Departamento de Informática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

Hybrid CNN+LSTM for Face Recognition in Videos

Proyecto de Tesis

Magister en Ciencias de la Ingeniería Informática

Alumno: Marco Bellantonio
Supervisor: Prof. Ricardo Nañculef
Co-Supervisor: Prof. Sergio Escalera

Valparaiso, 16/12/2016

Introduction

In the past 30 years:

- Powerful and low-cost computers.
- *high-performing* embedded computing systems.

↓

Automatic processing of digital images

↓

Automatic Face Analysis

Surveillance

Biometric Authentication

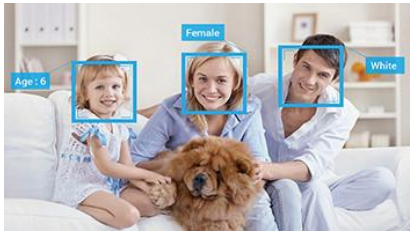
Human-Computer Interaction

Multimedia Management

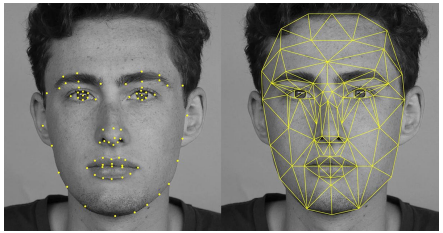
Introduction

Automatic Face Analysis

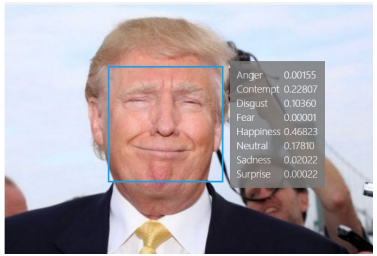
Face Detection & Tracking



Automatic Face Recognition



Expression/Emotion Recognition



Automatic Face Recognition

"*A facial recognition system is a computer application capable of identifying or verifying a person from a digital image or a video source.*"

Input type in computer vision:

- Raw images
- Videos
- Depth map
- Thermal images
- 3D Face models
- ...

Aim of this project

Design deep learning models tailored to exploit the temporal information contained in videos to perform video face recognition.

Overview:

- Exhaustive **review of recent papers** and **works** in the field of computer vision related to deep models for face recognition in videos.

  Lack of temporal models for video face recognition

- **Analysis** of the **most recent** and **efficient methods** along with the study of the performances reported and the databases used.

- **Definition of the architectures** involved, namely Convolution Neural Network and Long-Short Term Memory.

- **Choice** of the **dataset**. A **novel database** for video face recognition is also presented.

- Design of the experiments and conclusions.

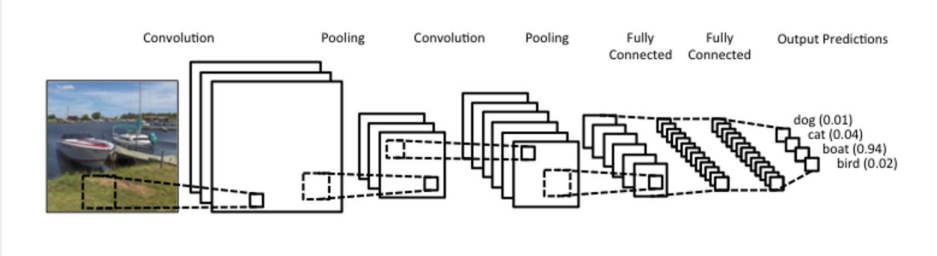
Spatial Models

Convolutional Neural Network

CNN are biologically-inspired variants of Multi Layer Perceptrons proposed by Yann LeCun in 1998.

Input: x_i is a $N \times N$ image
m: size of the convolutional kernel
w_{ab}: weights of the filter
σ: non-linear function
y_{ij}: output of the convolution

$$y_{ij} = \sigma\left(\sum_{a=0}^{m-1}\sum_{b=0}^{m-1} \omega_{ab} x_{(i+a)(j+b)}\right)$$

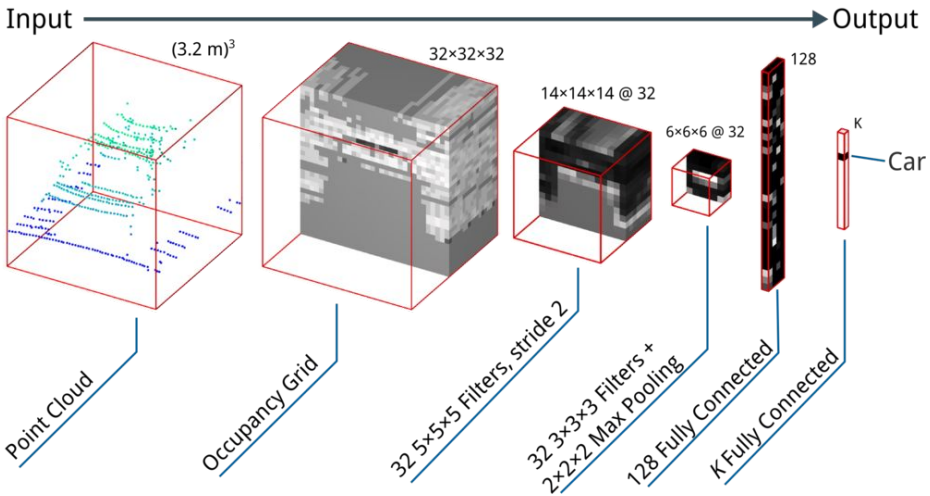


Convolutional layers are characterized by a set of learnable *filters* (kernels). Each filter is convolved across width and height of the input, producing a 2D activation map.

Pooling layers progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network.

Temporal model for Image processing

3D Convolutional Neural Network



Extension of the convolution along the **temporal axis**

↓

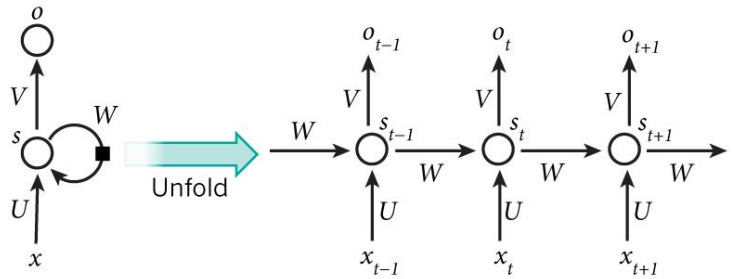
Input is a volume.

3D-CNN captures discriminative features along both spatial and temporal dimensions.

Not a “standard” temporal model!

Temporal Models

Recurrent Neural Network



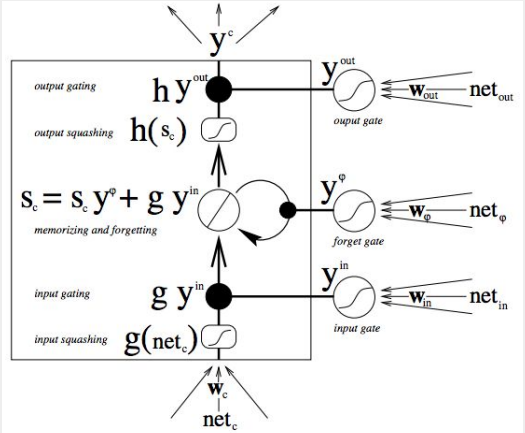
RNN is a class of artificial neural network where connections between units form a directed cycle.

$$S_t = f(W_x X_t + W_r S_{t-1})$$

Trained using Backpropagation Through Time (BBPT)

$$\frac{\partial \xi}{\partial S_{t-1}} = \frac{\partial \xi}{\partial S_t} \cdot \frac{\partial S_t}{\partial S_{t-1}} = \frac{\partial \xi}{\partial S_t} \cdot w_r$$

Long-Short Term Memory

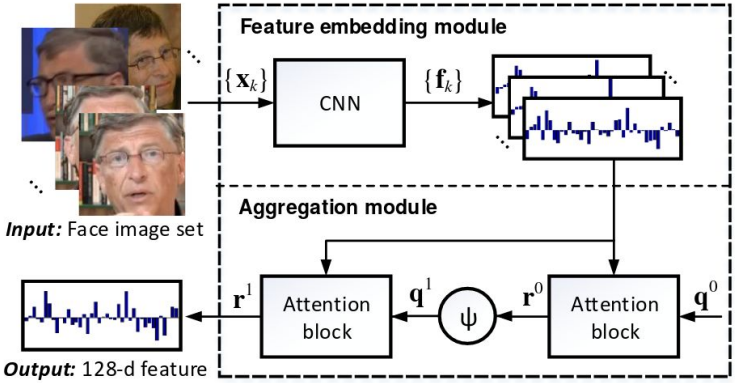


In LSTMs, information can be stored in, written to, or read from a **cell**.

The cell makes decisions about when to allow reads, writes and erasures via gates that open and close.
Those gates are called **input** gate, **forget** gate and **output** gate.

Deep Learning Methods Review - Spatial

Neural Aggregation Network for video face recognition



$\{\mathbf{x}_k\}$: input faces images.

$\{\mathbf{f}_k\}$: a set of feature representations.

$\mathbf{r}^1$: 128-dimensional vector representation for the input video faces.

Jiaolong Yang, Peiran Ren, Dong Chen, Fang Wen, Hongdong Li, Gang Hua

Inputs: face video or face image set of a person.

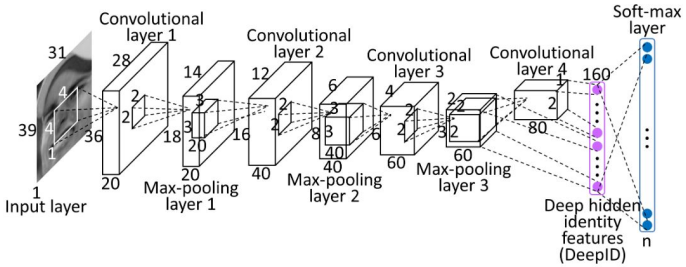
Output: compact and fixed dimension visual representation of that person.

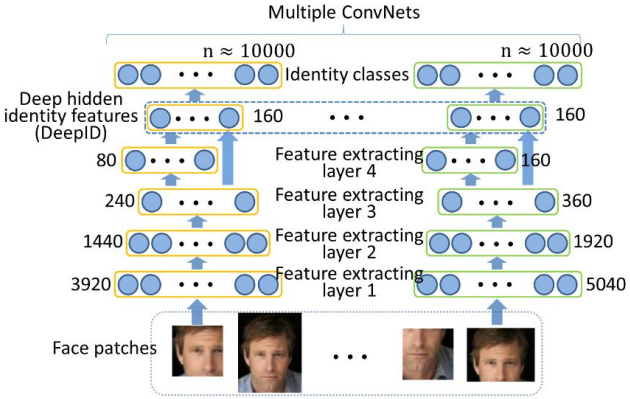
The whole network is composed of two modules:

1. **Feature embedding module**: a CNN which maps each face frame into a feature representation.
2. **Neural aggregation module**: two content-based attention blocks which are driven by a memory storing all the features extracted from the face video through the feature embedding module.
   The output of the first attention block adapts the second, whose output is adopted as the aggregated representation of the video faces.

Deep Learning Methods Review - Spatial

Deep Learning Face Representation from Predicting 10,000 Classes





Yi Sun, Xiaogang Wang, Xiaoou Tang

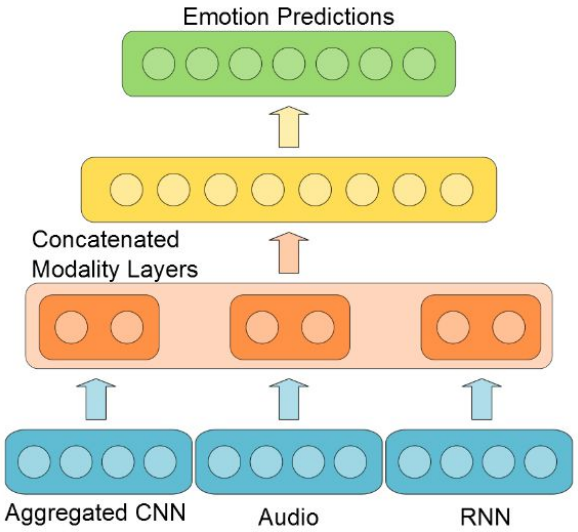
High-level feature representation extracted with a very deep CNN called *Deep hidden IDentity feature* (**DeepID**).

Feature are taken **from last hidden layer** neurons activation of the CNN and are extracted from various face regions.

Fusion of multiple CNNs achieve **97.45%** accuracy against LFW dataset.

Deep Learning Methods Review - Temporal

Recurrent Neural Networks for Emotion Recognition in Video



Hybrid CNN-RNN architecture for facial expression analysis and emotion recognition in videos.

Fusion of different modalities.

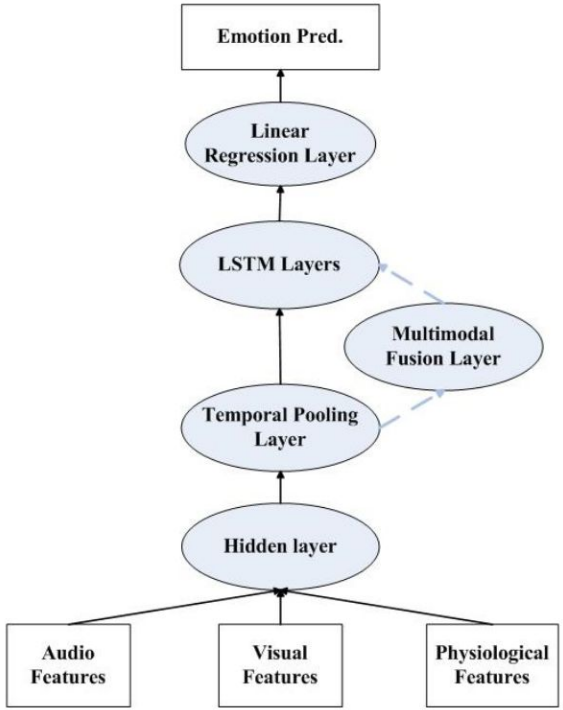
Aggregated CNN:

- Extract feature using a CNN
- Train a RNN to classify a video by feeding the features for each frame from the CNN sequentially to the network and using the last time-step softmax output as class prediction.

Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, Christopher Pal

Deep Learning Methods Review - Temporal

Long-Short Term Memory Recurrent Neural Network Based Multimodal Dimensional Emotion Recognition



Multi-modal (Audio-Visual-Physiology) approach to dimensional emotion recognition with a **LSTM-RNN** architecture.

Investigation of **ε-insensitive loss function L_ε**, instead of squared loss $(y - f(x, w))^2$.

$$L_\varepsilon = \begin{cases} 0 & if \ |y - f(x,w)| \leq \varepsilon \\ |y - f(x,w)| - \varepsilon & \text{otherwise} \end{cases}$$

ε-insensitive loss function is more **robust** to label noise and can ignore small errors to get stronger correlation between predictions and labels.

Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, Zhengqi Wen

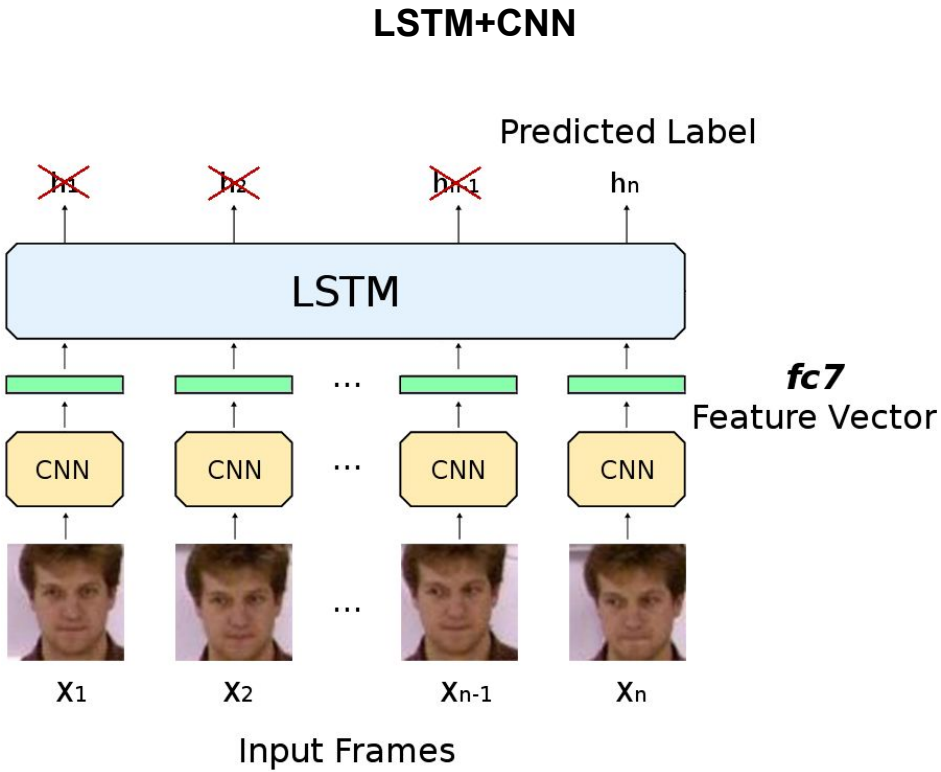
Lack of methods which use temporal models (RNN-LSTM) to perform Face Recognition <u>in videos</u>!

Moreover

“...spatio-temporal evolution of facial features is one of the strongest cues for emotion recognition...”

From “*Recurrent Neural Networks for Emotion Recognition in Video*”

The Proposed Method

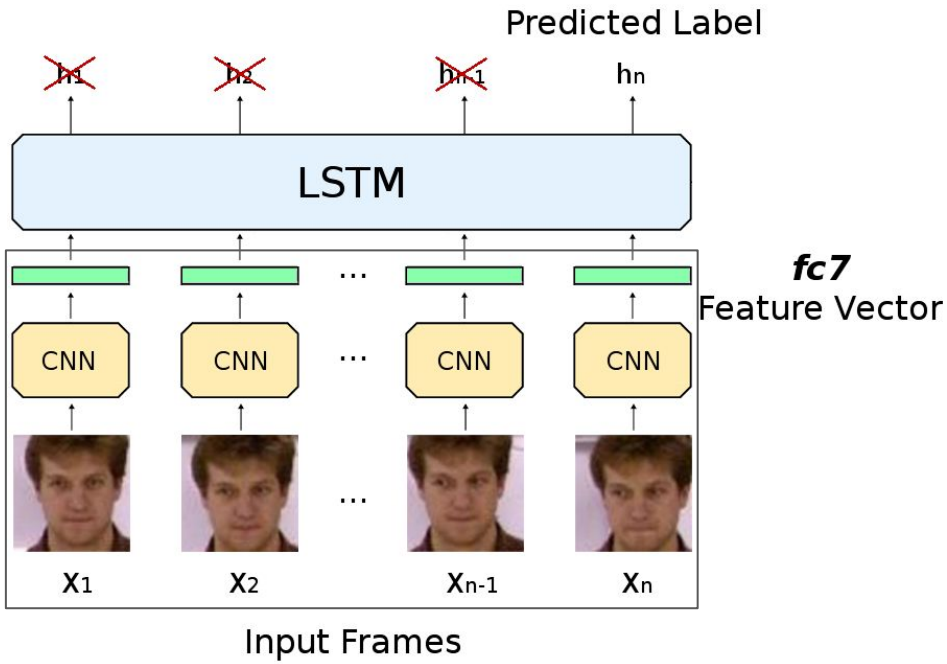


Combination of CNN and RNN for a **hybrid** framework to exploit both spatial and temporal information of face features for video face recognition.

Features are extracted via CNN and fed into a LSTM for prediction.

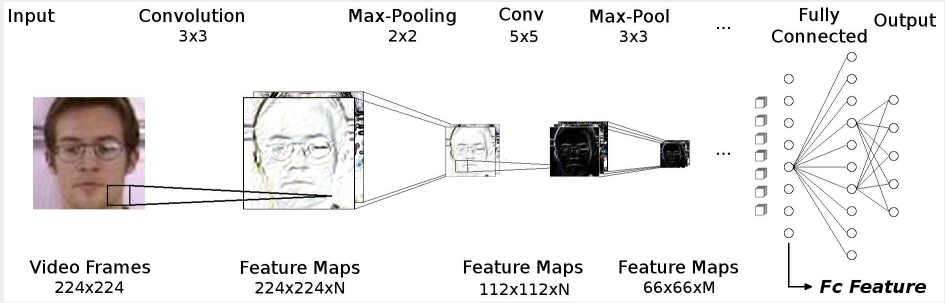
The Proposed Method

LSTM-on-CNN



fc7
Feature Vector

CNN



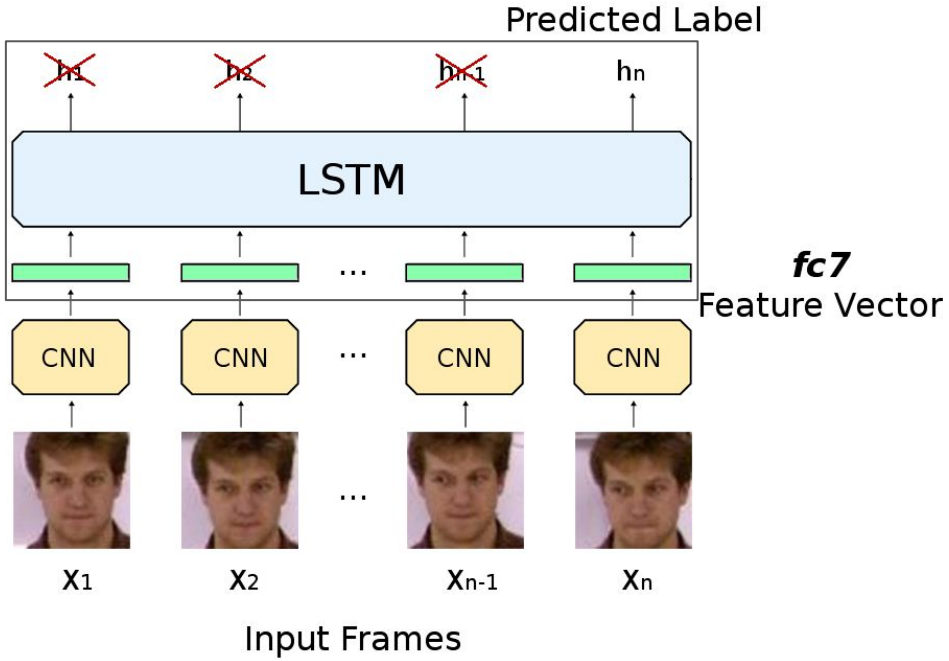
Input: X_i - $N \times N$ pixel's matrix.

Output: feature vector f_i extracted from one of the last fully connected layer.

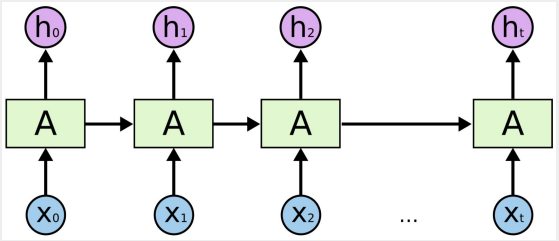
In the VGG-16 network, for instance, the layer usually used for feature extraction is the 7th fully-connected layer, called ***fc7***.

The Proposed Method

LSTM-on-CNN



LSTM



Input: f_j Feature vectors from the last CNN fully connected layer

Output: h_n LSTM prediction

Labels are predicted sequence-wise.
Given a sequence of n frames $X_i \in \{X_1, ..., X_n\}$, the target prediction is the face identity of the **last X_n** frame.

The **temporal window** defines the number of consecutive frames that have to be taken into account when predicting a target frame. Therefore the output of the LSTM is the **last frame of a defined temporal window.**

Architecture - CNN

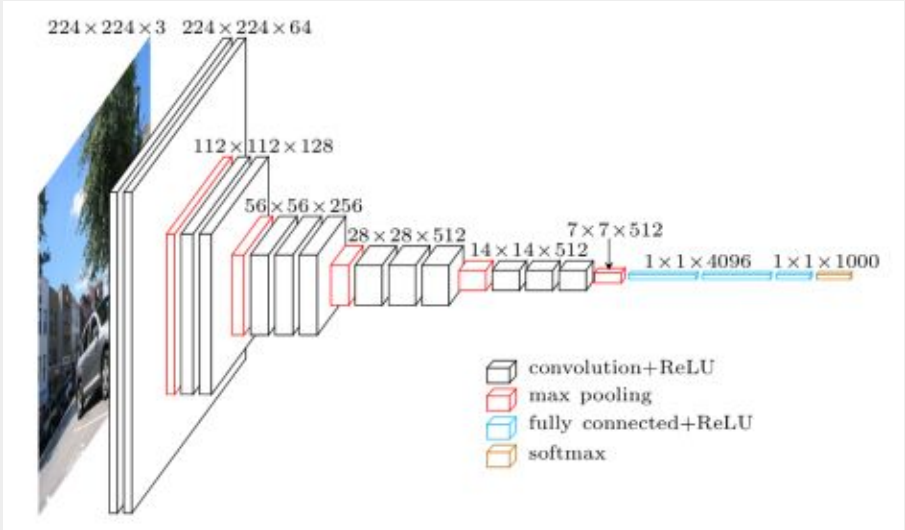
VGG-Very-Deep-16 CNN
Pre-trained Model

CNN architecture:

- **VGG-Very-Deep-16 CNN** (VGG-16).
- Pre-trained model from *Caffe ModelZoo* (*i*mported to work with Caffe library).
- Model trained from scratch using **2.6 Million images** of celebrities collected from the web (*VGG FACE*).

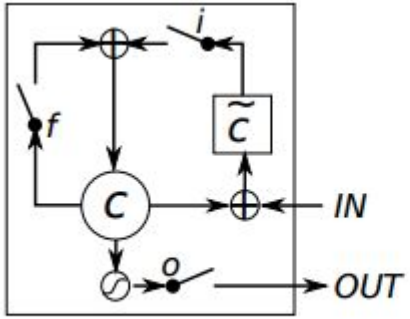
Specification:

- Input images size: 244 × 244
- Output: 226 classes
- ***fc7*** fully connected layer dimensionality: **4096**.



Architecture - LSTM

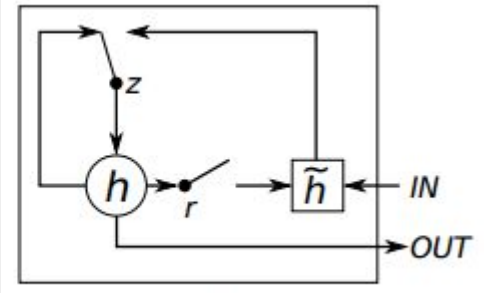
Long-Short Term Memory - LSTM



Input gate - **Forget** gate - **Output** gate

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \\
c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

Gated Recurrent Units - GRU



- **Forget** (*f*) + **Input** (*i*) = **Update** (*z*)
- *r* = **Reset** gate.
- Sometimes cell state and hidden state are also merged.

$$
\begin{aligned}
z_t &= \sigma\left(W_z \cdot [h_{t-1}, x_t]\right) \\
r_t &= \sigma\left(W_r \cdot [h_{t-1}, x_t]\right) \\
\tilde{h}_t &= \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right) \\
h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t
\end{aligned}
$$

Video Face Databases

Chosen Datasets:

- YouTube Face (YTF)
- CMU Motion of Body (MoBo)
- UNBC-McMaster Shoulder Pain

Other non-video face databases:

Labeled Faces in the Wild (LFW), IARPA Janus, Benchmark A(IJB-A), PaSC, Oxford Buffy db, ScFace, CMU-FIA, CameFace, Face96, MBGC, ND-Flip-QO, UMD ComCast10, ESOGU Face Videos, MAHNOB-HCI, MMSE-HR, Trailed Face Dataset.

Database	Year	Modalities	Details	Evaluation Metric
Celebrity 1000 (C1000)	2014	RGBv, face region, facial landmark	159726 videos 1000 subjects	os/cs protocol
Chokepoint	2011	RGBv, RGBi,	48 videos 54 subjects	V2V
CMU Motion of Body (MoBo)	2001	RGBi	600 videos 24 subjects	-
COX Face	2015	RGBi, RGBv	3000 videos 1000 subjects	V2V, V2S, S2V
Honda/UCSD	2005	B/Wv	75 videos 20 subjects	-
MOBIO	2010	Audio, RGBv	1824 a/v 152 subjects	-
PaSC	2013	RGBi, RGBv	2802 videos 293 subjects	S2S, V2V, S2V
UNBC-McMaster Shoulder Pain	2011	RGBi, FACs, AAMs	200 videos 25 subjects	S2S, V2V, S2V
vidTIMIT	2003	Audio, RGBv	430 a/v 43 subjects	-
WebV-Cele	2009	RGBv, coord, SIFT, CH	75073 videos 2427 subjects	-
YouTube Celebrities	2008	RGBv, BB	1910 videos 47 subjects	-
YouTube Face Dataset (YTF)	2011	RGBv Hand Pos	3425 videos 1595 subject	10-fold CV Pair-Match

YouTube Faces (YTF)

Database:

- Actions performed are naturally varied.
- Easier to acquire, thus allowing the baselines to be used by the research community at large.
- All subjects also have still images available in the Labeled Faces in the Wild (LFW) database, thus allowing baselines to be compared to the video to still image matching scenario.
- Low image quality: frames sequences of YouTube videos are generally worse than web photos, mainly because of motion blur or viewing distance.

Paper	Protocol	Metric	Result
DeepID2+ [44]	Standard protocol	ACC	93.2% (VR) 95% (IR)
DeepFace [47]	Standard protocol (unrestricted)	ACC	91.4% (CR)
		100%-EER	92.5%
Eigen-PEP for video face recognition [31]	Standard protocol	ACC	85.4%
Face Recognition in Movie Trailers via Mean Square Sparse Representation-based Classification [34]	Standard protocol	ACC	75.3%,
		AUC	82.9%
		EER	25.3%
Hierarchical-PEP model for real-world face recognition [29]	Not specifically defined	ACC	87%
MDLFace [16]	3M face images of 50K identities	ACC	97.9%
Neural Aggregations Networks [53]	100 frames for each video	ACC	96.5% (IR)
		AUC	98.7%
Sparsifying Neural Network Connections [45]	Train: 290K faces; Val: 47K faces; Test: 5K pairs of faces	ACC	93.5% (RR)
Unconstrained Face Recognition [7]	Own gallery (YTF+LFW) + fusion	ACC	**79%**

Notes: ACC: Accuracy, AUC: area under the curve, EER: Equal Error Rate, RR: Recognition Rate, IR: Identification
VR: Verification , Rate, CR: Classification Rate

★ 1595 people
★ 3425 videos (average of 2.15 videos for each subject).
★ Video lengths from 48 to 6070 frames (181.3 frames/video)
★ In total ~620,000 frames.

CMU Motion of Body (MoBo)

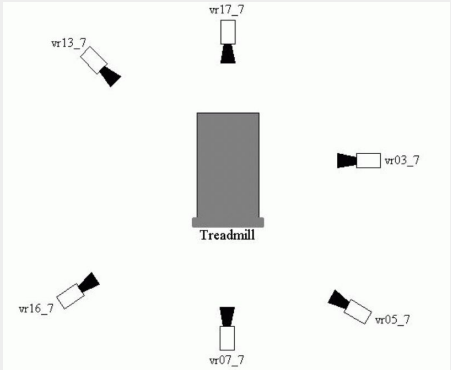
Database:

- 24 individuals.
- 4 activities: *slow walk, fast walk, incline, walk with a ball.*
- 6 high resolution cameras.
- 600 videos, 340 frames each.

Paper	Face Region	Protocol	Accuracy
Towards Large-Scale Face Recognition Based on Videos	-	1 train / 3 test	98.1% (CR)
Learning Personal Specific Facial Dynamics for Face Recognition From Videos	40x40	$\frac{1}{2}$ train / $\frac{1}{2}$ test	97.9%
Joint sparse representation for video-based face recognition	30x30	1 train / 3 test	96.5% (IR)
Face Recognition Based on Image Sets	40x40	1 train / 3 test	95.3, 98.1(CR)
From Still Image to Video-Based Face Recognition: An Experimental Analysis	40x40	1 train / 3 test	92.3% (RR)

Notes: RR: Recognition Rate, IR: Identification Rate, CR: Classification Rat

Cameras

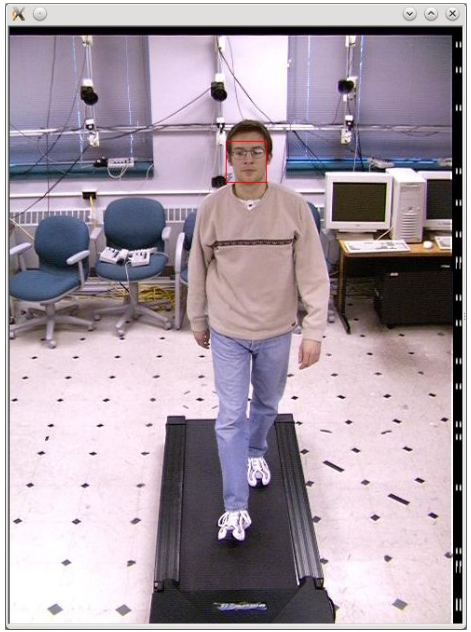


MoBo pre-processing

Face detector: ***dlib***

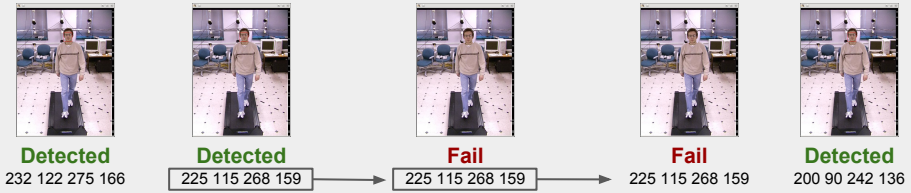
> python face_detector.py im02_19451807.jpg

processing file: im02_19451807.jpg
number of faces detected: 1
detection position left,top,right,bottom: 232 122 275 166



> convert im02_19451807.jpg -crop $position -resize 224x224
im02_19451807_cropped.jpg

dlib failure → *Interpolation*



After 15 missed faces → *dropping.*

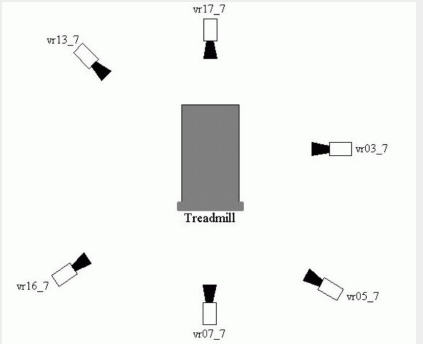
New “MoBo Face Database”

Frontal faces

Inclined faces

Total number of videos: (4 actions × 24 subjects) × 3 cameras = **288 videos**
Total number of frames: 23517 frames × 3 cameras = **70551 frames**

Inclined face camera angle: 45° (left and right)



UNBC-McMaster Shoulder Pain Expression Archive

Pain expression database collected by researchers at McMaster University and University of Northern British Columbia.

Details:

- 200 video sequences.
- 48398 FACS coded frames.

Videos shows patients who were suffering from shoulder pain while they were performing a series of active and passive range-of-motion tests.

¡ **NO** methods which performs face recognition !

Fine-Tuning Pre-trained VGG-16 Model

Fine-tuning takes an already learned model, adapts the architecture and resumes training from the already learned model weights on a different dataset.

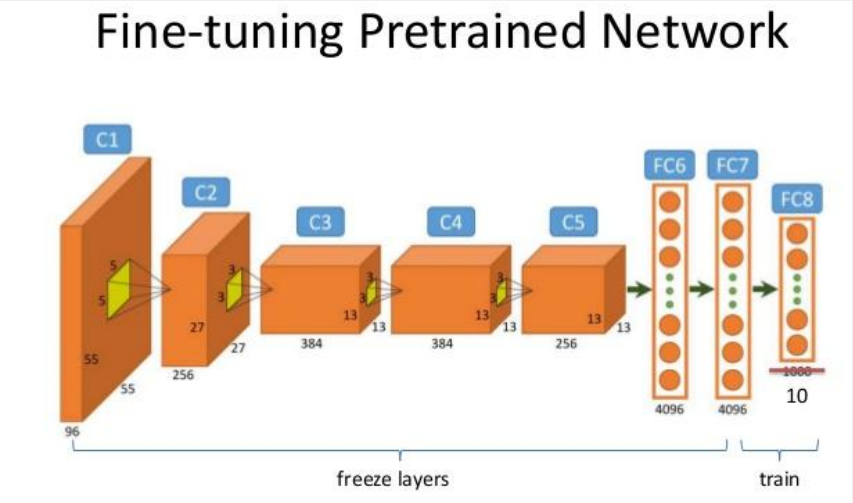
The steps to fine-tune the network are:

1. **Replace** the last layer of the CNN by a randomly initialized fully-connected layer with the correct number of face labels to recognize.

2. **Freeze** all the learning rates of the network.

3. **Set** the **learning rate** of the **fully connected layer** as ten times the learning rate of the rest of the CNN

4. **Set the global learning rate** to one tenth of the original one.

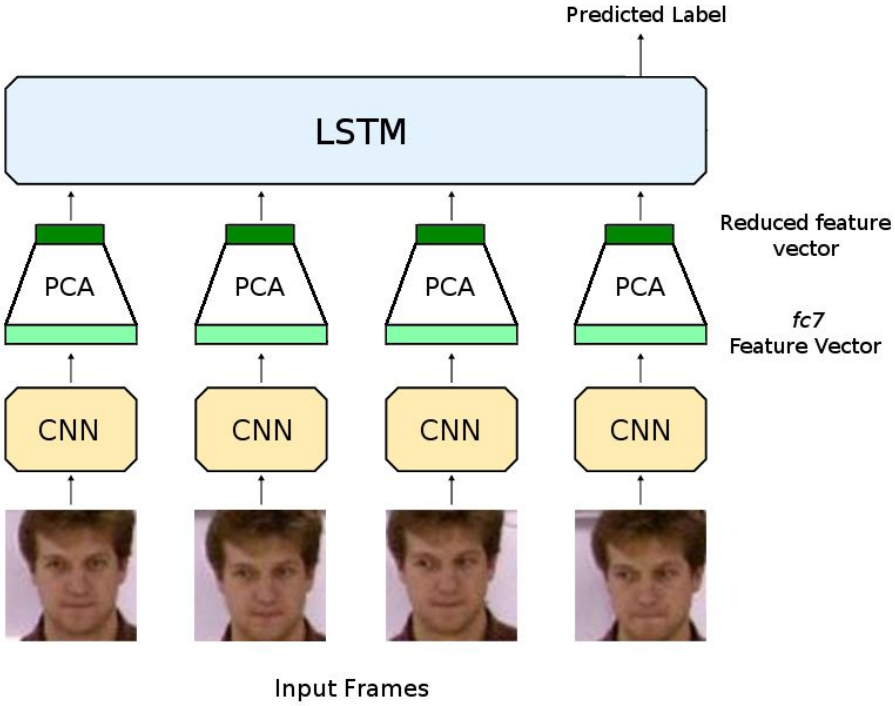


In Caffe:

caffe train --solver=$SOLVER --weights=$CAFFEMODEL

Dimensionality Reduction with Principal Component Analysis

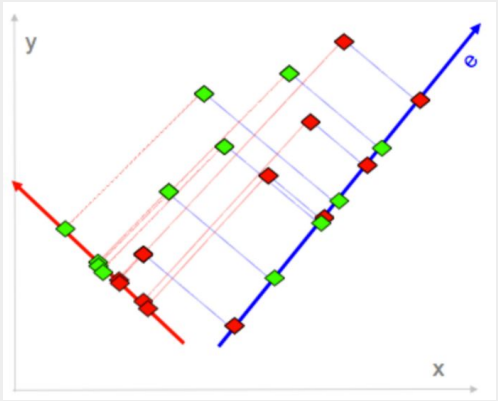
CNN + PCA + LSTM



Multivariate statistical procedure used to identify patterns in high-dimensional data or to reduce dimensionality.

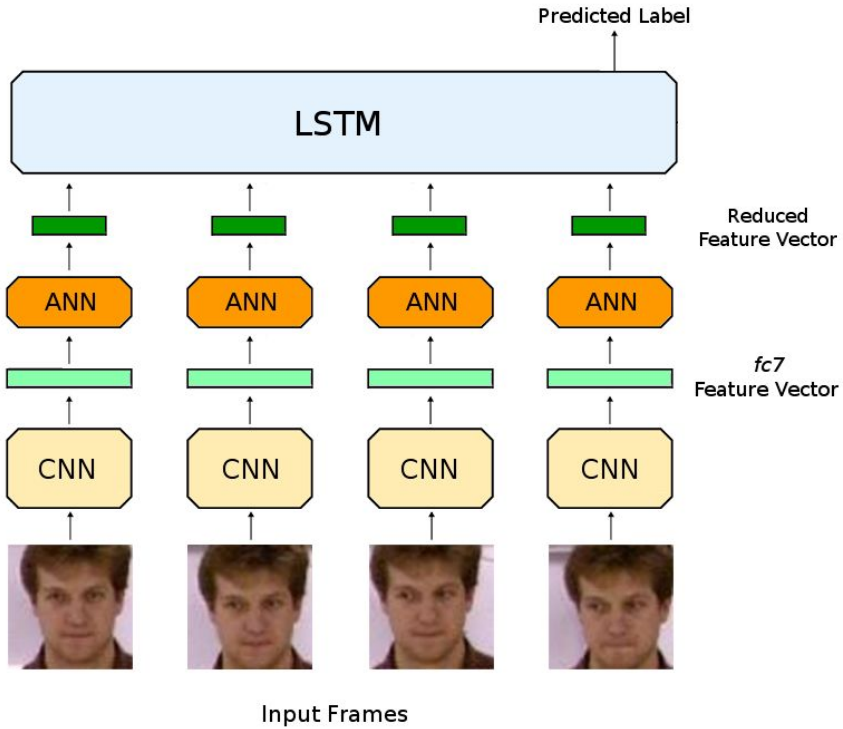
Principal components: new orthogonal coordinate axes along which the data varies the most.

Problem: PCA does not take into account class information when calculating the principal components.

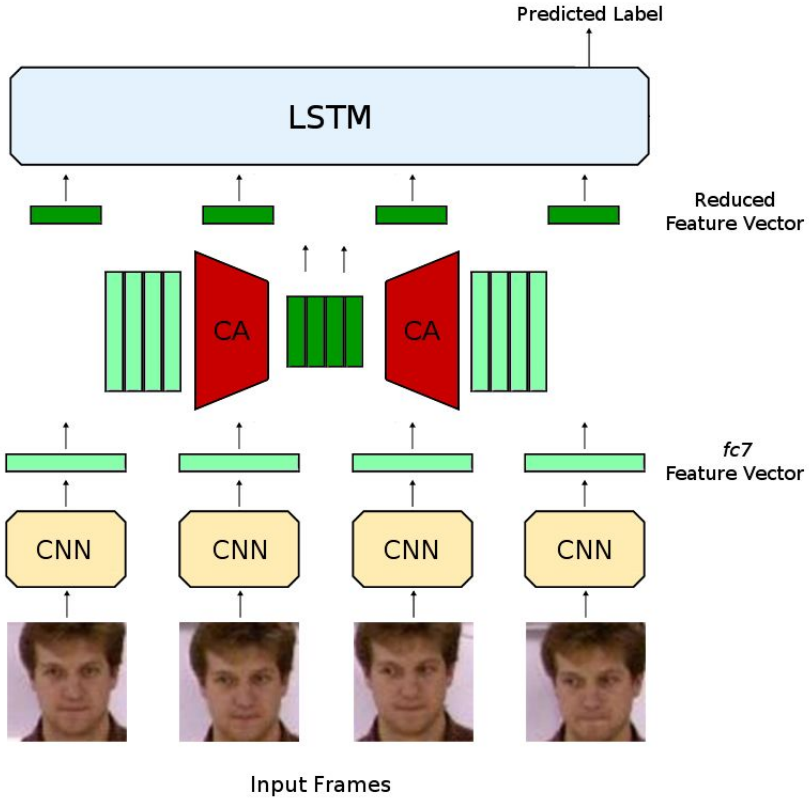


Dimensionality Reduction with Neural Network

CNN + ANN + LSTM



CNN + CA + LSTM



Evaluation metrics

Metric	Definition	Usage
Error Rate	$\frac{\#\text{ of misclassifications}}{\#\text{ samples in val set}}$	General accuracy evaluation
F1 Score	$\frac{2\times\text{true positive}}{(2\times\text{true positive})+\text{false negative}+\text{false positive}}$	Used to give a summary of the Precision-Recall (PR) curve.
ROC / PR curve	$\text{Precision} = \frac{\text{true positive}}{\text{true positive}+\text{false positive}}$ $\text{Recall} = \frac{\text{true positive}}{\text{true positive}+\text{false negative}}$	Used to show the overall performances of an algorithm as its discrimination threshold is varied.

Chosen evaluation metric:

- Accuracy = 1 − Error Rate
- F1 score → Confusion matrix

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Validation strategy

Validation	Definition and Usage
LpO CV	Leave-p-out cross-validation uses p obsarvation as the validation and the remaining observations as the training set.
LOOCV	Leave-*one*-out cross-validation is a particular case of LpO CV where $p = 1$
k-fold CV	In k-fold cross-validation the original sample is randomly partitioned into k equal sized sub-samples. The validation process is repeated k times, taking $k-1$ partitions as training and 1 as test
Monte Carlo CV	Repeater random sub-sampling cross validation, aklso known as Monte Carlo cross validation, randomly splits the dataset into training and validation data. Results are averaged over the splits.

Chosen validation strategy:

- *K*-fold cross-validation
- Depending on the chosen dataset for comparison

Conclusions

Objectives

- To **improve accuracy** of a CNN-based deep learning method for face recognition in videos.
- To **compare** the outcomes of the CNN alone with the CNN+LSTM system in order to investigate how temporal information affects the performances.
- To **build** a new publicly available framework for video face recognition.

Experiments

- **Train/Test** the **proposed CNN+LSTM** system against the databases and boost the model performances by fine-tuning its hyperparameters.
- **Train/Test** the **CNN alone** and the whole system **CNN+LSTM**.
- Design a methodology to investigate the **best LSTM temporal window**.

Contributions

- Detailed analysis of the performances of a **new hybrid CNN-LSTM deep temporal model** for **video face recognition.**
- Exhaustive investigation about **how** and **in which measure** temporal information can improve the performances of a CNN model.
- Provide a reasonable methodology to calculate the temporal window of a LSTM network for face recognition (perhaps for general face analysis tasks).

рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
faafetai lava
vinaka
спасибі
blagodaram
mersi
barka
kia ora
spas
welalin
tack
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
thank you
gracias
kiitos
dankie
nanni
nandri
dhanyavad
mauruuru
hvala
köszönöm
gràcie
bayarlalaa
bedankt
enkosi
dziękuję
akun
dankon
aciu
хвала
asante
manana
obrigada
tenki
djiere dieuf
tau
mochchakkeram
murakoze
дякую
mamnun
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
sobodi
dėkuji
sagolun
sukriya
kop khun krap
taiku
arigatô
takk
dakujem
trugarez
obrigado
mèsi
ありがとう
tanemirt
grazie
didi madloba
najis tuke
rahmet
shukriya
mercé
мерси
kam sah hamnida
rahmat
terima kasih
diolch
dhanyavadagalu
ευχαριστώ
তোমাকে ধন্যবাদ
감사합니다
xièxie
merci

Deep Learning Methods for (video) Face Recognition Comparison Summary Tables

Paper	Protocol	Metric	Result
DeepID2+ [44]	Standard protocol	ACC	93.2% (VR) 95% (IR)
DeepFace [47]	Standard protocol (unrestricted)	ACC	91.4% (CR)
		100%-EER	92.5%
Eigen-PEP for video face recognition [31]	Standard protocol	ACC	85.4%
Face Recognition in Movie Trailers via Mean Square Sparse Representation-based Classification [34]	Standard protocol	ACC	75.3%,
		AUC	82.9%
		EER	25.3%
Hierarchical-PEP model for real-world face recognition [29]	Not specifically defined	ACC	87%
MDLFace [16]	3M face images of 50K identities	ACC	97.9%
Neural Aggregations Networks [53]	100 frames for each video	ACC	96.5% (IR)
		AUC	98.7%
Sparsifying Neural Network Connections [45]	Train: 290K faces; Val: 47K faces; Test: 5K pairs of faces	ACC	93.5% (RR)
Unconstrained Face Recognition [7]	Own gallery (YTF+LFW) + fusion	ACC	**79%**

Notes: ACC: Accuracy, AUC: area under the curve, EER: Equal Error Rate, RR: Recognition Rate, IR: Identification
VR: Verification , Rate, CR: Classification Rate

Paper	Face Region	Protocol	Accuracy
Towards Large-Scale Face Recognition Based on Videos	-	1 train / 3 test	98.1% (CR)
Learning Personal Specific Facial Dynamics for Face Recognition From Videos	40x40	$\frac{1}{2}$ train / $\frac{1}{2}$ test	97.9%
Joint sparse representation for video-based face recognition	30x30	1 train / 3 test	96.5% (IR)
Face Recognition Based on Image Sets	40x40	1 train / 3 test	95.3, 98.1(CR)
From Still Image to Video-Based Face Recognition: An Experimental Analysis	40x40	1 train / 3 test	92.3% (RR)

Notes: RR: Recognition Rate, IR: Identification Rate, CR: Classification Rat

Work	Year	Database	Accuracy
DeepID2+	2014	LFW	99.47% (95%*)
		YTF	93.2
DeepFace	2014	LFW	97.4%
		YTF	91.4%
H-PEP	2015	LFW	91.1%
		YTF	87%
Sparse ConvNet	2015	YTF	99.55%
		LFW	89%
NAN	2016	IJB-A	92.5%
		YTF	95.7%

* Identification