Dynamic Texture Model for Eye Blinking Re-identification under Partial Occlusion

Cheng-You Hu[1], Shih-Kai Tai[1], Wei-Syuan Lee[1], Hsuan-Yu Liu[1], Yung-Hui Lin[1], and Huang-Chia Shih[1]

[1]Yuan Ze University

August 23, 2023

Abstract

In this study, an eye blinking re-identification system was proposed. A fast local binary pattern was used for feature extraction because its grayscale invariance and rotational invariance allow for the effective acquisition of feature information even in the presence of noise. Finally, a recurrent neural network and long short-term memory were used for model training. The results indicated that, compared with the model trained using static data, the models based on dynamic features were less affected by environmental noise in terms of accuracy. In addition, the model trained using the recurrent neural network was highly effective in identifying unenrolled users and achieved high overall accuracy.

Dynamic Texture Model for Eye Blinking Re-identification under Partial Occlusion

Cheng-You Hu, Shih-Kai Tai, Wei-Syuan Lee, Hsuan-Yu Liu, Yung-Hui Lin, and Huang-Chia Shih*
Department of Electrical Engineering, Yuan Ze University, Taiwan.
E-mail: *hcshih@saturn.yzu.edu.tw*

In this study, an eye blinking re-identification system was proposed. A fast local binary pattern was used for feature extraction because its grayscale invariance and rotational invariance allow for the effective acquisition of feature information even in the presence of noise. Finally, a recurrent neural network and long short-term memory were used for model training. The results indicated that, compared with the model trained using static data, the models based on dynamic features were less affected by environmental noise in terms of accuracy. In addition, the model trained using the recurrent neural network was highly effective in identifying unenrolled users and achieved high overall accuracy.

1. Introduction: A major challenge in face recognition for identification is compromised identification accuracy as a result of occluded features. For example, face masks conceal the lower half of the face, and bangs may conceal a quarter of the face. Such incomplete facial information may result in a lower accuracy and a higher false acceptance rate compared with when complete facial information is available. Therefore, this study used various data models and feature descriptions to enhance the available facial data surrounding the eyes. The analysis of eye blink artifacts, most studies are focused on the viewpoints of signal processing in scalp electroencephalogram (EEG) recording [1], [2]. Vision-based approaches also utilized to detect the eye blink event [3] and the eye state classifier for virtual reality headsets [4]. In biometric identification, reduced recognition accuracy due to missing feature information is considered the least desirable scenario and must be avoided. To increase the accuracy of identification with occlusion, contemporary researchers have increased the number of training and prediction levels through networks such as deep convolutional neural networks (CNNs), 16-layer visual geometry group networks, and 50-layer residual neural networks [5], [6]. Although only few studies have used dynamic time-series data to differentiate individual features, dynamic models have demonstrated superior performance in terms of classifying and detecting abnormalities compared with static models and have been applied in industrial classification and defect identification [7].

Based on our observations, current research focuses on feature enhancement and comparison using static data or emphasizes algorithm speed. This study proposes a method that will place more emphasis on the use of dynamic texture temporal data for achieving re-identification application. The proposed method in this study will mainly be applied to biometric unlocking, identify verification, and facial recognition. This study will prioritize the evaluation of differences in reliability and accuracy. Furthermore, considering the application on mobile devices, the proposed method will also focus on developing a method that maintains prediction confidence and accuracy without being affected by various environmental factors such as lighting intensity, lighting angles, day-night variations, color temperature, etc. Based on the above statements, the objectives of this study can be summarized as follows:

- Propose a model that uses temporal data capturing the dynamic texture variations around the eye during blinking behavior to enhance the accuracy of re-identification under mask-wearing conditions.
- Utilize the differences in blinking behavior as individual feature enhancements to reduce the probability of misidentifying unknown individuals as known individuals.
- Compare different feature descriptors to evaluate their impact on model decision-making under various environmental noise caused by changing factors.
- Based on the above points, compare and summarize the more efficient and accurate combination of features and the advantages of dynamic texture features over static ones in practical applications.

2. Methodology: In this study, OpenCV was used to convert facial information captured by cameras into a matrix. This matrix was then imported into the MediaPipe Face Mesh module, which uses machine learning to infer the three-dimensional facial surface. This module is capable of estimating 468 three-dimensional face landmarks in real time, even on mobile devices with only a camera input, and does not require a dedicated depth sensor.

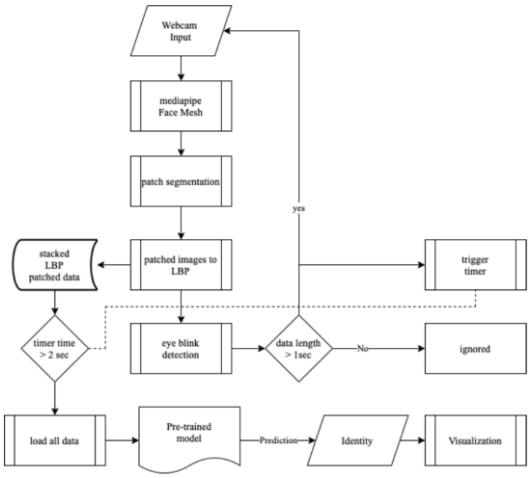


Fig. 1 Flowchart of the system.

Throughout the pipeline, a lightweight model architecture is used along with graphics processing unit acceleration to achieve real-time performance critical to live experiences [8]. However, face masks that partially cover these 468 landmarks may prevent data collection from certain features. Therefore, this study collected coordinate data from points on the face that considerably vary when the mask wearer blinks, are not covered by the mask, and are representative of the face. A total of 52 sets of coordinates were used as the reference point for feature collection. For each frame, features were collected using each of these points as the center to extract images within the 15 × 15-pixel area of each center point as the database for subsequent use.

2.1 Data Acquisition: Blinking was defined in accordance with the eye aspect ratio [9], [10]. For training purposes, the dynamic data length was set to 50 frames, meaning that each blink was expected to appear in the (25 ± 4)th frame of each group of dynamic data. Therefore, a matrix was established to continuously store the segmented images of the mask wearer's face. Each frame had a timestamp. To prevent memory overload, all data that entered the matrix for more than 2 seconds were deleted by the system. When the system detected a blink, the data obtained 1 second before and after the blink were extracted. Because of the variability of the frame rate in the hardware and camera, the system continuously deleted the first and then the final data points in each data group until the data length reached 50 frames. This process was conducted for each data group, which was then used in subsequent feature descriptions. Figure 2 depicts the overall data model.

2.2 Texture Feature Representation: The facial identification technology has become mature, with numerous edge-computing devices capable of performing standalone recognition. Regarding the applications in facial verification or re-identification, the mesh-based method can be employed to analyze the entire face if computational cost is not the major issue. However, for usage on mobile devices with limited computational power, lightweight algorithms are necessary. Furthermore, considering the issue of facial occlusion, especially when users are wearing masks, only the eye portion is visible. Therefore, this study proposes the use of motion texture features to enhance the feature dimensionality and maintain a low false positive rate. In this approach, each frame of video sequence is analyzed separately, and a model-based facial feature point detection method is employed instead of feature point matching to establish correlations between patches. Using feature point matching for adjacent frames would be more complex and chaotic, requiring additional anti-noise techniques such as random sample consensus (RANSAC) [11] algorithm and non-maximum suppression (NMS) [12] method to perform correspondence. Such a system would entail significant computational resources. Instead, this study adopts simple and well-known texture descriptors to extract texture features as the foundation for feature analysis. To eliminate external noise and irrelevant image information introduced during the segmentation of the original images, feature extraction was performed to reduce the dimensionality of the training

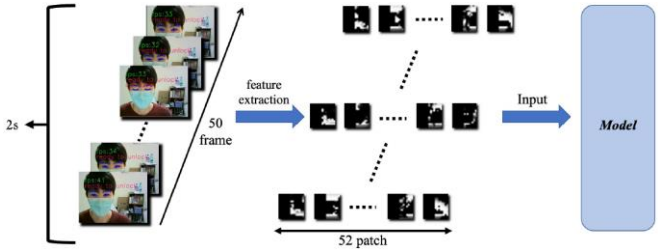
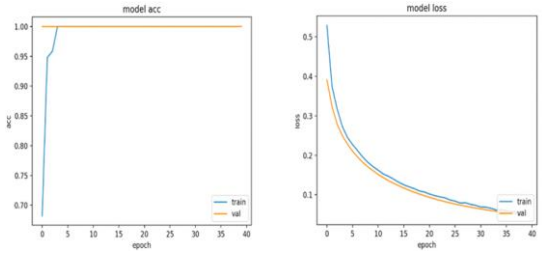
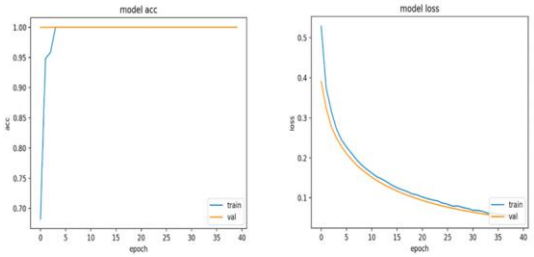


Fig. 2 *Depicts the overall data model.*



(a) (b)

Fig. 3 *RNN model training curve based on LBP feature description, (a) Accuracy training curve, (b) Loss training curve.*



(a) (b)

Fig. 4 *Bidirectional LSTM model training curve based on LBP feature description, (a) Accuracy training curve, (b) Loss training curve.*

data. The extraction procedure enhanced the critical information, reduced the hardware load, and accelerated the training process. In addition, local binary pattern (LBP) [13] and histogram of gradient (HoG) [14] descriptors were used to represent the texture features.

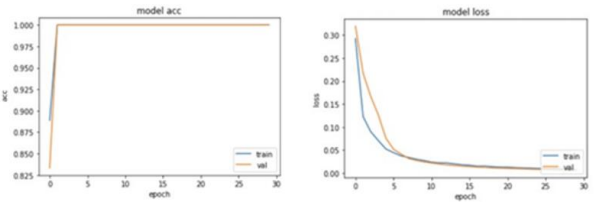
2.2.1 LBP descriptor: The LBP descriptor is originally used to describe the local feature operators of images. Its advantages include grayscale invariance and rotational invariance. In LBP feature description, a center pixel is used as a threshold, which is then compared to the neighboring pixels, with those greater than the center pixel value marked as 1 and those lower than the center pixel value marked as 0, to create a binary pattern. Finally, the established patterns are classified and rearranged into histograms, with each pattern corresponding to a single bin.

2.2.2 HoG descriptor: In HoG feature description, the appearance and shape of local targets in images are typically well described with the distributions of gradient or edge orientation densities even when the accurate relative gradient or edge positions are unknown. The translation invariance of the histogram contributes to its effectiveness in reflecting changes in lighting and shadows. HoG descriptors produce features that summarize the measurement distribution within an image area and are particularly effective in identifying textures with variable shapes.

2.3 Model Training: In this study, time-series data were used to reflect dynamic feature changes, and a recurrent neural network (RNN) and bidirectional long short-term memory (LSTM) were used to train a time-series model. For comparison, another model was established by training static data with the conventional CNN.

2.3.1 RNN training: During the time-series data training process, an RNN was selected because of its light weight. The training data set comprised 450 dynamic data points (150 for each of the three individuals whose identity was registered). The model training process was then conducted using the hyperparameter of a 0.0005 learning rate, a binary cross-entropy loss function, and an Adam optimizer. The training epoch and batch size were set to 50 and 8, respectively. Figure 3 depicts the accuracy and loss curve of the training process.

2.3.2 Bidirectional LSTM: During the bidirectional LSTM training process, the same data set and training parameters adopted in RNN



(a) (b)

Fig. 5 *CNN model training curve based on static data LBP feature description, (a) Accuracy training curve, (b) Loss training curve.*

(a) (b)

Fig. 6 *Results of identification, (a)enrolled user identification, (b)unenrolled user identification.*

Algorithm I: Identify Re-identification

Input : Data
Output : Verification
1 : D ← Data
2 : RP ← Registered Data
3 : S ← Shape of Data
4 : MD ← Model Prediction
5 : V ← Verification
6 : L ← Lengith (F)
*7 : D ← Reshape(Data, S[Frame], S_x * S_y* S_z)/255*
8 : MD ← Predict(D)
9 : if MD is registered AND MD confidence >0.9 then
10 : V ← RP[MD]
11 : else
12: V ← Fasle
13 : end if

training were used. Figure 4 shows the accuracy and loss curve of the training process.

2.3.3 Static CNN model: The CNN model, which is based on static data, was subjected to CNN training for two-dimensional reconstruction using LBP-segmented data. The training set comprised 750 images (250 for each of the three registered users). The model parameters used were the same as those employed in RNN training, and EarlyStopping callback was used to prevent overfitting. Figure 5 depicts the accuracy and loss curve of the training process.

2.4. Identification and Removal of Unenrolled Users: To achieve identity verification, the user is required to provide facial information input to the camera. This information is then subjected to image segmentation, with the data being accessed and the features being extracted. The model then provides a confidence value for each label, which the system uses to calculate the average confidence as a threshold to determine whether the user's identity has been registered in the database. In this study, the average confidence was 90%, indicating that any user with a label whose confidence value exceeded the 90% threshold was considered a registered user. By contrast, users with all of their label confidence values below the 90% threshold did not pass the identification step and were deemed unenrolled users. Figures 6 and Algorithm I depict the verification procedure and the computational details, respectively.

3. Experimental Results:

3.1. Identification Accuracy of Dynamic and Static Models with Different Feature Descriptors:

3.1.1. Experimental procedures: The purpose of this experiment was to determine how noise-containing data affect static and dynamic models and how feature descriptors handle these data. During the experiment, noise was created by changing the lighting intensity, lighting angle, and ambient color temperature and rotating the image. The experiment was

Table 1: Architecture of training model for CNN.

Model	Accuracy (%)	Average Confidence (%)
Dynamic LBP-RNN	95.6	92.1
Dynamic LBP-BILSTM	93.3	99.4
Dynamic HoG-BILSTM	71.1	98.8
Static LBP-CNN	72.6	86.4

Table 2: Test results based on 90 noise-containing and 90 optimal-environment data points.

Model	Accuracy (%)	Average Confidence (%)
Dynamic LBP-RNN	97.8	95.0
Dynamic LBP-BILSTM	96.7	99.6
Dynamic HoG-BILSTM	85.6	99.1
Ststic LBP-CNN	86.3	91.8

Table 3: Confusion matrix of the dynamic LBP-LSTM model.

	Forecast known	Forecast unknown
Actually known	88	2
Actually unknown	55	5

Sensitivity	0.98	Precision	0.62
Specificity	0.08	F1 score	0.76

Table 4: Confusion matrix of the dynamic LBP-RNN model.

	Forecast known	Forecast unknown
Actually known	80	10
Actually unknown	0	60

Sensitivity	0.89	Precision	1.00
Specificity	1.00	F1 score	0.94

Table 5: Confusion matrix of the static LBP-RNN model.

	Forecast known	Forecast unknown
Actually known	69	21
Actually unknown	24	36

Sensitivity	0.76	Precision	0.74
Specificity	0.60	F1 score	0.75

conducted on two stages. The first stage involved the collection of 90 noise-containing data points to test the ability of the model to accurately identify the three registered users and determine the average confidence in the presence of a large amount of noise. The second stage involved the collection of another 90 data points in an optimal environment into the original data set, with a total of 180 data points, for the same test.

3.1.2 Experimental results: As shown in Tables 1 and 2, the model identified features with dynamic data and effectively reduced the influence of noise on the identification results. Therefore, for face identifiers to be mounted on mobile devices, factors such as the angle, the intensity of ambient lighting, and day and night lighting must be considered. Accordingly, in variable environments, dynamic data are expected to outperform static data. According to the results from Table 4 and Table 5, it can be observed that the use of dynamic texture features with the LBP descriptor yields results approximately 20% higher than static features. The highest performance is achieved using the RNN model. However, when utilizing the HoG descriptor for dynamic texture features, the effectiveness seems to be worse than static features. This is likely due to the low resolution of facial images. The LBP method, which employs a more localized and dense texture representation, is better at preserving the original texture characteristics. On the other hand, using the HoG texture descriptor increases the likelihood of overlapping between patches, leading to distorted feature values, especially during moments of eye blinking.

3.2. Identifying Enrolled and Unenrolled User Data with Dynamic and Static Models:

3.2.1 Experimental procedures: During the experiment, two test data sets were used to compare the performance of the dynamic and static models. The first data set comprised 90 enrolled user data points, including both optimal-environment and noise-containing data. The second data set comprised 60 unenrolled user data points. The confidence threshold was set to 90%. Data with a model-predicted confidence value exceeding 90% were categorized as enrolled user data; otherwise, they were categorized as unenrolled user data. Both LBP-LSTM and LBP-RNN models, which demonstrated superior identification performance in the first experiment, were compared to the static LBP-CNN model. On the basis of the experimental results, data instantiation was then performed using confusion matrices for an objective evaluation of the differences and respective advantages of dynamic and static data.

3.2.2 Experimental results: According to the sensitivity results presented in Tables 3, 4, and 5, both dynamic models were less likely than the LBP-CNN model to misidentify enrolled users as unenrolled users. In terms of specificity, the LBP-LSTM model had a relatively high likelihood of misidentifying unenrolled users as enrolled users. However, compared with the LBP-CNN model, the LBP-RNN model made fewer such errors. The LBP-RNN model had the highest F1-score, at 0.94. Therefore, to prevent unenrolled users from passing the identity verification step and accessing the system, the LBP-RNN model was selected. In addition, taking into account both safety against unenrolled users and the accuracy of correctly identifying enrolled users, the LBP-RNN model, which demonstrated the highest specificity and second-highest accuracy, was selected as the model of choice.

4. Conclusions: This study proposed a method for verifying a user's identity when their forehead and facial features below the nose are both obscured and when noise is present. Dynamic data were used with a variety of descriptors and various training methods for model training. The established models were then evaluated and compared to a model constructed using static data. The experiments involved the models identifying users whose faces were partially obscured by relying on feature changes surrounding their eyes. In terms of identification accuracy by feature descriptors and the identification of both enrolled and unenrolled users, the dynamic data outperformed the static data. During the test that was entirely based on noise-containing data, the dynamic data effectively mitigated the influence of noise on the identification results. In addition, compared with the LBP-CNN model, the dynamic models, particularly the LBP-RNN model, had greater performance in terms of eliminating unenrolled users. In conclusion, given the ability of the dynamic-data-based LBP-RNN model to identify users correctly despite the presence of noise and partial occlusion and its ability to prevent unenrolled users from accessing the system, this model achieves optimal safety and precision.

Author contributions: Cheng-You Hu: Conceptualization, Investigation, Methodology. Shih-Kai Tai: Conceptualization, Investigation, Methodology. Wei-Syuan Lee: Validation, Visualization. Hsuan-Yu Liu: Formal analysis. Yung-Hui Lin: Resources, Writing-Original draft preparation. Huang-Chia Shih: Project administration, Supervision, Writing-review & editing.

Acknowledgments: This paper is partially supported by the National Science and Technology Council, grant number 111-2221-E-155-045 - MY2.

Conflict of Interest Statement: The authors declare no conflict of interest.

Data Availability Statement: Data available on request due to privacy/ethical restrictions

 Received: XX August 2023 Accepted: XX XXX 2023 doi: 10.1049/ellX.XXXXX

References

1. López-Ahumada, R.; Jiménez-Naharro, R.; Gómez-Bravo, F. A Hardware-Based Configurable Algorithm for Eye Blink Signal Detection Using a Single-Channel BCI Headset. Sensors 2023, 23, 5339.
2. Kong, W.; Zhou, Z.; Hu, S.; Zhang, J.; Babiloni, F.; Dai, G. Automatic and Direct Identification of Blink Components from Scalp EEG. Sensors 2013, 13, 10783-10801.
3. Al-gawwam, S.; Benaissa, M. Robust Eye Blink Detection Based on Eye Landmarks and Savitzky–Golay Filtering. Information 2018, 9, 93.

4. Alsaeedi, N.; Wloka, D. Real-Time Eyeblink Detector and Eye State Classifier for Virtual Reality (VR) Headsets (Head-Mounted Displays, HMDs). Sensors 2019, 19, 1121.
5. Yang L., Ma J., Zheng Y., and Liu H. Deep representation for partially occluded face verification. EUR-ASIP Journal on Image and Video Processing. 2018, 143, 1-10.
6. Hariri W. Efficient masked face recognition method during the COVID-19 pandemic. Signal, Image and Video Processing. 2022, 16, 605–612.
7. Chai Z. and Zhao C. Enhanced Random Forest with Concurrent Analysis of Static and Dynamic Nodes for Industrial Fault Classification. IEEE Transactions on Industrial Informatics. 2020, 16, 1, 54-66.
8. Lugaresi C., Tang J., Nash H., McClanahan C., Uboweja E., Hays M., Zhang F., Chang C. L., Yong M. G., Lee J., Chang W. T., Hua W., Georg M., andGrundmann M. MediaPipe: A Framework for Building Perception Pipelines. Google Research, 2019.
9. Soukupov T. and Cech J. Real-Time Eye Blink Detection using Facial Landmarks. In Proceedings of the 21st Computer Vision Winter Workshop, Rimske Toplice, Slovenia, 3–5 February, 2016.
10.Sathasivam S., Mahamad A. K., Saon S., Sidek A., and Som M. M., and Ameen H. A. Drowsiness Detection System using Eye Aspect Ratio Technique. In Proceedings of the IEEE Student Conference on Re-search and Development, Johor, Malaysia, 27-28, Sept. 2020.
11.M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Commun. ACM, vol. 24, no. 6, pp. 381–395, 1981.
12.J. H. Hosang, R. Benenson, and B. Schiele, “Learning non-maximum suppression,” arXiv 1705.02950, 2017.
13.T. Ojala, M. Pietikäinen, and T. T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary pattern,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, Jul. 2002.
14.N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886-893.