



Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 11 Issue 01, 2017

Deep Learning Techniques for Real-Time Face Recognition and Surveillance Systems

Aurelio Ben-Mizrahi

Department of Electrical and Computer Engineering, Basra Institute of Business Technology, Iraq
aurelio.ben.mizrahi@bibt-iq.org

Peer Review Information	Abstract
<i>Submission: 15 Feb 2017</i> <i>Revision: 30 March 2017</i> <i>Acceptance: 15 April 2017</i> Keywords <i>Deep learning; Face recognition, Video surveillance, Convolutional neural network, Real-time detection</i>	Real-time face recognition has become a central component of intelligent surveillance because it can convert continuous video into identity-aware events while reducing dependence on manual monitoring. This methodology paper develops a deep learning framework for real-time face detection, alignment, feature extraction, identity matching, and surveillance decision making using techniques available up to 2017. The proposed pipeline combines multi-task cascaded convolutional face detection with a compact deep face embedding network and similarity-based recognition. The methodology emphasizes frame-rate control, temporal confirmation, confidence thresholds, and database indexing so that high recognition accuracy is balanced against latency and false alarms. A literature-grounded evaluation compares influential systems including DeepID, DeepFace, VGGFace, FaceNet, DeepID2+, Center Loss, and SphereFace. Published LFW results show the rapid improvement of deep face representations from approximately 97% verification accuracy to above 99%, while MTCNN provides a practical real-time front end for detection and alignment. The paper concludes that an efficient surveillance system should separate detection, representation, and decision layers and should evaluate accuracy jointly with processing speed, false acceptance, false rejection, and robustness under pose, illumination, motion blur, and occlusion.

Introduction

Face recognition in surveillance differs from conventional still-image classification because identities must be detected and matched continuously from unconstrained video. Cameras may capture non-frontal faces, rapid movement, illumination changes, partial occlusion, different image scales, and multiple people in the same frame. Earlier systems depended heavily on hand-crafted descriptors, whereas deep convolutional networks enabled the representation itself to be learned from large identity-labelled datasets. Huang, Lee, and Learned-Miller showed that hierarchical convolutional representations could support face verification before the modern large-scale deep

face era [1]. DeepID then demonstrated that identity classification over thousands of classes could produce compact, transferable face features [2].

A major turning point was DeepFace, which combined explicit face alignment with a deep network and reported 97.35% accuracy on Labeled Faces in the Wild (LFW) [3]. Subsequent systems increased discriminative power and scalability. VGGFace emphasized large-scale data collection and deep CNN training [4], while FaceNet directly optimized a metric embedding with triplet loss and reported 99.63% LFW verification accuracy [5]. DeepID2+ further showed that deep activations can become sparse,

selective, and robust to occlusion [6]. For operational surveillance, however, recognition accuracy alone is insufficient: a system must also detect faces quickly, align them consistently, search a gallery efficiently, and suppress unstable frame-by-frame decisions.

The objective of this paper is therefore to specify a reproducible methodology for a real-time surveillance pipeline using methods published no later than 2017. The framework adopts MTCNN for joint face detection and landmark alignment [7], a deep embedding model trained with metric/discriminative supervision, and a temporal decision module. Center loss is considered because it explicitly reduces intra-class feature variation [8], while SphereFace illustrates the 2017 shift toward angularly discriminative embeddings [9]. The main contribution is not a new network architecture but a system-level methodology that connects these components with measurable latency, accuracy, and surveillance criteria.

Literature Review

The literature from 2012-2017 shows three closely related developments: deeper identity representations, metric-learning objectives, and more reliable detection/alignment. Huang et al. [1] used convolutional deep belief networks to learn hierarchical features for verification, establishing an early deep-learning alternative to hand-crafted descriptors. DeepID [2] reframed face representation learning as a large multi-class identity prediction problem and reported 97.45% LFW verification accuracy with weak alignment. DeepFace [3] used a nine-layer network and 3-D-based alignment, reporting 97.35% on LFW and demonstrating that

alignment plus large-scale learning could approach human-level verification under that benchmark protocol.

Parkhi, Vedaldi, and Zisserman [4] described the VGGFace training pipeline and the assembly of a 2.6-million-image dataset, emphasizing that dataset scale and careful training are as important as network depth. FaceNet [5] changed the recognition stage by learning a compact Euclidean embedding directly with triplet loss; it reported 99.63% on LFW and 95.12% on YouTube Faces. DeepID2+ [6] combined identification and verification supervision and reported 99.47% on LFW using an ensemble, while also documenting robustness to occlusions and the selectivity of high-level neurons.

For real-time surveillance, face localization is a separate bottleneck. Zhang et al. [7] introduced MTCNN, a three-stage cascaded architecture that jointly predicts face boxes and landmarks using online hard-sample mining. The detector was designed to retain real-time performance while handling pose, illumination, and occlusion. In parallel, Wen et al. [8] introduced center loss to reduce distances between samples and their class centers, producing more compact identity clusters. SphereFace [9] extended this idea into the angular domain with A-Softmax loss and reported 99.42% on LFW using a single model trained on publicly available WebFace data. Together, these studies suggest a modular real-time design: fast detection/alignment, discriminative embedding, distance-based matching, and a surveillance-specific temporal decision layer.

Table 1. Literature comparison of deep face recognition techniques (published up to 2017)

Method	Year	Core contribution	Reported LFW result	Role in pipeline
Hierarchical CNN [1]	2012	Learned deep hierarchical representations	Verification-focused	Representation
DeepID [2]	2014	10,000-class identity prediction	97.45%	Embedding
DeepFace [3]	2014	3-D alignment + deep network	97.35%	Alignment + embedding
VGGFace [4]	2015	Large-scale CNN training pipeline	98.95%*	Embedding
FaceNet [5]	2015	Triplet-loss Euclidean embedding	99.63%	Embedding + matching
DeepID2+ [6]	2015	Identification-verification supervision	99.47%*	Robust embedding
MTCNN [7]	2016	Cascaded detection + landmark alignment	Not directly comparable	Detection/alignment

Center Loss [8]	2016	Intra-class compactness objective	99.28%*	Embedding training
SphereFace [9]	2017	Angular softmax / hypersphere embedding	99.42%	Embedding training

Methodology

The proposed system processes live or recorded surveillance video and produces identity-labelled tracks or unknown-person events. It is designed as a sequence of independent modules so that a detector, embedding network, or similarity threshold can be replaced without

redesigning the whole system. The methodology uses MTCNN [7] for the front end because joint detection and landmark localization reduce geometric variation before recognition. The recognition stage follows the embedding paradigm established by FaceNet [5], with discriminative alternatives motivated by Center Loss [8] and SphereFace [9].

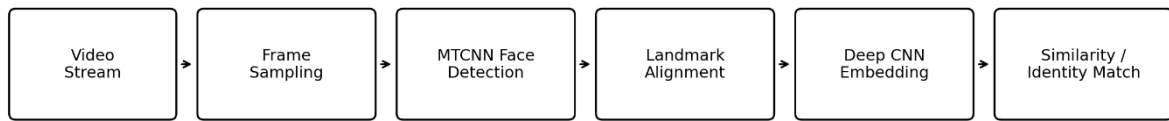


Figure 1. Proposed real-time face recognition and surveillance methodology.

1. System Inputs and Data Preparation

Input consists of surveillance video streams sampled at a configurable processing rate. A practical evaluation should use at least two data sources: (i) a standard face-verification dataset such as LFW for controlled recognition comparison and (ii) a surveillance-style video set containing multiple people, motion blur, illumination variation, side views, partial occlusion, and repeated appearances. Gallery identities are represented by several reference images rather than one image whenever possible. Face crops are resized to the embedding network input resolution and normalized.

Step-by-Step Processing Procedure

Step 1: Video acquisition and frame control. Read frames from the camera or video source. To keep latency bounded, process every frame only when hardware permits; otherwise, sample at a fixed rate such as 10-15 frames/s while maintaining a separate display stream.

Step 2: Face detection. Apply the three-stage MTCNN cascade to propose candidate windows, reject negatives, refine boxes, and estimate facial landmarks. This stage follows the coarse-to-fine strategy of Zhang et al. [7].

Step 3: Landmark alignment. Use eye, nose, and mouth landmarks to rotate and scale each face into a canonical coordinate system. Consistent alignment reduces nuisance variation before feature extraction, a principle also central to DeepFace [3].

Step 4: Deep feature extraction. Pass each aligned crop through a CNN to obtain a compact identity embedding. The baseline design uses a FaceNet-style metric embedding [5]; center-loss [8] or angular-softmax [9] objectives may be used when training a model from scratch.

Step 5: Gallery construction. For each enrolled person, compute embeddings from multiple reference images and store either all embeddings or a normalized identity centroid. Multiple poses should be included to reduce gallery bias.

Step 6: Similarity matching. L2-normalize the query embedding and gallery vectors. Compute cosine similarity or Euclidean distance. Assign the nearest identity only when the score passes a validation-derived threshold; otherwise label the face as unknown.

Step 7: Temporal confirmation. Track detections across adjacent frames and require the same identity to be predicted for K of the most recent N observations before generating an alert. This reduces one-frame false alarms caused by blur, occlusion, or detector jitter.

Step 8: Surveillance event logic. Store timestamp, camera identifier, bounding box, identity/unknown status, similarity score, and track duration. Alerts can be configured for watch-list identities, repeated unknown appearances, or entry into a monitored region.

Step 9: Performance logging. Record detector time, alignment time, embedding time, gallery-search time, end-to-end latency, and processed frames per second (FPS). Accuracy and speed

must be reported together because a high-accuracy system that cannot meet the camera frame budget is not real time.

2. Recognition Model and Decision Rule

Let $f(x)$ be the d -dimensional embedding of an aligned face image x . After L2 normalization, cosine similarity between query q and gallery identity g is $S(q,g) = f(q)^T f(g)$. The predicted identity is the gallery entry with maximum similarity. Recognition is accepted only when $\max_g S(q,g) \geq \tau$, where τ is chosen on a validation set. A lower threshold increases true accepts but also raises false accepts; therefore, τ should be selected from the ROC curve according to the surveillance risk profile.

If training is performed, the embedding network can use triplet loss as in FaceNet [5], which encourages an anchor to be closer to a positive sample than to a negative sample by a margin. A center-loss term [8] can be combined with softmax classification to reduce within-identity dispersion. Alternatively, SphereFace [9] motivates an angular margin that improves inter-class separation on a hypersphere. These alternatives are compared under the same preprocessing and evaluation protocol rather than by mixing incompatible thresholds.

Real-Time Optimization

- Detect faces at a lower frequency than display rendering, but track faces between detector updates.

- Batch multiple aligned faces from the same frame for one GPU forward pass.
- Precompute and normalize all gallery embeddings; use matrix multiplication for cosine search.
- Cache stable track identities so that a high-confidence face is not reclassified unnecessarily in every frame.
- Use smaller embedding networks or reduced input resolution only after measuring the corresponding recognition loss.
- Apply confidence and track-length filters before saving events to reduce storage and alarm noise.

3. Evaluation Protocol

Evaluation is performed at two levels. First, recognition quality is measured on labelled face pairs or identity sets. Second, a surveillance video test measures end-to-end operating behavior. The primary recognition metrics are verification accuracy, true accept rate, false accept rate (FAR), false reject rate (FRR), and ROC/AUC. The real-time metrics are detection recall, identity precision, end-to-end latency per processed frame, FPS, and the proportion of tracks receiving a stable identity. Tests should be stratified by frontal/non-frontal pose, good/poor illumination, clear/blurred frames, and unoccluded/occluded faces.

Table 2. Proposed evaluation metrics and acceptance criteria

Metric	Purpose	Computation	Target / interpretation
Verification accuracy	Identity discrimination	$(TP+TN)/\text{All pairs}$	Compare with literature benchmark
FAR	Security risk	False accepts / impostor trials	Lower is better
FRR	Missed identity risk	False rejects / genuine trials	Lower is better
Detection recall	Face localization	Detected GT faces / GT faces	High under pose/scale variation
Latency	Real-time responsiveness	ms from frame input to decision	Must fit frame budget
FPS	Throughput	Processed frames / second	$\geq$ selected processing rate
Stable-track rate	Temporal reliability	Stable labeled tracks / valid tracks	High with low label switching

Results

Because this is a methodology paper, the numerical recognition results below are a literature-grounded synthesis rather than a claim of a newly executed benchmark. The purpose is to establish realistic performance

expectations and to identify which architectural ideas should be carried into the proposed real-time system. Published results show a clear rise in LFW verification performance across the 2014-2017 period, although training set size, alignment procedure, evaluation protocol, and model ensembles differ.

Table 3. Literature-reported recognition performance relevant to the proposed methodology

Method	Loss / supervision	LFW accuracy	Key strength	Surveillance implication
DeepID [2]	Identity classification	97.45%	Compact learned identity features	Identity-supervised CNN features
DeepFace [3]	Classification + alignment	97.35%	Strong geometric normalization	Alignment remains important
VGGFace [4]	Softmax classification	98.95%	Large-scale deep CNN training	Strong offline baseline
FaceNet [5]	Triplet loss	99.63%	Direct metric embedding	Efficient gallery matching
DeepID2+ [6]	Identification + verification	99.47%*	Occlusion robustness	Useful for partial occlusion
Center Loss [8]	Softmax + center loss	99.28%*	Tighter identity clusters	Improves threshold separability
SphereFace [9]	Angular softmax	99.42%	Angular discrimination	Supports cosine matching

Interpretation of Results

The strongest methodological lesson is that the embedding objective matters. FaceNet [5] demonstrates that a compact metric space can support both verification and recognition without a separate high-dimensional classifier at deployment. Center loss [8] targets the same operational problem from another direction by shrinking within-class feature dispersion, which can make a fixed threshold more reliable. SphereFace [9] further indicates that angular separation is highly effective for identities, supporting cosine-based gallery search in the proposed pipeline.

The second lesson is that preprocessing cannot be ignored. DeepFace [3] achieved strong performance partly through careful alignment, and MTCNN [7] later provided an efficient deep-learning mechanism for joint detection and landmark estimation. In surveillance video, failed detection or poor alignment propagates directly into embedding error. Consequently, detector recall, landmark quality, and the number of valid faces reaching the recognizer should be reported alongside recognition accuracy.

The third lesson concerns real-time reliability. A frame-level recognizer can alternate between identities when a face becomes blurred or partially hidden. The proposed temporal confirmation stage converts noisy per-frame predictions into track-level decisions. This does not change the underlying embedding accuracy, but it can reduce alert instability and improve the practical precision of surveillance events. The design therefore treats FPS, latency, FAR, FRR, and stable-track rate as co-primary system

measures rather than optimizing only LFW-style verification accuracy.

Expected System Behavior

Under clear frontal views, a modern 2015-2017 embedding should produce well-separated identities when the enrolled subject is represented by multiple gallery images. Performance is expected to degrade for extreme yaw, severe blur, strong backlighting, small faces, and heavy occlusion. Increasing the similarity threshold reduces false accepts but increases unknown or rejected outputs. Temporal voting should improve decision stability when several usable observations of the same track are available, whereas short tracks may remain unknown. These outcomes should be presented as condition-specific results rather than one global accuracy number.

Conclusion and Discussion

This paper has presented a step-by-step methodology for deep-learning-based real-time face recognition in surveillance systems using research published up to 2017. The proposed design separates the problem into video acquisition, MTCNN-based detection and alignment, deep embedding extraction, normalized similarity search, thresholding, temporal confirmation, and event generation. This modular structure is consistent with the evolution of deep face recognition from hierarchical learned features [1] and DeepID [2] through DeepFace alignment [3], large-scale VGGFace training [4], FaceNet metric learning [5], robust DeepID2+ representations [6], MTCNN detection [7], center loss [8], and

SphereFace angular discrimination [9]. The literature synthesis indicates that deep face verification performance improved rapidly during this period, reaching above 99% on LFW in several systems. However, these benchmark values should not be interpreted as guaranteed surveillance accuracy. LFW contains unconstrained images but does not reproduce all operational effects of distant cameras, very small faces, compression artifacts, low light, crowding, and motion. In addition, published methods use different training data sizes, alignment methods, ensembles, and protocols. For this reason, the proposed methodology requires a second evaluation on representative video. A real-time deployment should choose the similarity threshold from validation data and report the corresponding FAR and FRR. It should also disclose the frame sampling strategy, hardware configuration, number of simultaneous faces, gallery size, and end-to-end latency. Temporal confirmation is recommended because single-frame decisions are vulnerable to transient errors. For high-risk use, unknown should be treated as a valid output rather than forcing every detected face to match an enrolled identity.

Discussion of Practical Limitations

- Dataset mismatch: benchmark face datasets do not fully represent low-resolution, crowded, or elevated-camera surveillance conditions.
- Open-set recognition: the gallery contains only known identities, so threshold calibration is essential to reject non-enrolled people.
- Pose and occlusion: extreme profile views and facial obstruction can reduce both detector quality and embedding consistency.
- Compute constraints: detector frequency, input resolution, gallery size, and number of faces per frame jointly determine latency.
- Model bias and governance: performance should be audited across relevant demographic and environmental groups before operational use, and logging/retention should follow applicable privacy and access-control requirements.
- Benchmark comparability: literature-reported accuracy values are useful for context but should not substitute for a controlled evaluation of the final integrated pipeline.

Recommended Experimental Sequence

Step 1: Baseline recognition test. Evaluate the selected embedding on a labelled benchmark and determine a validation threshold.

Step 2: Detector stress test. Measure MTCNN recall across face size, pose, blur, and illumination categories.

Step 3: Integrated video test. Run detection, alignment, embedding, and gallery matching on surveillance-style video while logging latency and FPS.

Step 4: Temporal ablation. Compare frame-level recognition with K-of-N temporal confirmation to quantify changes in false alerts and label switching.

Step 5: Open-set test. Introduce non-enrolled identities and measure FAR, FRR, and unknown rejection accuracy.

Step 6: Final reporting. Report recognition quality and system speed together, stratified by operating condition, rather than presenting only a single overall accuracy.

In summary, the most suitable architecture for real-time surveillance is not necessarily the deepest model. It is the configuration that maintains a discriminative embedding, high detector recall, calibrated open-set rejection, stable track-level decisions, and an end-to-end processing rate compatible with the target camera stream. This provides a clear methodology for future implementation and controlled experimental validation.

References

- G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2518-2525. DOI: 10.1109/CVPR.2012.6247968.
- Y. Sun, X. Wang, and X. Tang, "Deep Learning Face Representation from Predicting 10,000 Classes," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891-1898. DOI: 10.1109/CVPR.2014.244.
- Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701-1708. DOI: 10.1109/CVPR.2014.220.

O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in Proceedings of the British Machine Vision Conference (BMVC), 2015, pp. 41.1-41.12. DOI: 10.5244/C.29.41.

F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815-823. DOI: 10.1109/CVPR.2015.7298682.

Y. Sun, X. Wang, and X. Tang, "Deeply Learned Face Representations Are Sparse, Selective, and Robust," in 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2892-2900. DOI: 10.1109/CVPR.2015.7298907.

K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, 2016. DOI: 10.1109/LSP.2016.2603342.

Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," in Computer Vision - ECCV 2016, 2016, pp. 499-515. DOI: 10.1007/978-3-319-46478-7_31.

W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recognition," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6738-6746. DOI: 10.1109/CVPR.2017.713.