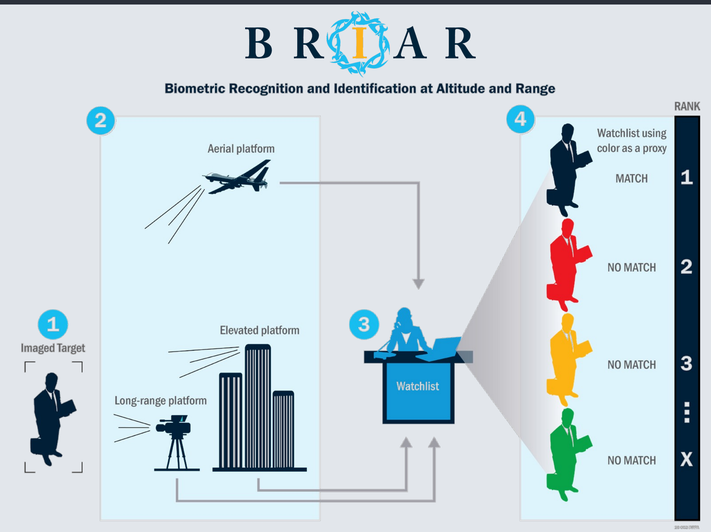
Learned Frame Feature Aggregation for Face Recognition with Low-Quality Video

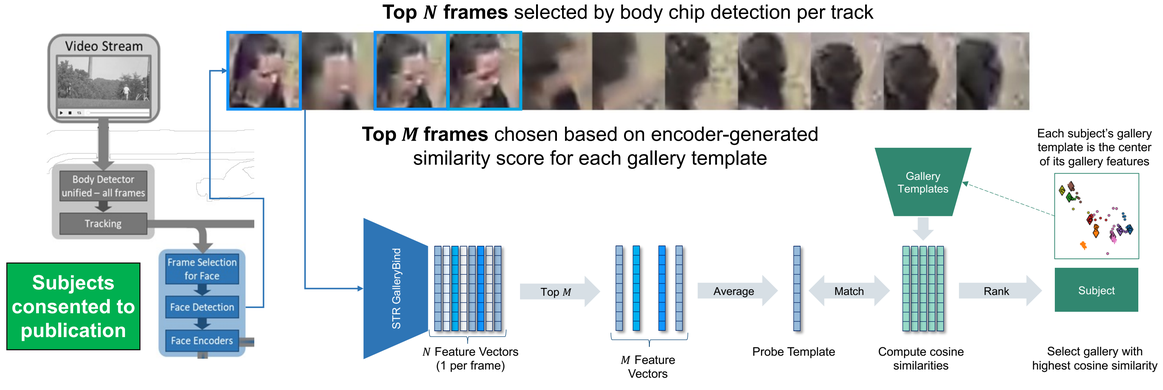
Allen Tu, Josh Gleason, Soraya Stevens, Matt Meyn, Nathan Shnidman, Jennifer Xu



- IARPA program that develops full-body biometric recognition systems for underconstrained scenarios.
- Footage from a wide range of sensor platforms like security cameras and UAVs in real-world environments.
- Challenges like extreme distance, camera pose, motion blur, occlusion, and atmospheric interference.
- STR's MORFI³US pipeline **fuses face, body, and gait recognition** for real-time, multimodal, opportunistic, and robust biometric recognition with incomplete identity information.

Background: Face Recognition

- **Gallery G:** High-quality images and footage of **known subjects** that can be identified.
  - **Gallery Template g_i:** For a given gallery subject i, (1) encode their face detections and then (2) use the center (mean) of the feature cluster to represent the subject as a single vector g_i.
- **Probe:** Real-world video (e.g., taken in the field by a UAV) of a subject that we want to identify
  - **Probe Template p_θ:** Generated for each video using **Top M Aggregation** (below).



Motivation

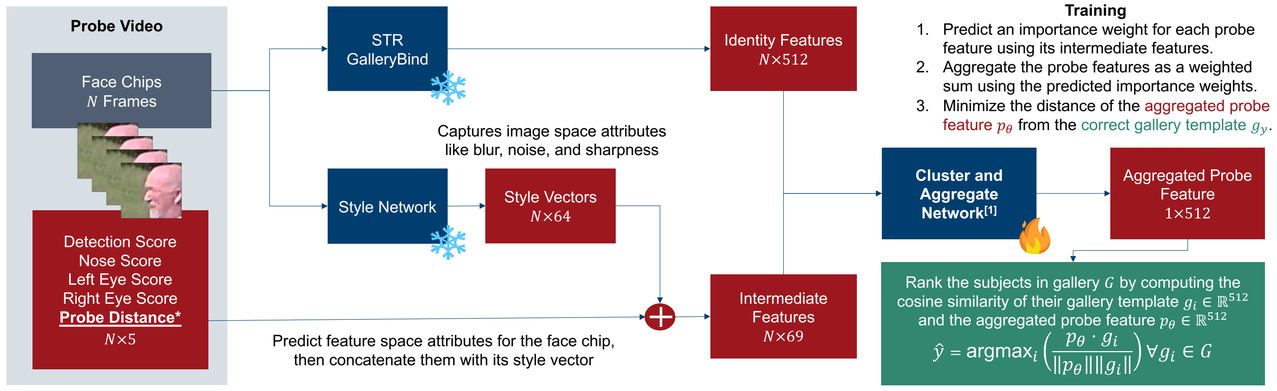
- Score-based probe template aggregation works very well most of the time. **However, it (1) limits the number of features used to create the probe template and (2) assigns equal importance to each one.**
- **Can we train a neural network to learn to aggregate a higher-fidelity probe template using image and feature space attributes of the probe features?**

[1] M. Kim, F. Liu, A. Jain, and X. Liu, 'Cluster and Aggregate: Face Recognition with Large Probe Set', in *Advances in Neural Information Processing Systems*, 2022.
[2] L. Kang, P. Ye, Y. Li, and D. Doermann, 'Convolutional Neural Networks for No-Reference Image Quality Assessment', in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1733–1740.

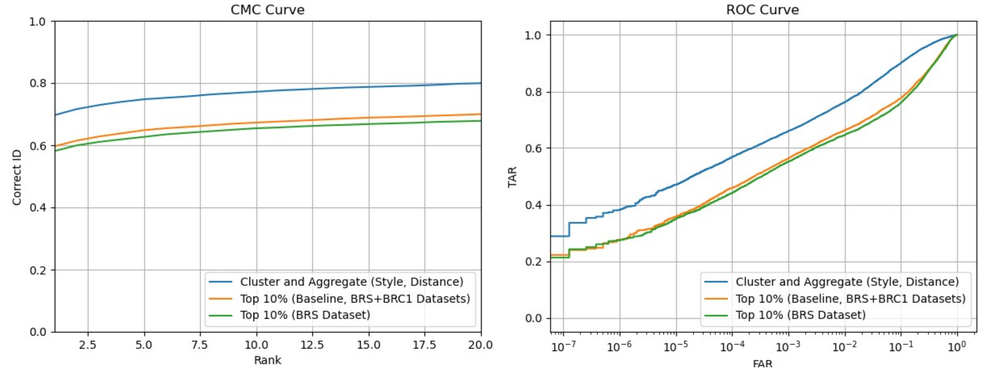
Approach

STR © 2024 • STR Proprietary

1. For a given probe, **produce attributes for each of the N face chips** such as visual style, face detection score, body keypoint scores, and predicted probe to gallery distance. Concatenate them into **N intermediate features.**
2. Train a **Cluster and Aggregate network**[1] to predict a **relative importance score** for each feature using the intermediate features. Features with undesirable attributes (e.g., the subject is looking away, poor visual fidelity) will have lower importance than those with desirable ones (e.g., the subject is in focus and looking at the camera).
3. Compute **a weighted sum of the probe features** using the relative importance scores**. This aggregated probe template p_θ (1) incorporates all N probe features and (2) considers their image and feature space attributes.**



Results



- Cluster and Aggregate **boosts performance by 10%** over the score-based baseline.
- Adding scores to the intermediates will most likely further improve performance, but rechipping the training data is time consuming.

- **Probe Distance*** is the cosine distance between the probe feature and its **correct gallery template $g_{\hat{y}}$**. However, **we do not know the correct subject y during evaluation**, so these results are an **upper bound for our next step**.

Next Steps

1. Train a neural network (i.e., CNNIQA[2]) to **predict the Probe Distance** given a face chip, filling in the missing piece that will **allow us to use our Cluster and Aggregate network during evaluation.**
2. Create an aggregation network that **reconstructs a high-quality feature** instead of predicting a weighted sum.