Accenture-MM1: A Multimodal Person Recognition Dataset

Kyle O'Brien Michelle Rybak Jiong Huang Adam Stevens
Madeline Fredriksz Michael Chaberski Danielle Russell Lindsey Castin
Michelle Jou Nishant Gurrapadi Marc Bosch
Accenture Federal Services
{kyle.obrien, michelle.rybak, jiong.huang, adam.stevens, madeline.fredriksz, m.chaberski,
danielle.russell, lindsey.castin, michelle.jou, nishant.gurrapadi, marc.bosch.ruiz}@afs.com

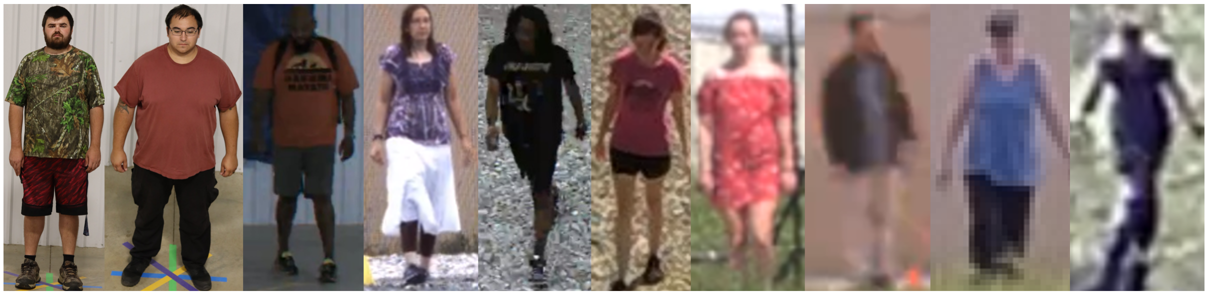

Figure 1. Example subject crops from ACC-MM1

Abstract

In this paper we present a new dataset to fuel multimodal research in uncooperative and surveillance scenarios. Accenture Multimodality 1 (ACC-MM1) is a large-scale multimodal biometric recognition dataset composed of imagery and video. The dataset includes challenges such as long ranges, high pitch angles, varied atmospheric conditions, and mixed image quality levels. Ultimately, a dataset containing 227 unique subjects, 303 hours of video, and 12,344 still images was captured in indoor and outdoor conditions. In addition to traditional modalities (face, gait, etc.), data for a novel biometric modality, activity gait, was collected. Covariates included appearance changes, walking with weighted loads, and body distortions. Furthermore, to enable standardized performance testing of ACC-MM1, an evaluation protocol was created. Baseline performance of popular and novel recognition algorithms is reported to encourage research in the challenging conditions present in ACC-MM1.

1. Introduction

Biometric datasets typically focus on a particular modality such as face (CelebA [1], CASIA-WebFace [2], etc.) or gait (OUMVLP [3], CASIA-B [4] , etc.), but do not necessarily consider multiple modalities *simultaneously*. Additionally, many common evaluation datasets are limited to constrained captures, few camera viewpoints, and lack environmental diversity. Focusing on this narrow set of scenarios has produced performance saturated benchmarks ([5], [6], [7]). To counter these shortcomings, we propose the Accenture Multimodality 1 (ACC-MM1) dataset. Note, our use of "multimodal" refers to biometric modalities, and not imagery types (RGB, thermal, etc.).

ACC-MM1 consists of four modalities captured across the same set of subjects. These modalities include face recognition (FR), gait recognition (GR), whole-body recognition (WBR), and activity gait recognition (AGR). FR and GR are well-established modalities and are, respectively, the recognition of an individual based on their face and walking gait. WBR is recognition via features such as body shape or other anthropometrics. WBR differs from GR in that it performs on both stationary and moving individuals. AGR is a novel modality and is the recognition of an individual based on how they perform daily activities, such as walking on staircases, opening doors, or texting on a mobile phone. AGR expands GR beyond walking, and enables recognition in scenarios where the subject may be performing distinct motions that are viable for biometric differentiation.

Including multiple modalities allows for testing and evaluation of algorithms which can leverage numerous complementary biometric signals. This encourages development

Condition	Description
Long Range (LR)	Captures from cameras positioned ≥150m from the subject
Aerial	Captures from unmanned aerial vehicle (UAV) mounted cameras
Elevated	Captures from elevated (>3m) or high pitch (≥30°) angle ground cameras
Atmopsheric	Captures in distortion-inducing weather and atmospheric conditions
Quality	Captures of mixed quality (resolution, compression, etc.)

Table 1. Description of challenging capture conditions

of more robust, and therefore more practical, recognition algorithms. In addition to considering multiple modalities, ACC-MM1 seeks to address a variety of challenging capture conditions, as specified in Table 1.

2. Background

Research in biometrics has traditionally been driven by the availability of datasets which can accurately capture a target problem space. For example, until pose-varied, 'in the wild'-style datasets became available, high performing facial recognition algorithms were limited to front-facing, passport-style photos [15]. The emergence of datasets like LFW and the IJB series ([16], [17], [18], [19]) helped drive the creation of algorithms capable of performing in more unconstrained scenarios. However, recognition in some of the most challenging conditions, as detailed in Table 1, has yet to be adequately solved, due in part to a lack of representative datasets. Improved recognition algorithm performance under these conditions enhances public safety and national security use cases.

2.1. Related Work

Biometric datasets have existed for decades as a means of training and evaluating biometric recognition algorithms [20]. However, most datasets typically only consider a subset of the modalities and conditions present in ACC-MM1. More recently, efforts have been made to add multiple modalities and challenging conditions. A summary of recent biometric datasets is detailed in Table 2.

UCCS and LRFID both target a limited set of modalities and do not consider aerial data. The IJB-S dataset focuses on FR and lacks atmospheric condition covariates. P-DESTRE and PRAI-1581 do not consider FR and GR, and lack LR, elevated, as well as atmospheric conditions. D4FLY only includes FR and does not consider challenging capture conditions. The BRIAR and MEVID datasets contain several of the modalities and conditions present in ACC-MM1. However, one key differentiator of ACC-MM1 is the inclusion of probe activity data to support AGR, as well as adding novel covariates, such as full body distortion (achieved by having the subject wear a disposable poncho).

2.2. Ethical Considerations

When creating datasets to support advanced biometric technology, the highest of ethical standards must be maintained during dataset collection and distribution. Two considerations are the importance of protecting subject safety and privacy. To meet these requirements, Human Subject Research (HSR) guidelines were strictly adhered to. All subjects signed an IRB-approved consent form and were compensated for their participation. Additionally, all subjects whose likeness appears in this paper have signed a separate consent form to appear in publications related to the collection. To ensure ethical downstream use, access to the dataset requires IRB review. Before access is granted, appropriate action must be taken to safeguard the data.

3. ACC-MM1 Data Collection

Data was collected in Indiana, U.S.A. over 26 days in July and August 2022. The dataset contains 227 unique subjects captured in a variety of different covariates, including an appearance change, a weighted load carry, and a full body distortion. In total, 303 hours of footage were captured, yielding a dataset 5.4TB in size.

Capture Station	# Cameras	# Media	# Covariates
Still Enrollment	2	12,344 Stills	3
Video Enrollment	4	4,003 Videos	4
Video Probe	11	6,455 Videos	4
Video Activity Probe	3	1,463 Videos	3

Table 3. Data Collection Overview

Each subject completed 4 capture stations throughout the collection, as shown in Table 3. A *still* is a standalone image capture, whereas a *frame* is an individual frame sourced from a video composed of many frames. Each of these sta-

Dataset	Year	# Subjects	# Media	Modalities	LR	Aerial	Elevated	Atmos.	Quality	Group
UCCS [8]	2017	4,362	75,738	FR, GR	✓		✓			✓
IJB-S	2018	202	6,208	FR	✓	✓	✓		✓	✓
LRFID [9]	2019	100	10,962	FR	✓			✓	✓	
P-DESTRE [10]	2021	253	105,518	WBR		✓			✓	✓
D4FLY [11]	2021	31	62	FR						
PRAI-1581 [12]	2021	1,581	39,461	WBR		✓			✓	
BRIAR [13]	2023	1,231	551,451	FR, GR, WBR	✓	✓	✓	✓	✓	
MEVID [14]	2023	158	∼17,700	FR, GR, WBR	✓	✓	✓	✓	✓	✓
ACC-MM1 (Ours)	2023	227	24,265	FR, GR, WBR, AGR	✓	✓	✓	✓	✓	✓

Table 2. Datasets related to ACC-MM1. Only RGB/visible spectrum data is considered in this table. The requirements to meet the LR, aerial, elevated, atmospheric, and quality conditions are as listed in Table 1. The group requirement means that some portion of probe videos contain more than one person.

tions served a different purpose. The still and video enrollment stations were designed to capture high quality data of the subject for the purpose of enrolling them into biometric galleries. The video probe and video activity probe captures sought to simulate challenging, real-world recognition scenarios across multiple biometric modalities.

Additionally, captured data was annotated and curated to enable training and evaluation of recognition algorithms. Body and face bounding box annotations were produced using a suite of automated and manual annotation tools. After data was annotated, videos, which were typically 1-2 minutes in length, were split into multiple short 1-7 second tracks. These tracks were then used to create evaluation subsets that allow for targeted performance assessments.

3.1. Still Enrollment Capture

The objective of the still enrollment capture was to obtain indoor imagery for FR and WBR algorithm enrollment. Imagery was captured with two digital single-lens reflex cameras positioned at a range of 4.85m from the subject. The camera properties and configuration are shown in Table 4 and Figure 2, respectively.

#	Range	Pitch	Yaw	Camera	HxW(px)
1	4.85m	0°	0°	Canon EOS Rebel T8i	6000x4000
2	4.85m	45°	0°	Canon EOS Rebel T8i	6000x4000

Table 4. Still enrollment camera properties. Range is relative to the center of the capture area.

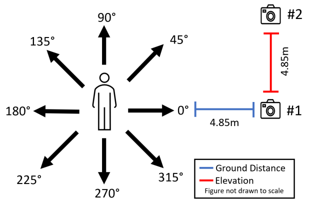

Figure 2. Still enrollment camera configuration

Upon starting the station, subjects were directed to face forward and keep their arms at their sides for each capture. Subjects were instructed to rotate through yaw angles in the range [0°, 360°] in 45° increments using an asterisk-shaped guide adhered to the floor. After each 45° rotation, both cameras captured a still. Subjects completed 3 trials with the following respective conditions for each trial; neutral pose, eyes closed, and appearance change. Example imagery is shown in Figure 3.

Trials were designed to capture covariates that have been known to challenge FR (pose [21], sensitivity to periocular region [22], etc.) and WBR algorithms (appearance changes [23]). The station yields 48 stills per subject.

3.2. Video Enrollment Capture

To capture enrollment data for video-based recognition algorithms, a multi-camera indoor walking course was established. Four 10m lengths were each offset by 45°, and all intersected with one another in the center to form an asterisk shape. This configuration allowed for multiple yaw angles to be captured. Four cameras (detailed in Table 5) were positioned about the structured walking course as shown in Figure 4.

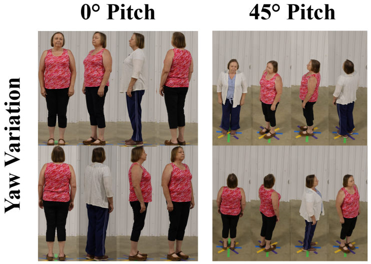

Figure 3. Enrollment stills taken from different angles and trials

#	Range	Pitch	Yaw	Camera	HxW(px)
1	10m	24°	90°	GoPro HERO9 Black	1520x2704
2	10m	0°	90°	FLIR BFS-PGE-50S5C-C	1080x2448
3	10m	24°	180°	GoPro HERO9 Black	1520x2704
4	10m	0°	180°	FLIR BFS-PGE-50S5C-C	1080x2448

Table 5. Video enrollment cameras. Yaw is relative to position "A" set as 0° and camera range measurements are relative to the center of the walking course.

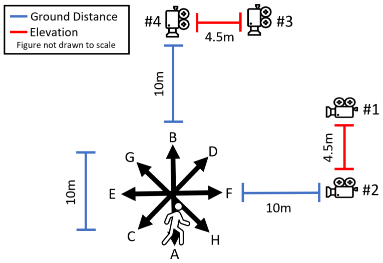

Figure 4. Video enrollment camera configuration

In reference to Figure 4, subjects were instructed to walk on a flat cement floor along the 10m lengths in the order A-B-A-C-D-C-E-F-E-G-H-G. In some cases, due to time limitations, subjects only walked the lengths A-B-A-E-F-E. Subjects faced forward and paused for ∼1 second at the end of each 10m length. Subjects completed 4 trials of the walking course with changes made for each trial. The conditions for the four walking trials were baseline, carrying of a 2-4kg backpack, a shoe change, and an appearance change. Covariates were selected to introduce variations that have historically impacted GR performance (footwear, load, etc. [24]). In total, the video enrollment station yields 16 videos per subject.

3.3. Video Probe Capture

The video probe capture collected outdoor video data in unconstrained conditions known to disrupt biometric recognition algorithms. These conditions included LR [25], aerial captures, high surveillance-style pitch angles [26], atmospheric conditions [27], and data quality [28]. The LR, aerial, and pitch angle conditions are achieved by the camera configuration described later in this section. The atmospheric conditions were captured by performing the collection in an area which experiences warm summer months. This reliably introduced atmospheric turbulence into the captured imagery. To add conditions related to quality, mixed compression levels were used and cameras were selected to obtain in-frame subject heights which varied widely between 25 and 400 pixels.

Subjects were instructed to complete a structured walking course like that described in Section 3.2. However, the environment and terrain changed from an indoor cement floor to an outdoor graveled environment. This change simulated a real-world recognition setup, where the subject is unlikely to be walking on the same surface or in the same conditions during both enrollment and probe captures.

Table 6 details the cameras used in this station, and Figure 5 shows the camera configuration. The combination of cameras varied throughout the collection, and not all subjects were filmed with every camera. The camera configuration includes ground level cameras, elevated surveillance cameras, camcorders, and cameras mounted on both a fixed-wing (Camera #4) and multirotor (Camera #5) UAV. The fixed-wing UAV circled the capture area, while all other cameras remained stationary during capture.

#	Range	Pitch	Yaw	Camera	HxW(px)
1	10m	0°	90°	Vivotek IZ9361-EH	1080x1920
2	10m	30°	90°	FLIR BFS-PGE-50S5C-C	1080x2448
3	10m	30°	90°	Axis P1455-LE	1080x1920
4	15m	34°	N/A	FlightWave Edge	720x1280
5	15m	34°	180°	Anafi Parrot	1080x1920
6	100m	0°	225°	FLIR BFS-PGE-50S5C-C	1080x2448
7	100m	0°	225°	Vivotek IZ9361-EH	1080x1920
8	300m	0°	270°	FLIR BFS-PGE-50S5C-C	1080x2448
9	300m	0°	270°	Canon EOS R5	2160x4096
10	300m	0°	270°	Canon XF400	1080x1920
11	500m	0°	315°	Canon EOS R5	2160x4096

Table 6. Probe video camera properties. Yaw is relative to position “A” set as 0°, and camera range is relative to the center of the walking course.

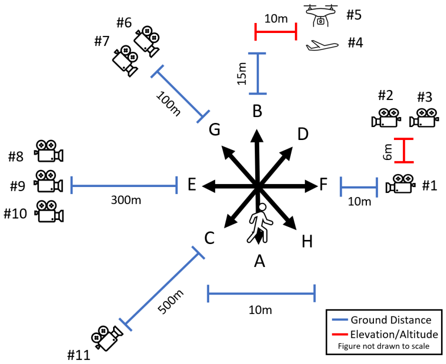

Figure 5. Probe video camera configuration

Subjects completed 4 trials of the course, with modifications being made for each trial. These trials included a baseline walk, appearance change, simultaneous appearance change and carrying of a 2-4kg backpack, and finally a body distortion using a poncho and foot booties. A notable modification from the covariates used during video enrollment (Section 3.2), was the addition of body distorting accessories (such as ponchos and foot booties). This covariate was designed to disrupt recognition algorithms which consider body shape as a biometric feature. Example probe video imagery is shown in Figure 6.

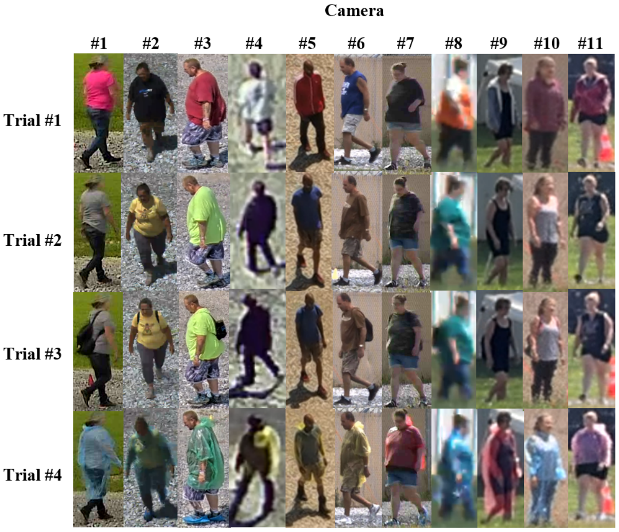

Figure 6. Crops from the probe video capture. Camera numbers correspond to Table 6.

3.4. Video Activity Probe Capture

To support the AGR modality, an activity course was constructed to capture probe videos of subjects performing common activities. The activities are described in Table 7. Subjects were instructed to navigate the course by walking to an assigned area, pausing for ∼1 second, and then performing the designated activity. The cameras that were utilized are detailed in Table 8. An overview of the course and camera configuration is shown in Figure 7. Subjects completed 3 trials. These included a baseline course navigation, navigation with body distortion via a poncho, and a group navigation consisting of multiple subjects completing the activity course simultaneously.

3.5. Annotations

ACC-MM1 was annotated with bounding boxes using a combination of automated and semi-automated annotation tools. For indoor enrollment captures, YOLOX [29] was used for body detection, and MogFace [30] was used for face detection. Imagery was first passed through YOLOX to identify the body region. Thereafter, if YOLOX returned

Activity #	Description
Walk	Walking on flat ground
Enter Car	Entering a car
Exit Car	Exiting a car
Stairs (Up)	Walking up a 6-stair staircase
Stairs (Down)	Walking down a 6-stair staircase
Pull Door	Pulling a door open and entering a building
Push Door	Pushing a door open and exiting a building
Text	Texting while walking

Table 7. Activity Descriptions

#	Range	Pitch	Yaw	Camera	HxW(px)
1	15m	22°	45°	Axis P1455-LE	1080x1920
2	25m	0°	0°	FLIR BFS-PGE-50S5C-C	1080x2448
3	100m	0°	315°	FLIR BFS-PGE-50S5C-C	1080x2448

Table 8. Activity probe video camera properties

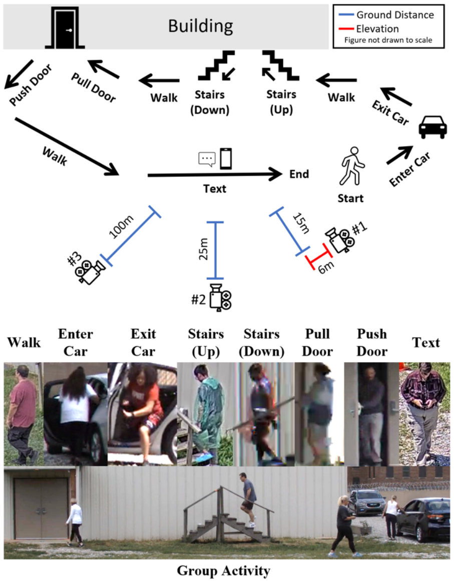

Figure 7. Activity course, camera configuration, and example imagery for the activity probe capture. Subjects performed a series of common activities. Though not discernable from the figure, all activities were in the field of view of all cameras.

a person detection, the frame would be cropped, and MogFace would automatically detect the highest-probability face within the YOLOX detection region.

Probe videos were processed by running YOLOX in the same manner as the enrollment captures. However, before running MogFace, body bounding boxes were manually reviewed for accuracy with a semi-automated pipeline that utilized the Computer Vision Annotation Tool (CVAT) [31]. Among other annotation capabilities, the CVAT interface allows manual annotators to create and modify bounding boxes for video data. CVAT was preloaded with the automated annotations from YOLOX, and annotators were tasked with (1) drawing bounding boxes in cases where YOLOX failed to detect the subject, (2) deleting false positive bounding boxes, and (3) adjusting bounding boxes to better localize the subject. These manual review steps were taken due to the lower quality of the probe data. CVAT was also used to produce activity annotations for the probe activity videos. Annotators were shown full videos and marked the start and end frame of each predefined activity.

The high volume of probe videos collected in Section 3.3 necessitated an automated activity labeling approach. To accomplish this, segments of the video probe capture were labeled as either “standing” or “walking” using bounding box information and known capture conditions. Since subjects were instructed to stand during the beginning and end of videos, standing tracks were obtained from those portions of the recording. Walking tracks were labeled such that they ideally contained one unidirectional segment of the subject’s walk. Put another way, each walking track should contain the subject walking no more than 10 meters in one direction. Identifying walking automatically was challenging, as the subject’s structured walk required them to change direction multiple times as they walked through different yaw angles relative to the camera. However, by considering temporal bounding box information, it is possible to extract unidirectional walking tracks. As a subject reaches the end of a length, their bounding box location in the horizontal direction typically reaches a local minimum or maximum before they change direction. The horizontal direction, rather than the vertical direction, was utilized since it is a less noisy signal. The noise in the vertical direction is caused by the subject’s natural tendency to oscillate vertically [32] as they walk.

To produce walking tracks using bounding box information, peak and valley detection [33] was run on the signal produced by the subject’s X-coordinate position through time. The peaks and valleys represent the points in the video when the subject changes direction. Candidate walking tracks were then identified by selecting frames which fell between two direction changes. Candidate walking tracks were further filtered based on a minimum (4 second) and maximum (7 second) duration. Figure 8 shows a plot of a subject’s movement in the X-direction labeled with events of interest. One limitation of this automated approach is that in cases where the subject is walking directly at the camera (resulting in little movement in the horizontal direction), it can be difficult to accurately segment the unidirectional walking track. This approach is also limited to stationary cameras, which resulted in the need to manually label walking tracks from the circling, fixed-wing UAV.

In addition to bounding box and activity annotations, subject demographic information, environmental conditions, and camera conditions are provided to allow for targeted training or evaluation.

3.6. Evaluation Datasets

Establishing standardized evaluation datasets ensures fair comparison of recognition algorithms. It also allows

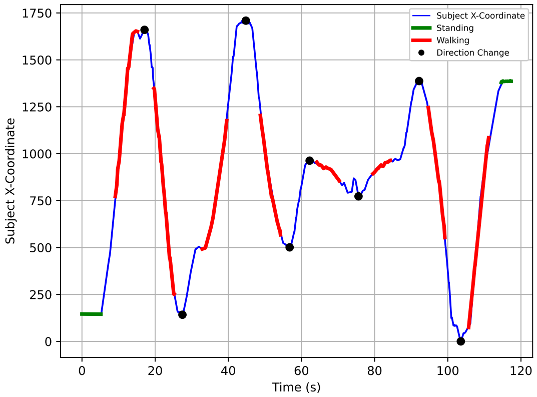

Figure 8. Plot of a subject's X-coordinate versus time. This video was trimmed into 2 standing and 8 walking tracks.

for especially challenging conditions to be targeted for research. To create evaluation datasets, probe videos were broken into multiple, 1-7 second tracks, depending on the activity the subject was performing. Generating activity-specific tracks ensures that each probe track contains limited information. This processing mimics challenging real-world recognition scenarios where, for a given video, a person of interest may appear only briefly. Enrollment, however, is treated as being controlled and cooperative. As such, enrollment stills and videos are available in their entirety for gallery creation.

The probe tracks were then sub-selected based on the conditions and activities present in their source videos to create three targeted evaluation subsets (1) *ACC-MM1-Standard*, (2) *ACC-MM1-Challenge*, and (3) *ACC-MM1-Activity*. From here onward, they will be referred to as *Standard*, *Challenge*, and *Activity*. The tracks in the *Standard* and *Challenge* subsets were sourced from videos captured in Section 3.3. The *Standard* subset included tracks from videos which were captured at <150m, and from low pitch angles (<30°). The *Challenge* subset is composed of footage captured from ≥150m, as well as elevated and UAV cameras yielding pitch angles ≥30°. *Activity* includes tracks taken from the videos described in Section 3.4. *Activity* omits the group walk trial to focus the evaluation scope.

4. Baseline Evaluation

A series of recognition algorithms were evaluated against the *Standard*, *Challenge*, and *Activity* subsets. This testing was conducted to provide baseline performance against the evaluation subsets and gain an understanding of how different modalities perform in the conditions broadly present in ACC-MM1. Algorithms are detailed in Table 9.

The selection of algorithms gives an overview of recognition performance across recent algorithmic approaches. GR-MVIT2 and WB-SWINB were created for the purposes of this evaluation and utilize popular modern backbones

Algorithm	Year	Modality	Preprocessing	Backbone	Speed (ms/frame)
ArcFace [34]	2019	FR	SCRFD [35]	ResNet-50 [36]	320.5
GaitSet [37]	2019	GR	Mask2Former [38]	GaitSet	90.5
GaitPart [39]	2020	GR	Mask2Former	GaitPart	91.2
PartialFC [40]	2021	FR	MTCNN [41]	iResNet-100 [42]	75.8
OSNet [43]	2021	WBR	-	OSNet	80.0
GaitGL [44]	2021	GR	Mask2Former	GaitGL	91.0
Centroids-ReID [45]	2021	WBR	-	ResNet-50	65.0
AGW-Body [46]	2022	WBR	-	ResNet50-NL [47]	69.0
AdaFace [48]	2022	FR	MTCNN	ResNet-101	65.1
CFSM [49]	2022	FR	MTCNN	iResNet-50	53.6
GaitBase [50]	2022	GR	Mask2Former	ResNet-9	90.7
GR-MVIT2	2023	GR	-	MViTv2 [51]	8.4
WB-SWINB	2023	WBR	-	Swin-B [52]	86.4

Table 9. Baseline algorithms. Speed is reported as the average per-frame inference time on 1080x1920 (HxW) RGB videos using an NVIDIA V100 GPU. To initially localize the subject, all algorithms utilized YOLOX running at ∼5.6 ms/frame.

trained on the BRIAR dataset. All subjects in the ACC-MM1 dataset were considered viable for testing, as no algorithm had been trained on ACC-MM1.

4.1. Aggregation and Scoring

The recognition algorithms in Table 9 operate by producing template representations for each identity of interest. Ideally, when compared, templates produced by media containing the same individual will be similar, and templates produced by media of different individuals will be dissimilar. In some cases, the algorithms shown in Table 9 expect multiple frames as input, producing a template for a batch of frames. Others, however, produce a single template for each still or frame. The evaluation protocol requires each probe track to be compared against gallery subjects, each of whom is represented by multiple enrollment videos and still media. Therefore, for practical computation purposes, a template aggregation method was utilized to avoid a combinatorial explosion during template similarity comparison. Template aggregation was performed as a multi-step mean aggregation across all media templates belonging to a given probe track or enrollment subject, as shown in Figure 9.

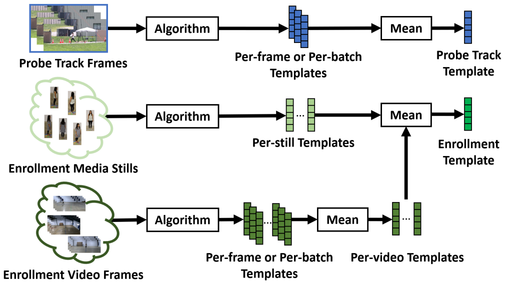

Figure 9. Baseline template aggregation method. Probe track templates are averaged across the video. Enrollment still templates are obtained on a per-still basis. Enrollment video templates are first averaged across videos, before being broadly averaged with the enrollment still templates to create the enrollment template.

Enrollment templates for each subject are stored to create the gallery. Cosine similarity was used to calculate the similarity of probe track templates to each enrollment template in the gallery. The aggregation and scoring methodology described above can be performed a myriad of ways [53], and the approach in Figure 9 only represents a naïve implementation for baseline evaluation purposes.

4.2. Baseline Results

Algorithms were first evaluated on the *Standard* and *Challenge* subsets. Performance is reported on biometric verification, which is the task of verifying whether an input probe subject is the gallery subject they claim to be. For evaluation, receiver operator characteristic (ROC) curves are plotted to show the tradeoff between true positives and false positives across a sweep of thresholds. Here, true positives are represented by the *True Accept Rate* (TAR), and false positives are represented by the *False Accept Rate* (FAR). The ROC curves for the top performing algorithms are shown in Figure 10. Results for all algorithms are shown in Table 10.

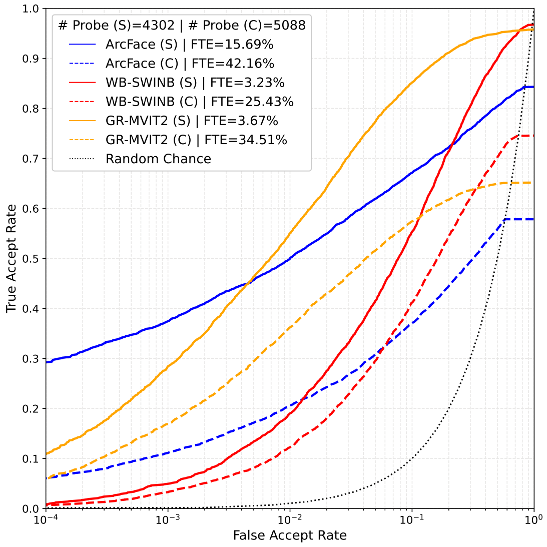

Figure 10. ROC curves for the top performing algorithms of each modality on the *Standard* (S) and *Challenge* (C) subsets. The abrupt horizontal lines on the righthand side of the plots indicate that the algorithm failed to enroll (FTE) a subset of probe tracks. This is typically caused by missed face or body detections.

Considering TAR@FAR=10^{-2}, the top performing algorithms on *Challenge* were GR-MVIT2, ArcFace, and WB-SWINB. Most algorithms failed to perform above 0.1 TAR@FAR=10^{-2} on *Challenge*, demonstrating its difficulty. The results also show that among top performing algorithms FR and GR outperformed WBR. This is unexpected since all tracks include the whole body, but not all tracks include visible faces or walking sequences. Lower WBR performance may be attributable to the nascency of WBR research relative to FR and GR.

Another task is closed-set ranked retrieval (CSRR). For CSRR, a probe subject is known to exist in the gallery.

Algorithm	Modality	TAR@FAR=10^{-3}	TAR@FAR=10^{-2}	TAR@FAR=10^{-1}
ArcFace	FR	**0.374** | ***0.111***	***0.501*** | ***0.206***	***0.671*** | *0.370*
GaitSet	GR	0.048 | 0.031	0.169 | 0.099	0.411 | 0.285
GaitPart	GR	0.040 | 0.019	0.139 | 0.075	0.367 | 0.266
PartialFC	FR	*0.192* | 0.037	*0.318* | 0.086	0.510 | 0.205
OSNet	WBR	0.014 | 0.010	0.053 | 0.049	0.231 | 0.204
GaitGL	GR	0.019 | 0.011	0.061 | 0.042	0.207 | 0.172
Centroids-ReID	WBR	0.031 | 0.025	0.133 | 0.099	0.435 | 0.342
AGW-Body	WBR	0.039 | *0.038*	0.147 | 0.114	0.418 | 0.335
AdaFace	FR	0.177 | *0.038*	0.295 | 0.081	0.477 | 0.192
CFSM	FR	0.149 | 0.024	0.262 | 0.067	0.451 | 0.185
GaitBase	GR	0.050 | 0.034	0.178 | 0.114	0.471 | 0.328
GR-MVIT2	GR	***0.285*** | **0.169**	**0.550** | **0.362**	**0.852** | **0.574**
WB-SWINB	WBR	0.050 | 0.033	0.190 | *0.122*	*0.554* | ***0.412***

Table 10. Algorithm results at key FAR values. Results for each data subset are shown as *Standard* | *Challenge*. In this table, and all following results tables, top performing algorithms at each rate are bolded, second-best algorithms are bolded and italicized, and third-best algorithms are italicized.

At runtime, the probe is scored against all gallery subjects and a ranked list of the gallery subjects is returned sorted by descending similarity to the probe. To measure CSRR performance across many probes, identification rate (IDR) is plotted at increasing rank values. For example, if the IDR@Rank=5 is 0.6, it would indicate that for 60% of probe tracks, the probe identity appears in the top 5 most similar identities returned by the query. CSRR results are shown in Figure 11 (top performing) and Table 11 (all).

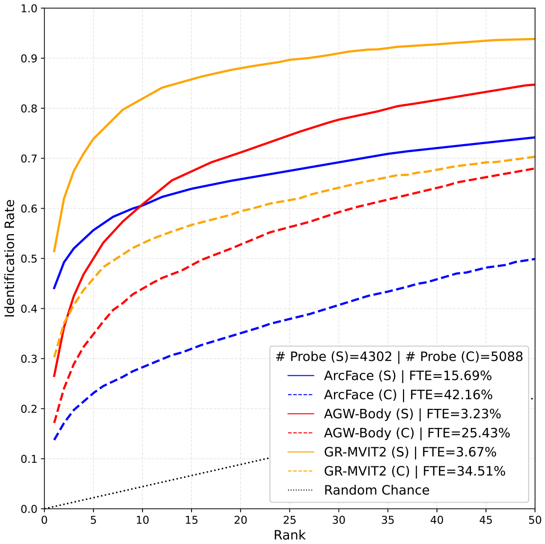

Figure 11. CSRR of top performing algorithms from each modality on *Standard* (S) and *Challenge* (C) subsets.

For the CSRR task, GR-MVIT2, AGW-Body, and ArcFace were the top performing algorithms for each modality based on IDR@Rank=5 on *Challenge*. On this metric, the top WBR algorithm outperformed the top FR algorithm. However, most algorithms failed to achieve results above 0.5 IDR@Rank=5 on either the *Standard* or *Challenge* datasets, which potentially limits their use in many real-world applications.

Algorithm	Modality	IDR@Rank=1	IDR@Rank=5	IDR@Rank=20
ArcFace	FR	***0.441*** | *0.137*	***0.556*** | 0.231	*0.659* | 0.348
GaitSet	GR	0.196 | 0.103	0.379 | 0.222	0.561 | 0.369
GaitPart	GR	0.182 | 0.089	0.355 | 0.198	0.548 | 0.349
PartialFC	FR	*0.272* | 0.051	0.373 | 0.101	0.482 | 0.192
OSNet	WBR	0.078 | 0.051	0.180 | 0.144	0.360 | 0.300
GaitGL	GR	0.088 | 0.049	0.193 | 0.113	0.380 | 0.244
Centroids-ReID	WBR	0.173 | 0.108	0.372 | 0.266	0.601 | *0.446*
AGW-Body	WBR	0.265 | ***0.171***	*0.499* | ***0.350***	***0.712*** | ***0.528***
AdaFace	FR	0.248 | 0.049	0.344 | 0.095	0.457 | 0.180
CFSM	FR	0.225 | 0.039	0.313 | 0.083	0.430 | 0.172
GaitBase	GR	0.242 | 0.134	0.430 | *0.270*	0.620 | 0.410
GR-MVIT2	GR	**0.515** | **0.303**	**0.739** | **0.461**	**0.881** | **0.595**
WB-SWINB	WBR	0.134 | 0.076	0.329 | 0.204	0.595 | 0.404

Table 11. All algorithm results shown as *Standard* | *Challenge*

Lastly, results are shown for the Activity subset, designed for AGR algorithms. Since AGR is a novel modality, no baseline algorithms for this approach were available. Instead, the baseline algorithms from Table 9 were run against the Activity subset. As shown in Table 12 the top performing algorithm across all activities at key metric values was GR-MVIT2. The high performance of this GR algorithm may be due to elements of walking being present in many of the activities. These results show that research is needed to improve AGR performance in unconstrained conditions. Additionally, the top two algorithms, GR-MVIT2 and ArcFace, were designed for GR and FR, respectively. This indicates that research into multimodal fusion is likely to boost performance.

Results show that for verification and CSRR tasks, the *Challenge* subset provides a significant level of difficulty for state-of-the-art recognition algorithms. Improving performance on the conditions present in the *Challenge* subset may be possible by training algorithms on data which incorporates many of the conditions found in ACC-MM1. Improvement on the *Activity* subset could be realized by targeted training or enrollment of specific activities. Although the ACC-MM1 evaluation subsets are challenging, future algorithms which take advantage of multiple gallery media, more intelligently aggregate templates across clips and frames, use novel distance metrics, or fuse modalities may improve performance and real-world viability.

5. Conclusion

This paper introduced the ACC-MM1 dataset and evaluation subsets to fuel multimodality biometric research under unconstrained and challenging conditions. ACC-MM1 includes multiple modalities and targets challenging capture conditions, which are not typically present in most large-scale video recognition datasets. By providing an evaluation protocol, researchers can compare recognition algorithms in a standardized manner.

Areas of improvement for ACC-MM1 are increasing variation in environments, subject demographics, camera models, and the addition of other capture covariates. Further, the dataset could better mimic challenging capture conditions with the addition of occlusions and less constrained walking patterns. Future collections in the ACC-MM series will expand upon these areas to create a more comprehensive recognition dataset. Ultimately, this comprehensiveness will ensure real-world generalizability of the models trained and evaluated on the ACC-MM series.

Acknowledgments: We would like to thank the following individuals who contributed to the planning, collection, and annotation of the dataset: Daniel Cuervo, Richard Brown, Jake Frommeyer, Larry Clayton, Mary Pagnutti, Bob Ryan, and Jeff Beaubien.

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100001]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Algorithm	Activity							
	Walk	Enter Car	Exit Car	Stairs (Up)	Stairs (Down)	Pull Door	Push Door	Text
ArcFace	***0.377*** | 0.448	***0.178*** | 0.201	***0.270*** | *0.344*	***0.274*** | 0.324	***0.280*** | 0.315	0.062 | 0.055	***0.268*** | ***0.342***	***0.290*** | 0.353
GaitSet	0.181 | 0.459	0.071 | 0.223	0.082 | 0.202	0.116 | 0.278	0.093 | 0.294	*0.087* | 0.178	0.037 | 0.164	*0.147* | 0.403
GaitPart	0.128 | 0.349	0.045 | 0.164	0.060 | 0.145	0.066 | 0.270	0.073 | 0.263	0.076 | 0.175	0.056 | 0.138	0.100 | 0.393
PartialFC	0.125 | 0.189	0.093 | 0.108	0.131 | 0.156	0.104 | 0.147	0.100 | 0.163	0.004 | 0.011	*0.086* | 0.100	0.130 | 0.200
OSNet	0.064 | 0.174	0.037 | 0.152	0.039 | 0.121	0.039 | 0.135	0.042 | 0.159	0.029 | 0.105	0.022 | 0.104	0.057 | 0.203
GaitGL	0.096 | 0.221	0.037 | 0.134	0.028 | 0.138	0.046 | 0.189	0.024 | 0.121	0.025 | 0.127	0.052 | 0.100	0.047 | 0.173
Centroids-ReID	0.096 | 0.335	0.033 | *0.294*	0.060 | 0.277	0.069 | 0.293	0.087 | 0.291	0.055 | 0.189	0.056 | 0.212	0.107 | 0.350
AGW-Body	0.149 | ***0.541***	0.089 | ***0.442***	0.064 | ***0.390***	0.069 | ***0.371***	0.087 | ***0.436***	0.044 | ***0.298***	0.048 | *0.249*	*0.147* | ***0.520***
AdaFace	0.139 | 0.196	0.100 | 0.130	0.113 | 0.170	0.108 | 0.151	0.100 | 0.135	0.007 | 0.015	0.071 | 0.071	0.123 | 0.173
CFSM	0.085 | 0.174	0.071 | 0.115	0.103 | 0.156	0.066 | 0.112	0.093 | 0.125	0.000 | 0.007	0.078 | 0.078	0.063 | 0.157
GaitBase	*0.231* | *0.505*	0.048 | 0.201	0.067 | 0.227	0.108 | *0.355*	0.090 | *0.349*	***0.098*** | *0.229*	0.071 | 0.186	0.133 | *0.477*
GR-MVIT2	**0.548** | **0.762**	**0.431** | **0.587**	**0.390** | **0.518**	**0.390** | **0.587**	**0.422** | **0.561**	**0.371** | **0.502**	**0.368** | **0.483**	**0.547** | **0.730**
WB-SWINB	0.181 | 0.352	*0.112* | 0.223	*0.135* | 0.230	*0.124* | 0.232	*0.149* | 0.266	*0.087* | 0.160	*0.086* | 0.201	0.143 | 0.273

Table 12. Algorithm results on the *Activity* subset. Results are shown as TAR@FAR=10^{-2} | IDR@Rank=5.

References

[1] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[2] D. Yi, Z. Lei, S. Liao, and S. Li, "Learning Face Representation from Scratch," *ArXiv*, vol. abs/1411.7923, 2014.

[3] X. Li, Y. Makihara, C. Xu, and Y. Yagi, "Multi-View Large Population Gait Database With Human Meshes and Its Performance Evaluation," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pp. 234–248, 2022.

[4] S. Yu, D. Tan, and T. Tan, "A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition," *18th International Conference on Pattern Recognition (ICPR 2006)*, vol. 4, pp. 441–444, 2006.

[5] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," October 2007.

[6] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The First Manually Collected, In-the-Wild Age Database," *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vols*, 2017.

[7] S. Sengupta, J. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to Profile Face Verification in the Wild," *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, 2016.

[8] M. Gunther, P. Hu, C. Herrmann, C. Chan, M. Jiang, S. Yang, A. R. Dhamija, D. Ramanan, J. Beyerer, J. Kittler, M. A. Jazaery, M. I. Nouyed, C. Stankiewicz, and T. E. Boult, "Unconstrained Face Detection and Open-Set Face Recognition Challenge," *International Joint Conference on Biometrics*, 2017.

[9] K. Miller, B. Preece, T. D. Bosq, and K. Leonard, "A Data-constrained Algorithm for the Emulation of Long-range Turbulence-degraded Video," vol. 11001, 2019.

[10] S. V. A. Kumar, E. Yaghoubi, A. Das, B. S. Harish, and H. Proenca, "The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, and Short/Long-Term Re-Identification From Aerial Devices," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1696–1708, 2021.

[11] L. Chen, J. Boyle, A. Danelakis, J. Ferryman, S. Ferstl, D. Gicic, A. Grudzien, A. Howe, K. Marcin, K. Mierzejewski, and T. Theoharis, "D4FLY Multimodal Biometric Database: Multimodal Fusion Evaluation Envisaging On-the-move Biometric-based Border Control," *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 1–8, 2021.

[12] S. Zhang, Q. Zhang, Y. Yang, X. Wei, P. Wang, B. Jiao, and Y. Zhang, "Person Re-Identification in Aerial Imagery," *IEEE Transactions on Multimedia*, vol. 23, pp. 281–289, 2021.

[13] D. Cornett, J. Brogen, N. Barber, D. Aykac, S. Baird, N. Burchfield, C. Dukes, A. Duncan, R. Ferrell, J. Goddard, G. Jager, M. Larson, B. Murphy, C. Johnson, I. Shelley, N. Srinivas, B. Stockwell, L. Thompson, M. Yohe, R. Zhange, S. Dolvin, H. J. Santos-Villalobos, and D. S. Bolme, "Expanding Accurate Person Recognition to New Altitudes and Ranges: The BRIAR Dataset," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 593–602, 2023.

[14] D. Davila, D. Dawei, B. Lewis, C. Funk, J. V. Pelt, R. Collins, K. Corona, M. Brown, S. McCloskey, A. Hoogs, and B. Clipp, "MEVID: Multi-view Extended Videos with Identities for Video Person Re-Identification," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1634–1643, 2023.

[15] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face Recognition: A Literature Survey," *ACM Comput. Surv., vol. 35, no*, vol. 35, pp. 399–458, Dec. 2003.

[16] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Janus Benchmark A," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1939, 2015.

[17] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother, "IARPA Janus Benchmark-B Face Dataset," *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 592–600, 2017.

[18] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, T. W. Niggel, J. Anderson, J. Cheney, and P. Grother, "IARPA Janus Benchmark-C:

Face Dataset and Protocol," *International Conference on Biometrics (ICB)*, pp. 158–165, 2018.

[19] N. D. Kalka, B. Maze, J. A. Duncan, K. O'Connor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain, "IJB-S: IARPA Janus Surveillance Video Benchmark," *IEEE 9th International Conference on Biometrics Theory, Applications and Systems*, pp. 1–9, 2018.

[20] I. D. Raji and G. Fried, "About Face: A Survey of Facial Recognition Evaluation," 2021.

[21] NIST, "Ongoing Face Recognition Vendor Test (FRVT)," 2023.

[22] S. Karahan, M. Yıldırım, K. Kırtac, F. Rende, G. Butun, and H. Ekenel, "How Image Degradations Affect Deep CNN-based Face Recognition," *2016 International Conference of the Biometrics Special Interest Group*, pp. 1–5, 2016.

[23] X. Gu, H. Chang, B. Ma, S. Bai, S. Shan, and X. Chen, "Clothes-Changing Person Re-identification with RGB Modality Only," 2022.

[24] M. H. Khan, M. S. Farid, and M. Grzegorzek, "Vision-Based Approaches Towards Person Identification Using Gait," *Comput. Sci*, Nov. 2021.

[25] F. W. Wheeler, R. L. Weiss, and P. H. Tu, "Face Recognition at a Distance System for Surveillance Applications," *Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–8, 2010.

[26] S. Ruan, C. Tang, Z. Xu, Z. Jin, S. Chen, H. Wen, H. Liu, and D. Tang, "Multi-Pose Face Recognition Based on Deep Learning in Unconstrained Scene," *Applied Sciences*, vol. 10, 2020.

[27] W. Robbins and T. Boult, "On the Effect of Atmospheric Turbulence in the Feature Space of Deep Face Recognition," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1617–1625, 2022.

[28] T. Marciniak, A. Chmielewska, R. Weychan, M. Parzych, and A. Dabrowski, "Influence of Low Resolution of Images on Reliability of Face Detection and Recognition," *Multimedia Tools and Applications*, vol. 74, no. 12, pp. 4329–4349, 2015.

[29] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," 2021.

[30] Y. Liu, F. Wang, J. Deng, Z. Zhou, B. Sun, and H. Li, "MogFace: Towards a Deeper Appreciation on Face Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4093–4102, 2022.

[31] B. Sekachev, N. Manovich, M. Zhiltsov, A. Zhavoronkov, D. Kalinin, B. Hoff, D. K. Tosmanov, A. Zankevich, D. Sidnev, and M. Markelov, "OpenCV/CVAT," *M. Chenuet, a-andre, telenachos, A. Melnikov, J. Kim, L. Ilouz, N. Glazov, Priya4607, R. Tehrani, S. Jeong, V. Skubriev, S. Yonekura, V. Truong, zliang7, lizhming and T. Truong, opencv/cvat: v1. 1*, vol. 1, no. 0, 2020.

[32] F. Massaad, T. M. Lejeune, and C. Detrembleur, "The Up and Down Bobbing of Human Walking: A Compromise Between Muscle Work and Efficiency," *The Journal of Physiology*, vol. 582, pp. 789–799, 2007.

[33] P. Du, W. A. Kibbe, and S. M. Lin, "Improved Peak Detection in Mass Spectrum by Incorporating Continuous Wavelet Transform-based Pattern Matching," *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065, 2006.

[34] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019.

[35] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, "Sample and Computation Redistribution for Efficient Face Detection," *ArXiv*, 2021.

[36] K. He, X. Zhang, R. Shaoqing, and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[37] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition," *Proceedings of the AAAI conference on artificial intelligence*, pp. 8126–8133, 2019.

[38] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention Mask Transformer for Universal Image Segmentation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1280–1289, 2022.

[39] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "GaitPart: Temporal Part-based Model for Gait Recognition," *Proceedings of the IEEE/CVF Conference on Computer Cision and Pattern Recognition*, pp. 14225–14233, 2020.

[40] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, and Y. Fu, "Partial FC: Training 10 Million Identities on a Single Machine," *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1445–1449, 2021.

[41] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, pp. 1499–1503, 2016.

[42] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Improved Residual Networks for Image and Video Recognition," *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 9415–9422, 2021.

[43] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Learning Generalisable Omni-Scale Representations for Person Re-Identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5056–5069, 2022.

[44] B. Lin, S. Zhang, and X. Yu, "Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14648–14656, 2021.

[45] M. Wieczorek, B. Rychalska, and J. Dabrowski, "On the Unreasonable Effectiveness of Centroids in Image Retrieval," *Neural Information Processing: 28th International Conference, ICONIP 2021*, pp. 212–223, 2021.

[46] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. Hoi, "Deep Learning for Person Re-Identification: A Survey and Outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2022.

[47] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local Neural Networks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

[48] M. Kim, A. K. Jain, and L. Xiaoming, "AdaFace: Quality Adaptive Margin for Face Recognition," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18729–18738, 2022.

[49] F. Liu, M. Kim, A. Jain, and X. Liu, "Controllable and Guided Face Synthesis for Unconstrained Face Recognition," *ArXiv*, 2022.

[50] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu, "OpenGait: Revisiting Gait Recognition Toward Better Practicality," *ArXiv*, 2022.

[51] Y. Li, C. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4794–4804, 2022.

[52] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021.

[53] S. N. Garg, R. Vig, and S. Gupta, "A Survey on Different Levels of Fusion in Multimodal Biometrics," *Indian Journal of Science and Technology*, vol. 10, Nov. 2017.