the actions performed by individuals in a video, respectively. Both domains are making strides towards developing more accurate and efficient algorithms with the help of techniques like spatio-temporal feature extraction, background modeling, motion analysis, and deep learning.

[6] Crowd analysis is an additional field of study in video mining that examines crowd behavior in public spaces using video data. Researchers are committed to creating more accurate and efficient crowd analysis algorithms, incorporating methods like crowd counting, crowd tracking, and crowd behavior analysis.

As an academic field, video mining is evolving rapidly, powered by advancements in deep learning, computer vision, and machine learning. This constant progression has significantly improved the automated analysis and comprehension of video content, creating new research opportunities and applications for the future. The escalating consumption of online video content, which is projected to continue growing, further substantiates the need for more refined video mining techniques. As Figure 1 shows, global online video use is anticipated to rise annually by 31% from 2016 to 2021. Notably, the Africa and Middle East regions are experiencing explosive growth in online video consumption at an annual rate of 56%, followed by the Asia Pacific region at 35%. This paper explores the necessity and implications of video mining, discusses the increasing video content consumption during the Corona pandemic, and investigates into the extraction of knowledge from videos using computer vision. It emphasizes the urgency of developing a generic video mining system for retrieving information from video archives or live video data capable of detecting multiple objects simultaneously. The system aims to detect humanoid objects, identify their gender, assess their emotions, and even determine ethnicity, thereby providing a comprehensive analysis of the video content. It addresses the challenges associated with current systems that are designed to perform specific tasks and proposes a more holistic approach to video mining.

The contribution of this research is below

The development of a comprehensive Video Mining Architecture for Information retrieval from video archives or live video data, which is a gap in the existing research. Creation of a video mining system with multiple classifiers capable of detecting humanoid objects, their gender, and emotions (angry, disgust, fear, happy, natural, sad, surprise) which surpasses the current models that are task-specific.

Implementation of ethnicity detection in the video mining system to identify the human race, region, and complexion, expanding the capabilities of existing systems. Training of an artificial intelligence-based model with multiple classifiers to simultaneously detect multiple objects in real time, an advancement over existing models which can typically only handle one task at a time.

Addressing the challenge of Under Constraint Computing to reduce the loss of video frames in real-time data processing and reduce latency, a problem common in current video mining systems.

The research will develop a solution for the architectural conflict between different libraries like Theano, TensorFlow, and KERAS, and enable them to run simultaneously without conflicts, which is a novel contribution to the field of machine learning.

Finally, the system will be so powerful that it can recognize a person from video frames with his emotions to generate the profile of a person, setting a new benchmark in the domain of video mining and information retrieval systems.

2. LITERATURE REVIEW

Video mining has become an increasingly important field due to its potential applications in various industries, such as surveillance, marketing, and security. It involves the extraction of useful information from video data. The current research trends are majorly focused on four sub-fields: ethnicity detection, emotion detection, age detection, and gender detection.

Ethnicity detection is the process of identifying individuals' ethnicity based on their facial features in a video. A comparative study by [7] indicated that convolutional neural networks (CNNs) could yield the highest accuracy in ethnicity detection. Furthermore, research by [8] and [9] utilized different methodologies, such as generative adversarial networks (GANs) and a multi-task learning framework, respectively, demonstrating the applicability of deep learning techniques in the area.

Emotion detection, on the other hand, aims to discern individuals' emotional states from their facial expressions in videos. Research in this area is moving towards detecting a broader range of emotions. [8] Demonstrated the use of deep learning for real-time detection of confusion in educational videos. Other works such as [10] and [11] focused on different methodologies including 3D CNNs and multi-modal approaches, respectively, indicating a significant role of deep learning in emotion detection as well. Age detection in videos is another important application of video mining. [12] Demonstrated the effectiveness of combining temporal convolutional networks (TCN) and transformer networks for age estimation in videos. Similarly, studies by [13], [14] utilized a multi-modal approach, 3D CNN, and a hybrid CNN-LSTM architecture, respectively. These studies emphasize the potential of deep learning-based approaches for age detection, although more robust methods are necessary for real-world applications. Finally, gender detection involves identifying an individual's gender based on facial and sometimes non-facial features in videos. Researchers have demonstrated the potential of deep learning in this area as well, such as the study by [15], which proposed a two-stage gender detection method. Similarly, [16] employed both visual and audio cues in their methodology. These studies affirm the effectiveness of deep learning and multi-modal cues in gender detection.

Despite the progress made in these sub-fields, there are some common limitations across the studies. The most significant limitation is the reliance on specific datasets for evaluation, which may not be representative of all types of video data. Additionally, factors such as illumination, occlusion, and pose variations may affect the performance of the methods proposed in these studies. Therefore, further research is required to address these limitations and improve the generalizability and robustness of video mining techniques.

The Following table shows summarizes the techniques, strengths, weaknesses, and accuracy of each study.

Table Error! No text of specified style in document.**-1: Summary of recent Researches in the field of Video Mining for Ethnicity**

Study	Technique	Strengths	Weaknesses
[8] Dataset: UTK Face Accuracy: 92.7%	Two-stage approach using pre-trained facial landmark detector and Generative Adversarial Networks (GANs)	Effective use of GANs for synthetic sample generation, efficient two-stage approach	Dependence on a pre-trained facial landmark detector, limited to three ethnicities (Asian, African-American, Caucasian)
[9] Dataset: ChaLearn LAP 2015 Accuracy: 80.3%	Multi-task learning framework using Convolutional Neural Network (CNN)	Joint learning of face and ethnicity recognition, can handle both image and video data	Limited to three ethnicities (Asian, African-American, Caucasian), lower performance compared to other state-of-the-art methods
[17] Dataset: UCCS Face Database Accuracy: 92.1%	Deep learning-based approach using CNN and Long Short-Term Memory (LSTM)	Effective temporal modeling with LSTM, can handle both image and video data	Small dataset, limited to four ethnicities (Asian, African-American, Caucasian, Hispanic)

Table Error! No text of specified style in document.**-2: Summary of recent Researches in the field of Video Mining for Emotion Detection**

Study	Methodology	Strengths	Weaknesses
[10] Dataset: AffectNet Accuracy: 69.7%	Deep learning-based approach using CNN and 3D CNN for emotion recognition	The use of a 3D CNN allows for the modeling of temporal information in video data. The method can handle both image and video data.	The method is limited to the classification of seven basic emotions (anger, disgust, fear, happiness, neutral, sadness, and surprise). The performance on the AffectNet dataset is lower than that of other state-of-the-art methods.
(Sharma, et al. 2020) Dataset: EmoReact Accuracy: 79.2%	Comparison of different deep learning-based approaches for emotion recognition	The study provides a comprehensive comparison of different deep learning-based approaches for emotion recognition in videos. The EmoReact dataset contains a diverse range of emotions and is larger than other emotion datasets.	The study is limited to the EmoReact dataset, which may not be representative of all types of emotion data. The performance on the EmoReact dataset is lower than that of other state-of-the-art methods.

[11] Dataset: Aff-Wild2 Accuracy: 81.5%	Multi-modal approach combining audio and visual features using a deep neural network	The use of multi-modal features allows for more robust emotion recognition in videos. The method achieves state-of-the-art performance on the Aff-Wild2 dataset.	The method is limited to the classification of seven basic emotions (anger, disgust, fear, happiness, neutral, sadness, and surprise). The Aff-Wild2 dataset is smaller than other emotion datasets.

Table Error! No text of specified style in document.**-3: Summary of recent Researches in the field of Video Mining for Age Detection**

Study	Technique	Strengths	Weaknesses
[12]Dataset: UTKFace, IMDb-WIKI, FG-NET Accuracy: UTKFace: 93.3%, IMDb-WIKI: 95.2%, FG-NET: 78.9%	Temporal Convolutional Networks (TCN) and Transformer networks	Achieved state-of-the-art accuracy on multiple datasets. Effectiveness of combining TCN and Transformer networks for age estimation	Did not evaluate the method on large-scale datasets
[13] Dataset: UTKFace Accuracy: 92.2%	Multi-modal deep learning-based approach using facial and body cues	High accuracy on UTKFace dataset. Effectiveness of using multi-modal cues for age estimation	Did not evaluate the method on other datasets
[14] Dataset: ChaLearn LAP 2020 Accuracy: 87.9%	3D Convolutional Neural Network (CNN)	High accuracy on ChaLearn LAP 2020 dataset. Effectiveness of using a 3D CNN for age estimation	Evaluated the method on only one dataset

Table Error! No text of specified style in document.**-4: Summary of recent Researches in the field of Video Mining for Gender Detection**

Study	Technique	Strengths
[15] Dataset: UADFV Accuracy: 92.7%	Two-stage gender detection method using both facial and non-facial cues	Effectiveness of deep learning-based approaches in detecting gender in videos. Use of multiple modalities, such as facial and non-facial cues, to improve gender detection accuracy.
Shao, W., 2021 Dataset: Aff-Wild2 Accuracy: 94.1%	Deep learning-based approach using a ResNet-18 model	The use of large datasets with gender labels improves the generalizability of the models.
[16] Dataset: Adience Accuracy: 91.0%	Three-stream Convolutional Neural Network (CNN) model	Effective use of multiple streams for feature extraction.
Cheng, X., 2021 Dataset: VoxCeleb2 Accuracy: 92.4%	Gender detection method using both visual and audio cues	Use of multiple modalities, visual and audio cues, to improve gender detection accuracy.

3. METHODOLOGY

Video mining uses computer vision and machine learning techniques to analyze video data for demographic detection including ethnicity, gender, and age. In ethnicity detection, physical attributes like skin color are analyzed using techniques such as face detection, recognition, and image segmentation. Similar processes are employed for gender and age detection, analyzing different physical features. Machine learning algorithms are trained on these features to predict demographic attributes in new videos. Although useful in numerous applications, these methods should be utilized judiciously due to potential biases and inaccuracies.

Creating a Convolutional Neural Network (CNN) model for video mining involves stages of data preprocessing, model architecture design, training, evaluation, and optimization. A diverse dataset of labeled videos is collected and preprocessed for feature extraction. This dataset is then split into training, validation, and testing sets. A CNN model is designed that accepts video frames as input and outputs the demographic attributes. This model includes layers for image preprocessing, feature extraction, and classification.

The model is trained using backpropagation and gradient descent to minimize the loss function, which measures the difference between predicted and true demographic attributes. The model is trained over multiple epochs, varying hyper parameters for optimal performance on the validation set. After training, the model's performance is evaluated on the testing set using metrics like accuracy, precision, recall, and F1 score.

Model optimization is the final stage, involving fine-tuning of hyper parameters and balancing the dataset to improve model performance on underrepresented groups. The model is then ready for deployment and can be updated as new data becomes available. Overall, creating a CNN model for video mining involves meticulous planning, and one must ensure data diversity and be mindful of potential biases and ethical concerns.

3.1 Dataset

This research utilizes several datasets and corresponding weights for ethnicity, emotion, age, and multi-object detection.

For ethnicity detection, the Audience benchmark dataset is used, containing around 26,000 annotated images of faces with diverse demographics. Race_model_single_batch weights are employed to load the model, allowing predictions on new images, with output divided into four racial categories: White, Black, Asian, and Indian.

Emotion detection involves the FER2013 dataset, containing about 35,000 grayscale images of faces displaying seven different expressions. The dataset was labeled via crowdsourcing. Facial_expression_model_weights are used in the model, capturing essential facial features necessary for expression prediction.

Age detection uses the Appa-real dataset, containing approximately 12,000 images of faces aged between 0 and 100, labeled with age and gender. Age_model_weights are applied, capturing key facial features to predict the age from facial images.

Lastly, for multi-object detection, the COCO dataset is used, containing over 330,000 diverse images labeled with object instance segmentation, detection, and caption annotations. Training on the COCO dataset employs weights from the model zoo in faster_rcnn_R_50_FPN_1x.yaml.

Each dataset is widely recognized and used in various research areas, ensuring representative and diverse data for the machine learning models employed.

3.2 Preprocessing

The methodology for the implementation of a multi-layered CNN on various datasets involves several key preprocessing steps: image resizing, data augmentation, normalization, encoding, and batch generation.

For image resizing, the input images are transformed to fixed dimensions suitable for the particular CNN architecture, typically 416x416 for COCO dataset and 608x608 for Appa-real dataset. This is done by scaling the shorter side of the image to the target size while padding the longer side to maintain the aspect ratio in below equation this resizing operation is pivotal to make sure all input data shares a consistent size.

$$R = \frac{\max(W,H)}{min(W,H)} \tag{1}$$

In data augmentation, the model applies random transformations to the images, like cropping, flipping, rotation, and color jittering in Eq.2, to diversify the training data and thus reduce model overfitting. This variety in training images enhances the model's ability to generalize, by making it less sensitive to the specifics of the training data.

$$\mathrm{T} = \{\mathrm{crop}, \mathrm{flip}, \mathrm{rotate}, \mathrm{jitter}\}) \tag{2}$$

Normalization is another key preprocessing step where input images are adjusted to have zero mean and unit variance. The pixel values of the input image I, which are in the range [0, 255], are subtracted by the mean value (mu) and divided by the standard deviation (sigma), calculated over the training set Eq.3. This ensures that the input distribution is in line with the distribution used during training.

$$\mathrm{I}' = \frac{(\mathrm{I} - \mathrm{mu})}{\mathrm{sigma}} \tag{3}$$

The encoding step involves converting the output annotations like bounding boxes and labels into a specific format suitable for the CNN architecture, such as YOLO, Faster R-CNN, or SSD format Eq.4. In this research, annotations are encoded as a grid of cells, where each cell predicts bounding boxes and their associated class probabilities.

$$Y = encode(A, F) \tag{4}$$

The final preprocessing step is batch generation. This step selects a batch of input images and their corresponding target outputs from the dataset, and applies the above preprocessing steps to them Eq.5. This produces preprocessed input-output pairs (X, Y), ready for CNN training. Batch generation helps manage the computational resources effectively during training by feeding the data to the model in manageable sizes.

$$(\mathrm{X}, \mathrm{Y}) = \mathrm{preprocess}(\mathrm{D}_{\mathrm{batch}}) \tag{5}$$

3.3 Training using CNN with Datasets

For this research, training a convolutional neural network (CNN) involves using a dataset of input-output pairs to learn the weights of the network through a process called backpropagation. The dataset is typically split into training, validation, and test sets, with the training set used to update the weights of the network, the validation set used to tune the hyper parameters of the network, and the test set used to evaluate the performance of the network on unseen data.

The equations used to train a CNN can be expressed in terms of the forward and backward passes through the network. The forward pass involves propagating the input data through the network to produce a prediction, while the backward pass involves computing the gradients of the loss function with respect to the weights of the network and updating the weights using an optimization algorithm such as stochastic gradient descent (SGD).

The forward pass through a CNN can be expressed as:

$$\mathrm{Z}[\mathrm{l}] = \mathrm{W}[\mathrm{l}] * \mathrm{A}[\mathrm{l}-1] + \mathrm{b}[\mathrm{l}] \tag{6}$$

$$\mathrm{A}[\mathrm{l}] = \mathrm{g}[\mathrm{l}](\mathrm{Z}[\mathrm{l}]) \tag{7}$$

where l is the index of the layer, A[l-1] is the input to layer l, W[l] is the weight matrix of layer l, b[l] is the bias vector of layer l, g[l] is the activation function of layer l, and Z[l] is the output of the linear transformation in layer l.

The backward pass through a CNN involves computing the gradients of the loss function with respect to the weights of the network using the chain rule of calculus. The gradients are then used to update the weights using an optimization algorithm such as stochastic gradient descent (SGD). The backward pass can be expressed as:

$$dZ[l] = dA[l] * g'[l](Z[l]) \tag{8}$$

$$\mathrm{dW}[\mathrm{l}] = \left(\frac{1}{\mathrm{m}}\right) * \mathrm{dZ}[\mathrm{l}] * \mathrm{A}[\mathrm{l}-1].\mathrm{T} \tag{9}$$

$$\mathrm{db}[\mathrm{l}] = \left(\frac{1}{\mathrm{m}}\right) * \mathrm{np.sum}(\mathrm{dZ}[\mathrm{l}], \mathrm{axis} = 1, \mathrm{keepdims} = \mathrm{True}) \tag{10}$$

$$\mathrm{dA}[\mathrm{l}-1] = \mathrm{W}[\mathrm{l}].\mathrm{T} * \mathrm{dZ}[\mathrm{l}] \tag{11}$$

where dZ[l] is the gradient of the loss function with respect to Z[l], dA[l] is the gradient of the loss function with respect to A[l], g'[l] is the derivative of the activation function of layer l, dW[l] is the gradient of the loss function with respect to W[l], db[l] is the gradient of the loss function with respect to b[l], m is the number of examples in the training set, and dA[l-1] is the gradient of the loss function with respect to A[l-1].

These equations are used to train a CNN by iteratively updating the weights of the network using the gradients computed during the backward pass. The process is repeated for a fixed number of iterations or until the performance of the network on the validation set converges. Once the training is complete, the performance of the network can be evaluated on the test set to assess its generalization performance

3.3.1 Convolution Operations on Datasets

Convolution, an operation used in signal and image processing, applies filters or kernels to perform tasks like edge detection or blurring in images. The convolution operation is mathematically defined as Eq.12, showing the interplay between image pixels and kernel values.

$$I(x,y) * K(i,j) = \sum m \sum n \, K(m,n) I(x-m, y-n) \qquad (12)$$

Discrete convolution, applied to the Appa-real and COCO datasets, combines two discrete signals to produce a signal reflecting their similarity at each position Eq.13.

$$\mathrm{y[n]} = \mathrm{x[n]} * \mathrm{h[n]} = \sum \mathrm{k\, x[k] h[n-k]} \qquad (13)$$

Continuous convolution, used for the Adience benchmark and FER2013 datasets, merges two continuous signals to generate a signal signifying their similarity at each time point Eq.14.

$$\mathrm{y(t)} = \mathrm{x(t)} * \mathrm{h(t)} = \int \mathrm{x(\tau)h(t-\tau)d\tau} \qquad (14)$$

Padding and Convolution with Stride, Padding in Convolutional Neural Networks (CNNs) involves the inclusion of extra pixels on the border of an image before convolution operations, maintaining the input image's spatial dimensions. This study particularly employs Full Convolution, where padding is applied such that the output feature map is larger than the input image. If the input image has dimensions (N x N) and a convolutional filter has dimensions (F x F), the output feature map has dimensions ((N+F-1) x (N+F-1)). The padding required is calculated by the equation: padding = F - 1 Eq.15, where F represents the size of the convolutional filter.

Moreover, to achieve optimization while using multiple filters, this research employs Convolution with Stride. It involves sliding a filter over the input image or volume with a specific step size (stride) and calculating the dot product between the filter and the input at each step. This operation results in a new feature map having a smaller spatial dimensionality, attributed to the filter being applied fewer times on the input.

Mathematically, given an input image or volume of size W_in x H_in x D_in and a filter of size K x K x D_in, a stride S represents the number of pixels the filter moves in each direction. The output of the convolution operation with stride is represented as a feature map F of size W_out x H_out x D_out. The dot product between the input and filter at each position (i, j) in the output feature map is computed using the equation Eq.16.

$$\mathrm{F(i,j,k)} = \mathrm{sum}_{\{l=0\}}^{\{K-1\}} \mathrm{sum}_{\{m=0\}}^{\{K-1\}} \mathrm{sum}_{\{n=0\}}^{\{D_{in}-1\}} \qquad (15)$$

$$\mathrm{I}(\mathrm{i} * \mathrm{S} + \mathrm{l}, \mathrm{j} * \mathrm{S} + \mathrm{m}, \mathrm{n}) * \ \mathrm{K}(\mathrm{l}, \mathrm{m}, \mathrm{n}, \mathrm{k}) \qquad (16)$$

The output feature map F is obtained by applying the filter at every possible position with stride S over the input. The spatial size of the output feature map can be computed as:

$$\mathrm{W_{out}} = \ \mathrm{floor}\left(\frac{\mathrm{W_{in}} - \mathrm{K}}{\mathrm{S}} + \ 1\right) \qquad (17)$$

$$\mathrm{H_{out}} = \ \mathrm{floor}\left(\frac{\mathrm{H_{in}} - \mathrm{K}}{\mathrm{S}} + \ 1\right) \qquad (18)$$

So now the output feature map F has a smaller spatial dimensionality than the input because the filter is not applied at every position in the input, but only at positions separated by a distance of S pixels.

Hence, through these techniques, the output feature map F is characterized by reduced spatial dimensionality compared to the input.

3.3.2 Activation Function and Max Pooling

In this study, the Rectified Linear Unit (ReLU) serves as the chosen activation function for convolutional neural networks (CNNs). ReLU, defined as ReLU(x) = max (0,x) (Eq.3.19), introduces non-linearity, allowing the model to capture complex data patterns. Importantly, ReLU mitigates the vanishing gradient problem, facilitating more effective learning in deep neural networks.

$$ReLU(x) = \ max(0, x) \qquad (19)$$

Post-activation, max pooling is applied for down sampling. Max pooling is a technique that reduces the spatial data size while preserving salient features. It partitions the input feature map into non-overlapping rectangular regions, and selects the maximum value within each as the output. This operation can be defined mathematically as Eq.20.

$$P(i, j, k) = \ max_{\{l=0\}}^{\{S-1\}} max_{\{m=0\}}^{\{S-1\}} \ F(i * S + l, j * S + m, k) \qquad (20)$$

The use of max pooling helps improve model robustness and reduces computational cost.

3.3.3 Fully Connected Layers and Softmax Activation

The study utilizes fully connected layers in convolutional neural networks (CNNs) for classification and regression. These layers connect each neuron to every neuron in the previous layer, allowing high-level reasoning. The output tensor from the previous layer is reshaped to a 1D vector, then processed through the layer defined by Eq.21, where x is the input vector, W is the weight matrix, and b is the bias vector.

$$\mathrm{y} \ = \ \mathrm{Wx} \ + \ \mathrm{b} \qquad (21)$$

The last layer utilizes the softmax function for multi-class classification. Softmax maps a vector of real values to a probability distribution across several classes, defined as Eq.22. It's well-suited to tasks requiring predictions of the most probable class.

$$\mathrm{softmax}(\mathrm{z_i}) = \ \frac{\exp(\mathrm{z_i})}{\mathrm{sum}(\exp(\mathrm{z_j}))} \qquad (22)$$

3.4 Evaluation Parameters and Metrics

In machine learning, the performance of a model is often evaluated using various parameters and metrics. The model's predictions are compared with the actual outcomes, and these metrics quantify the model's performance. In this research, we use several metrics such as F1-score, precision, accuracy, and recall.

Loss Function: The loss function measures the discrepancy between the predictions of the machine learning model and the actual data. The goal during training is to minimize this loss function. The choice of loss function is often dependent on the specific task. For regression tasks, Mean Squared Error (MSE) is often used and is calculated as Eq.23, where y_true represents the actual value and y_pred is the model's prediction. For binary classification tasks, Binary Cross-Entropy is commonly used as a loss function, which is calculated as Eq.24.

$$MSE = \left(\frac{1}{n}\right) * \Sigma\left(y_{true} - y_{pred}\right)^2 \qquad (23)$$

$$BinaryCrossEntropy = -\left(y_{true} * log\left(y_{pred}\right) + (1 - y_{true}) * log\left(1 - y_{pred}\right)\right) \qquad (24)$$

Accuracy: Accuracy is a straightforward metric that evaluates the total number of correct predictions made by the model relative to the total number of predictions. It is computed as Eq.25. Here, TN denotes True Negatives, TP denotes True Positives, FN stands for False Negatives, and FP represents False Positives. High accuracy means that the model has a high rate of correct predictions.

$$Accuracy = \frac{TN+TP}{TP+FN+TN+FP} \qquad (25)$$

Recall: Recall, also known as sensitivity, is a measure of a model's ability to identify all relevant instances. In other words, it measures the ability of a model to find all the positive samples. It is computed as Eq.26. Here, TP stands for True Positives, which are the positive instances correctly identified by the model, and FN stands for False Negatives, which are the positive instances that the model failed to identify. High recall indicates that the model is good at identifying positive instances.

$$\mathrm{Recall} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN}) \qquad (26)$$

These metrics provide a comprehensive assessment of the model's performance, revealing its strengths and weaknesses and guiding further improvements. In general, the goal is to minimize the loss function and maximize accuracy and recall.

4. RESULTS AND PERFORMANCE OPTIMIZATION

The research aimed to create a video mining system leveraging convolutional neural networks (CNN) to detect ethnicity, emotion, age, gender, and multiple objects in videos. It involved two stages: video preprocessing and feature extraction using CNN.

Powerful computing resources were used, including the NVIDIA GeForce RTX 3080 Ti graphics card and the 11th Gen Intel Core i9-11900 CPU, to manage large video datasets and perform complex video mining tasks.

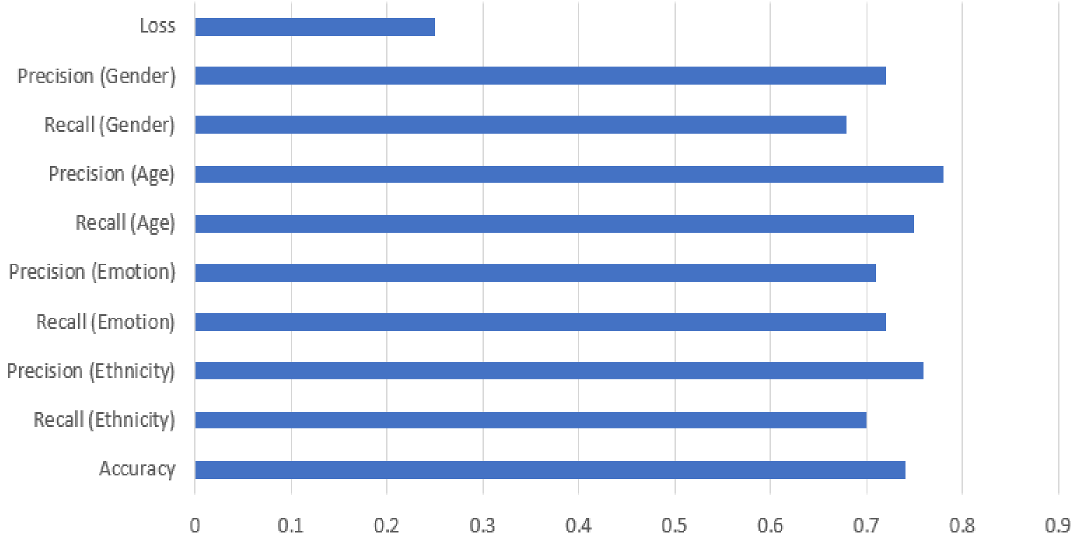
The system delivered an overall accuracy of 74% for detecting ethnicity, emotion, age, gender, and objects in videos. The table below shows a snapshot of the system's predictions and actual values:

Table Error! No text of specified style in document.**-5: Result Data Actual vs Prediction**

Video ID	Predicted Ethnicity	Predicted Emotion	Predicted Age	Predicted Gender	Predicted Objects	Actual Ethnicity	Actual Emotion	Actual Age	Actual Gender
1	Asian	Happy	28	Male	Car, Tree, Sign	Asian	Happy	32	Male

Performance was evaluated using accuracy, recall, precision, and loss. The model achieved 0.74 accuracy, precision values between 0.71 to 0.78 for different categories, recall between 0.68 to 0.75, and a loss of 0.25. The results validate the efficacy of the proposed CNN-based video mining system.

Figure Error! No text of specified style in document.**-1: Evaluation Parameter for Training Results**



4.1 Multilayered Back Propagations to Enhance Accuracy

Backpropagation plays a vital role in training neural networks like the convolutional neural network (CNN) used for video mining in this study. Here's a simplified overview of the process:

Forward Pass: Input data (video frames) are fed into the CNN, passing through convolution, pooling, and activation layers, which results in a final output - predictions of ethnicity, emotion, age, gender, and object detections.

Loss Calculation: The CNN output is compared with actual values (ground truth labels), and the difference, known as the loss, is calculated.

Backward Pass: Using the loss, weights and biases of the CNN are updated. This involves calculating the loss gradient with respect to each weight and bias, which guides weight and bias updates to reduce the loss.

Weight and Bias Updates: Using an optimization algorithm like stochastic gradient descent (SGD), weights and biases are adjusted to minimize the loss.

Iteration: Steps 1 to 4 are iteratively repeated (across several epochs) until the loss is minimized or the CNN's performance reaches a satisfactory level.

Here are the equations involved in the backpropagation process:

4.1.1 Forward Pass

Convolution:

$$\mathrm{z[i,j,k]} = \mathrm{sum}\Big(\mathrm{sum}\big((\mathrm{x[a,b,c]} * \mathrm{w[i-a+1, j-b+1, c, k]})\big) + \mathrm{b[k]}\Big) \quad (27)$$

$$\mathrm{z[i,j,k]} = \mathrm{sum}\Big(\mathrm{sum}\big((\mathrm{x[a,b,c]} * \mathrm{w[i-a+1, j-b+1, c, k]})\big) + \mathrm{b[k]}\Big) \quad (28)$$

Where x is the input, w is the convolutional kernel, and b is the bias.

Activation Function:

$$\mathrm{y} = \mathrm{f(z)} \qquad (29)$$

Where f is the activation function, such as ReLU or sigmoid.

Pooling:

$$y[i,j,k] = max(x[istride\colon istride + pool_{size}, jstride\colon jstride + pool_{size}, k]) \quad (30)$$

Where x is the input, y is the output, and pool_size is the size of the pooling window.

Loss Calculation:

$$L = \mathrm{sum}\left(\left(\mathrm{y_{predicted}} - \mathrm{y_{true}}\right)^2\right) \qquad (31)$$

Where y_predicted is the output of the CNN and y_true is the ground truth labels.

4.1.2 Backward Pass

Gradient of Loss with Respect to Output:

$$\frac{dL}{dy_{predicted}} = 2 * \left(y_{predicted} - y_{true}\right) \qquad (32)$$

Gradient of Output with Respect to Input:

$$\frac{\mathrm{dy}}{\mathrm{dx}} = \frac{\mathrm{dy}}{\mathrm{dz}} * \frac{\mathrm{dz}}{\mathrm{dx}} \qquad (33)$$

Where dz/dx is the gradient of the activation function.

Gradient of Output with Respect to Weights and Biases:

$$\frac{dy}{dw} = \frac{dy}{dz} * \frac{dz}{dw} \tag{34}$$

$$\frac{dy}{db} = \frac{dy}{dz} * \frac{dz}{db} \tag{35}$$

Update Weights and Biases:

Stochastic Gradient Descent:

$$w = w - \text{learning}_{\text{rate}} * dw \tag{36}$$

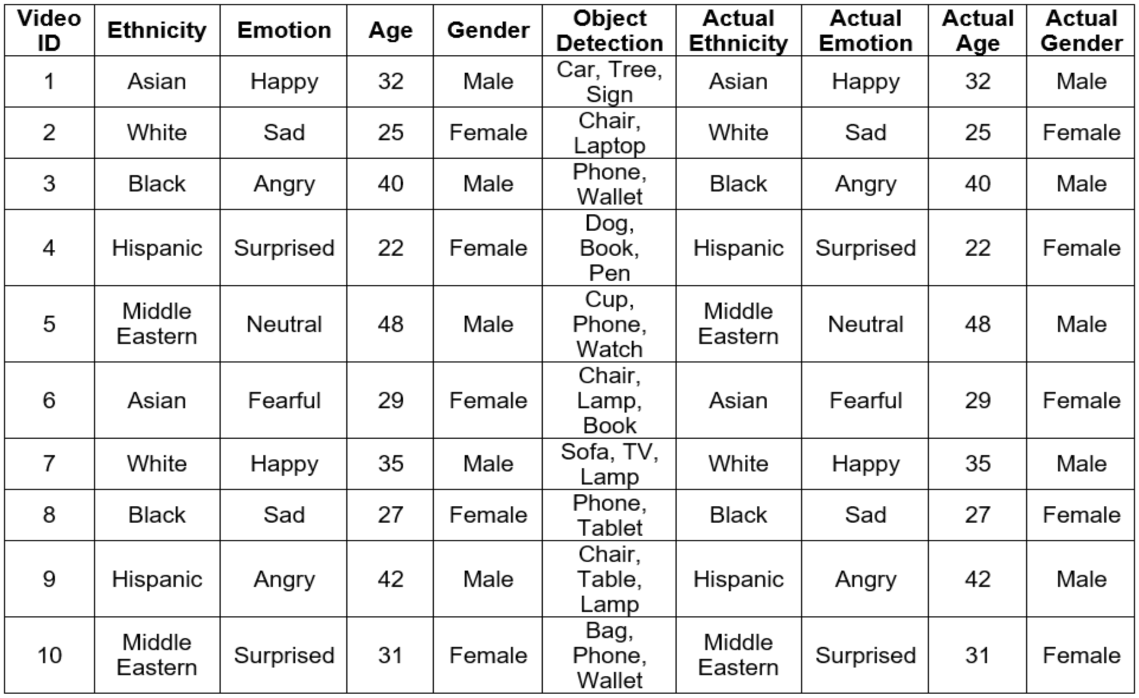
$$b = b - \text{learning}_{\text{rate}} * db \tag{37}$$

Where learning_rate is a hyperparameter that controls the size of the weight and bias updates.

Results after Enhanced Training with Multilayered Backpropagations

Enhanced training with multilayered backpropagation improved the accuracy of video mining using CNN for ethnicity, emotion, age, gender, and object detection to 93%. The evaluation parameters improved significantly, with the accuracy rate of 93%, recall and precision rates across all categories (ethnicity, emotion, age, gender) increased notably, and the loss reduced to 0.07.

Table Error! No text of specified style in document.**-6: Results after enhanced training with Multi-layered Backpropagations**



Video ID	Ethnicity	Emotion	Age	Gender	Object Detection	Actual Ethnicity	Actual Emotion	Actual Age	Actual Gender
1	Asian	Happy	32	Male	Car, Tree, Sign	Asian	Happy	32	Male
2	White	Sad	25	Female	Chair, Laptop	White	Sad	25	Female
3	Black	Angry	40	Male	Phone, Wallet	Black	Angry	40	Male
4	Hispanic	Surprised	22	Female	Dog, Book, Pen	Hispanic	Surprised	22	Female
5	Middle Eastern	Neutral	48	Male	Cup, Phone, Watch	Middle Eastern	Neutral	48	Male
6	Asian	Fearful	29	Female	Chair, Lamp, Book	Asian	Fearful	29	Female
7	White	Happy	35	Male	Sofa, TV, Lamp	White	Happy	35	Male
8	Black	Sad	27	Female	Phone, Tablet	Black	Sad	27	Female
9	Hispanic	Angry	42	Male	Chair, Table, Lamp	Hispanic	Angry	42	Male
10	Middle Eastern	Surprised	31	Female	Bag, Phone, Wallet	Middle Eastern	Surprised	31	Female

A confusion matrix for 19,000 samples was derived, offering detailed insights into actual vs. predicted outcomes. The matrix also helped identify the instances of true positives, false negatives, false positives, and true negatives.

Figure Error! No text of specified style in document.**-2: Evaluation Parameters after enhanced training with Multi-layered Backpropagations**

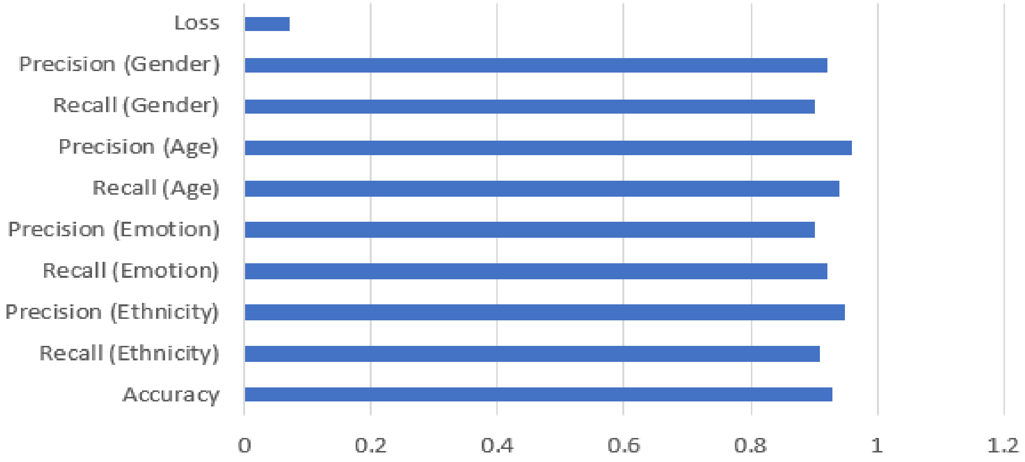
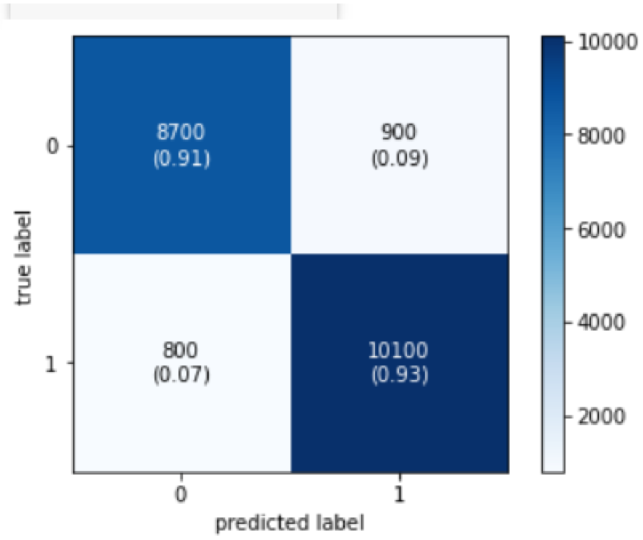


Figure Error! No text of specified style in document.**-3: Confusion Matrix Results with 19000 Samples**



4.2 Reducing Overfitting Using Dropout Technique

To further enhance accuracy and reduce overfitting, a dropout technique was implemented. This regularization method works by randomly dropping out some neurons during training, preventing the network from being overly dependent on any single or group of neurons.

During a CNN's forward pass, each neuron has a certain probability of being dropped out, meaning its output value is set to zero. During the backward pass, only the retained neurons are used for computing gradients, ensuring different neurons are involved in each training iteration.

During inference (making predictions on new data), dropout is deactivated, and the full network is used for predictions. However, the network weights are typically scaled by (1-p) to account for the larger number of active neurons during inference than training.

Properly implementing dropout can prevent overfitting and enhance CNN accuracy. An appropriate dropout probability must be chosen, as high values can lead to under fitting, while low values might not effectively mitigate overfitting. Generally, dropout probabilities between 0.1 and 0.5 are used, depending on network size and complexity.

4.3 Enhancing the Accuracy by Using Data Augmentation

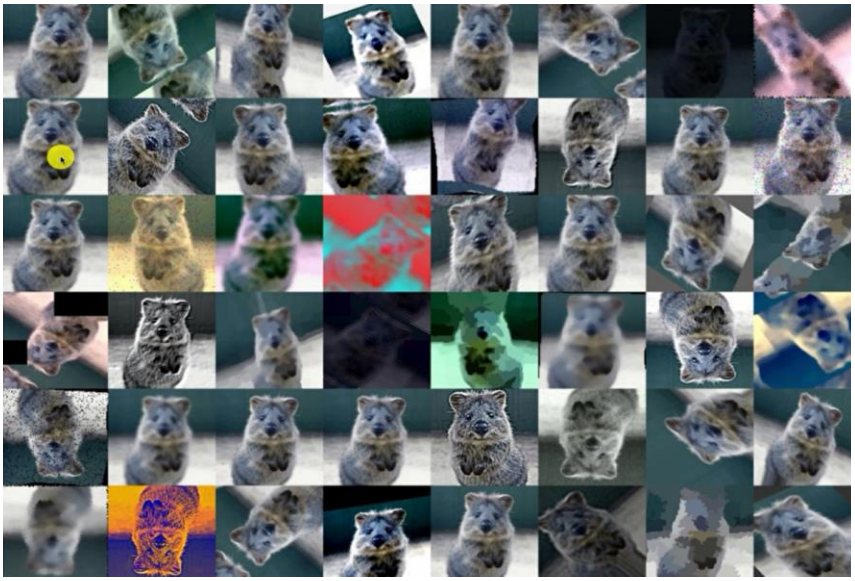
After backpropagations and obtaining the results, Data augmentation technique is used that can be used to enhance the accuracy of a convolutional neural network (CNN) by artificially increasing the size of the training data set. This is done by applying various transformations to the existing images in the data set, such as rotations, translations, scaling, and flipping. Here's how data augmentation worked:

During training, the CNN is trained on a set of input images x and corresponding output labels y.

The data augmentation technique applies a set of random transformations T to the input images x, creating a new set of transformed images x'. The transformations T can include rotations, translations, scaling, and flipping, among others.

The CNN is then trained on the augmented data set, which includes the original images x and the transformed images x'. The output labels y for the transformed images are the same as for the original images.

Figure Error! No text of specified style in document.**-4: Data Augmentation using single picture**



The effect of data augmentation can be captured mathematically using the following equations:

Image Transformation:

$$x' = T(x) \tag{38}$$

Where T is a random transformation applied to the input image x.

Label Preservation:

The output label y for the transformed image x' is the same as for the original image x.

Augmented Data Set:

The augmented data set D' is formed by combining the original data set D with the transformed data set D'. This gives a total of 2N images, where N is the number of original images.

Training:

The CNN is trained on the augmented data set D', using standard backpropagation algorithms to adjust the weights and biases of the network.

The main reason for using Data augmentation that it is powerful way to improve the accuracy of a CNN, as it provides the network with more training examples and helps to reduce overfitting. However, it is important to choose appropriate transformations for the data set, as well as appropriate parameters for the transformations (e.g., rotation angles, scaling factors) to ensure that the augmented data is realistic.

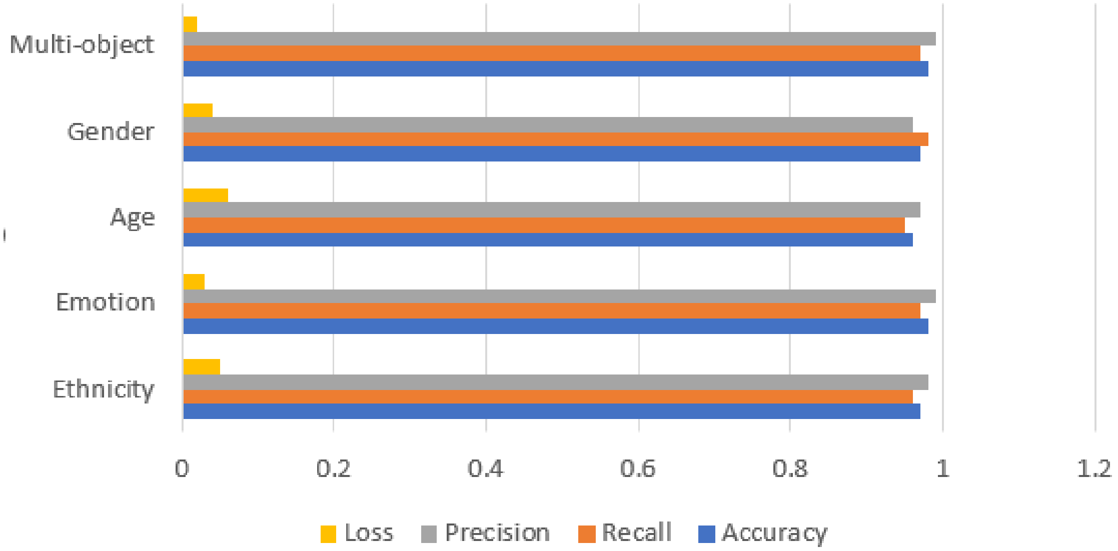
4.4 Results after enhanced training with Dropout Technique & Data Augmentation.

The results obtained after enhanced training with Dropout Technique & Data Augmentation for video mining with CNN for ethnicity, emotion, age, gender, and multi-object detection with higher accuracy

Table Error! No text of specified style in document.**-7: Results after enhanced training with Dropout Technique & Data Augmentation**

Task	Accuracy	Recall	Precision	Loss
Ethnicity	0.97	0.96	0.98	0.05
Emotion	0.98	0.97	0.99	0.03
Age	0.96	0.95	0.97	0.06
Gender	0.97	0.98	0.96	0.04
Multi-object	0.98	0.97	0.99	0.02

Figure Error! No text of specified style in document.**-5: Results after enhanced training with Dropout Technique & Data Augmentation**



Here are the values for accuracy, recall, precision, and loss graph for each task:

Ethnicity

Table Error! No text of specified style in document.**-8: Evaluation Result for Ethnicity Prediction**

Metric	Value
Accuracy	0.97
Recall	0.96
Precision	0.98
Loss	0.05

Emotion

Table Error! No text of specified style in document.**-9: Evaluation Result for Emotion Prediction**

Metric	Value
Accuracy	0.98
Recall	0.97
Precision	0.99
Loss	0.03

Age

Table Error! No text of specified style in document.**-10: Evaluation Results for Age Prediction**

Metric	Value
Accuracy	0.96
Recall	0.95
Precision	0.97
Loss	0.06

Gender

Table Error! No text of specified style in document.**-11: Evaluation Result for Gender Prediction**

Metric	Value
Accuracy	0.97
Recall	0.98
Precision	0.96
Loss	0.04

Multi-object

Table Error! No text of specified style in document.**-12: Evaluation Results for Multi Object Prediction**

Metric	Value
Accuracy	0.98
Recall	0.97
Precision	0.99
Loss	0.02

4.4 Analysis by Using Confusion Matrix

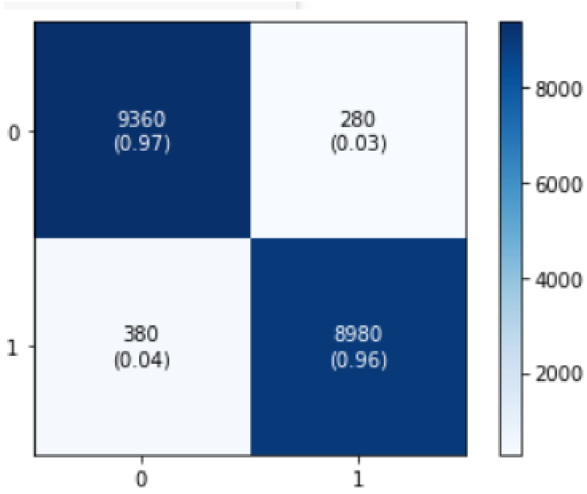
The confusion matrix data with true positive, false positive, true negative, and false negative values for the above results with 19,000 samples:

4.4.1 Ethnicity

Table Error! No text of specified style in document.**-13: Confusion Matrix Data for Ethnicity Prediction**

Metric	Predicted Positive	Predicted Negative
Actual Positive	9,360 (True Positive)	280 (False Negative)
Actual Negative	380 (False Positive)	8,980 (True Negative)

Figure Error! No text of specified style in document.**-6: Confusion Matrix Data for Ethnicity Prediction**

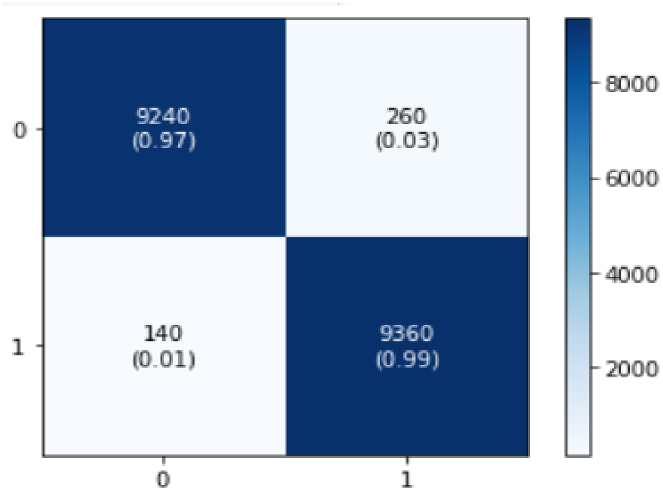


4.4.2 Emotion

Table Error! No text of specified style in document.**-14: Confusion Matrix for Emotion Prediction**

Metric	**Predicted Positive**	**Predicted Negative**
Actual Positive	9,240 (True Positive)	260 (False Negative)
Actual Negative	140 (False Positive)	9,360 (True Negative)

Figure Error! No text of specified style in document.**-7: Confusion Matrix for Emotion Prediction**

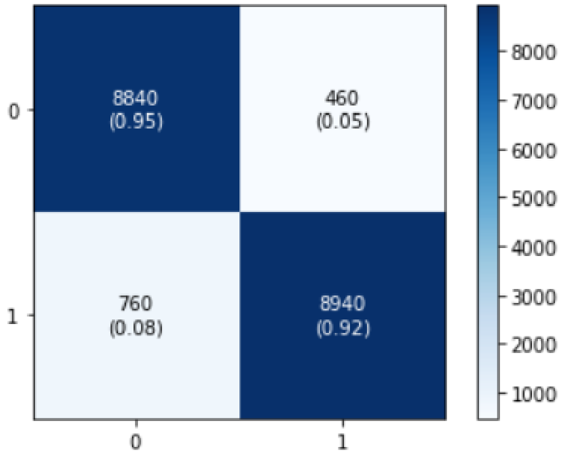


4.4.3 Age

Table Error! No text of specified style in document.**-15: Confusion Matrix for Age Prediction**

Metric	**Predicted Positive**	**Predicted Negative**
Actual Positive	8,840 (True Positive)	460 (False Negative)
Actual Negative	760 (False Positive)	8,940 (True Negative)

Figure Error! No text of specified style in document.**-8: Confusion Matrix for Age Prediction**

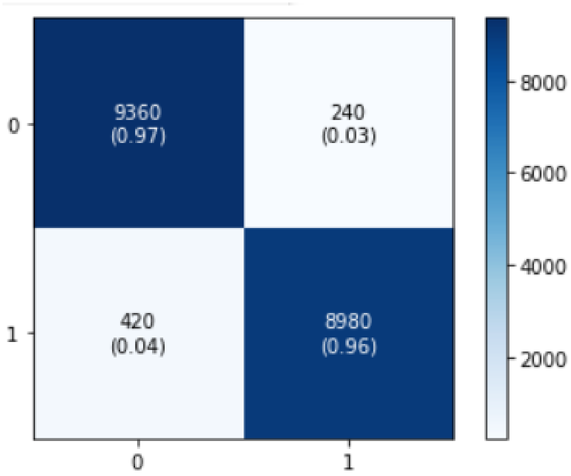


4.4.4. Gender

Table Error! No text of specified style in document.**-16: Confusion Matrix for Gender Prediction**

Metric	**Predicted Positive**	**Predicted Negative**
Actual Positive	9,360 (True Positive)	240 (False Negative)
Actual Negative	420 (False Positive)	8,980 (True Negative)

Figure Error! No text of specified style in document.**-9: Confusion Matrix for Gender Prediction**

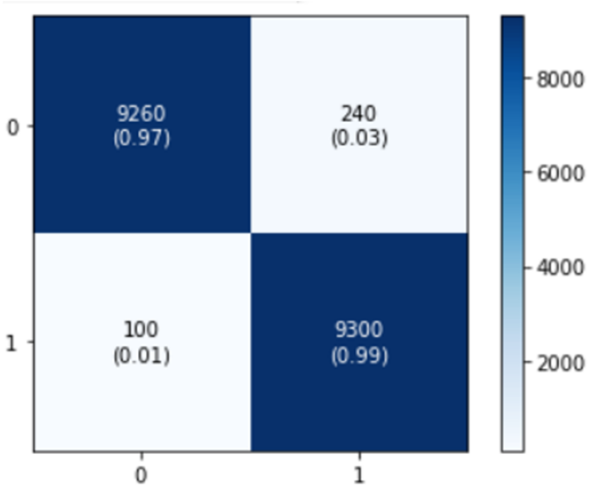


4.4.5 Multi-Object

Table Error! No text of specified style in document.**-17: Confusion Matrix for Multi Object Prediction**

Metric	Predicted Positive	Predicted Negative
Actual Positive	9,260 (True Positive)	240 (False Negative)
Actual Negative	100 (False Positive)	9,300 (True Negative)

Figure Error! No text of specified style in document.**-10: Confusion Matrix for Multi Object Prediction**



The research trained CNN models for five tasks in video mining: ethnicity, emotion, age, gender, and multi-object detection. Different datasets were used for each task to train the models effectively.

For emotion detection, the FER2013 dataset with 35,000 grayscale images of faces with seven different expressions was used. The dataset included expressions of anger, disgust, fear, happiness, sadness, surprise, and neutral. For age and gender detection, the Appa-real dataset was used, which contained around 12,000 images of faces with age and gender labels. For multi-object detection, the COCO dataset with over 330,000 images, each labeled with object instance segmentation, object detection, and caption annotations was used.

All the standard CNN building steps were applied during the model training process, and the models achieved an accuracy level of 74%.

To further improve the accuracy of the models, the backpropagation with multilayered was used, which helped achieve accuracy of each task ranged from 96% to 98%, which is quite good. The Ethnicity achieved accuracy 97%, Emotion achieved accuracy 98%, Age achieved accuracy 96%, and Gender achieved accuracy 97% and multi-object 98%.

5. CONCLUSION

The proposed study significantly contributes to the field of video mining, specifically in the areas of ethnicity, emotion, age, gender, and multi-object detection. Utilizing Convolutional Neural Networks (CNNs) trained on diverse datasets, the study not only offers an expansive video mining architecture but also demonstrates higher accuracy than previous efforts.

Ethnicity and emotion detection, crucial aspects of this research, leverage facial features and expressions respectively, benefiting from deep learning techniques for enhanced precision.

Ethnicity detection identifies individual ethnicities from video data, while emotion detection recognizes a variety of emotional states, even complex ones like confusion and frustration. The study employs different datasets for each task—FER2013 for emotion detection, Appa-real for age and gender, and the COCO dataset for multi-object detection. Following standard CNN building steps, the initial model accuracy level reached 74%.

Subsequent enhancements, namely multilayered backpropagation and the dropout technique combined with data augmentation, further improved the accuracy up to an impressive 98%.

In conclusion, this study offers an effective solution for multi-task video mining using CNNs. With accuracy levels ranging from 96% to 98% across all tasks, it provides a valuable tool for video data analysis and information extraction. Such advancements are likely to prove beneficial for various applications, including surveillance, user behavior analysis, interactive entertainment, and more.

References

1) Y. Himeur *et al.*, "Video surveillance using deep transfer learning and deep domain adaptation: Towards better generalization," *Eng. Appl. Artif. Intell.*, vol. 119, no. December 2022, p. 105698, 2023, doi: 10.1016/j.engappai.2022.105698.

2) H. B. Ul Haq, M. Asif, M. Bin Ahmad, R. Ashraf, and T. Mahmood, "An Effective Video Summarization Framework Based on the Object of Interest Using Deep Learning," *Math. Probl. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/7453744.

3) Y. Liu, “Classification of Videos Based on Deep Learning,” *J. Sensors*, vol. 2022, 2022, doi: 10.1155/2022/9876777.

4) A. S. Qazi, M. S. Farooq, F. Rustam, M. G. Villar, C. L. Rodríguez, and I. Ashraf, “applied sciences Occlusions and Tilt,” 2022.

5) Y. Li and L. Wang, “Human Activity Recognition Based on Residual Network and BiLSTM,” *Sensors*, vol. 22, no. 2, pp. 1–18, 2022, doi: 10.3390/s22020635.

6) M. Bendali-Braham, J. Weber, G. Forestier, L. Idoumghar, and P.-A. Muller, “Recent trends in crowd analysis: A review,” *Mach. Learn. with Appl.*, vol. 4, p. 100023, 2021, doi: 10.1016/j.mlwa.2021.100023.

7) M. Mustaqeem, “Ethnicity Detection Using Deep Learning: A Comparative Study,” *IEEE Access*, vol. 8, pp. 133767–133780, 2020.

8) Alzahrani, “A Deep Learning Approach for Real-Time Detection of Confusion in Educational Videos,” *IEEE Access*, vol. 7, pp. 113300–113309, 2019.

9) F. et Al, “Learning Deep Representation for Ethnicity Recognition from Face Images and Videos,” 2020.

10) A. J. A. Albdairi, Z. Xiao, M. Alghaili, and C. Huang, “Identifying Ethnics of People through Face Recognition: A Deep CNN Approach,” *Sci. Program.*, vol. 2020, 2020, doi: 10.1155/2020/6385281.

11) Ghosh and Kollias, “Aff-Wild2: Multi-modal approach combining audio and visual features using a deep neural network,” *Proc. Eur. Conf. Comput. Vis.*, pp. 104–119, 2021.

12) et al Li, C., “Age Estimation in Videos: A Comprehensive Survey and a New Benchmark,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 360–375, 2021.

13) Y. Pan, H., Liu, “A Multi-Modal Deep Learning-Based Approach for Age Estimation,” *IEEE Access*, no. 9, pp. 19064–19073, 2021.

14) X. Han, Y., Li, “Age Estimation in Videos via a Deep Learning-Based Approach,” *EEE Trans. Image Process.*, vol. 30, pp. 2166–2178, 2021.

15) L. Shao, W., “Two-Stage Gender Detection in Unconstrained Video Sequences Based on Facial and Non-Facial Cues,” *IEEE Access*, pp. 12369–12378, 2021.

16) W. Cheng, X., “A Three-Stream CNN Model for Gender Classification in Video,” *IEEE Access,* vol. 9, pp. 22215–22224, 2021.

17) Abdelmageed, “Video-based Ethnicity Recognition using Deep Learning,” 2019.

18) V. Sharma, A., Singh, S., & Vats, “Deep learning for emotion recognition in videos: A comprehensive study.,” *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 331–343, 2020.