Detecting-Staff-Tags-in-Video-Using-Deep-Learning

1.0 Introduction

This report evaluates two methods for detecting staff presence in a video: a Convolutional Neural Network (CNN) model and K-Means clustering. The purpose of the analysis is to determine which method more accurately identifies frames containing staff members, based on labeled frame data from a sample video.

2.0 Frame Identification

The frames that contain staff presence are explicitly identified in the range from Frame 0 to Frame 1340, as indicated by the detection outcomes. This means that staff is present for a significant portion of the clip, providing consistent monitoring coverage throughout the video. The goal here was to ensure comprehensive identification of staff frames to evaluate the performance of the detection models against a manually validated reference.

3.0 Methods

3.1 Convolutional Neural Network (CNN)

The CNN model was trained to detect the presence of staff members in video frames. The model architecture included several convolutional layers followed by max-pooling, flattening, dense layers, and a dropout layer to reduce overfitting. The final output layer uses a sigmoid activation function for binary classification (staff present or not).

Model Training Details:

- **Input Size:** 224x224 pixels
- **Optimizer:** Adam with a learning rate of 0.001
- **Loss Function:** Binary Crossentropy
- **Metrics:** Accuracy
- **Training Duration:** 10 epochs
- **Data Split:** 80% for training, 20% for validation

The dataset was created by extracting frames from a video, labeling them as "staff_tag" or "no_staff_tag," and splitting them into training and validation sets. Data augmentation was applied to the training images to improve generalization.

3.2 K-Means Clustering

K-Means clustering was used as an unsupervised learning method to detect frames with staff members. Frames were resized to 128x128 pixels and converted to grayscale before applying the K-Means algorithm to cluster the frames into two groups: one with staff and one without.

Clustering Details:

- **Number of Clusters:** 2
- **Frame Representation:** Each frame was reshaped into a one-dimensional vector for clustering.

4.0 Coordinates of Staff Presence

In addition to identifying the frames, the exact coordinates of the staff presence within each frame were also extracted. The coordinates were consistently at (480, 360) across all frames. This suggests that the staff member's position remained static throughout the video, or alternatively, the camera was fixed on a specific area, allowing the staff to be consistently centered. This static positioning has implications for model evaluation, particularly in assessing the accuracy of staff presence detection.

5.0 Model Detection Results and Implications
5.1 CNN Model Results

The CNN model detected staff presence in 960 frames, with a training accuracy of 87.78% and a validation accuracy of 81.25%. While the training accuracy was relatively high, the validation accuracy indicates that the model still faced some challenges in generalizing to unseen data, though it performed better compared to the initial report. This suggests that the model may have struggled to accurately detect staff in frames that differed significantly from the training data, which could lead to false positives or missed detections.

5.2 K-Means Clustering Results

The K-Means clustering method identified 874 frames with staff presence. As an unsupervised method, K-Means does not rely on labeled data and instead clusters frames based on inherent similarities. The clustering approach provided fewer detections than the CNN model, which might imply a lower rate of false positives but also a potential for missed detections due to the simplistic nature of clustering.

5.3 Frame Detection Analysis

Manual analysis of frames revealed that staff presence was confirmed in frames ranging from Frame 0 to Frame 1340. This comprehensive detection serves as a baseline reference to compare against the automated methods used. Given the consistent location of staff at (480, 360), models such as CNN and K-Means should theoretically have an easier time detecting the staff. However, discrepancies in model outcomes suggest potential challenges in model robustness, overfitting, and sensitivity to frame variations.

6.0 Comparison of Accuracy Between CNN and K-Means

6.1 CNN Accuracy

The CNN model demonstrated a training accuracy of 87.78% and a validation accuracy of 81.25%. This relatively high accuracy during training indicates that the model was effective at learning the features of the training data. However, the drop in validation accuracy suggests some overfitting, as the model struggled to generalize to unseen frames. This overfitting issue often arises due to the model being too complex or the training data not being diverse enough. Despite these challenges, the CNN model still outperformed K-Means in terms of the total number of frames detected with staff presence (960 frames).

6.2 K-Means Accuracy

The K-Means clustering approach, being an unsupervised method, does not have a straightforward accuracy metric comparable to CNN's training and validation accuracy. Instead, its performance can be evaluated based on the number of correctly clustered frames. K-Means identified staff presence in 874 frames, which was fewer compared to CNN. The advantage of K-Means lies in its simplicity and its ability to provide insights without requiring labeled data. However, the reduced number of detections might indicate that K-Means missed some frames where staff were present, potentially due to the simplistic clustering criteria that may not capture the nuanced features of the frames.

6.3 Summary of Accuracy Comparison

Overall, the CNN model demonstrates higher detection capability with higher training (87.78%) and validation accuracy (81.25%) compared to K-Means, making it more capable of effectively detecting frames with staff presence. However, the CNN model is prone to overfitting and generalization issues, particularly with unseen data, requiring careful parameter tuning and

more diverse training data to improve robustness. In contrast, K-Means offers a simpler, unsupervised approach that avoids overfitting and does not require labeled data, making it useful for exploratory analysis. However, the K-Means model detected fewer frames with staff presence, suggesting that it may have missed some important patterns due to its simplicity and lack of nuanced feature detection, limiting its effectiveness for detailed tasks.