Adaptive Face Tracking Based on Online Learning

Aasim Khurshid
Instituto de informatica, UFRGS, Brazil, and
SIDIA Instituto de Ciencia e Tecnologia (SIDIA), Manaus, Brazil
Emails: aasim.khurshid@sidia.com

Jacob Scharcanski
Instituto de informatica,
Universidade Federal do Rio Grande do Sul, Brazil
Emails: jacobs@inf.ufrgs.br

***Abstract*—Object tracking can be used to localize objects in scenes, and also can be used for locating changes in the object's appearance or shape over time. Most of the available object tracking methods tend to perform satisfactorily in controlled environments but tend to fail when the objects appearance or shape changes, or even when the illumination changes (e.g., when tracking non-rigid objects such as a human face). Also, in many available tracking methods, the tracking error tends to increase indefinitely when the target is missed. Therefore, tracking the target objects in long and uninterrupted video sequences tends to be quite challenging for these methods. This work proposes a face tracking algorithm that contains two operating modes. Both the operating modes are based on feature learning techniques that utilize the useful data accumulated during the face tracking and implements an incremental learning framework. To accumulate the training data, the quality of the test sample is checked before its utilization in the incremental and online training scheme. Furthermore, a novel error prediction scheme is proposed that is capable of estimating the tracking error during the execution of the tracking algorithm. Furthermore, an improvement in the Constrained Local Model (CLM), called weighted-CLM (W-CLM) is proposed that utilize the training data to assign weights to the landmarks based on their consistency. These weights are used in the CLM search process to improve CLM search optimization process. The experimental results show that the proposed tracking method (both variants) perform better than the comparative state of the art methods in terms of Root Mean Squared Error (RMSE) and Center Location Error (CLE). In order to prove the efficiency of the proposed techniques, an application in yawning detection is presented.** [1]

Keywords: Face Tracking, Facial landmarks tracking, Incremental Learning, Dictionary Learning, Tracking error predictor, Yawning detection.

I. Introduction

Object tracking essentially deals with locating, identifying, and determining the dynamics of the moving (possibly deformable) target(s). The target(s) could be a single object or parts of an object. In fact, object tracking may become quite challenging when there are changes in the appearance or shape of the target, when the scene illumination changes, temporary occlusions and/or tracking conditions are altered in time. Similarly, noise and different lighting conditions during the day may affect the local illumination in various ways [1]. Numerous algorithms have been proposed in the literature for object tracking in video sequences such as incremental learning for robust visual tracking [2], Multiple Instance Learning (MIL) discriminative classifier based tracking [3], Appearance and shape models for face detection and facial landmark tracking [4]–[6], and Continuously Adaptive Mean Shift (CAMShift) tracker [7]. However, most methods available in the literature tend to perform well over short time spans and under controlled conditions. Furthermore, in most of these methods, when the object tracking method misses the target, the tracking error tends to increase indefinitely. This work proposes to minimize this difficulty by using online learning scheme that utilizes the data received during tracking to update the appearance model of the object (i.e., face). The appearance model is updated after checking the quality of the tracked target object samples before utilizing this sample to update the appearance. Also, a resyncing scheme is introduced that corrects the tracking process once the tracking error is estimated to be high.

This work proposes a face tracking method which contains two operating modes: Multi-Model Dictionary Learning Face Tracking with dictionaries Update (MMDL-FTU) and Adaptive Face Tracking using Resyncing Mechanism with Weighted CLM search (AFTRM-W). The proposed tracking algorithm has two components, which are the motion model and the appearance model. **Motion model** is responsible for handling the motion parameters of the face and estimation of the candidate target face samples. **Appearance model** is utilized to estimate the tracked target face among the candidate target face samples.

MMDL-FTU operating mode models the appearance of the tracked target using novel incremental learning of a multi-model K-Singular Value Decomposition (K-SVD) dictionary [8]. This method performs well; however, in complex scenarios like a rapid movement of the face, it tends to fail.

For this reason, the AFTRM-W operating mode adds another component to make the tracking process robust with an additional cost of time. AFTRM-W uses incremental Singular Value Decomposition (SVD) as an appearance model that increases the speed at the cost of tracking quality. Furthermore, in this tracking mode, a resyncing scheme is used to improve the tracking process when the proposed tracking predictor indicates high tracking error. A weighted Constrained Local Model (W-CLM) scheme is proposed as a resyncing scheme that improves the tracking performance.

Face tracking based on facial features is relevant for a number of applications, such as yawning detection, expression analysis, human computer interfaces, and face recognition [9]–[12]. Furthermore, image-based measurements can provide

[1] Aasim khurshid, PhD thesis

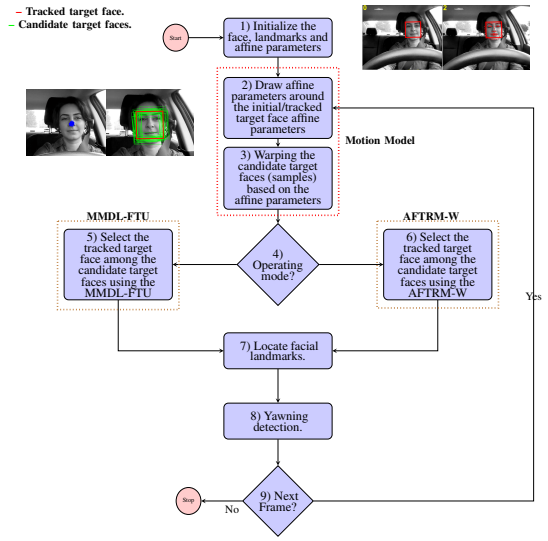

Fig. 1: Block diagram of the proposed face tracking method.

cost effective solutions for fatigue and vigilance systems if the detected facial features are accurate [13]. In our experiments, the proposed method is evaluated in a face tracking application: yawning detection in the context of a driving scenario.

A. Contributions

The main contributions of the proposed method include:

- An incremental Multi-Model Dictionary Learning (MMDL), which combines two dictionaries (a classification and a reconstruction dictionary) and MMDL is used for face tracking which tends to improve tracking robustness.
- A smart approach to update MMDL incrementally, such that our approach updates dictionaries efficiently and reduces tracking error.
- An error prediction scheme to evaluate the correctness of the tracking process during face tracking.
- Utilization of a resyncing mechanism based on CLM.
- An improvement in the classical CLM approach, so-called Weighted CLM (W-CLM) is proposed.
- An improvement in an application of the facial analysis (i.e., yawning detection).
- An adaptive mean of the tracked target face is proposed and plays an important role in the current face tracking. The mean face $\mu(t)$ at time t is updated as follows and employs a forgetting factor f as: $\mu(t) = \frac{f \cdot n \cdot \mu_n + m \cdot \mu_m}{m + f \cdot n}$, where $\mu(t)$ is the updated mean, μ_n represents the mean of the older data (X_n), μ_m is the mean of the newly added observations (X_m) and $t = m + n$.

II. Proposed Method

Figure 1 shows the block diagram of the proposed tracking method and is explained below:

- Block 1: In the first frame, the initial target face, the affine parameters ($\chi(t)$) and the facial landmarks are provided by a face landmark localization method [5];
- Block 2: In the subsequent frames, a finite number (η) of affine parameters are drawn around the affine parameters of the initial/tracked target face using a Gaussian distribution (see Eq. 2);
- Block 3: To locate the tracked target face in the frame at time t, the candidate target face samples (u $\times$ u) are warped according to the computed affine parameters to be compared with the tracked target face (see the example in Fig. 1 at the left of Block 2: in red, the tracked target face from the previous frame; in green, candidate target face samples). See details in Section II-A;
- Block 4-6: Among the candidate target face samples, the tracked target face is selected using one of the operating modes, i.e., MMDL-FTU or AFTRM-W operating mode. (II-B and II-C);
- Block 7: The facial landmarks of the tracked target face are located (Section II-B1);
- Block 8: Yawning is detected. (III-A for details);
- Block 9: Finally, if there are more frames to process the affine parameters of the current tracked target face are used in the next frames, and the process re-starts from Block 2.

Both the operating modes of the tracking scheme shares the same motion model, explained in Section II-A, and differs by how they model the appearance of the tracked target face, explained in Sections II-B and II-C.

A. Motion Model and Sampling

For face tracking, state of the tracked target face is described by an affine parameters variable $\chi(t)$ which describes the location of the face at time t. Furthermore, $\chi(t)$ of the tracked target face are used to estimate the face landmarks, detect yawning and calculate the tracking error (see Eq. 4). For a set of tracked target face samples at time t, $\mathcal{I}(t)$=$\{\mathbf{I}(1),\mathbf{I}(2)...,\mathbf{I}(T)\}$, the face tracker estimates the hidden state variable $\chi(t)$ using:

$$p(\chi(t)|\mathcal{I}(t)) \propto p(\mathbf{I}(t)|\chi(t)) \times \int p(\chi(t)|\chi(t-1)) p(\chi(t-1)|\mathcal{I}(t-1)) d\chi(t-1). \quad (1)$$

The candidate target face samples that may contain the tracked target face are sampled following the motion model between two states $p(\chi(t) \mid \chi(t-1))$, assuming a Gaussian distribution around the tracked target face location in the previous frame. At time t, the state of the target face in a video sequence is described by the affine parameters $\chi(t)$=$(x(t), y(t), s(t), \theta(t), \beta(t), \phi(t))$, where $x(t)$ and $y(t)$ represent the translation, $s(t)$ is the scale, whereas $\theta(t)$, $\alpha(t)$ and $\phi(t)$ are the rotation angle w.r.t the horizontal axis, the aspect ratio, and the skew direction of the tracked target face, respectively. The dynamics of each parameter in $\chi(t)$ is modeled independently by a Gaussian distribution centered at $\chi(t-1)$, and going from $\chi(t-1)$ to $\chi(t)$ is given by (for details on Gaussian distribution, see Chapter 3 in thesis [14]):

$$p(\chi(t)|\chi(t-1)) = \mathcal{N}(\chi(t); \chi(t-1), \psi(t)), \quad (2)$$

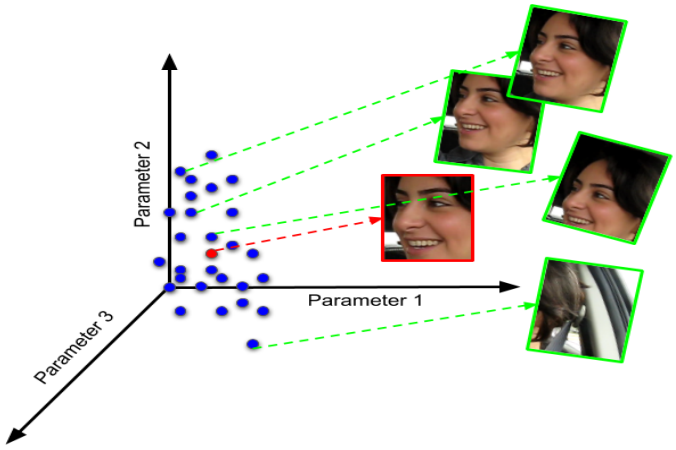

Fig. 2: Motion model example ($p(\chi(t)|\chi(t-1))$) in image space.

where $\psi(t)$ is a diagonal matrix with each element representing the variance of its corresponding affine parameters element, and $\mathcal{N}$ represents a Gaussian distribution. These affine parameters are used to warp the candidate target face samples that may contain a face in the current frame. These candidate target face samples are tested for quality using the appearance model, and one of them is selected as the tracked target face in the current frame at time t using the techniques explained in Sections II-B and II-C.

Figure 2 show an example of how the motion model works. The affine parameters $\chi(t)$ are represented by a point in affine parameter space; the affine parameter space is a six-dimensional space, and only three dimensions are shown in figure 2. The red point in the figure 2 represent the affine parameters of the tracked target face in the previous frame. Numerous affine parameters are computed using the Gaussian distribution centered around the affine parameters associated with the tracked target face in the previous frame using eq. 2, and these affine parameters are shown as blue points in figure 2. Furthermore, these affine parameters are used to warp the candidate target face samples which may contain the tracked target face in the current frame, shown in green color faces in figure 2. One of these candidate target faces is selected to be the tracked target face by using MMDL-FTU operating mode or AFTRM-W operating mode.

B. Multi-Model Dictionary Learning for Face Tracking (MMDL-FT) and MMDL-FT with Update Test (MMDL-FTU) Operating Mode

In MMDL-FTU operating mode, two techniques are used for the training data collection to build and update the dictionaries incrementally. In the MMDL-Face Tracker (MMDL-FT), the dictionaries are updated using the tracked target face samples collected without checking their quality, as proposed by Ross et al. [2]. In the MMDL-Face Tracker with dictionaries Update (MMDL-FTU), only those tracked target face samples are collected which has reconstruction error smaller than a specific threshold ε.

Dictionary learning has been explored in object tracking, but the proposed dictionary learning methods usually are based on static dictionaries that are not updated during object tracking [15]. Most of the methods that use dictionaries for object tracking are focused on the representation of the target [16], or on the discrimination between the target and the background [17]. In this work, a new approach called Multi-Model Dictionary Learning (MMDL) is proposed for face tracking that builds and updates in parallel two dictionaries, i.e., a reconstruction dictionary (D_p) and a classification dictionary (D_c).

The **reconstruction dictionary** (D_p) is used to estimate the appearance difference between the reconstructed sample using $\varepsilon_r = ||\mathbf{I}_c - D_p\alpha_j||_2^2$, where $\mathbf{I}_c$ is the patch matrix of the candidate target face samples, and α_j are the D_p sparse coefficients.

The **classification dictionary** (D_c) is utilized to discriminate the candidate target face from the background. The classification error is given by $\varepsilon_c = ||Y_i - W\alpha_i||^2$ where $Y_i \in [0\ 1]$ is the label indicator, and $W \in \mathbb{R}_{\times}\neg$ is the linear classification parameters learned with a labeled dictionary.

These two dictionaries are combined into a single multi-model, which tends to improve the tracking robustness. The proposed method learns the face appearance using dictionary atoms constructed from patches, that are taken from positive and negative samples of the training data. Furthermore, a smart approach is proposed to update incrementally and efficiently the dictionaries using SVD, making the application of our method to realistic tracking scenarios feasible [14]. Furthermore, the proposed method collects training samples to update the two dictionaries during face tracking using a proposed scheme [8], [14]. The quality of the samples (i.e., reconstruction error) is assessed before utilizing them to update the dictionaries, which is an aspect that other methods that implement incremental learning seem to miss [2]. Both the dictionaries are initialized using Singular Value Decomposition (SVD), which is more efficient than initializing the process by combining some random training samples as proposed by Elad et al. [18]. For details on both the dictionaries update and the pseudo code, please refer to Chapter 4 and Section 4.2 of the thesis [14].

The candidate target face samples are sampled using the motion model explained in Section II-A. To obtain the combined probability $p(\mathbf{I}(t) \mid \chi(t))$ of the candidate target face to be the tracked target face, the reconstruction and classification probabilities are combined as follows:

$$p(\mathbf{I}(t)|X(t)) = \Omega p_r(\mathbf{I}(t)|\chi(t)) + (1-\Omega)p_c(\mathbf{I}(t)|\chi(t). \quad (3)$$

where, $p_r(\mathbf{I}(t)|\chi(t))$ is the reconstruction dictionary (D_p) probability and $p_c(\mathbf{I}(t)|\chi(t)$ is the classification dictionary (D_c) probability of the candidate target face to be the tracked target face, whereas Ω is a weight associated to the classification and reconstruction dictionaries, and indicates the trade-off between reconstruction and classification dictionaries probability of the candidate to be the tracked target face.

1) Facial Landmarks Localization: The candidate face sample that has higher combined probability is selected to be the tracked target face, and the associated affine parameters $\chi(t)$ are used to estimate landmarks on the tracked target face:

$$\Lambda_T(t) = \chi(t) \times [\Lambda(1); \vec{1}], \tag{4}$$

where, $\Lambda(1)$ are the landmark locations in the initial target face and $\vec{1}$ is an unitary vector of length Z (total number of landmarks). This tracked target face is used to update the two dictionaries depending on the reconstruction error (ε_r).

C. Adaptive Face Tracker with Resyncing Mechanism (AFTRM) and AFTRM Weighted (AFTRM-W) Operating Mode

The proposed approach improves on a well-known object tracking method based on the incremental PCA [2]. The proposed scheme learns from the data generated during face tracking and corrects the tracking mistakes with a resyncing mechanism. Also, a dynamic tracking error predictor is proposed to estimate how accurately the target face is being tracked. Furthermore, the tracking error predictor adapts itself in time and tends to be consistent in long video sequences (see Section. II-C1). Consequently, if the estimated tracking error is increasing, the tracking process is corrected by a resyncing mechanism based on CLM. In addition, it is also proposed an improvement of CLM named Weighted CLM (W-CLM) that utilizes the training data to assign a weight to each landmark (feature point) based on its consistency in time. One of the possible applications of the proposed tracking method is the face and facial landmarks tracking, where Constrained Local Models (CLM) or Weighted CLM (W-CLM) can be used to re-adjust the facial features locations (landmarks) when there is a potential tracking failure. AFTRM variant of this operating mode uses classic CLM as a resyncing mechanism, whereas, AFTRM-W utilizes W-CLM to resync important features. For details on this, please refer to Chapter 4 and Section 4.3 in the thesis [14].

The proposed methodology models the appearance of the tracked target face using a probabilistic PCA. A candidate target face sample $\mathbf{I}(t)$ that is warped using the affine parameters $\chi(t)$ is assumed to be generated from the subspace of the target face spanned by the eigenbases U and centered at the mean $\mu(t)$. The probability p of a candidate target face being generated from this subspace is inversely proportional to its distance δ from the reference point (i.e., mean ($\mu(t)$)) of the subspace. This distance is comprised of the distance to the subspace (δt) and within the subspace distance (δw) of the projected sample to the subspace center ($\mu(t)$). The likelihood of a candidate target face sample being the tracked target face is given by the combined probability of its distance from the subspace $p_{\delta t}$ and within space distance $p_{\delta w}$:

$$\begin{aligned} p(\mathbf{I}(t)|\chi(t)) &= p_{\delta t}(\mathbf{I}(t)|\chi(t)) p_{\delta w}(\mathbf{I}(t)|\chi(t)) \\ &= \mathcal{N}(\mathbf{I}(t); UU^T + \epsilon I)\mathcal{N}(\mathbf{I}(t); \mu, U\Sigma^{-2}U^T), \end{aligned} \tag{5}$$

where Σ is the singular value matrix, $p_{\delta t}(\mathbf{I}(t)|\chi(t)) = exp(-||(\mathbf{I}(t) - \mu(t)) - UU^T(\mathbf{I}(t) - \mu(t))||^2)$ and $p_{\delta w}(\mathbf{I}(t)|\chi(t)) = exp(-||(\mathbf{I}(t) - \mu(t))^T U\Sigma^{-2}U^T(\mathbf{I}(t) - \mu(t))||^2)$. The candidate target face sample that has the highest probability of being the tracked target face is selected and its associated affine parameters $\chi(t)$ are used to estimate the facial landmarks using Eq. 4.

1) Tracking Error Predictor and Resyncing Mechanism: Visual tracking is prone to failure if the object changes, does a quick motion or changes appearance, and so on. Therefore, often tracking methods fail, and the tracking error keeps on increasing, and the tracking fails indefinitely. Most of these methods fail to provide a self assessment of tracking [2], [3], [19]–[21]. The proposed method is based on an error predictor which tries to estimate the tracking error at runtime. It was found in the experiments that a relevant measure to predict the tracking error is the tracking difference in the landmarks (feature points) represented by ($\Delta(t)$) at time t. This is verified using the correlation (ρ) with the tracking error (ε), and $\Delta(t)$ at time t is given by:

$$\Delta(t) = \frac{1}{Z}\sum_{i=1}^{Z} ||\Lambda_T^{(i)}(t) - \Lambda_T^{(i)}(t-1)||^2, \tag{6}$$

where $\Lambda_T^{(i)}(t)$ is the location (x, y) of the landmark i at time t estimated by the proposed method. To further improve the tracking error prediction, median filter is applied to the $\Delta(t)$ noisy estimates.

In the next stage, the tracking error is predicted if the value of $\Delta(t)$ in Eq. 6 is higher than a certain threshold. A dynamic threshold Γ_T ($\Gamma_T = Median(\Delta(T))$) is proposed which can auto-adjust to different environments. Resyncing flag $\Psi(t)$ is used to indicate if resyncing is required and is computed as:

$$\Psi(t) = \begin{cases} 1, & \text{if } \Delta(t) \geq \Gamma_T, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where $\Delta(T) = \{\Delta(1), ..., \Delta(t)\}$ (see details in Chapter. 4 and 5 of the thesis [14]). When the tracking predictor indicates a higher tracking error, the resyncing of the features using W-CLM is called to correct the tracking process by re-adjusting the tracked landmarks.

2) CLM Weighted Search: The W-CLM search process combines the shape and patch models to detect the facial landmarks of a face. For details on training shape and patch model, refer to [14]. Given a set of initial facial landmarks, the cropped patch around the current position of each landmark is processed by the SVM based patch model, while preserving the shape constraints. Both these goals are combined using the following objective function with the corresponding weights of the landmarks as:

$$f(S_t) = \sum_{i=1}^{Z} \hat{w}_i \imath_i(x_i, y_i) - \beta \sum_{j=1}^{o} \frac{-h_j^2}{\lambda_j}, \tag{8}$$

where $\imath_i(x_i, y_i)$ is the shape model response of c^{th} feature template and $\hat{w}_c$ represents the weight of the landmark i. The weight $\hat{w}_i$ describes how much effect this particular landmark will have in the fitting process. The term $\sum_{j=1}^{o} \frac{-h_j^2}{\lambda_j}$ is the

shape constraint, whereas the parameter $\beta \in [0\ 1]$ is a bias determining the compromise between shape fit and the SVM based patch model.

III. Experimental Evaluation

The proposed tracking algorithms were implemented in Matlab 2015a on an IBM PC compatible with 3.40GHz i7-6700 CPU with 16GB internal memory. For experimental evaluation, the YawDD dataset [22], which contains videos of 119 participants who belongs to different race and color of all ages, performing various facial expressions such as normal, talking and yawning and in various illumination conditions.

The proposed face tracking algorithms are quantitatively evaluated using Center Location Error (CLE), that measures the distance between center locations of the tracked target face with the manually labeled center location of the target face that is used as the groundtruth. Furthermore, for detailed evaluation, six videos have been annotated manually, which includes the target face and landmarks ($Z = 68$) on the face, nose and the eyes. The error was measured by the root mean squared error (RMSE) between the estimated landmark locations (Λ_T) and the manually-labeled groundtruth (Λ_G) locations of the landmarks as follows:

$$\varepsilon(t) = \frac{1}{Z}\sum_{i=1}^{Z}||\Lambda_G^{(i)}(t) - \Lambda_T^{(i)}(t)||_2, \tag{9}$$

where $\varepsilon(t)$ represents the tracking error of the current frame at time t, whereas i is the i^{th} landmark and $\Lambda_G^{(i)}$, and $\Lambda_T^{(i)}$ represent the ground truth and estimated location in (x, y) of the i^{th} landmark.

Figure 3 shows some examples of the proposed tracking method. It can be seen that the proposed tracking method performs well in different illumination conditions, occlusion, or the visual angles.

Table I and Table II provide quantitative comparison of the proposed MMDL-FT, MMDL-FTU, AFTRM and AFTRM-W (AFTRM with the weighted CLM), with the Incremental Learning for Robust Visual Tracking (ILRVT) [2], incremental learning tracking based on Independent Component Analysis (ILICA), Incremental Cascaded Continuous Regression (iCCR) [23] and Approximate structured output learning for CLM [24].

Table I and Table II suggest that proposed MMDL-FT, MMDL-FTU, AFTRM and AFTRM-W outperform the other methods, and AFTRM-W has much-improved performance than AFTRM. This is due to the weighting mechanism, as consistent landmarks receive higher weight and thus improves the quality of the resyncing mechanism using W-CLM search process. The methods proposed by zheng et al. [24] and sanchez et al. [23] have close results to the proposed MMDL-FT, MMDL-FTU, AFTRM method for some videos, whereas, AFTRM-W has performed better than all the other methods on five videos out of six videos and had much smaller tracking error. For detailed analysis, look at the thesis chapter 05 of the thesis [14].

TABLE I: Average RMSE comparison of MMDL-FT, MMDL-FTU, AFTRM and AFTRM-W with comparative methods (the best results are in bold).

Video	1	2	3	4	5	6	Average
Terissi et al. [19]	38.43	26.93	50.38	66.44	66.12	16.75	34.24
Ross et al. [2]	21.43	10.56	183.72	30.12	6.23	12.17	44.04
Zheng et al. [24]	33.93	11.46	12.41	17.05	12.26	14.02	16.86
Sanchez et al. [23]	16.42	11.48	10.33	22.07	14.49	9.84	14.10
MMDL-FT	10.12	7.19	**7.63**	22.02	8.06	30.75	14.29
MMDL-FTU	*9.73*	*6.50*	*7.76*	16.62	7.76	19.37	11.29
AFTRM	15.01	9.22	13.78	*15.31*	*5.91*	*7.53*	*11.12*
AFTRM-W	**6.54**	**3.56**	10.65	**5.27**	**4.85**	**3.62**	**5.65**

TABLE II: Center Location Error (CLE) comparison of MMDL-FT, MMDL-FTU, AFTRM and AFTRM-W with comparative methods on YawDD dataset [22] (the best results are in bold).

Video	Male videos	Female videos	Average
Terissi et al. [19]	25.92	18.37	22.15
Ross et al. [2]	14.74	11.33	13.03
Zheng et al. [24]	13.02	10.14	11.58
Sanchez et al. [23]	14.11	10.17	12.14
MMDL-FT	10.61	8.70	9.65
MMDL-FTU	10.36	8.68	9.52
AFTRM	*8.81*	*7.54*	*8.18*
AFTRM-W	**5.31**	**4.24**	**4.78**

A. Evaluation of the Proposed Face Tracking Method in Yawning Detection

In the experiments, yawning detection is used as a case study to evaluate the correctness and effectiveness of the proposed tracking method in a real facial analysis problem, where the local face appearance is changing. The proposed method improves the method in Omidyeganeh et al. [9] in two ways. Firstly, the proposed method uses only the pixels in the lips to measure the mouth openness in a binary image, as compared to [9] which uses a rectangular mouth block and includes some pixels outside the lips to detect yawning. Secondly, Yawning is detected in each video frame if the following three conditions are satisfied:

$$\frac{NBC}{NBR} > \Gamma_1, \frac{NBC}{NWC} > \Gamma_2, \frac{VD}{HD} > \Gamma_3, \tag{10}$$

where NBC and NBR are the total number of black pixels in the current and the reference frames mouth respectively, whereas, NWC is the number of white pixels in the current frames mouth, HD is the horizontal distance between mouth corners and VD is the vertical distance between the center points of lips. The first frame is used as a reference in the proposed scheme and is assumed to contain a closed mouth. The thresholds are selected by using a ROC curve from the training set [9].

The proposed yawning detection is evaluated in terms of; the True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), False Negative Rate (FNR) and Correct Detection Rate ($CDR = \frac{TPR+TNR}{TPR+TNR+FPR+FNR}$).

(a) Distance to the camera (b) Partial occlusion; (c) Visual angle; (d) Illumination change.

Fig. 3: Results of the proposed MMDL-FTU method, red = tracked landmarks, yellow = ground-truth landmarks.

Table III exhibits a comparison on the YawDD dataset [22] of the proposed method using data provided by MMDL-FT, MMDL-FtU, AFTRM and AFTRM-W, with state of the art methods in yawning detection, including Chiang et al [25], Bouvier et al. [26] and Omidyeganeh et al. [9]. Table III suggests that the proposed method outperforms the comparative methods. Furthermore, the proposed method has the lowest FPR, which indicates the effectiveness of the proposed method. The threshold values for Γ_1, Γ_2 and Γ_3 are set to 1, 0.5 and 2.5, respectively.

TABLE III: Yawning Detection Results (the best result are in bold).

Method	TPR	TNR	FPR	FNR	CDR
Chiang et al [25]	0.3990	0.4562	0.6010	0.5438	0.4276
Bouvier et al. [26]	0.6764	0.5437	0.3236	0.4563	0.6101
Omidyeganeh et al. [9]	0.6578	**0.7733**	0.3419	**0.2266**	0.7155
MMDL-FT	0.7342	0.6435	0.2658	0.3565	0.6888
MMDL-FTU	0.7913	0.7432	0.2087	0.2568	0.7672
AFTRM	*0.8120*	0.7222	*0.1879*	0.2777	0.76703
AFTRM-W	**0.9307**	0.7551	**0.0693**	0.2449	**0.8429**

IV. Conclusions

This work proposes an adaptive face and facial landmark tracking scheme. The proposed face tracker contains two operating modes: MMDL-FTU and AFTRM-W. The operating mode selects the tracked target face among the candidate target face samples given by the motion model; they are based on feature learning techniques that accumulate face samples during tracking, and update the model incrementally to adapt to the current appearance of the tracked target face over time. To accumulate the training data, the quality of the test sample is checked before being used in the incremental and online training scheme. The MMDL-FTU operating mode represents the appearance of the face using a novel multi-model dictionary learning scheme for robust face tracking. The AFTRM-W operating mode uses the SVD subspace to model the appearance of the tracked target face, and uses a resyncing scheme in case of tracking failure. The tracking error of the proposed method is estimated using a novel error prediction scheme based on tracked landmark differences. The proposed resyncing scheme is called W-CLM, an improvement of classical CLM. W-CLM uses training data to assign weights to each landmark based on the consistency of the texture information and these weights are used to facilitate the W-CLM search process. Furthermore, an improvement in the yawning detection method is proposed, which uses the facial landmarks to estimate the features for yawning detection.

The proposed face tracker is evaluated using CLE of the tracked target face and RMSE of the facial landmarks with the comparative methods which are representative of the state-of-the-art. The experimental results show that both the operating modes (MMDL-FTU and AFTRM) of the proposed face tracker provide competitive face tracking results in comparison to methods that are representative of the state-of-the-art. Furthermore, the proposed improvement in yawning detection presents higher TPR and CDR than the comparative methods representative of the state-of-the-art.

V. Publications

- Mona Omidyeganeh, Shervin Shirmohammadi, Shabnam Abtahi, Aasim Khurshid, and Muhammad Farhan, Jacob Scharcanski, Behnoosh Hariri, Daniel Laroche, Luc Martel, "Yawning detection using embedded smart cameras". IEEE Transactions on Instrumentation and Measurement, IEEE, v. 65, n. 3, p. 570582, 2016.
- A. Khurshid and J. Scharcanski, "Incremental multi-model dictionary learning for face tracking," 2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Houston, TX, USA, 2018, pp. 1-6. [8].

VI. Submitted Article

- Aasim khurshid, Jacob Scharcaski, "A New Adaptive Face Tracker with Applications", submission: Transaction on Instrumentation and Measurement.

VII. Thesis Approved

- Thesis approved on Nov 29,2018 with Concept **A**.

Acknowledgment

The authors would like to thank thesis evaluation committee for their valuable suggestions to improve the work and CAPES, Brazil and SIDIA, Brazil for financial support.

REFERENCES

[1] C. Jung and J. Scharcanski, "Wavelet transform approach to adaptive image denoising and enhancement," *Journal of Electronic Imaging*, vol. 13, no. 2, pp. 278–285, 2004.

[2] D. A. Ross, J. Lim, R. S. Lin, and M. H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[3] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 983–990.

[4] T. F. Cootes, G. J. Edwards, C. J. Taylor *et al.*, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[5] S. Lucey, Y. Wang, M. Cox, S. Sridharan, and J. F. Cohn, "Efficient constrained local model fitting for non-rigid face alignment," *Image and Vision Computing*, vol. 27, no. 12, pp. 1804–1813, 2009.

[6] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *BMVC*, vol. 1, no. 2, 2006, pp. 929–938.

[7] G. R. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on*. IEEE, 1998, pp. 214–219.

[8] A. Khurshid and J. Scharcanski, "Incremental multi-model dictionary learning for face tracking," in *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, May 2018, pp. 1–6.

[9] M. Omidyeganeh, S. Shirmohammadi, S. Abtahi, A. Khurshid, M. Farhan, J. Scharcanski, B. Hariri, D. Laroche, and L. Martel, "Yawning detection using embedded smart cameras," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 3, pp. 570–582, 2016.

[10] S. Vater, R. Ivancevic, and F. P. Len, "Integration of precise iris localization into active appearance models for automatic initialization and robust deformable face tracking," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sept 2017, pp. 2617–2621.

[11] J. Soldera, G. Schu, L. R. Schardosim, and E. T. Beltrao, "Facial biometrics and applications," *IEEE Instrumentation Measurement Magazine*, vol. 20, no. 2, pp. 4–30, April 2017.

[12] J. Soldera, K. Dodson, and J. Scharcanski, "Face recognition based on geodesic distance approximations between multivariate normal distributions," in *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*, Oct 2017, pp. 1–6.

[13] S. Shirmohammadi and A. Ferrero, "Camera as the instrument: the rising trend of vision based measurement," *IEEE Instrumentation & Measurement Magazine*, vol. 17, no. 3, pp. 41–47, 2014.

[14] A. Khurshid and J. Scharcanski, "Adaptive face tracking based on online learning," Ph.D. dissertation, Universidade Federal do Rio Grande do Sul, 2018.

[15] H. Liu, S. Li, and L. Fang, "Robust object tracking based on principal component analysis and local sparse representation," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 11, pp. 2863–2875, 2015.

[16] X. Cheng, N. Li, T. Zhou, L. Zhou, and Z. Wu, "Visual tracking via sparse representation and online dictionary learning," in *International Workshop on Activity Monitoring by Multiple Distributed Sensing*. Springer, 2014, pp. 87–103.

[17] Y. Xie, W. Zhang, C. Li, S. Lin, Y. Qu, and Y. Zhang, "Discriminative object tracking via sparse representation and online dictionary learning," *IEEE Transactions on Cybernetics*, vol. 44, no. 4, pp. 539–553, 2014.

[18] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[19] D. Terissi and J. C. Gómez, "Facial motion tracking and animation: An ica-based approach," in *Proceedings of 15th European Signal Processing Conference, Poznan, Poland, September*, 2007, pp. 3–7.

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–511.

[21] Y. Yuan, S. Emmanuel, W. Lin, and Y. Fang, "Visual object tracking based on appearance model selection," in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–4.

[22] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, "Yawdd: a yawning detection dataset," in *Proceedings of the 5th ACM Multimedia Systems Conference*. ACM, 2014, pp. 24–28.

[23] E. Sánchez-Lozano, B. Martinez, G. Tzimiropoulos, and M. Valstar, "Cascaded continuous regression for real-time incremental face tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 645–661.

[24] S. Zheng, P. Sturgess, and P. Torr, "Approximate structured output learning for constrained local models with application to real-time facial feature detection and tracking on low-power devices," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[25] C.-C. Chiang, W.-K. Tai, M.-T. Yang, Y.-T. Huang, and C.-J. Huang, "A novel method for detecting lips, eyes and faces in real time," *Real-time Imaging*, vol. 9, no. 4, pp. 277–287, 2003.

[26] C. Bouvier, A. Benoit, A. Caplier, and P.-Y. Coulon, "Open or closed mouth state detection: static supervised classification based on log-polar signature," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2008, pp. 1093–1102.