# Deep Convolutional Neural Network using Triplet of Faces, Deep Ensemble, and Score-level Fusion for Face Recognition

Bong-Nam Kang, Yonghyun Kim, Daijin Kim
***POSTECH, South Korea***

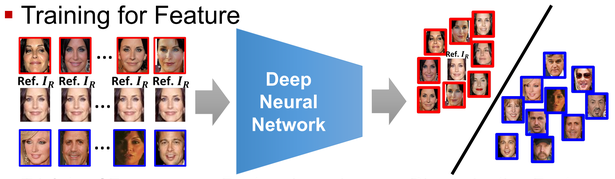
IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 HONOLULU 2017

## 1. Objective

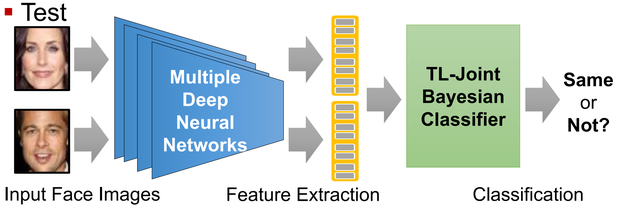
- Improve the accuracy of face recognition when the amount of training data is insufficient to train deep neural network models.

## 2. Overview

- Training for Feature




- Test




## 3. Discriminative Feature Learning

- Triplet of Face for Deep Neural Network Learning
  - Given a face $I_r$, triplet consists of a form of $(I_r, I_p, I_n)$
  - Positive face $I_p$: face image with a same identity
  - Negative face $I_n$: face images with different identity
- Loss Functions for Feature Learning

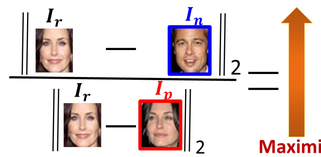
$$L_{total} = L_{triplets} + L_{pairs} + L_{identity}$$

$$L_{triplets} = max\left(0, 1 - \frac{\|F(I_r) - F(I_n)\|_2}{\|F(I_r) - F(I_p)\|_2 + m}\right)$$

$$L_{pairs} = \|F(I_r) - F(I_p)\|_2^2$$

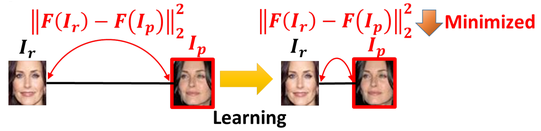
$$L_{identity} = -\sum_{i=1}^{m} \log \frac{e^{F(I^i)}}{\sum_{j=1}^{n} e^{F(I^j)}}$$

1) $L_{triplets}$




&

2) $L_{pairs}$



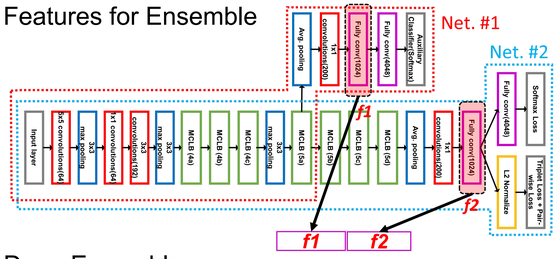

&

3) $L_{identity}$

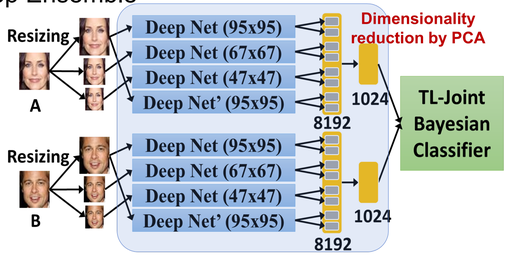
- Use negative log-likelihood loss with ***Softmax.***
- Reflect characteristics for each identity.
- Encourage the separability of features.

## 4. Description using Deep Ensemble

- Features for Ensemble



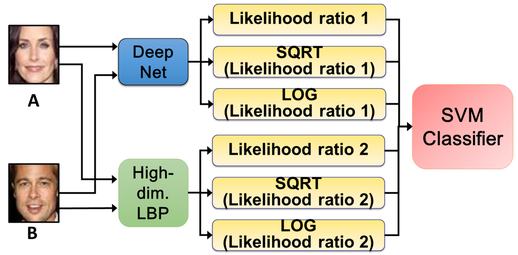

- Deep Ensemble




*Feature extraction from multiple Neural Networks and deep ensemble generation*

## 5. Fusion

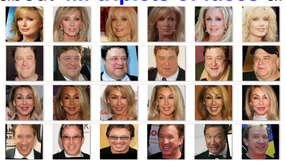
- Score-level Fusion




- Use 6 different types of similarities of DCNN ensemble and high-dim. LBP as features.
- Use Support Vector Machine (SVM) as a classifier (recognizer)

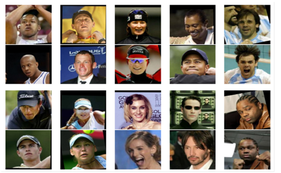
## 6. Experiments

- Training Data
  - **4,048 subjects** with more than 10 images (**198,018**).
  - 396,036 images (horizontal flipped) are used to generate about ***4M* triplets of faces** and train DCNNs.



$T = ($Ref. $I_r$, $I_p$, $I_n$$)$

- Test Data – LFW (Labeled Faces in the Wild)
  - Each of **10 folders** consists of **300 intra pairs** and **300 extra pairs** (Total: 6,000 pairs).
  - 10-fold cross validation



- Results of loss functions on validation set

| | Accuracy (%) | Error reduce |
|---|---|---|
| DNN+$L_{identity}$ (baseline) | 88.17 | - |
| DNN+$L_{triplet}$ + $L_{identity}$ | 91.32 | 26.62% |
| ***DNN+$L_{triplet}$+$L_{pairs}$+$L_{identity}$*** | ***93.45*** | ***44.63%*** |

- Comparison of No. of images, No. of DNNs, feature dimensionality, and accuracy

| Method | No. of images | No. of DNNs | Feature dim. | Accuracy (%) |
|---|---|---|---|---|
| *Human* | - | - | - | *97.53* |
| *Joint Bayesian* | *99,773* | - | *8,000* | *92.42* |
| *Fisher vector face* | *N/A* | - | *256* | *93.03* |
| *Tom-vs-Pete classifier* | *20,639* | - | *5,000* | *93.30* |
| *High-dim. LBP* | *99,773* | - | *2,000* | *95.17* |
| *TL-Joint Bayesian* | *99,773* | - | *2,000* | *96.23* |
| *DeepFace* | *4M* | *9* | *4,096 x 4* | *97.25* |
| *DeepID* | *202,599* | *120* | *150 (PCA)* | *97.45* |
| *DeepID3* | *300,000* | *50* | *300 x 100* | *99.53* |
| *FaceNet* | *200M* | *1* | *128* | *99.63* |
| *Learning from Scratch* | *494,414* | *2* | *320* | *97.73* |
| ***Proposed Method (+Joint Bayesian)*** | ***198,018*** | ***4*** | ***1,024 (PCA)*** | ***96.23*** |
| ***Proposed Method (+TL-Joint Bayesian)*** | ***198,018*** | ***4*** | ***1,024 (PCA)*** | ***98.33*** |
| ***Proposed Method (Score-level Fusion)*** | ***198,018*** | ***4*** | ***6*** | ***99.08*** |

## 7. Conclusion

- Proposed **Loss functions** to learn a discriminative feature is **effective**.
- **The proposed method is more efficient**
  - **Small number of data** – only 198,018 images
  - **Only 4 different** deep network models are used
  - Accuracy: **99.08%** (Score-level Fusion)
- The proposed method is useful when **the amount of training data is insufficient** to train DCNNs.