IEEE 2017 Conference on Computer Vision and Pattern Recognition

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

Bong-Nam Kang*, Yonghyun Kim[†], and Daijin Kim[†]
[*]Dept. of Creative IT Engineering,
[†]Dept. of Computer Science & Engineering
{bnkang, gkyh0805, dkim}@postech.ac.kr

POSTECH

Introduction

- **Face recognition in unconstrained environments is very challenging problem.**
  - Inter-class variations

Same person?

Jennifer Carpenter / Hugh Grant

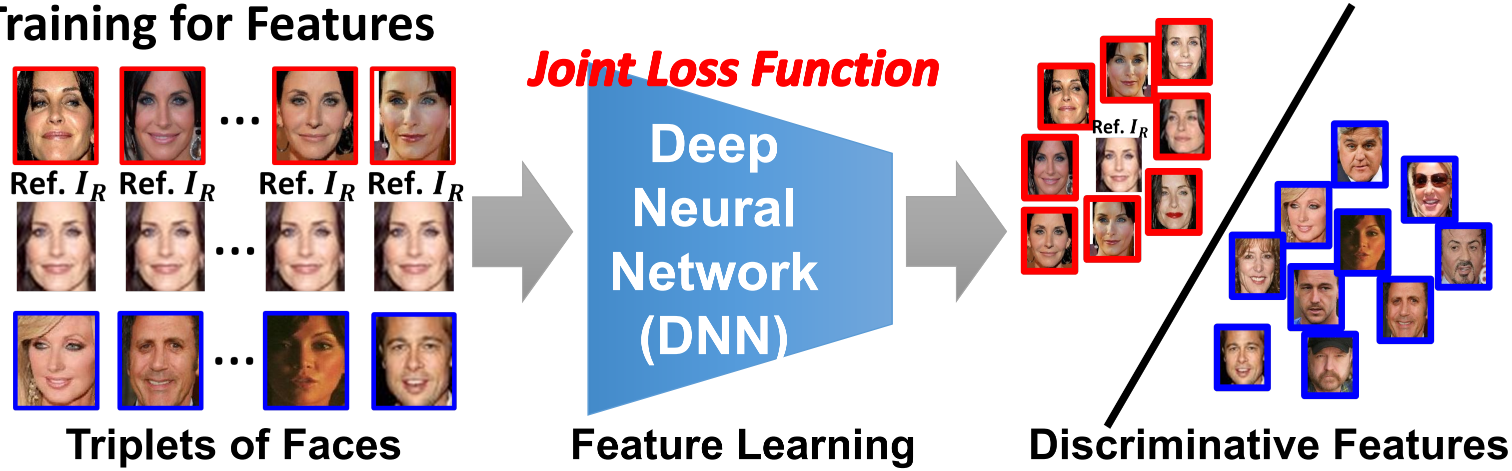
Phil Mickelson / Hope Solo

    - Facial poses, expressions, and illumination changes cause problems to misidentify faces of different identities as the same identity.
  - Intra-class variations

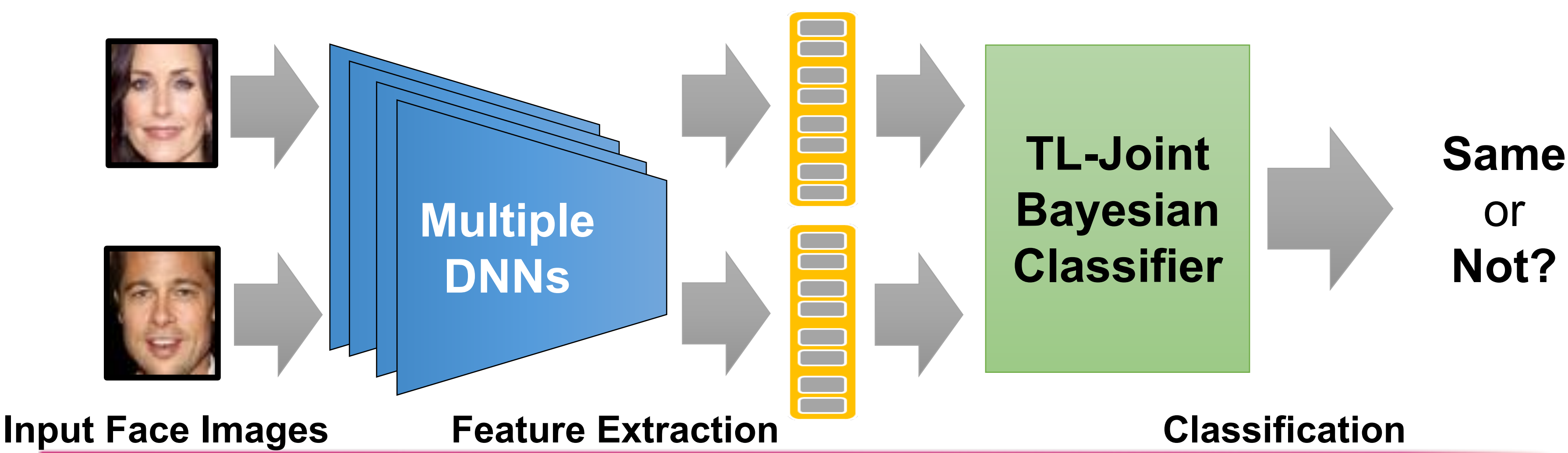
    - Such variations within the same identity could overwhelm the variations due to identity differences and make face recognition challenging.

Overview

- Training for Features



- Test



Discriminative Feature Learning (1/2)

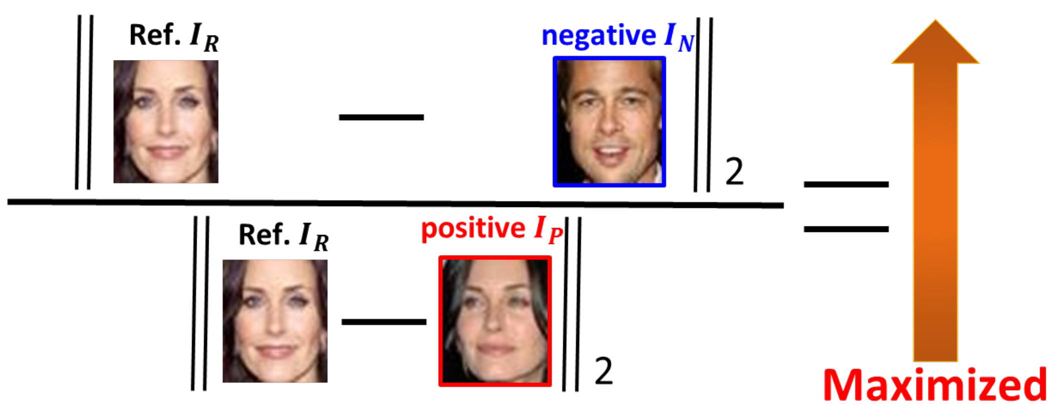
- **Joint Loss Function for Feature Learning**

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

- **Triplet Loss $L_{triplets}$**

$$L_{triplets} = max\left(0, 1 - \frac{\|F(I_R) - F(I_N)\|_2}{\|F(I_R) - F(I_P)\|_2 + m}\right)$$

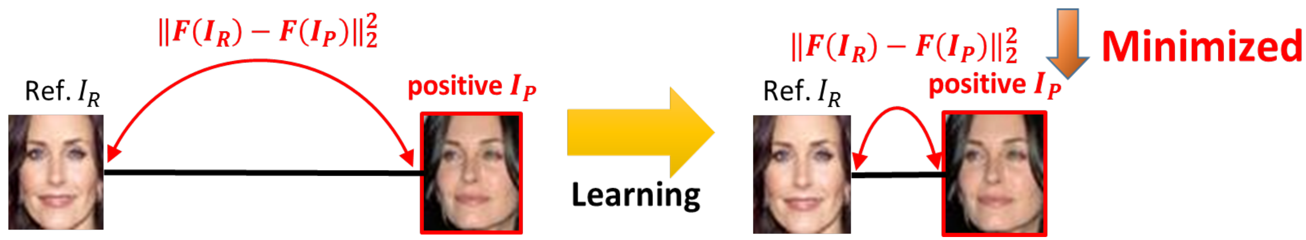
  - The output of network is represented by $F(I) \in R^d$.
  - m is a margin: define the minimum ratio between the negative pairs and the positive pairs in the Euclidean space



Discriminative Feature Learning (2/2)

- **Pairwise Loss L_{pairs}**
  - Minimize the absolute distances between the positive data in the triplets T.

$$L_{pairs} = \sum_{(I_R, I_P) \in \forall T} \| F(I_R) - F(I_P) \|_2^2$$



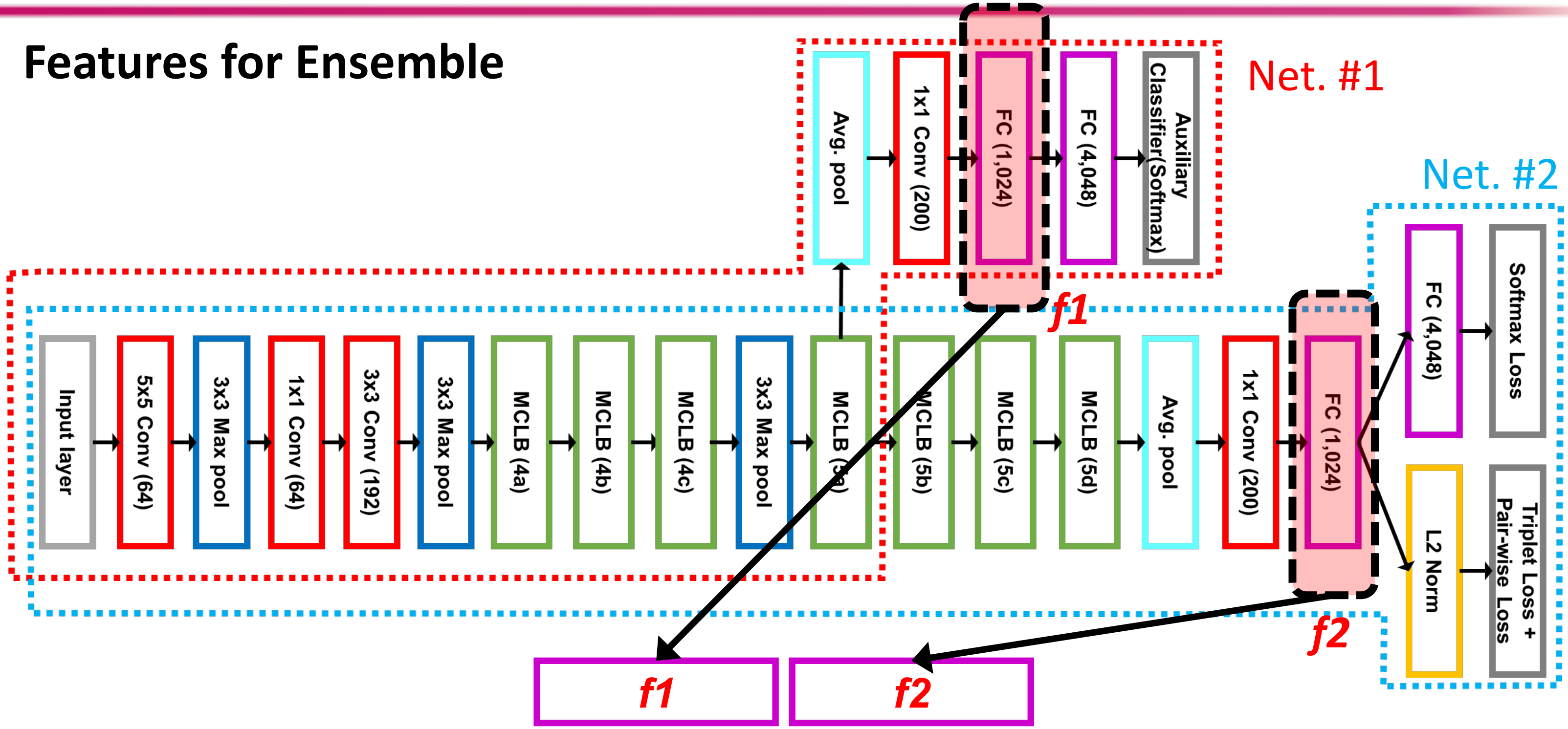
- **Identity loss $L_{identity}$**

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

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

Description using Deep Ensemble (1/2)

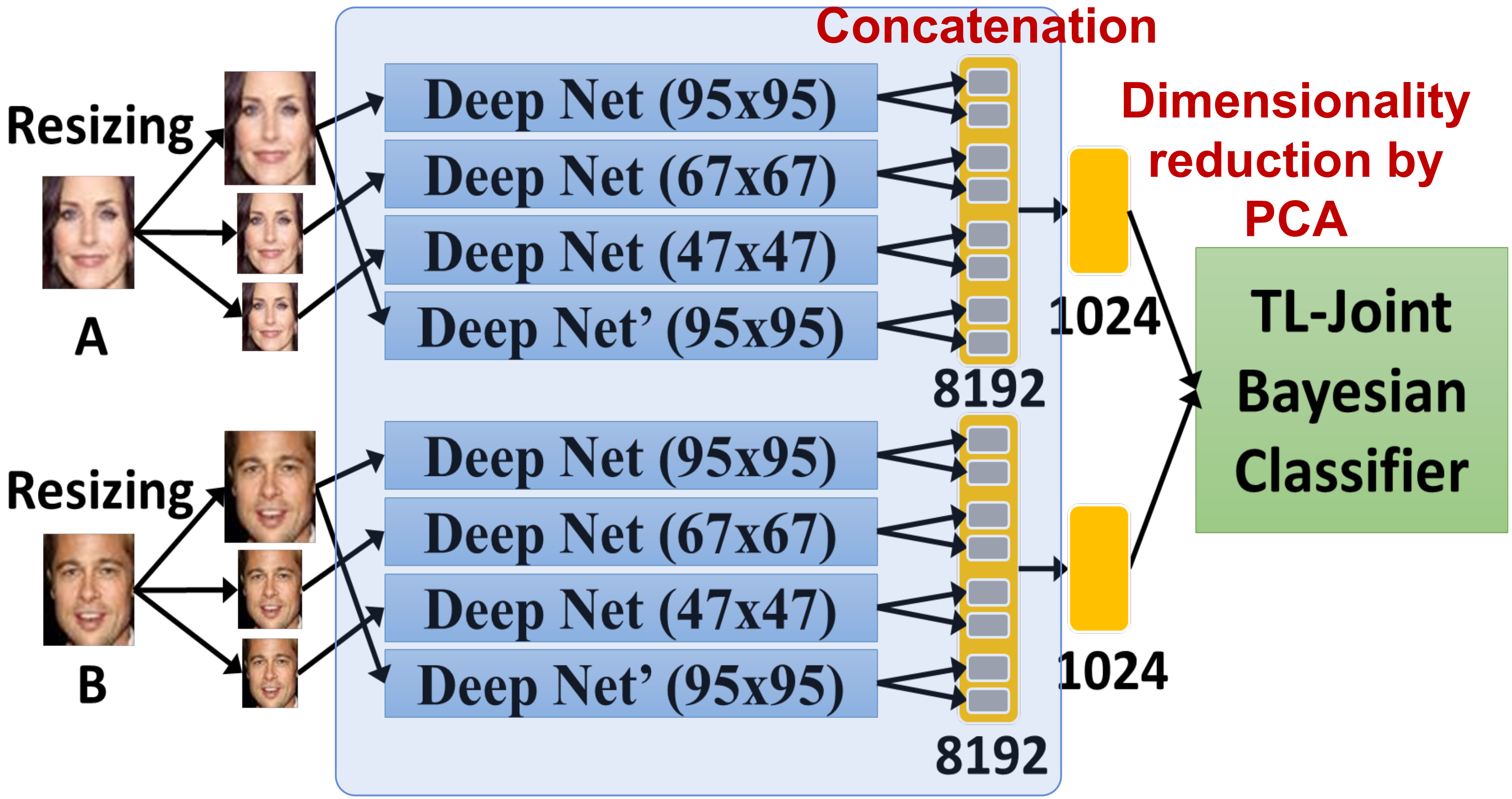
- **Features for Ensemble**



  - In conventional applications of DCNN, the output of the last fully connected layer ***f2*** is used only as a feature.
  - Use DNN features taken from ***f1*** and ***f2*** fully connected layers.
    ➔ Multi-scale feature effect

Description using Deep Ensemble (2/2)

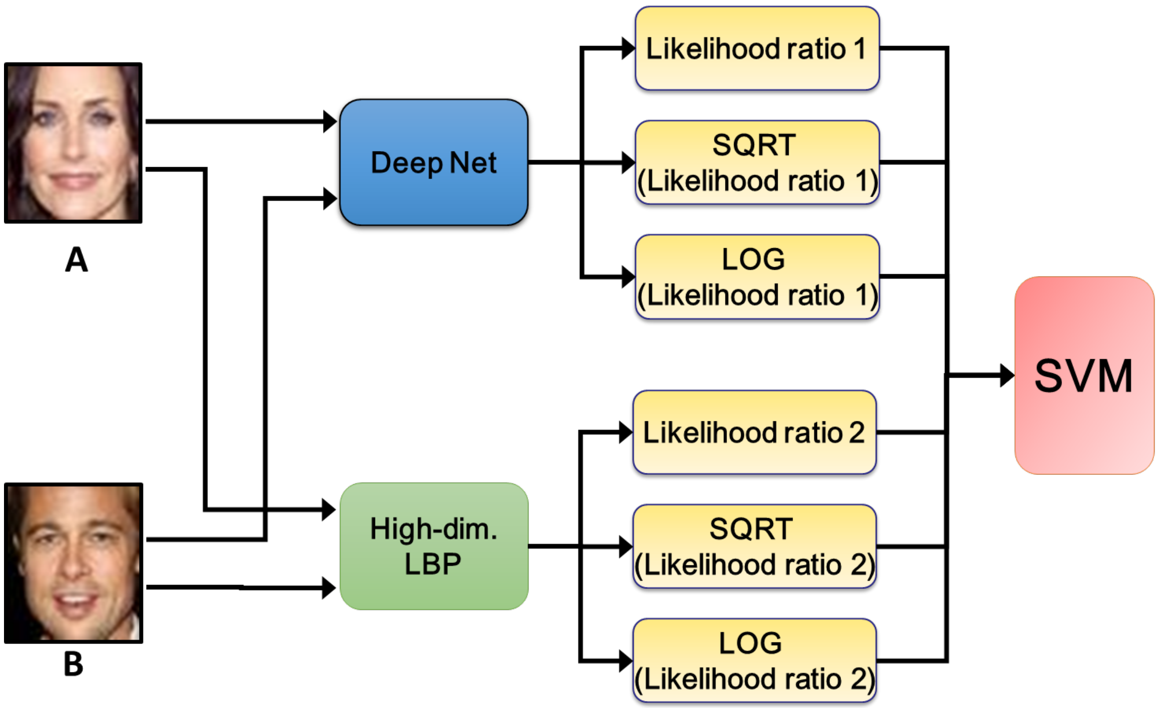
- **Deep Ensemble**



Feature extraction from multiple Neural Networks and deep ensemble generation

Fusion

- **Score-level fusion**



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

Experimental Results (1/2)

- **Training data**
  - 4,048 subjects with more than equal 10 images (198,018).
  - 396,036 face images (horizontal flipped) are used to generate about ***4M* triplets** of faces for training.

- **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.

Experimental Results (2/2)

- **Results on LFW**

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 (Fusion)	***198,018***	***4***	***6***	***99.08***

Discussion & Conclusion

- Joint Loss Function to learn a discriminative feature is effective

- The proposed method is more efficient
  - Small number of data - only 198,018 training images
  - Only 4 different deep network models used
  - Accuracy: 99.08% (Score-level Fusion)

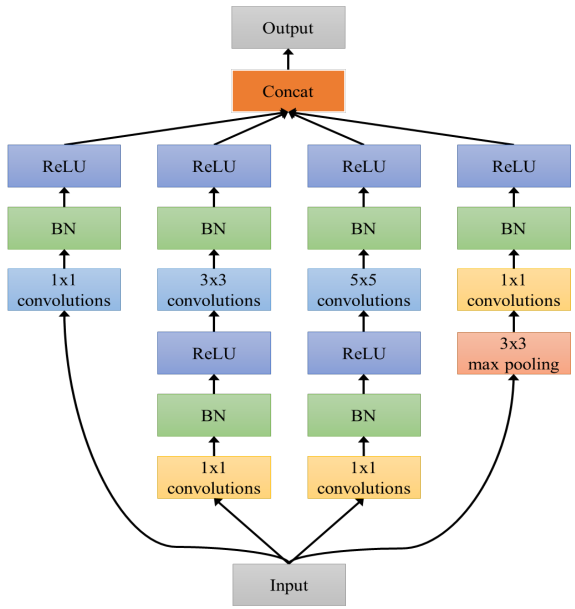
- The proposed method is useful when
  - The amount of training data is insufficient to train deep neural networks.

Thank you !!!

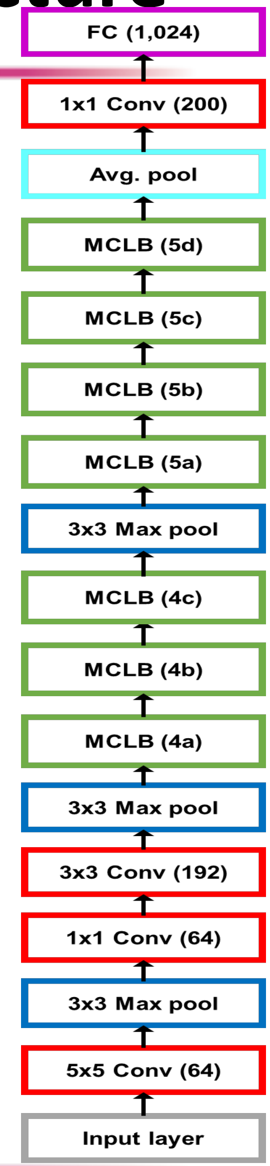
Appendix

Deep Convolution Neural Network Architecture

- **Multi-scale Convolution Layer Block (*MCLB*)**
  - Consists of 1x1, 3x3, 5x5 convolution layers, and 3x3 max pooling layer.



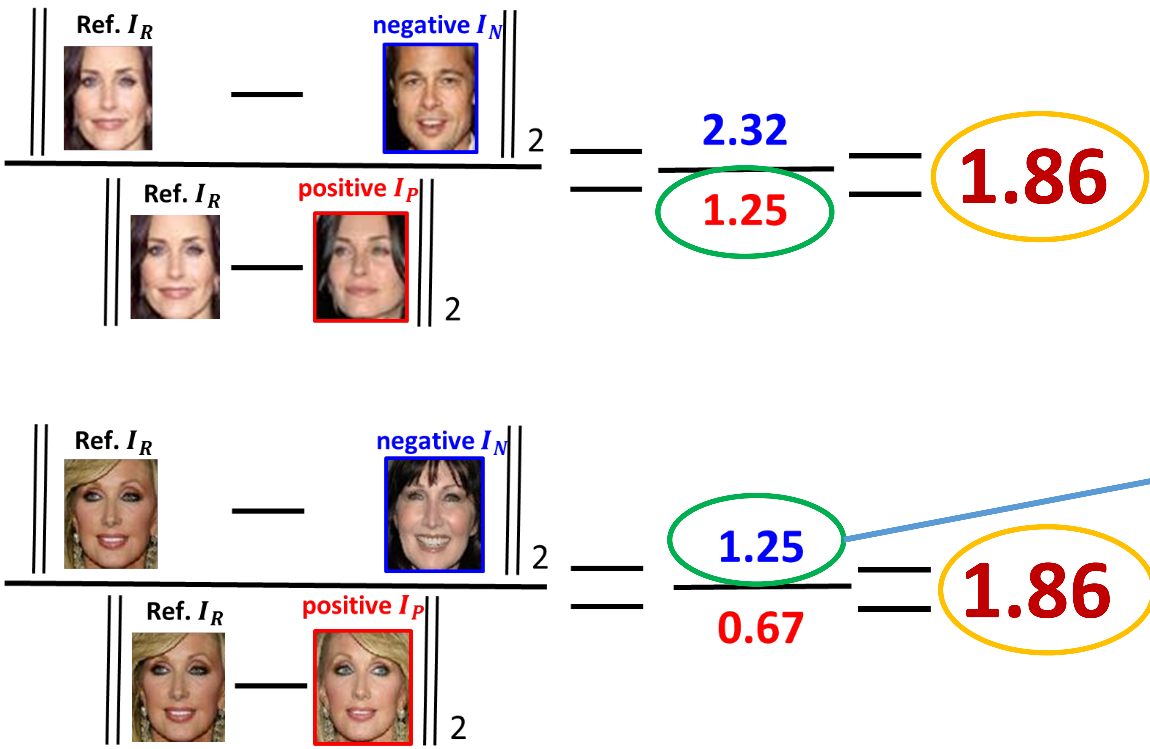
- **Structure of Deep Convolution Neural Network**
  - Constructed by stacking MCLBs on top of each other (24 layers deep).
  - All convolutions and fully connected layers use the ReLU non-linear activation.
  - Average pooling takes the average of each feature map and sums out the spatial information.
  - Dropout is only applied to the last fully connected layer for regularization.



POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Discriminative Feature Learning

- After training with only triplet loss, we observed that the range of distances between each pair data was not within the certain range.
  - Although the ratio of the distances was within the certain range, the range of the absolute distances was not within the certain range.



Ratios of distances are same.

If distance ≤ 1.25, positive (same identity).

These two image are with same identity. Acceptable?

Experimental Results

- **Results of Joint Loss Function on Validation Set**
  - 55,747 face images are used as a validation set.

	Accuracy (%)	Error reduction
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%