Human Recognition in Surveillance Settings

Rodrigo Silva
Universidade da Beira Interior
rd.silva@ubi.pt

***Abstract*—This project presents a dual-system approach for human recognition in surveillance, combining ResNet50-based similarity scoring with interpretable outputs from vision-language models. A 6-channel input from instances is processed by a ResNet50-based model, with the results reasoning from LLaVA and GPT-4o. Trained on 30,000 instances, the system delivers classification results paired with human-readable explanations.**

***Index Terms*—Human Recognition, Similarity, Surveillance, ResNet, LVLM, LLaVA, GPT-4o, Explainability**

I. Introduction

Human recognition in surveillance environments poses challenges such as varying lighting, low image quality, diverse camera angles, and pose variations. While deep learning models often achieve high accuracy, they typically lack explainability, which hinders trust and interpretability [1], [2]. Explainability is crucial in surveillance, where automated decisions impact privacy, security, and individual rights, making transparent models a need for building user trust [3]. For this project, it is proposed a dual-system approach combining a convolutional neural network and multimodal vision-language models (LVLMs) to balance accuracy and explainability. Our contributions include: (1) a data preparation pipeline; (2) training a ResNet50-based classifier; and (3) integrating LVLMs (LLaVA and GPT-4o) to generate natural language explanations of predictions. The ResNet50 architecture was selected based on professor recommendation and its strong performance in image classification benchmarks [4]. For the explainability component, the models used are LLaVA and GPT-4o, two multimodal LVLMs capable of processing both visual and textual inputs. These models generate natural language explanations of the classification results, offering insight into the visual features influencing the predictions.

This document is structured as follows: Section II outlines the methodology, Section III presents the results and analysis, and Section IV concludes the work

A. Motivation

In the surveillance scope, it is not enough for a system to identify individuals accurately, it is necessary to provide the reasoning that led to its results. The issue is that the models behind these systems, typically, do not explain their outputs. The motivation behind this project is to train a model able to identify pairs of images that belong or do not belong to the same person and enhance the explainability of its results. By combining the performance of an image classification model with the explainability offered by vision-language models, the aim is to create a more transparent system that not only makes decisions but also explains them in natural language.

B. Related Work

Models like ResNet have been widely used for image classification, showing high accuracy when trained on large datasets [4], [5]. However, most of these systems act as black boxes, providing no explanation for their outputs [6]. To address this problem, the use of explainable AI (XAI) methods is increasing [7], [8], aiming to make model decisions more transparent, trustworthy and understandable to humans. By combining the output from a deep learning method with the use of an LVLM, one can generate natural language explanations that help interpret the model's decisions. Recent models like GPT-4o demonstrate strong capabilities in aligning visual inputs with descriptive text, making them suitable tools for explainability in vision-based tasks [9].

II. Methodology

A. System Overview

The proposed system combines a deep learning model for image pair classification with a vision-language model for explainability. It takes as input a pair of RGB frames from the dataset and predicts whether they belong to the same person or not. The image pairs are fed to a modified ResNet50 model to compute a similarity score. To generate human-readable justifications, the output is fed into a vision-language model (LLaVA or GPT-4o), which produces textual explanations based on both image content and model predictions. The full architecture is shown in Figure 3.

B. Dataset Preparation

The dataset used in this project was built from a collection of Region of Interest (ROI) videos[1], where each video contained footage of a specific person. For each video, frames were extracted at a regular interval, 1 frame every 5 frames, to reduce redundancy and memory usage. Each extracted frame was then resized and padded to a fixed resolution of **128×256 pixels**, preserving the original aspect ratio, to ensure consistency in the model input. To construct the dataset, a total of **30,000 image pairs** were created: **15,000 genuine pairs**, where both images belong to the same person but come from different ROI videos, and **15,000 impostor pairs**, where the images show different individuals. This constraint on genuine pairs ensures the model learns identity features rather than just memorizing frames from a single video of that person. Each pair was formatted as a single training instance

[1]Dataset obtained from: https://socia-datasets.di.ubi.pt:50004/hugomcp/CV_24_25_data.zip

by stacking the two RGB images in-depth, resulting in an input with 6 channels (i.e., two concatenated RGB images). The final dataset was randomly split into 80% training, 10% validation, and 10% testing subsets, with proper seeding to guarantee the reproducibility of this project. Figure 1 shows an example of genuine image pairs from the dataset, where both images belong to the same person, but under different conditions (pose, lightning, camera angle). Figure 2 shows an example of impostor image pairs from the dataset.

Fig. 1. Example of genuine image pairs used in the dataset.

Fig. 2. Example of impostor image pairs used in the dataset.

C. Architecture

The system is composed of two main components (see figure 3): (1) ResNet50-based model for quantitative similarity scoring; and (2) Large Vision-Language Models (LVLMs) for qualitative analysis. ResNet50 is a deep convolutional neural network architecture known for the use of residual connections (with ResNet being a short for Residual Networks and 50 for 50 hidden layers), which help mitigate the vanishing gradient problem in very deep networks [10]. In this project, the ResNet50 is adapted to accept a 6-channel input, representing a stacked pair of RGB images (i.e., two frames), and outputs a similarity score between the two using a sigmoid activation

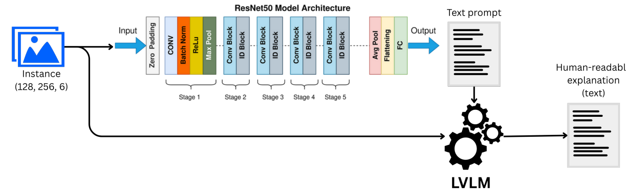

Fig. 3. System Architecture

function. The base model was initialized without pretrained weights and excludes the default classification layers. A custom classification head was added, consisting of a Global Average Pooling layer followed by a dense layer with 256 units, ReLU activation, and L2 regularization ($\lambda = 1 \times 10^{-4}$). A dropout layer with a rate of 0.5 is used to prevent overfitting, and a final sigmoid-activated dense layer with one unit performs the binary classification. The model is compiled with the Adam optimizer (learning rate of 0.0001) and trained using binary cross-entropy loss. Metrics tracked include accuracy, precision, recall, AUC, and confusion matrix components. For the explainability purposes, two LVLMs are integrated into the system: `llava-1.5-7b-hf`[2] and `gpt-4o`[3]. Both models process visual and textual inputs (multimodal) to generate natural language explanations of the classification results. LLaVA [11] is an open-source model that extends the capabilities of a language model with visual perception, trained through visual instruction tuning. GPT-4o, released by OpenAI in May 2024, is a state-of-the-art multimodal model that unifies vision, text, and audio processing with high speed and low latency. It is particularly effective in generating fluent and semantically rich image-based explanations, making it well-suited for real-time or interactive applications. Both are accessed via API; however, LLaVA is free (up to a certain limit), while GPT-4o requires a payment through the OpenAI API. Each image pair is passed to these models along with a fixed prompt. The output is a natural language explanation that justifies whether the two images belong to the same person or not, always based on the classification result. GPT-4o was selected for subsequent processes (described in the following sections) because it identifies more details in the images and delivers clearer, more accurate explanations compared to the other model.

D. Training Process

The ResNet50-based model was trained to classify genuine versus impostor image pairs using the aforementioned prepared dataset. Training was performed on a MacBook with an M1 processor and lasted approximately **15 hours**, with each epoch taking around **48 minutes**. The model was trained for up to 30 epochs with a batch size of 16, but the best performance was reached at **epoch 19** based on validation

[2]https://huggingface.co/llava-hf/llava-1.5-7b-hf
[3]https://platform.openai.com/docs/models/gpt-4o

loss monitored through early stopping with a patience of 10 epochs. Binary Cross-Entropy as the loss function and the Adam optimizer (learning rate 0.0001) were used during training. Model checkpoints were saved based on maximum validation accuracy. Lastly, performance was evaluated using accuracy, precision, recall, AUC (ROC Curve) and confusion matrix components. As shown in the Figure 4, both training accuracy and loss steadily improved, but starting at epoch 19, where the best model was trained upon, a clear divergence between the training and validation curves is easily observed, which meant that the model was starting to overfit. Training

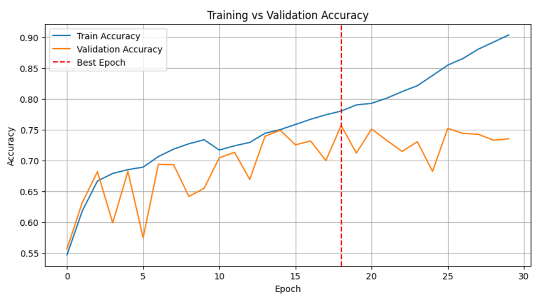

Fig. 4. Training and validation accuracy/loss over training epochs.

deep learning models on a CPU-based laptop poses challenges due to limited processing power and memory[4]. Unlike GPUs, which speed up training through parallel computation, the MacBook's CPU results in slower training and smaller batch sizes, as was experienced first hand in this project. Despite these limitations, the model was successfully trained using careful hyperparameter tuning and early stopping to avoid respective overfitting.

E. Explainability Pipeline

To enhance the transparency and explainability of the project, LVLMs were integrated to generate natural language explanations based on the outputs of the ResNet50-based classification model. As mentioned before, two models were evaluated for this purpose: LLaVA and GPT-4o. GPT-4o was ultimately selected due to its superior coherence, fluency, and semantic accuracy in the test scenarios. To reduce randomness and improve the consistency of the generated explanations, each input pair was sampled 50 times using an identical prompt. This approach created multiple candidate explanations per input. Each explanation, in textual form, was subsequently converted into a high-dimensional vector representation using sentence embeddings. These embeddings were computed via the OpenAI API using the `text-embedding-3-large`[5] model. Sentence embeddings are dense, fixed-size numerical representations that capture the semantic content of text, enabling meaningful comparisons between explanations. The embedding model maps semantically similar sentences to nearby points in a continuous vector space. This means that similar sentences should have similar embeddings; therefore, they should be closer to each other in the vector space. Given the high dimensionality of the embeddings, Principal Component Analysis (PCA) was employed to project them into a lower-dimensional space, in this case two dimensions. This step facilitates both computational efficiency and interpretability while clustering. After the dimensionality reduction process, KMeans clustering with $k = 3$ was applied to group the reduced embeddings into clusters, each representing a coherent trend or theme in the generated explanations. KMeans is an unsupervised learning algorithm that partitions data into k clusters by minimizing the variance within each cluster. It iteratively assigns points to the nearest cluster centroid and updates the centroids based on the current assignments. The dominant cluster was identified as the one containing the largest number of explanations. To determine a final, representative explanation, the embedding closest to the centroid of the dominant cluster was selected. This approach ensures that the chosen explanation is not only semantically central but also representative of the majority of generated outputs, thereby enhancing consistency and reliability in model explanation.

III. Results and Analysis

A. Evaluation Metrics

To evaluate the performance of the trained model, the following metrics were computed: accuracy, precision, recall, F1-score, and AUC, using a threshold of 0.5 on the similarity score. This threshold meant that equal or above 0.5, the image pair was genuine, otherwise, under 0.5, the image pair was impostor. Each of these metrics highlights a different aspect of the model's performance. Accuracy measures the proportion of correct predictions over the total number of predictions. Precision measures the proportion of correct positive predictions, indicating how often the model is right when it predicts a positive. Recall indicates the proportion of actual positives the model correctly detects, reflecting its ability to avoid missing relevant instances. The F1-score balances precision and recall, making it helpful when both types of errors are important. Lastly, the AUC measures the model's ability to distinguish between classes across all thresholds, with highest values the better. As shown in Table I, the model achieved an accuracy of 75.37%, meaning that fairly amount of predictions matched the ground truth. The precision of 71.06% suggests that a good proportion of positive predictions were correct, while the recall of 85.60% highlights the model's effectiveness at identifying true positives. The F1-score reached 77.65%, reflecting a good overall performance in balancing precision and recall. Lastly, the AUC of 0.8417 (see Figure 5 for the corresponding ROC curve) demonstrates the model's strong ability to distinguish between classes (across different decision thresholds).

The confusion matrix in Figure 6 provides further insight into the distribution of true positives, true negatives, false positives, and false negatives. The high number of true positives suggests that the model rarely misses actual positive instances, which is consistent with the high recall score. However, the presence of a notable number of false positives is reflected in the moderate precision. In the context of surveillance, false

[4]https://www.ibm.com/think/topics/cpu-vs-gpu-machine-learning
[5]https://platform.openai.com/docs/models/text-embedding-3-large

TABLE I
EVALUATION METRICS

Metric	Value
Accuracy	0.7537
Precision	0.7106
Recall	0.8560
F1 Score	0.7765
AUC	0.8417

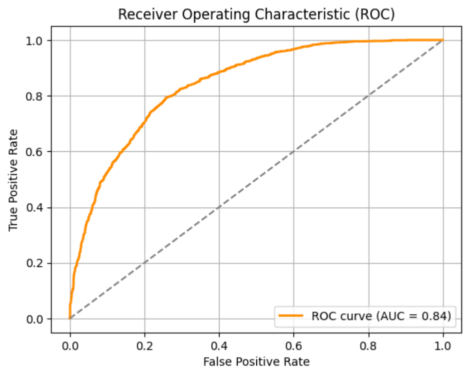

Fig. 5. ROC Curve

negatives—failing to detect a person of interest—tend to be more costly than false positives, which only raise a false alarm. Therefore, achieving high recall is generally more critical than attaining high precision.

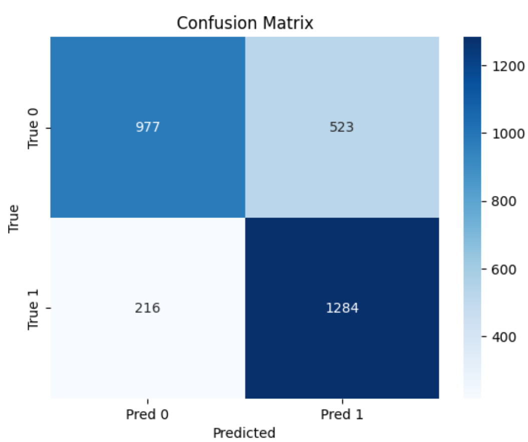

Fig. 6. Confusion matrix of the model.

Overall, these results demonstrate that the proposed approach provides a reliable classification pipeline, with particularly strong performance in recall and class separability, as indicated by the AUC.

B. Explainability Results

To guide the selection of the most representative explanation, the cluster-based approach proposed in class was applied and visualized using PCA. This method aims to identify the most coherent explanation among those generated by the LVLM. Figure 7 presents a 2D projection of sentence embeddings from 50 candidate explanations generated by GPT-4o, where each point represents an individual explanation and colors denote different clusters. The final explanation was selected as the point nearest to the centroid of the largest cluster, ensuring semantic centrality and consistency while reducing the impact of noise and outliers. It is possible to see that the best response is inside a considerably dense cluster, in contrast to the other two clusters. The best response is present fully in the appendix A.

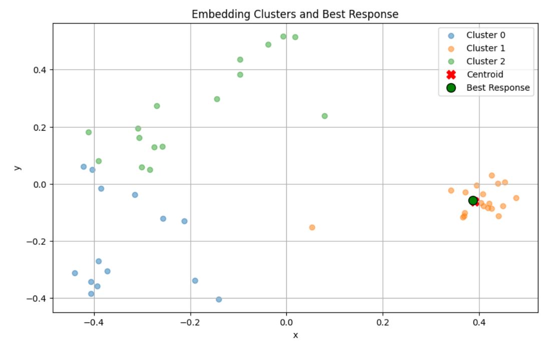

Fig. 7. 2D PCA projection of sentence embeddings from 50 explanations; Colors represent KMeans clusters ($k = 3$).

This method proved effective in filtering variability and producing consistent explanations with the real explanation. GPT-4o generated focused and contextually rich descriptions. It matched not only the colors of the clothing and the environment present in images but also the poses of the person and camera angles (from top, front, etc.). LLaVA was initially considered but ultimately discarded due to its inability to generate sufficiently detailed or relevant explanations for the selected samples, often mistaking colors and clothing.

IV. CONCLUSION

This work presents an automatic and explainable method for human recognition in surveillance. The system combines a custom ResNet50 architecture for image pair classification with GPT-4o for human-readable justifications. A PCA and clustering approach was proposed to consolidate diverse natural language explanations into a consistent semantic core. This enables both accurate identification and explainable output. Future directions include extending this to a few-shot approach.

REFERENCES

[1] T. Ahmad, P. Katari, A. K. Pamidi Venkata, C. Ravi, and M. Shaik, “Explainable ai: Interpreting deep learning models for decision support,” *Advances in Deep Learning Techniques*, vol. 4, no. 1, pp. 80–108, 2024.

[2] E. ŞAHiN, N. N. Arslan, and D. Özdemir, “Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning,” *Neural Computing and Applications*, vol. 37, no. 2, pp. 859–965, 2025.

[3] B. Sasikala and S. Sachan, “Decoding decision-making: embracing explainable ai for trust and transparency,” *Exploring the frontiers of artificial intelligence and machine learning technologies*, vol. 42, 2024.

[4] S. Mascarenhas and M. Agarwal, “A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification,” in *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)*, vol. 1. IEEE, 2021, pp. 96–99.
[5] Y. Pan, J. Liu, Y. Cai, X. Yang, Z. Zhang, H. Long, K. Zhao, X. Yu, C. Zeng, J. Duan *et al.*, “Fundus image classification using inception v3 and resnet-50 for the early diagnostics of fundus diseases,” *Frontiers in Physiology*, vol. 14, p. 1126780, 2023.
[6] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, “Explainable ai: A brief survey on history, research areas, approaches and challenges,” in *Natural Language Processing and Chinese Computing*, J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, Eds. Cham: Springer International Publishing, 2019, pp. 563–574.
[7] V. Kamakshi and N. C. Krishnan, “Explainable image classification: The journey so far and the road ahead,” *AI*, vol. 4, no. 3, pp. 620–651, 2023.
[8] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan *et al.*, “Explainable ai (xai): Core ideas, techniques, and solutions,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, 2023.
[9] T. Les and B. Sawicki, “Textual explanations for image classification using multimodal llm.”
[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015. [Online]. Available: https://arxiv.org/abs/1512.03385
[11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

Appendix A
GPT-4o Generated Output

This appendix shows the generated output for the image present in the figure 8. **1. Describe Each Person Individually**

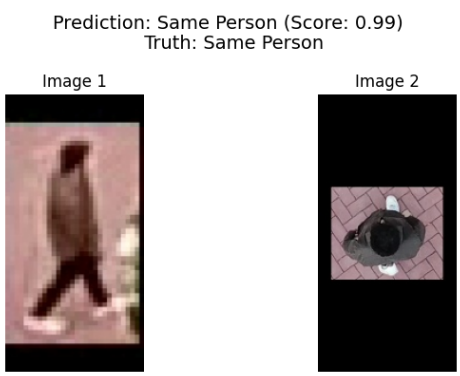

Fig. 8. Example of the output of the ResNet model with the image pair

Image A:
- **Gender**: Indeterminate
- **Approximate age range**: Indeterminate
- **Skin tone**: Not clearly visible
- **Hair color and style**: Not clearly visible
- **Facial structure**: Not clearly visible
- **Facial expression**: Not applicable
- **Clothing**:
  – Appears to be a dark-colored jacket
  – Dark pants
  – Light-colored shoes
- **Accessories**: None visible

Image B:
- **Gender**: Indeterminate
- **Approximate age range**: Indeterminate
- **Skin tone**: Not clearly visible
- **Hair color and style**: Possibly dark hair, short or close-cropped
- **Facial structure**: Not visible
- **Facial expression**: Not applicable
- **Clothing**:
  – Appears to be a dark-colored jacket
  – Dark pants
  – Light-colored shoes
- **Accessories**: None visible

2. Identify Key Features for Comparison
- **Skin tone**: Not visible
- **Hair**: Dark, short or close-cropped in Image B; not visible in Image A
- **Clothing**: Both images show a dark-colored jacket, dark pants, and light-colored shoes
- **Accessories**: None visible in either image

3. Justification for Identifying as the Same Person

Both images depict a person wearing similar clothing: a dark jacket, dark pants, and light-colored shoes. The angles and perspectives are different, with Image A showing a side view and Image B showing a top-down view. The hair in Image B appears dark and short, but Image A does not provide enough information about the hair. The lack of visible facial and distinguishing features makes it challenging to confirm identity solely from these images.

4. Summary of the Analysis

The conclusion that the individuals could be the same person relies on the similarity in clothing—specifically the dark jacket, dark pants, and light-colored shoes visible in both images. However, due to the lack of visible facial features and the different perspectives, this observation is not definitive.