Modular Multi-Face Tracking Geared Toward Face Recognition in Surveillance Videos

Cássio B. Nascimento
Department of Computing
Federal University of Paraná - UFPR
Curitiba - PR - Brazil
Email: cbnascimento@inf.ufpr.br

Adolfo J. Neto
EBTS Empresa Brasileira de Tecnologias
e Sistemas Ltda
Curitiba - PR - Brazil
Email: adolfo.neto@ebts.com.br

Luciano Silva
Department of Computing
Federal University of Paraná - UFPR
Curitiba - PR - Brazil
Email: luciano.ufpr.br@gmail.com

***Abstract*—Face recognition has achieved great accuracy when used in controlled conditions, however, these results aren't usually carried over to video surveillance scenarios. To facilitate the use of face recognition for video surveillance, face selection can be employed as an intermediate step. This dissertation presents a study of face selection where we rework a multi-face tracking pipeline and with few changes manage to increase tracking and reconnection capabilities. Through experimentation with different face detection models, random parameter search and a simpler face quality measure, we achieved an increase of 10.1% in Multiple Object Tracking Precision (MOTP) and 9% more in the IDF1 metric. All experiments were conducted on a public multi-face tracking dataset, which we also expanded through manual video annotations. [1].**

I. Introduction

Closed-Circuit Television (CCTV) systems play an important role in everyday security as a situational crime prevention method, which involves the manipulation of an environment towards reducing the opportunities for crime to occur while increasing a potential offender's perception of risk [1]. CCTV systems act as a deterrent to possible illegal activities and as a source of visual evidence for investigative purposes, as a consequence, the availability of CCTV recordings is linked to substantial increases in the probability of a crime being solved [2].

However, due to the vast amount of visual data that CCTV systems can gather, especially in crowded environments, it becomes increasingly difficult for law enforcement and CCTV operators to process the more fine-grained information. Therefore, it is imperative to incorporate active monitoring in the form of event-based notifications, firearm detection and wanted persons detection, seeing as these measures are directly correlated with increased effectiveness of video surveillance systems [3].

With respect to face recognition, there are many requirements and challenges involved when gathering facial information from CCTV videos in crowded and unrestricted environments. For example, low resolution, poor light distribution, frequent face occlusion and irregular head pose. For these reasons, it is ill-advised to choose a straightforward approach such as employing face recognition on every detected face, since this method is computationally expensive and might deteriorate the face association accuracy [4].

To advance the use of face recognition in crowded CCTV videos, one approach is to use a processing pipeline to detect, cluster and assess the faces and identities before the face recognition step. Namely, face selection concatenates face detection, tracking and quality assessment, so that face recognition is employed only in a set of images that best represents the face of each person captured by a video surveillance camera.

We rework the pipeline proposed in [5] to extract face tracks and the set of highest-quality face images associated with each track. Through a newer face detection model, simplified and more strict face quality assessment, parameter tuning and added dataset, we achieve an overall increase of 10.1% and 9% to two important multi-object tracking metrics. In summary, the contributions of the present work are the following:

- We introduce three, manually annotated, publicly available, unconstrained videos of difficult surveillance scenarios for multi-face tracking evaluation.
- Newer face detection models were evaluated in regard to their average precision and speed, with the best one being chosen for further experimentation.
- Improved tracklet reconnection due to proposed changes on the face quality assessment step, resulting in higher identification metric results.
- Further exploration and tuning of system's parameters.
- We present the results through common multi-object tracking metrics, which allows for standardized and easier comparison between future works.
- Our implementation and dataset annotations are publicly available for further improvement and experimentation in [6].

II. Related Work

Early works for face selection utilize the Viola-Jones [7] model to detect faces in videos and, afterward, analyze each face according to quality metrics such as pose, brightness, resolution and detection confidence. These methods [4], [8], [9] demonstrate that choosing faces based on quality metrics increases recognition accuracy and super-resolution quality [10].

[1] This work relates to Cássio B. Nascimento's M.Sc. dissertation

The use of pose as a quality metric for face image selection is supported in [11], where a model for face pose estimation was developed and tested in surveillance videos to support face recognition. Thus, showing that the selection of frames with the best face orientation can improve face recognition's effectiveness.

In cases with multiple simultaneous faces in the same video sequence, multi-face tracking separates faces into identities before quality assessment and selection. Usually, tracking is done by instantiating a generic object tracker upon face detection. As is the case in [12], where a particle filter-based tracking algorithm tracks the faces and the best quality image per identity is stored.

A common approach to face selection is to design a pipeline comprised of a series of models for each step [13] (face detection, tracking, quality assessment and association) . In [14], the KCF tracker [15] is used to track detected faces and to improve processing speed. After obtaining the face tracks, face selection is applied based on detection confidence, image size and blur. A face representation is created with the best images for each identity, resulting in a face retrieval accuracy of 84%.

Another example, the Long-Term Face Tracking (LTFT) system [5] presents a multi-face tracking pipeline composed of 4 modules, with the difference being a tracklet reconnection module based on face recognition.

Similarly [16] proposes a method for unconstrained multi-face tracking and clustering that also focuses on providing longer tracklets of each person. The generation of tracklets relies both on the face and multiple body parts. Tracklets with a strong association are recursively connected providing clusters of identities, which are then used for outlier detection and reconnection.

An important point to be highlighted is the scarcity of publicly accessible datasets specifically aimed at developing multi-face tracking algorithms in surveillance scenarios [5]. Some of the most common datasets for tracking such as PETS2017 [17], CAVIAR [18], i-LIDS [19], VIRAT [20], CVBASE [21], VOT [22] and MOTChallenge [23], refer to tasks such as pedestrian tracking, vehicle tracking and person re-identification.

Other datasets aimed at the study of face images do not necessarily provide annotations (ground-truth) that can be used for simultaneous multi-face tracking, for example, 300-VW's [24] annotations pertain to fiducial points for a single person by video sequence; the Chokepoint dataset [25], despite being recorded in a surveillance scenario that is relevant to the task at hand, only offers annotations of people's eyes positions.

To test and evaluate multi-face tracking methods a dataset must annotate the positions of all faces in each frame, by bounding box or segmentation masks for example, and all bounding boxes that represent the same face throughout the video sequence must be assigned a single numeric identifier for each identity.

III. Methodology

As discussed previously, face recognition's performance is consistently improved when assisted by face selection, especially in multi-face scenarios. This section introduces the LTFT system [5], chosen as a baseline due to it providing a dataset for multi-face tracking evaluation, its modularity and the results of its pipeline being compatible with the expected results from a face selection algorithm: a collection of face tracks where each one has a set of best images associated to it that can then be utilized for face recognition.

We hypothesize that modular multi-face tracking-by-detection systems can be improved with few significant changes, allowing for an ever-adaptive pipeline. Subsections III-B and III-C present the proposed changes to the LTFT pipeline and their motivations. We also identify a dataset bias in the original work, for which we propose to add new manually annotated videos for evaluation, although this discussion is reserved to subsection IV-B.

A. The LTFT System Architecture

The LTFT pipeline can be described as a rank-based verification system for long-term multi-face tracking in crowded environments. We implemented the system from scratch in Python and it is available at [6]. For further details, the reader is also directed to [5].

The LTFT system is a tracking-by-detection approach and its pipeline, illustrated in Figure 1, is composed of 4 modules. The first module is the tracking module, which is responsible for predicting the positions of face tracklets where there was no face detection matched to their current locations. To this end, an instance of the KCF tracker [15] is initialized for each of these lost tracklets and their locations are predicted for T_{max} frames or, until they can be associated with a new face detection.

The second module is the face association module. It propagates the identity of each tracklet by solving a data association problem, which involves trying to link the position of each tracklet on the previous frame, with the position of each new face detection on the current frame. The Faceboxes [26] model is used for face detection and the association is solved using the Hungarian method [27].

The third module is the face-based tracklet reconnection module. Whenever a face is detected its image quality is checked and separated into one of three categories: enrollable, verifiable, or discarded. To extend tracks, this module uses enrollable and verifiable face images to generate face templates for each tracklet. Therefore, each tracklet is composed of an identifier and the detections associated with it, which are separated into two sets of face images, one verifiable and one enrollable.

Afterward, allow $\mathcal{T}$ to represent the set of all tracklets obtained until the current frame. For each tracklet $T_k \in \mathcal{T}$ that has an assigned detection on the current frame, take all tracklets $T_i \in \overline{\mathcal{T}}$ where $\overline{\mathcal{T}} = \mathcal{T} \setminus \{T_k\}$. For each tracklet T_i, calculate the average of its enrollable face templates, $\overline{E_{T_i}}$. And

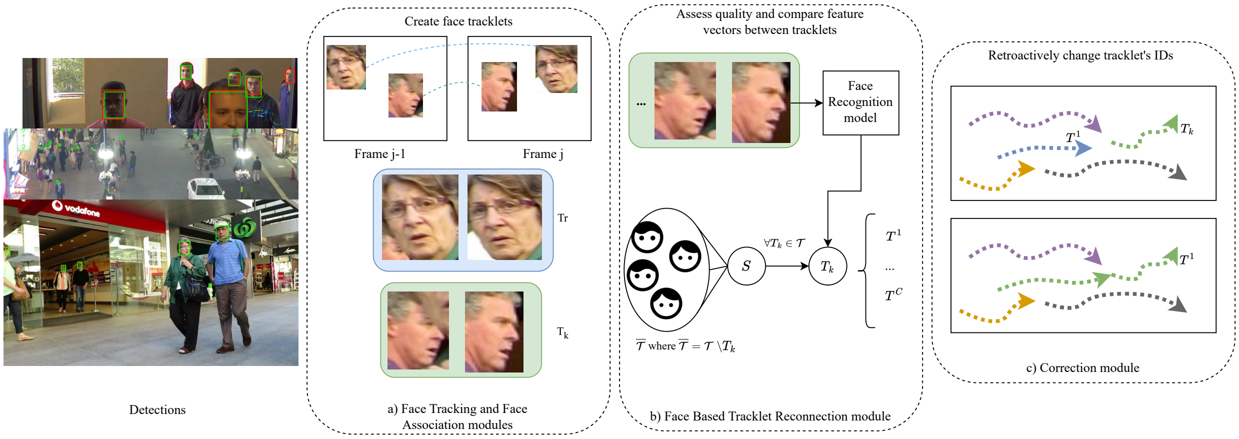

Detections
Create face tracklets
Frame j-1
Frame j
Tr
Tk
a) Face Tracking and Face Association modules
Assess quality and compare feature vectors between tracklets
...
Face Recognition model
S
$\forall T_k \in \mathcal{T}$
T_k
T^1
...
T^C
$\overline{\mathcal{T}}$ where $\overline{\mathcal{T}} = \mathcal{T} \setminus T_k$
b) Face Based Tracklet Reconnection module
Retroactively change tracklet's IDs
T^1
T_k
T^1
c) Correction module

Fig. 1. Illustration of the LTFT tracking pipeline.

for each tracklets T_k, calculate the average of its verifiable face templates, $\overline{V_{T_k}}$.

Let S be a similarity function between two face templates. T^R is defined as the rank-R candidate tracklet, that is the tracklet in position R, after having sorted all candidates from highest to lowest similarity to the tracklet T_k, as described in Equation (1).

$$T^R = \underset{T_i \in \overline{\mathcal{T}} \setminus \{T^j\}_{1 \leq j < R}}{argmax} S(\overline{E_{T_i}}, \overline{V_{T_k}}). \tag{1}$$

For tracklet T_k to be reconnected to T^1 (T^1 being the most similar tracklet to T_k), there are still two conditions to be verified. First, the similarity value between T_k and T^1 must be higher than a threshold λ_{FBTR}, as shown in Equation (2), where $0 \leq \lambda_{FBTR} \leq 1$,

$$S(\overline{E_{T^1}}, \overline{V_{T_k}}) \geq \lambda_{FBTR}. \tag{2}$$

Furthermore, the similarity value between T_k and T^1 must also be higher than the average of the next C-highest ones by a margin of $1/\epsilon$:

$$S(\overline{E_{T^1}}, \overline{V_{T_k}}) \geq \frac{1}{\epsilon} . \frac{1}{C} \sum_{r=2}^{C+1} S(\overline{E_{T^r}}, \overline{V_{T_k}}), \tag{3}$$

where $0 < \epsilon \leq 1$ and $C \in \mathbb{N} \geq 1$.

Subsequently, whenever T^1 verifies both conditions imposed by Equations (2) and (3), the tracklets T_k e T^1 are marked for reconnection. A pair$\langle T_k, T^1 \rangle$ is generated and kept on a list of pairs to be reconnected by the next module.

The fourth module is the correction module that acts upon the pairs associated by the tracklet reconnection module, where for each pair (T_k, T^1) the correction module changes the numeric identifier from T_k's to T^1's, therefore transferring all detections previously assigned to T_k to T^1.

B. Face Detection Models

Tracking-by-detection methods strongly rely upon the chosen object detection model [28]. The LTFT system [5] employs the Faceboxes [26] single shot detection network for face detection.

Due to its initial convolution layers, called rapidly digested convolutional layers, the spatial size of the input image is quickly decreased. Naturally, this leads the model to present difficulties when detecting tiny faces when tested in datasets with high pose and scale variations, such as the hard subset of the WIDER FACE dataset [29].

Therefore, it is fundamental to test other face detection models that may be suitable to replace Faceboxes on the LTFT pipeline and, possibly, obtain better tracking results. Models such as DSFD [30], RetinaFace [31], SRN [32], YOLO5Face [33] and TinaFace [34] were selected for further benchmarking.

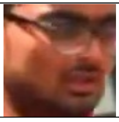
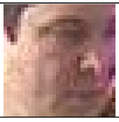
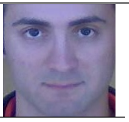
C. The Face Quality Assessment Process

In the LTFT system, three scores are employed to assess the quality of each detected face: detection confidence, head angles and sharpness measure. The detection confidence is a value in the interval of $[0, 1]$, given by the face detection model; the head angles are used to measure pose, being obtained through the use of the head pose estimation model 3DDFA [35] as a set of three values that represent the angular rotation of a face along the three dimensions axis; the sharpness is calculated as the average of the modified Laplacian filter's [36] response. The **Table I** below, illustrates four face images and their respective quality scores.

The employed sharpness measure assigns very small values and with little variation for the majority of detected faces (lesser than 0.1, these values are expected to be distributed between $[0.0, 1.0]$), we argue that this hinders the reconnection process performed by the face-based tracklet reconnection module. Thus, we propose to replace the modified Laplacian filter [36] by the method presented in [4] and described by Equation (4),

$$Sh_{X_i} = avg(abs(X_i - lowpass(X_i))), \tag{4}$$

TABLE I
EXAMPLES OF SCORES UTILIZED FOR FACE QUALITY ASSESSMENT BY THE LTFT SYSTEM.

Face Image				
Confidence	1.0	0.91	1.0	1.0
Angles (in º)	[-32.36, 10.33, -4.34]	[-54.06, 4.95, 2.04]	[-31.02, -2.22, 1.85]	[-3.48, 13.77, -0.5]
Sharpness	0.0125	0.0124	0.0188	0.031

where X_i is the i-th face image detected on video, Sh_{X_i} is the final sharpness score of image X_i, avg is the average of all pixels, abs means absolute value and $lowpass$ is a simple mean filter of kernel size 3x3. The modified Laplacian filter [36] requires the location of fiducial points along the detected face in an attempt to crop the face region before sharpness evaluation. Therefore, this replacement also has the added benefit of not necessitating a model for fiducial points detection.

Another recurring image quality measure is the face image resolution. In [37], bounding boxes with resolutions lower than 32x32 were found to degrade the results of face recognition, additionally [38] artificially lowered face image resolutions and showed a correlation with lower face recognition accuracy. Consequently, we present two modifications to the face quality assessment step: replace the sharpness measure (I) and add resolution as another measure of quality (II).

IV. EVALUATION

Both the baseline and the proposed system's changes are evaluated on two datasets, the LTFT dataset and the proposed dataset expansion. The datasets contain challenging sequences, with multiple faces annotated with their positions and a unique identifier for each identity. More details about each dataset are given in subsections **IV-A** and **IV-B**.

Furthermore, well-known and relevant metrics for multi-object tracking are utilized to validate the contributions of our proposal with respect to accuracy, precision, identification and tracking consistency. These metrics are discussed in subsection **IV-C**.

A. The LTFT Dataset

The LTFT dataset [5], contains 10 semi-automatically annotated videos, that total 8 minutes and 54 seconds, where bounding boxes are assigned to all faces and a unique numeric identifier for each identity. Out of the 10 annotated videos, 8 were obtained from the Youtube platform in various resolutions and the other 2 originate from the Chokepoint dataset [25]. Figure 2 displays 4 frames from the datasets utilized in this work, and Table II details the characteristics of each video sequence.

B. Proposed Dataset Expansion

As conveyed previously, the LTFT dataset was annotated through a semi-automatic process. Specifically, the Faceboxes face detection model [26] was used to obtain bounding boxes. Then, said bounding boxes were verified and annotated with numeric identifiers to represent the trajectory of each face. However, the same Faceboxes model is then used as the face detector on the LTFT system, consequently, it is beneficial to introduce new annotated videos to analyze the system's performance within a less biased scenario.

Three videos were selected from the MOT17 dataset [23] and annotated for multi-face tracking purposes, where the face positions were manually assigned bounding boxes and unique numeric identifiers for each identity. The sequences MOT17-01, MOT17-04 and MOT17-09 were chosen due to their environments being different from one another and the camera staying still throughout the recording, such as surveillance cameras.

Each video's characteristics are detailed in **Table II**, marked by the + symbol. Altogether the proposed expansion contains, by itself, 22575 faces annotated with bounding boxes (denoted as "BBoxes"), 2025 total frames and 81 different identities to be tracked.

TABLE II
CHARACTERISTICS OF THE VIDEO SEQUENCES AND ANNOTATIONS OF THE LTFT DATASET [5] AND THE EXPANSION PROPOSED IN THIS WORK. THE "BBOXES" COLUMN MEANS THE TOTAL NUMBER OF BOUNDING BOXES ANNOTATED IN THAT SEQUENCE AND THE "N° IDS" COLUMN MEANS THE TOTAL NUMBER OF DIFFERENT IDENTITIES ANNOTATED PER VIDEO.

Video	Resolution	Duration	BBoxes	N° IDs
Choke1	800x600	1'24"	7964	24
Choke2	800x600	1'11"	8710	26
Terminal1	1920x1080	1'18"	13722	148
Terminal2	1920x1080	1'15"	11551	140
Terminal3	1920x1080	26"	4255	59
Terminal4	1920x1080	35"	6756	126
Sidewalk	1920x1080	27"	8433	34
Bengal	1920x1080	40"	6953	36
Street	1920x1080	1'8"	4883	31
Shibuya	3840x2160	30"	8058	91
MOT17-09 +	1920x1080	18"	1805	19
MOT17-01 +	1920x1080	15"	3833	14
MOT17-04 +	1920x1080	35"	16937	48

C. Metrics

All the metrics utilized in this work are well known in the literature and used by benchmarks such as the MOT Challenge [23], which enables us to put the tracking results into the context of other works. The threshold for bounding box association in every metric was IoU = 0.5.

We utilize the Clear MOT metrics [39], namely, the Multiple Object Tracking Precision (MOTP) and Multiple Object Tracking Accuracy (MOTA) to measure tracking precision and accuracy, respectively.

To evaluate the assignment of identifiers and how well the tracking system maintains them, the IDF1 [40] identification metric rewards the system based on the amount of time it

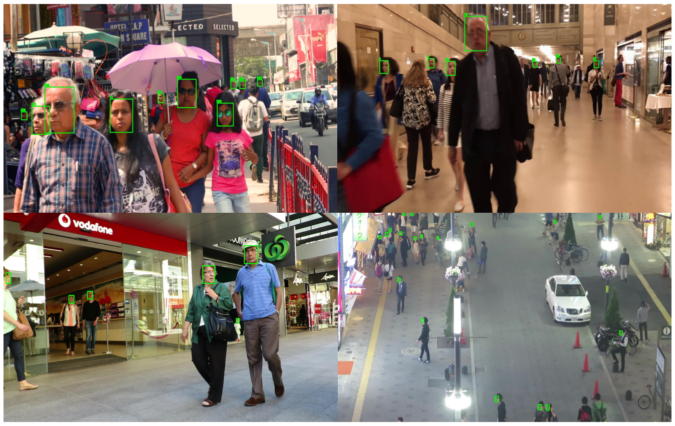


Fig. 2. Four annotated frames of the dataset. From left to right: the first two are from the Bengal and Terminal2 sequences from the LTFT dataset, the last two are from the MOT17-09 where occlusions and faces with different sizes can be seen, and a frame from the Terminal2 sequence that displays occlusions and unaligned faces.

correctly identifies the target. For the reasons stated in [41], IDF1 is considered the more meaningful metric for identity association when tuning parameters.

Mostly Tracked (MT) refers to the number of targets that were tracked for, at minimum, 80% of their total trajectory. ID Switches (IDS) represent the number of times when the tracking system wrongly changes a tracklet's numeric identifier. Other metrics are also used for more fine-grained information, such as False Positives (FP) and False Negatives (FN).

V. Experiments and Results

In this Section we present the results of the changes proposed to the LTFT system [5]. In subsection V-A all chosen face detection models are evaluated to find the most suitable one. Subsection **V-B** displays the results obtained through the proposed changes to the baseline system. All experiments were done through the Google Colab Pro platform, in a machine equipped with an Intel(R) Xeon(R) CPU at 2.00GHz and a NVIDIA Tesla P100 GPU.

A. Comparative Analysis of Face Detection Models

This experiment compares different face detection models to find the best candidate to replace the Faceboxes [26] model originally used by the baseline. The chosen models were: DSFD [30], RetinaFace [31], SRN [32], YOLO5Face [33] and TinaFace [34].

The AP$_{.50}$ (*Average Precision at IoU = .50*) is a commonly used metric for the evaluation of object detection models, that represents the area under the precision-recall curve. In the AP$_{.50}$ case, for a detection to be considered successful, there must be an IoU of at least 50% between the predicted bounding box and the ground-truth bounding box.

All models were evaluated on the same dataset. Both the LTFT dataset and the proposed dataset were united as a single face detection dataset. The results are presented in **Table III**.

TABLE III
Scores obtained by the face detection models tested on the union of the LTFT dataset with the proposed dataset. Results are shown using the Average Precision metric with an IoU threshold of 0.5 and Frames per Second (FPS) to represent each model's speed.

Model	AP$_{.50}$	FPS
SRN	**0.786**	0.9
YOLO5Face(s)	0.767	14.5
YOLO5Face(n)	0.764	**15.3**
RetinaFace	0.731	5.3
DSFD	0.696	1.0
TinaFace	0.692	3.0
Faceboxes	0.689	14.9

The best AP$_{.50}$ result belongs to the SRN model [32] with an AP$_{.50}$ of 0.786, however, it lacks speed when compared to the baseline's detector, Faceboxes. Since processing speed is of importance given the nature of CCTV systems, a viable replacement must surpass the detection performance, while being as fast as, or faster, than the Faceboxes model.

Conversely, both YOLO5Face [33] model variations display AP$_{.50}$ results of approximately 8 percentage points above Faceboxes, all the while offering similar inference speeds, or higher in the YOLOV5Face(n) model's case.

Among the models that surpass Faceboxes on the AP$_{.50}$ metric, the RetinaFace [31] model displays an AP$_{.50}$ equal to 0.731, approximately 4.2 percentage points above Faceboxes. However, its speed is 2.8x slower, thus its choice is not justified as a replacement.

Given the previous face detection results, the model chosen to replace Faceboxes in the LTFT system's pipeline was the YOLO5Face(s) model, due to its superior detection metric and similar inference speed.

B. Comparative Study

This section presents the proposed experiments and quantifies their results. The five experiments are described below. We also compare our results against the SORT [42] tracker, which was ranked the best open-source multi-object tracker at the time of its publication.

- (I) LTFT: the LTFT system in its baseline state, using the Faceboxes model as face detector, original parameters and with $T_{max} = 5$;
- (II) LTFT+thresholds: same as the previous baseline experiment, this time changing the sharpness measure's thresholds, where $n_E = 0.06$ and $n_V = 0.04$;
- (III) LTFT+Yolo5s: this experiment evaluates the contribution of changing the face detection model, where $n_E = 0.06$, $n_V = 0.04$ and $T_{max} = 5$.
- (IV) LTFT+Yolo5s+FQA: in addition to replacing the face detection model, the modified Laplacian filter by [36] is replaced by the face sharpness measure presented by [4] and the face image resolution is added to the FBTR module as an extra face quality assessment measure.
- SORT: the SORT tracker applied with the Yolo5s detections.

For experiments LTFT+Yolo5s and LTFT+Yolo5s+FQA, parameters like detection threshold (λ_{det}), enrollable face confidence threshold (d_E), verifiable face confidence threshold (d_V), enrollable face resolution (res_E), verifiable face resolution (res_V), enrollable face sharpness score (n_E) and verifiable face sharpness score (n_V) were explored for 250 iterations at random in specific intervals.

Table IV shows each experiment's results when evaluated in both datasets together, that is, the union of the LTFT dataset with the additional dataset proposed in this work.

TABLE IV
RESULTS OF EXPERIMENTS CONDUCTED ON THE UNION OF BOTH DATASETS. ↑ MEANS HIGHER IS BETTER AND ↓ MEANS LOWER IS BETTER. THE BEST RESULTS IN EACH COLUMN ARE SHOWN IN BOLD.

Experiment	IDF1↑	MOTA↑	MOTP↑	MT↑	FP↓	FN↓	IDS↓
I (Baseline)	38.1%	33.9%	68.6%	452	28220	37746	2634
II	44.5%	32.5%	68.6%	435	27998	39566	2503
III	44.8%	33.2%	**78.7%**	483	27649	39650	2051
IV (Ours)	**47.1%**	33.5%	**78.7%**	492	**27647**	39356	**2038**
SORT	42.3%	**36.0%**	75.0%	**615**	40806	**26686**	2136

The utilization of the YOLO5Face [33] model for face detection raises the MOTP metric by 10.1% (from 68.6% to 78.7%), as well as gradually improving FP, IDS, MT and IDF1, where both experiments that utilize YOLO5Face present the best results for these metrics, besides MT which is lead by the SORT tracker.

The LTFT+Yolo5s+FQA experiment attests to the efficacy of the proposed changes to the face quality assessment step. It displays the best results of IDF1, MOTP, FP and IDS. These metrics indicate that the modifications improve the length of each track with fewer identity switches when compared to the baseline. Nonetheless, the SORT tracker's data association provides better accuracy and a greater MT value without the need to reconnect lost tracklets.

Figure 3 shows detected faces on the experiment LTFT+Yolo5s+FQA separated by their quality into one of three groups. The first and second row, represent enrollable faces and verifiable faces, respectively. It is expected that these face images display satisfactory sharpness, resolution and pose orientation for the task of face recognition and tracklet reconnection. The last row shows discarded faces, whose quality is deemed inadequate and its use may add noise and interfere with the recognition process.

Table V reports the experiments only on the LTFT dataset. It shows that the experiments which utilize Faceboxes show the best results for accuracy metrics, MOTA, FP and FN. This is expected since Faceboxes was also utilized in the labeling of this dataset. However, one can argue that the difference of 2.4% in MOTA is overcome by an increase of 10% in MOTP and 10.3% in IDF1.

VI. Conclusion

This dissertation presented a study of face selection and its benefits for video face recognition. The LTFT system and dataset [5] were further expanded upon. Through simple changes such as a new face detection model, a simpler approach to quality assessment, parameter tuning and multi-face tracking dataset enrichment, we were able to achieve better tracking and recognition capabilities, where faces are tracked for longer, and fragmented tracklets are more easily reconnected.

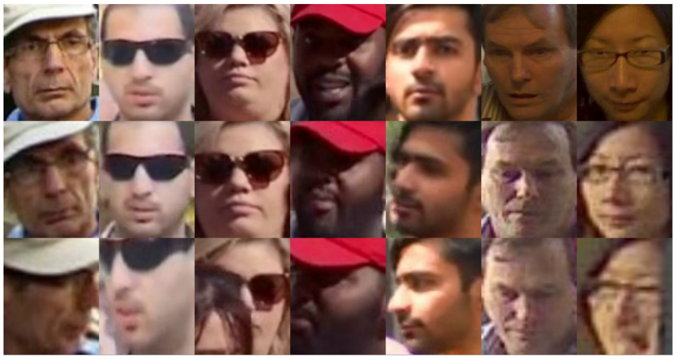


Fig. 3. Faces detected and separated by quality on the LTFT+Yolo5s+FQA experiment. The first, second and third row represent, respectively, enrollable, verifiable and discarded faces. For means of presentation, all detections were resized to the same resolution.

TABLE V
EXPERIMENTS TESTED ON THE LTFT DATASET [5]. ALL VIDEOS WERE SEMI-AUTOMATICALLY ANNOTATED BY FIRST EXTRACTING FACE BOUNDING BOXES USING THE FACEBOXES [26] MODEL. ↑ MEANS HIGHER IS BETTER AND ↓ MEANS LOWER IS BETTER. THE BEST RESULTS OF EACH COLUMN ARE SHOWN IN BOLD.

Experiment	IDF1↑	MOTA↑	MOTP↑	MT↑	FP↓	FN↓	IDS↓
I (Baseline)	42.0%	**42.5%**	68.8%	445	26643	17577	2547
II	49.4%	40.7%	68.7%	428	**26421**	19397	2416
III	49.7%	39.7%	**78.8%**	472	27093	19973	1940
IV (Ours)	**52.3%**	40.1%	**78.8%**	481	27091	19679	1927
SORT	44.5%	37.6%	77.0%	**593**	35145	**13814**	**1786**

For future works, overall system complexity can be reduced since the current proposition depends on multiple deep models. There's a need to unify all the face selection pipeline's steps into a single model. A suggestion is the use of Long Short Term Memory (LSTM) networks, seeing as these architectures are capable of comprehending the temporal aspect of data, inherent to video recordings.

Acknowledgment

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001.

References

[1] R. V. Clarke, "Situational crime prevention: Its theoretical basis and practical scope," *Crime and Justice*, vol. 4, pp. 225–256, 1983. [Online]. Available: https://doi.org/10.1086/449090

[2] M. P. J. Ashby, "The value of cctv surveillance cameras as an investigative tool: An empirical analysis," *European Journal on Criminal Policy and Research*, vol. 23, no. 3, pp. 441–459, Sep 2017. [Online]. Available: https://doi.org/10.1007/s10610-017-9341-6

[3] E. Piza, B. Welsh, D. Farrington, and A. Thomas, "Cctv surveillance for crime prevention: A 40-year systematic review with meta-analysis," *Criminology & Public Policy*, vol. 18, pp. 135–159, 03 2019.

[4] K. Nasrollahi and T. B. Moeslund, “Extracting a good quality frontal face image from a low-resolution video sequence,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 10, pp. 1353–1362, 2011.

[5] G. Barquero, I. Hupont, and C. Fernández Tena, “Rank-based verification for long-term face tracking in crowded scenes,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 495–505, 2021.

[6] B. Cássio, “Ltft-implementation,” https://github.com/Cassio-4/LTFT-Implementation, 2022.

[7] P. Viola and M. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, pp. 137–154, 05 2004.

[8] B. F. Momin and Y. Jere, “Mining visitors in video surveillance system,” in *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2015, pp. 1–4.

[9] S. Vignesh, K. M. Priya, and S. S. Channappayya, “Face image quality assessment for face selection in surveillance video using convolutional neural networks,” in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2015, pp. 577–581.

[10] K. Nasrollahi and T. B. Moeslund, “Hybrid super resolution using refined face logs,” in *2010 2nd International Conference on Image Processing Theory, Tools and Applications*, 2010, pp. 435–440.

[11] P. Barra, S. Barra, C. Bisogni, M. De Marsico, and M. Nappi, “Web-shaped model for head pose estimation: An approach for best exemplar selection,” *IEEE Transactions on Image Processing*, vol. 29, pp. 5457–5468, 2020.

[12] A. Del Bimbo, F. Dini, and G. Lisanti, “A real time solution for face logging,” in *3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009)*, 2009, pp. 1–6.

[13] J. Zheng, R. Ranjan, C.-H. Chen, J.-C. Chen, C. D. Castillo, and R. Chellappa, “An automatic system for unconstrained video-based face recognition,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 3, pp. 194–209, 2020.

[14] Y. Cai and H. Gan, “An online face clustering algorithm for face monitoring and retrieval in real-time videos,” in *2019 IEEE Intl Conf on Parallel Distributed Processing with Applications, Big Data Cloud Computing, Sustainable Computing Communications, Social Computing Networking (ISPA/BDCloud/SocialCom/SustainCom)*, 2019, pp. 825–830.

[15] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-speed tracking with kernelized correlation filters,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[16] C.-C. Lin and Y. Hung, “A prior-less method for multi-face tracking in unconstrained videos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] L. Patino, T. Nawaz, T. Cane, and J. Ferryman, “Pets 2017: Dataset and challenge,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 2126–2132.

[18] R. Fisher, “Caviar: Context aware vision using image-based active recognition,” http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/, 2003, acessado em 15/02/2021.

[19] i-LIDS Team, “Imagery library for intelligent detection systems (i-lids); a standard for testing video based detection systems,” in *Proceedings 40th Annual 2006 International Carnahan Conference on Security Technology*, 2006, pp. 75–80.

[20] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, “A large-scale benchmark dataset for event recognition in surveillance video,” in *CVPR 2011*, 2011, pp. 3153–3160.

[21] J. Pers and D. R. Magee, “CVBASE ’06 - Workshop on Computer Vision Based Analysis in Sport Environments,” http://vision.fe.uni-lj.si/cvbase06/, 2006, acessado em 21/03/2021.

[22] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Cehovin, G. Nebehay, G. Fernandez, T. Vojir, A. Gatt, A. Khajenezhad, A. Salahledin, A. Soltani-Farani, A. Zarezade, A. Petrosino, A. Milton, B. Bozorgtabar, B. Li, C. S. Chan, C. Heng, D. Ward, D. Kearney, D. Monekosso, H. C. Karaimer, H. R. Rabiee, J. Zhu, J. Gao, J. Xiao, J. Zhang, J. Xing, K. Huang, K. Lebeda, L. Cao, M. E. Maresca, M. K. Lim, M. El Helw, M. Felsberg, P. Remagnino, R. Bowden, R. Goecke, R. Stolkin, S. Y. Lim, S. Maher, S. Poullot, S. Wong, S. Satoh, W. Chen, W. Hu, X. Zhang, Y. Li, and Z. Niu, “The visual object tracking vot2013 challenge results,” in *2013 IEEE International Conference on Computer Vision Workshops*, 2013, pp. 98–111.

[23] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, “MOT16: A benchmark for multi-object tracking,” *arXiv:1603.00831 [cs]*, Mar. 2016, arXiv: 1603.00831. [Online]. Available: http://arxiv.org/abs/1603.00831

[24] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, “The first facial landmark tracking in-the-wild challenge: Benchmark and results,” in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015, pp. 1003–1011.

[25] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, “Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition,” in *IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, June 2011, pp. 81–88.

[26] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, “Faceboxes: A cpu real-time face detector with high accuracy,” in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 1–9.

[27] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109

[28] Y. Park, L. M. Dang, S. Lee, D. Han, and H. Moon, “Multiple object tracking in deep learning approaches: A survey,” *Electronics*, vol. 10, no. 19, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/19/2406

[29] S. Yang, P. Luo, C. C. Loy, and X. Tang, “WIDER FACE: A face detection benchmark,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, “Dsfd: Dual shot face detector,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5055–5064.

[31] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-shot multi-level face localisation in the wild,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5202–5211.

[32] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Li, and X. Zou, “Selective refinement network for high performance face detection,” in *Association for the Advancement of Artificial Intelligence (AAAI)*, 07 2019.

[33] D. Qi, W. Tan, Q. Yao, and J. Liu, “Yolo5face: Why reinventing a face detector,” 2021. [Online]. Available: https://arxiv.org/abs/2105.12931

[34] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong, “Tinaface: Strong but simple baseline for face detection,” 2020. [Online]. Available: https://arxiv.org/abs/2011.13183

[35] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, “Face alignment in full pose range: A 3d total solution,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 78–92, 2019.

[36] Nikitin, A. Konushin, and V. Konushin, “Face quality assessment for face verification in video,” in *The 24th International Conference on Computer Graphics and Vision (GraphiCon2014)*, 2014, pp. 111–114.

[37] B. Boom, G. Beumer, L. Spreeuwers, and R. N. J. Veldhuis, “The effect of image resolution on the performance of a face recognition system,” in *2006 9th International Conference on Control, Automation, Robotics and Vision*, 2006, pp. 1–6.

[38] 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, Jun 2015. [Online]. Available: https://doi.org/10.1007/s11042-013-1568-8

[39] K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: The clear mot metrics,” *EURASIP Journal on Image and Video Processing*, vol. 2008, 01 2008.

[40] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, pp. 17–35.

[41] R. Henschel, T. von Marcard, and B. Rosenhahn, “Simultaneous identification and tracking of multiple people using video and imus,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 780–789.

[42] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3464–3468.