Facial Expression Recognition summary

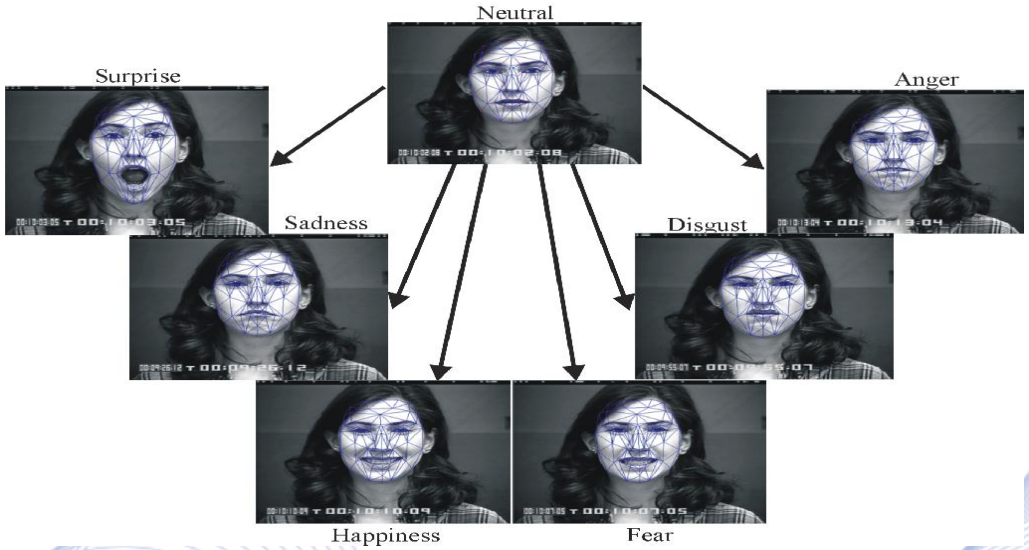
C. Aslanidou, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 3.0

Artificial Intelligence & Information Analysis Lab

Facial Expression Recognition

- **Classical Facial Expression Recognition**
  - **Grid-Based Methods**
  - **Subspace methods**
- DNN Facial Expression Recognition
  - DNN Facial Expression Recognition on static images
  - DNN Facial Expression Recognition on videos
- 3D Facial Expression Recognition
- Facial Expression Recognition datasets

Facial Expression Analysis



Problem statement:

- To identify a facial expression

in a facial image or 3D point cloud.

- Input: a face ROI or a point cloud
- Output: the facial expression label

(e.g. neutral, anger, disgust, sadness, happiness, surprise, fear).

Informative Content of Facial Expressions

- Human communication is mainly performed by nonverbal means (gestures and facial actions).
- Facial actions: important source for understanding human emotional state and intention.
- Key importance to various fields e.g. human behavior analysis, affective video content description, psychology, HCI, ambient intelligence, entertainment etc.

Facial Expression Analysis

- Latent facial image data dimensionality

  Facial image space dimensionality is much higher than required.

  Necessitates the use of efficient dimensionality reduction methods.

  Reduce complexity and boost performance of expression analysis algorithms.

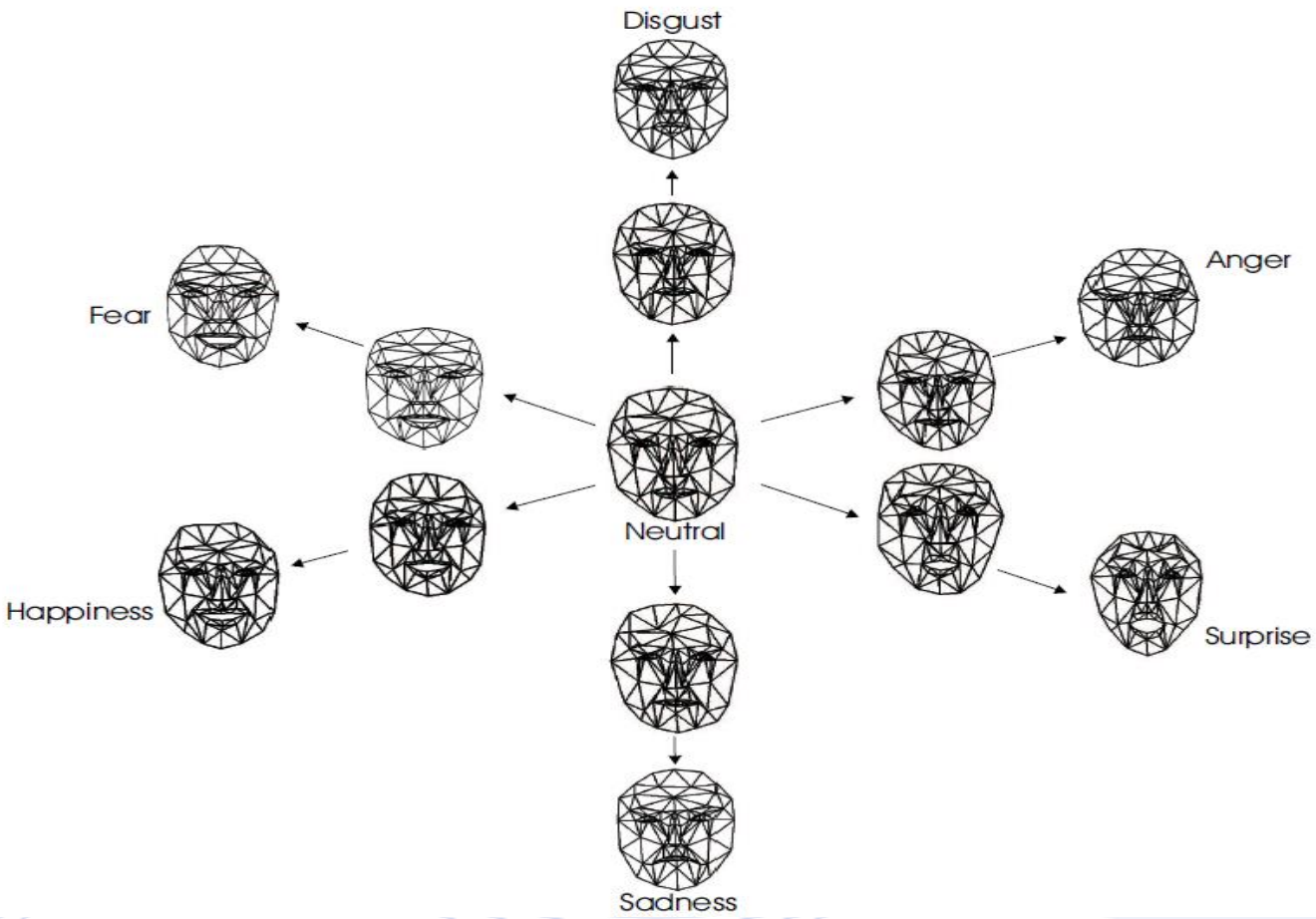
- Two popular approaches to handle facial expression data high dimensionality:

  Grid-Based Methods.

  Subspace Learning Methods.

Grid-Based Methods

- A facial grid is a parameterized face mask specifically developed for model-based coding of human faces.
- A popular facial wireframe model is the Candide grid.
- Facial expression information extraction is performed by facial feature point tracking.

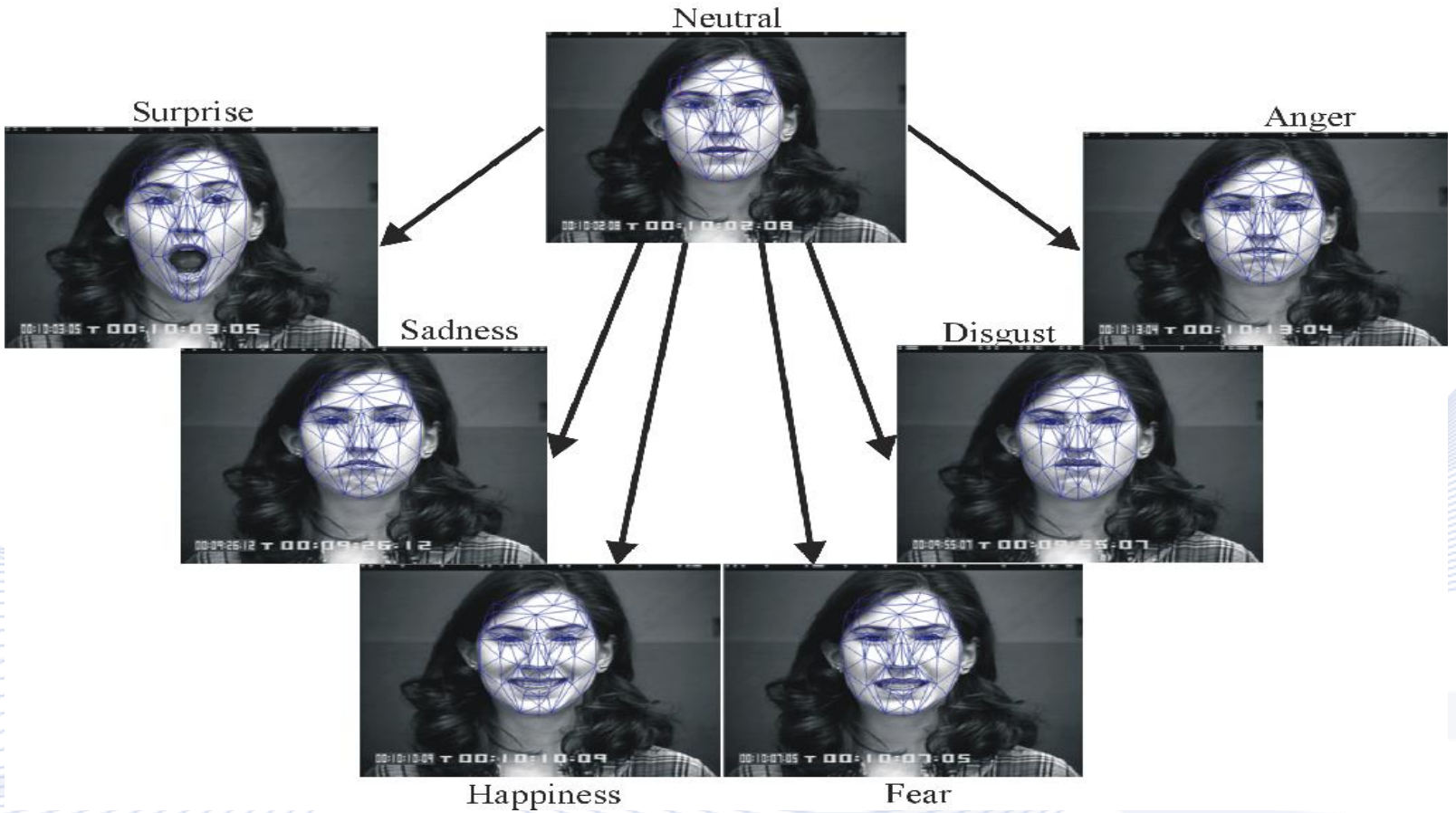


Grid-Based Methods

- Expression recognition:

  Track geometric positions of the grid nodes corresponding to fiducial points on a face;

  Grid nodes displacements are used as classification features.



Subspace Learning Methods

- Aim to discover latent facial features by projecting linearly or nonlinearly the facial image to a low dimensional subspace where a certain criterion is optimized.
- Popular dimensionality reduction techniques:

Principal Components Analysis (PCA)

Linear Discriminant Analysis (LDA)

Independent Components Analysis (ICA)

Singular Value Decomposition (SVD)

Non-negative Matrix Factorization (NMF)

Non-Negative Matrix Factorization (NMF)

- NMF is an unsupervised matrix decomposition method that requires both the decomposed data and the derived factors to contain non-negative elements.
- NMF Problem Formulation:

  Given the non-negative matrix whose columns are F-dimensional feature vectors obtained by scanning row-wise each of the L facial images, NMF considers the factorization:

$$X \approx ZH \qquad X \in R_{+}^{F*L}$$

$Z \in R_{+}^{F*M}$: basis images matrix.

$H \in R_{+}^{M*L}$: coefficients matrix.

Non-Negative Matrix Factorization (NMF)

- Original facial images are reconstructed using only additive combinations of the resulting basis images.
- Combination weights: coefficients in **H**.

- Consistent with the psychological intuition regarding the objects representation in the human brain (i.e. combining parts to form the whole).

Non-Negative Matrix Factorization (NMF)

- A popular approach to measure the quality of the approximation is the matrix Frobenious norm:

$$\mathcal{D}_{NMF}(\boldsymbol{X}||\boldsymbol{ZH}) \triangleq \left|\left|\boldsymbol{X}-\boldsymbol{ZH}\right|\right|_F^2 = \sum_{j=1}^{L}\sum_{i=1}^{F}(x_{i,j}-[\boldsymbol{ZH}]_{i,j})^2$$

- NMF training:

  Minimize the cost function, subject to nonnegativity constraints.

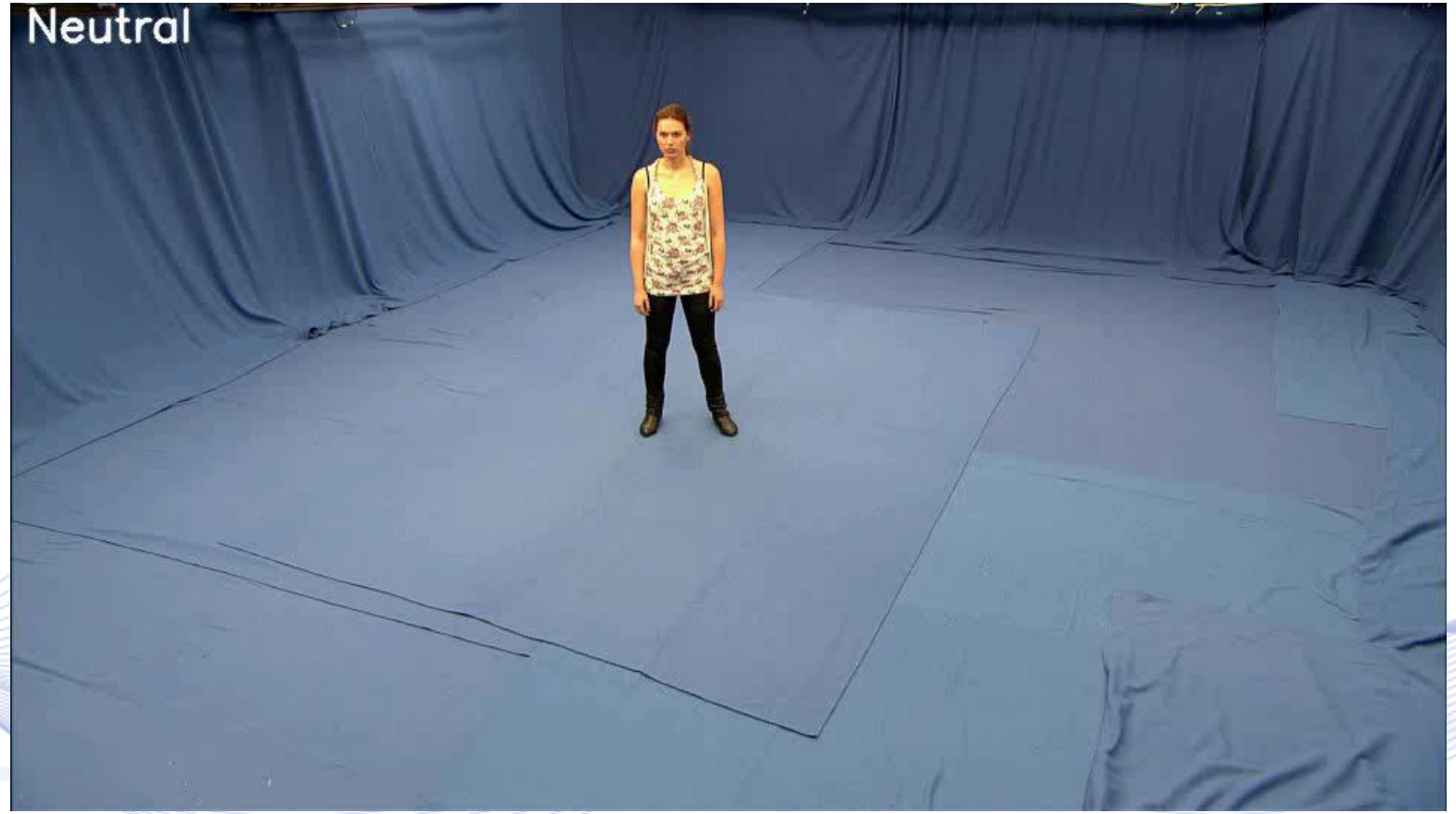
  Identify different basic facial parts (basis images).

  Approximate the appropriate weights to reconstruct the original facial images.

Facial Expression Subclasses

- Facial Expression clustering in Cohn-Kanade Database
  - Cohn-Kanade database images depict subjects of different racial background under severe illumination variations.
  - We partitioned each expression class into three subclasses and computed the mean expressive image for the two more distant clusters of each class.

Experimental Results



Facial Expression Recognition

- Classical Facial Expression Recognition
  - Grid-Based Methods
  - Subspace methods
- **DNN Facial Expression Recognition**
  - DNN Facial Expression Recognition on static images
  - DNN Facial Expression Recognition on videos
- 3D Facial Expression Recognition
- Facial Expression Recognition datasets

Deep facial expression recognition

There are the three main steps that are common in automatic deep FER:

- Pre-processing
- Deep feature learning and
- Deep feature classification.

It follows a summary of the widely used algorithms for each step.

Pre-processing

The input image to the FER may contain noise and have variation in illumination, size, color, background and head poses.

To get accurate and faster results on the algorithm, some preprocessing operations were done on the image.

The preprocessing strategies used are conversion of image to grayscale, normalization, and resizing of image.

Pre-processing

- **Normalization**: Normalization of an image is done to remove illumination variations and obtain improved face image
- **Grayscaling**: Grayscaling is the process of converting a colored image input into an image whose pixel value depends on the intensity of light on the image. Grayscaling is done as colored images are difficult to process by an algorithm.
- **Resizing**: The image is resized to remove the unnecessary parts of the image. This reduces the memory required and increases computation speed.

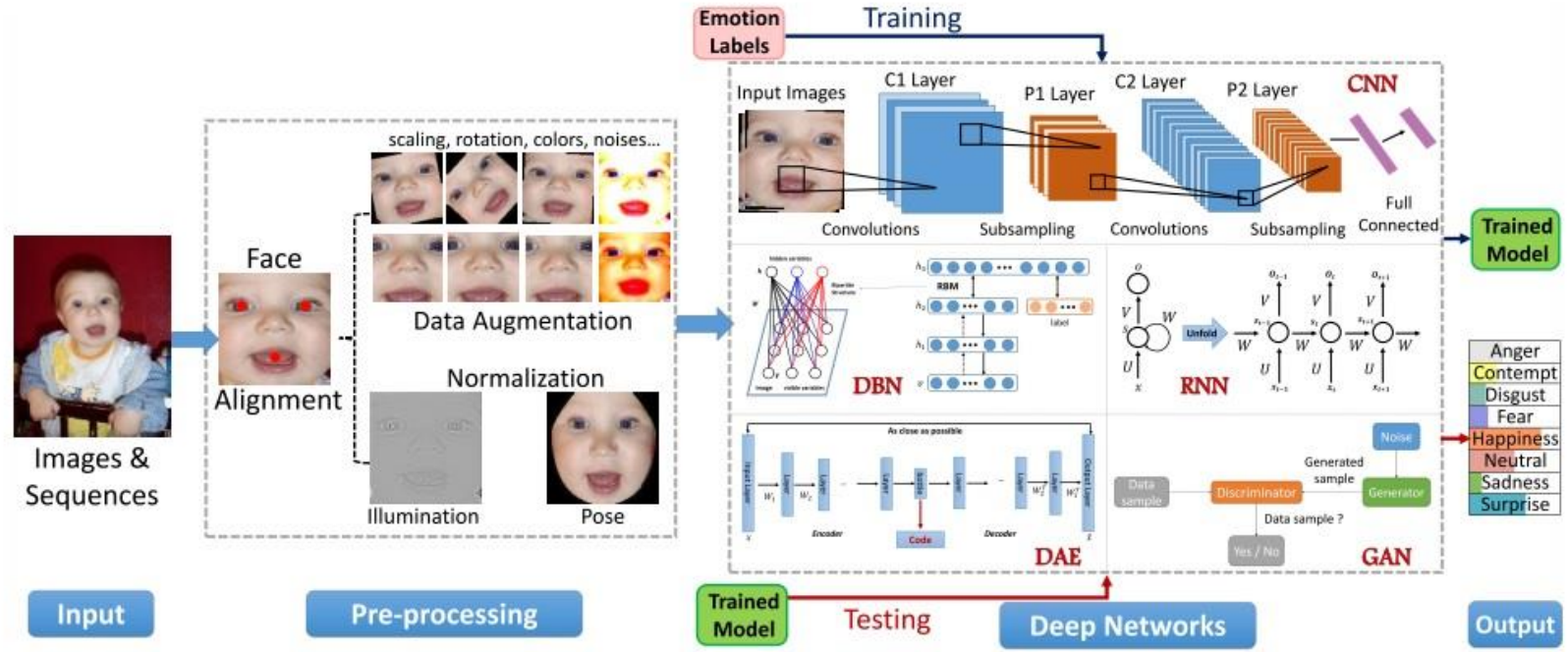
Deep feature learning

Deep learning attempts to capture high-level abstractions through hierarchical architectures of multiple nonlinear transformations and representations.

The traditional architectures of these deep neural networks are:

- Convolutional neural network (CNN)
- Deep belief network (DBN)
- Deep autoencoder (DAE)
- Recurrent neural network (RNN)
- Generative Adversarial Network (GAN)

Pre-processing, Deep feature learning & classification



The general pipeline of deep facial expression recognition systems (Image from Semantic Scholar).

Deep FER networks

There are novel deep neural networks designed for FER and related training strategies proposed to address expression-specific problems.

There are two main groups depending on the type of data:

- Deep FER networks for static images and
- Deep FER networks for videos.

Facial Expression Recognition

- Classical Facial Expression Recognition
  - Grid-Based Methods
  - Subspace methods
- DNN Facial Expression Recognition
  - **DNN Facial Expression Recognition on static images**
  - DNN Facial Expression Recognition on videos
- 3D Facial Expression Recognition
- Facial Expression Recognition datasets

Deep FER networks for static images

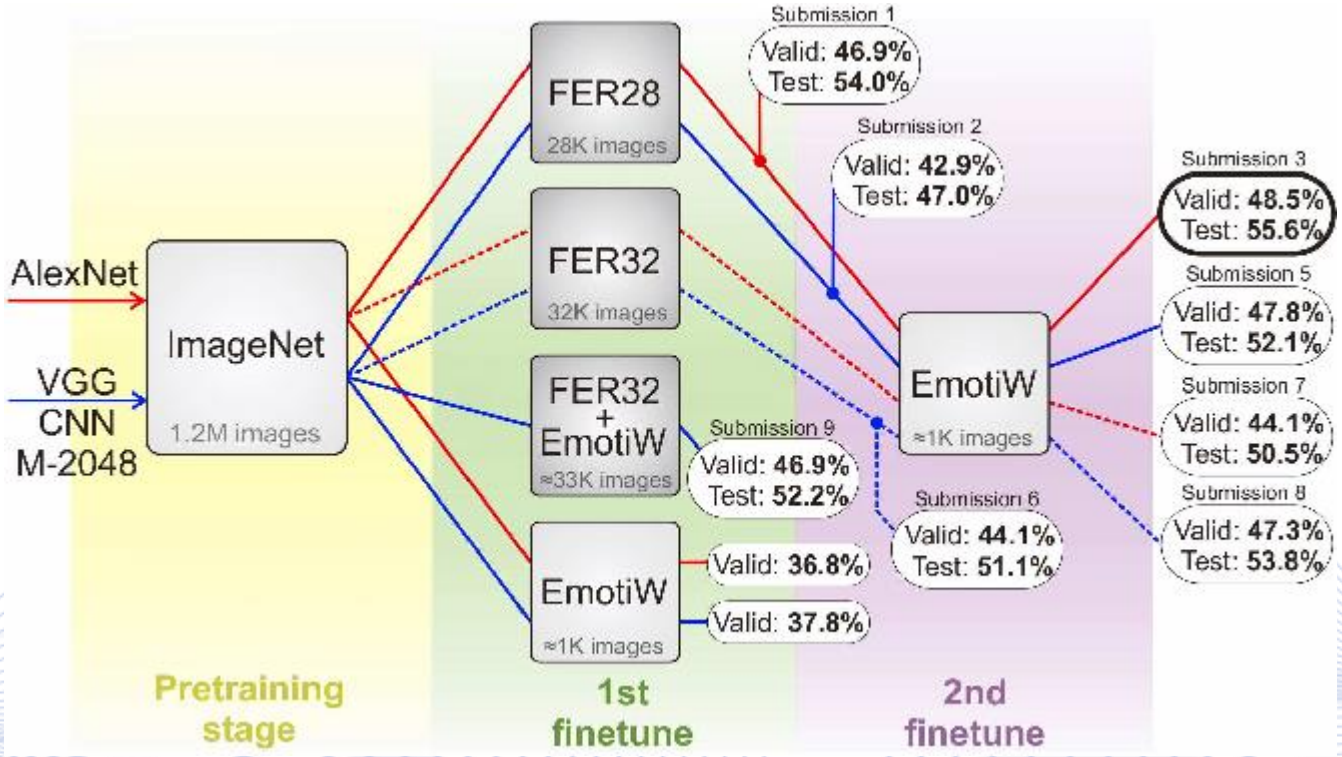
A large volume of the existing studies conducted expression recognition tasks based on static images without considering temporal information due to the convenience of data processing and the availability of the relevant training and test material.

We'll see first the specific pre-training and fine-tuning skills for FER and then the novel deep neural networks.

Deep FER networks for static images

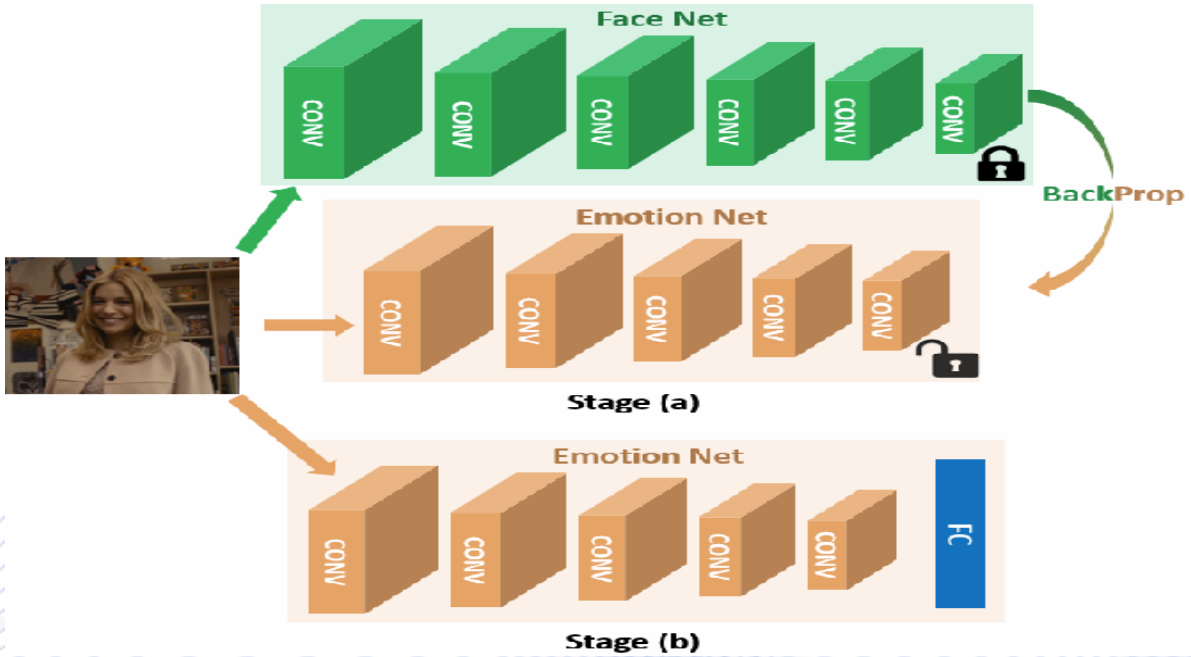
- Pre-training and fine-tuning
- Diverse network input
- Auxiliary blocks & layers
- Network ensemble
- Multitask networks
- Cascaded networks
- Generative adversarial networks (GANs)
- Comparison of different types of methods for static images

Deep FER networks for static images



Flowchart (Image from Semantic Scholar)

Deep FER networks for static images



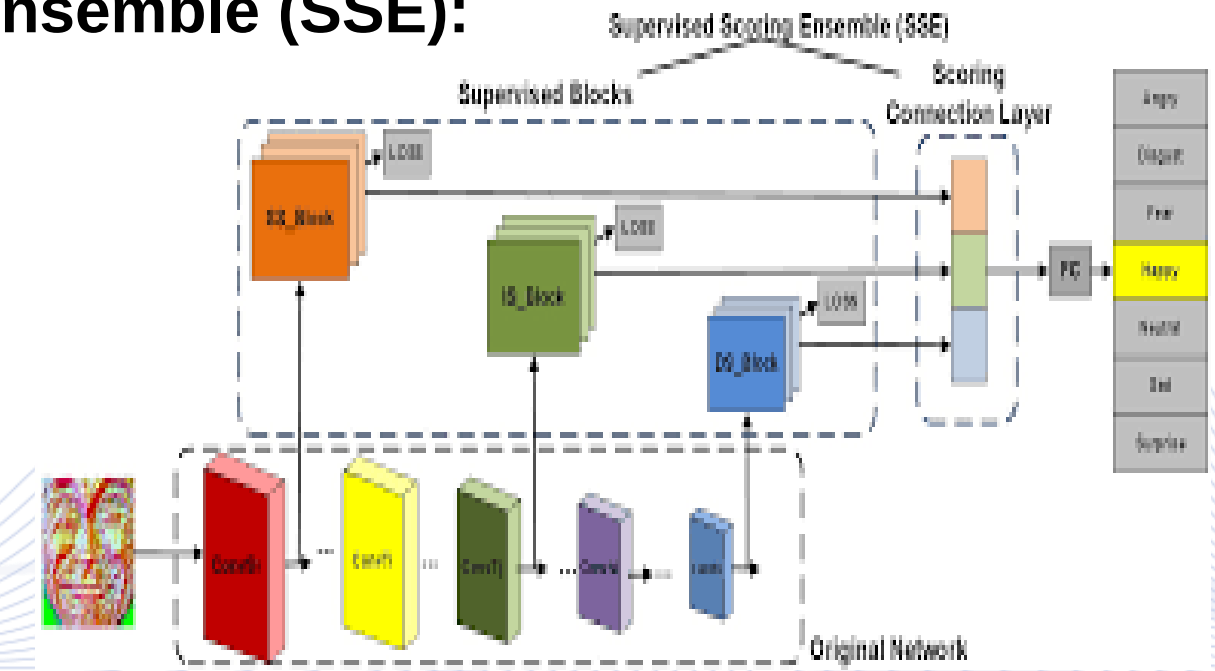
Two-stage training flowchart. (Image from Semantic Scholar)

Deep FER networks for static images

Image intensities (left) and LBP codes (middle). Proposed mapping these values to a 3D metric space (right) as the input of CNNs (Image from arXiv Vanity).

Deep FER networks for static images

Scoring Ensemble (SSE):

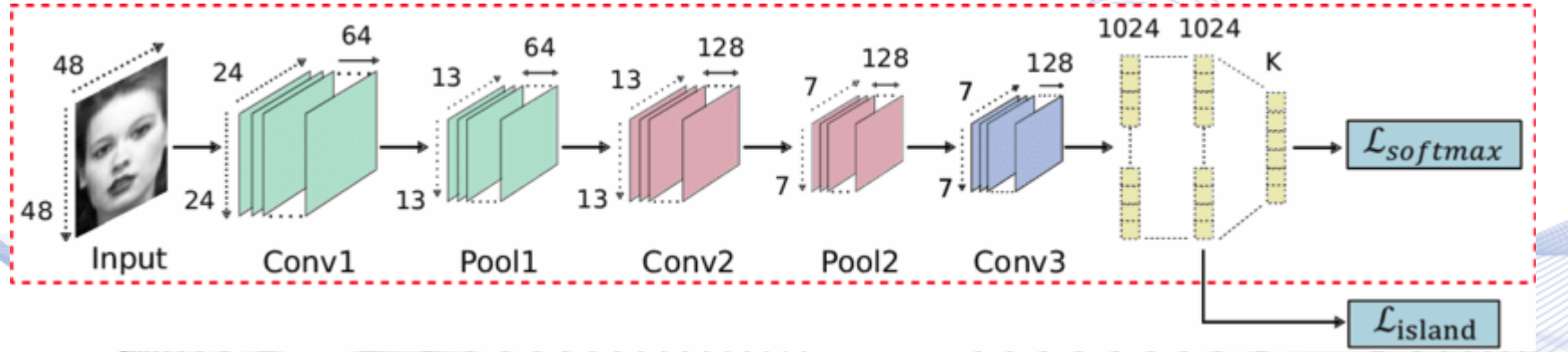


(Image from arXiv)

Deep FER networks for static images

Island loss layer:

The island loss calculated at the feature extraction layer and the softmax loss calculated at the decision layer are combined to supervise the CNN training.

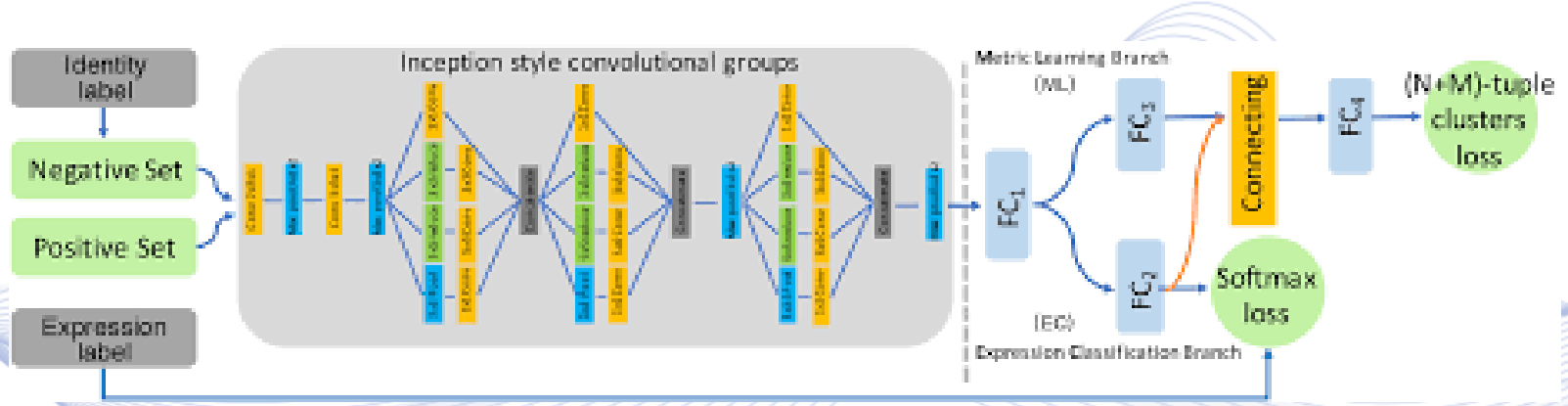


(Image from ResearchGate)

Deep FER networks for static images

(N+M)-tuple clusters loss layer:

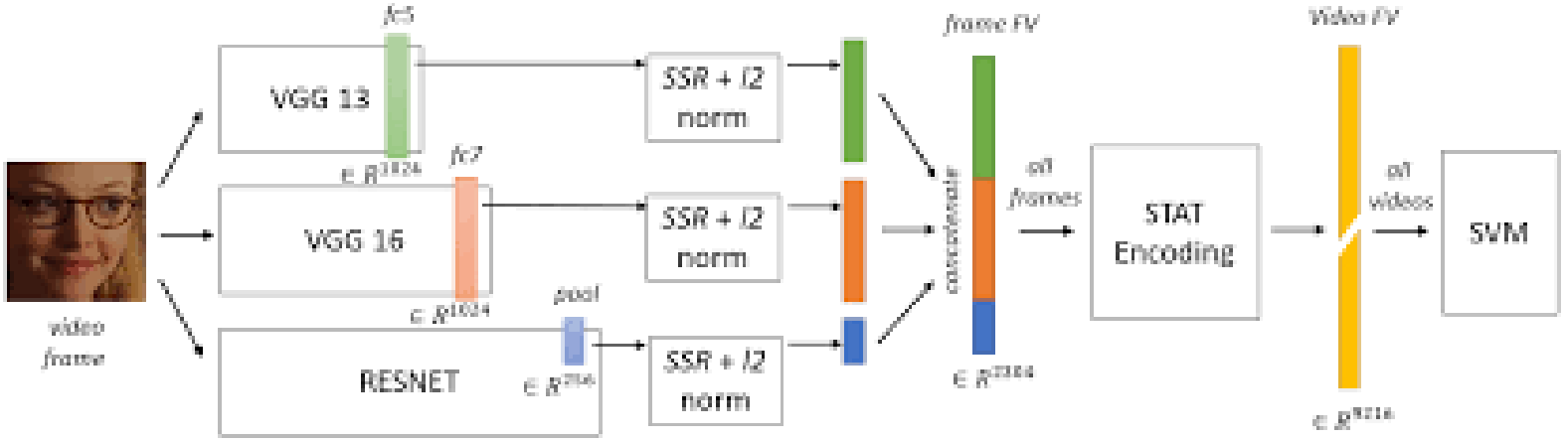
In (N+M)-tuple clusters loss layer, during training, the identity aware hard-negative mining and online positive mining schemes are used to decrease the inter-identity variation in the same expression class.



(Image from arXiv)

Deep FER networks for static images

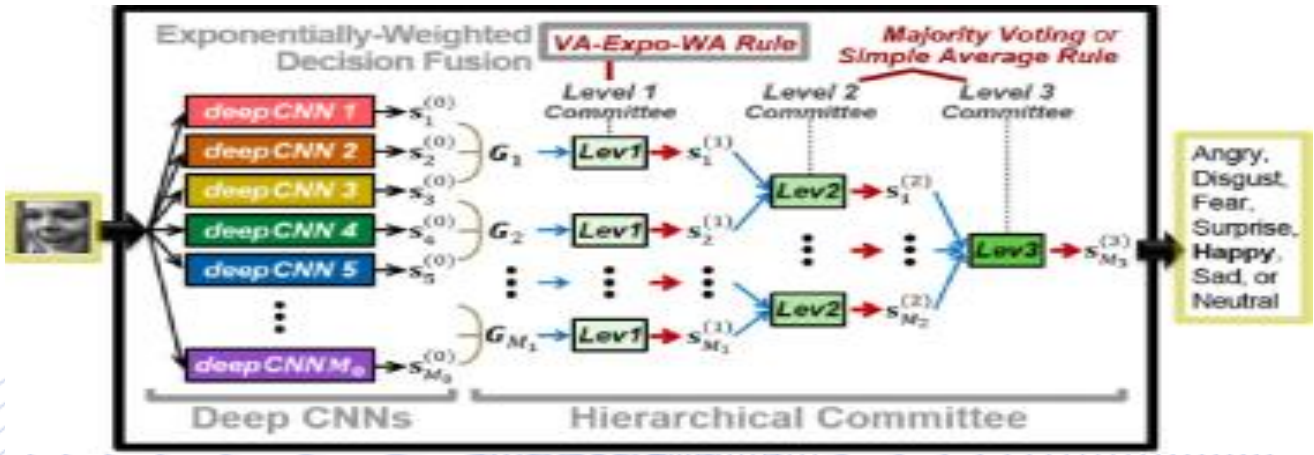
Feature-level ensemble:



(Image from arXiv)

Deep FER networks for static images

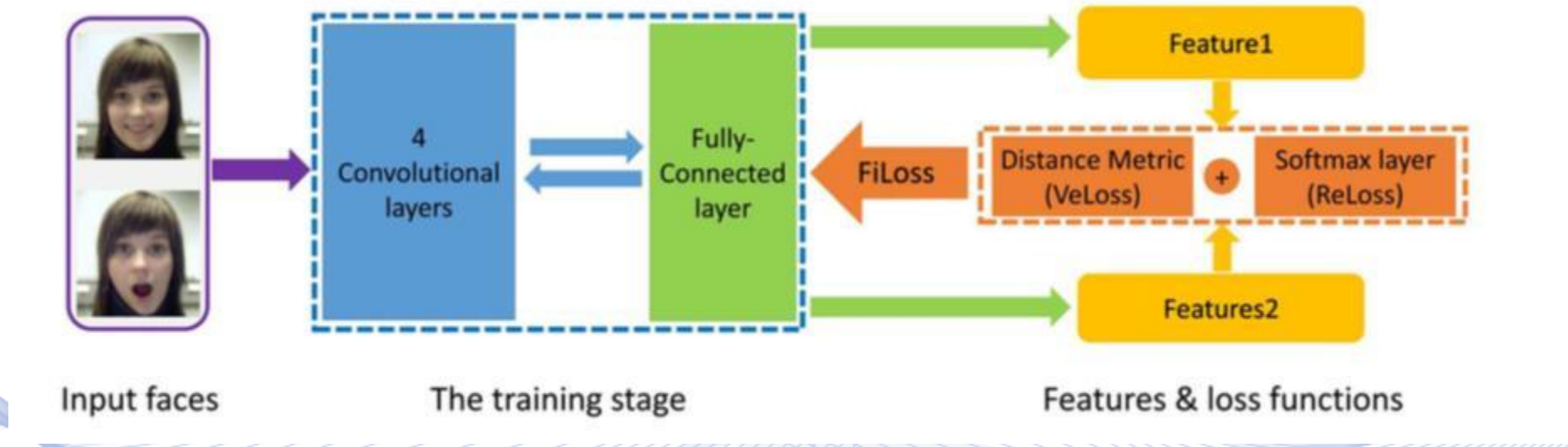
Decision-level ensemble:



(Image from GroundAI)

Deep FER networks for static images

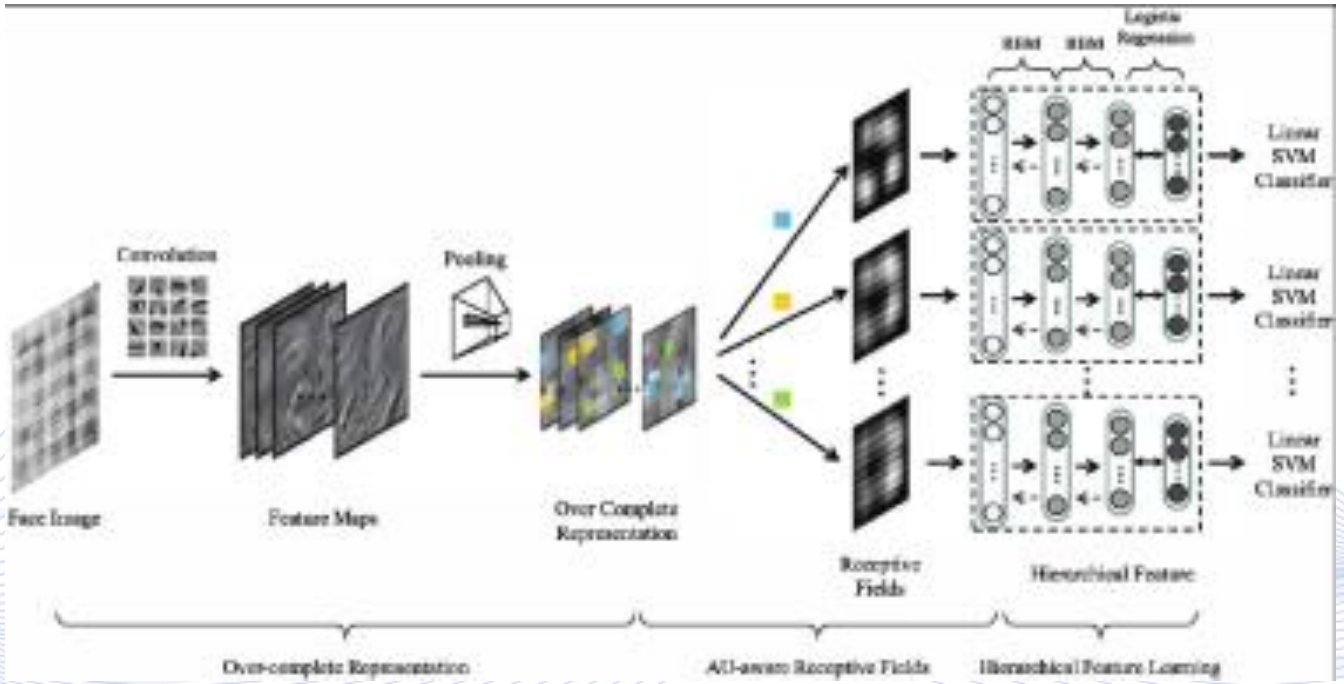
Representative multitask network for FER:



(Image from Programmer Shout)

Deep FER networks for static images

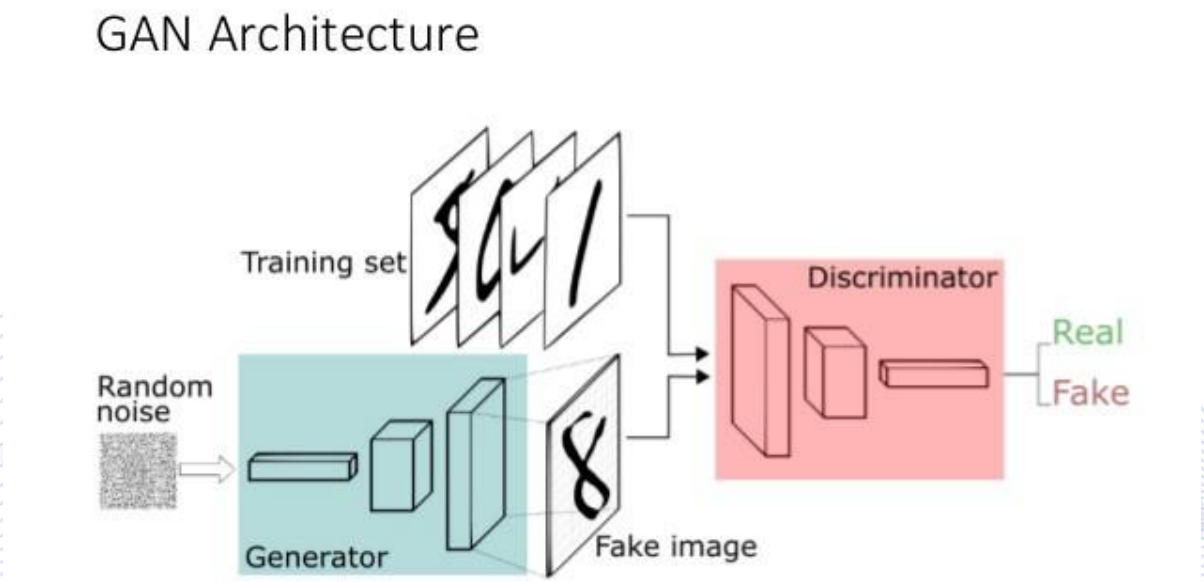
Representative cascaded network for FER:



(Image from GroundAI)

Deep FER networks for static images

Generative adversarial networks (GANs):



(Image from mc.ai)

Facial Expression Recognition

- Classical Facial Expression Recognition
  - Grid-Based Methods
  - Subspace methods
- DNN Facial Expression Recognition
  - DNN Facial Expression Recognition on static images
  - **DNN Facial Expression Recognition on videos**
- 3D Facial Expression Recognition
- Facial Expression Recognition datasets

Deep FER networks for videos

Most of the previous models focus on static images, facial expression recognition can benefit from the temporal correlations of consecutive frames in a sequence.

Considering that in a videostream people usually display the same expression with different intensities, we further review methods that use images in different expression intensity states for intensity-invariant FER.

Deep FER networks for videos

Frame aggregation:

Because the frames in a given video clip may vary in expression intensity, directly measuring per-frame error does not yield satisfactory performance.

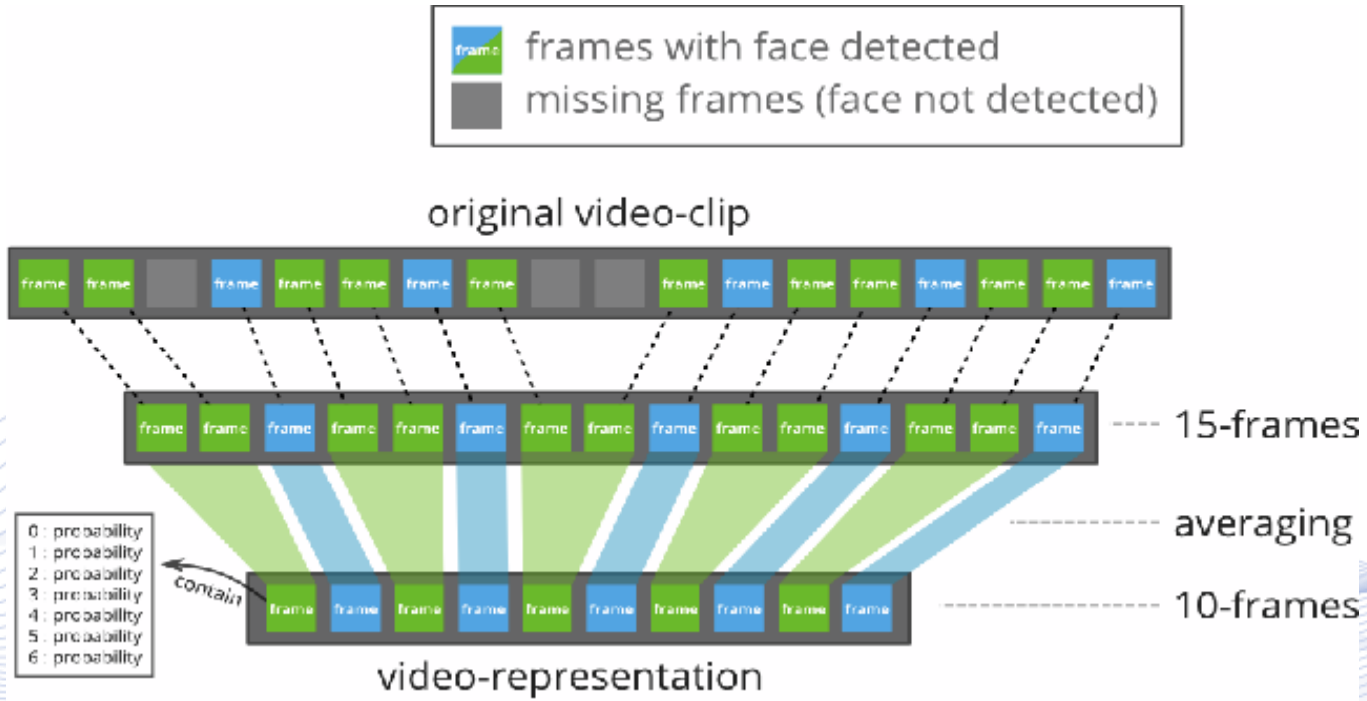
Various methods have been proposed to aggregate the network output for frames in each sequence to improve the performance.

These methods are divided into two groups:

- decision-level frame aggregation and
- feature-level frame aggregation.

Deep FER networks for videos

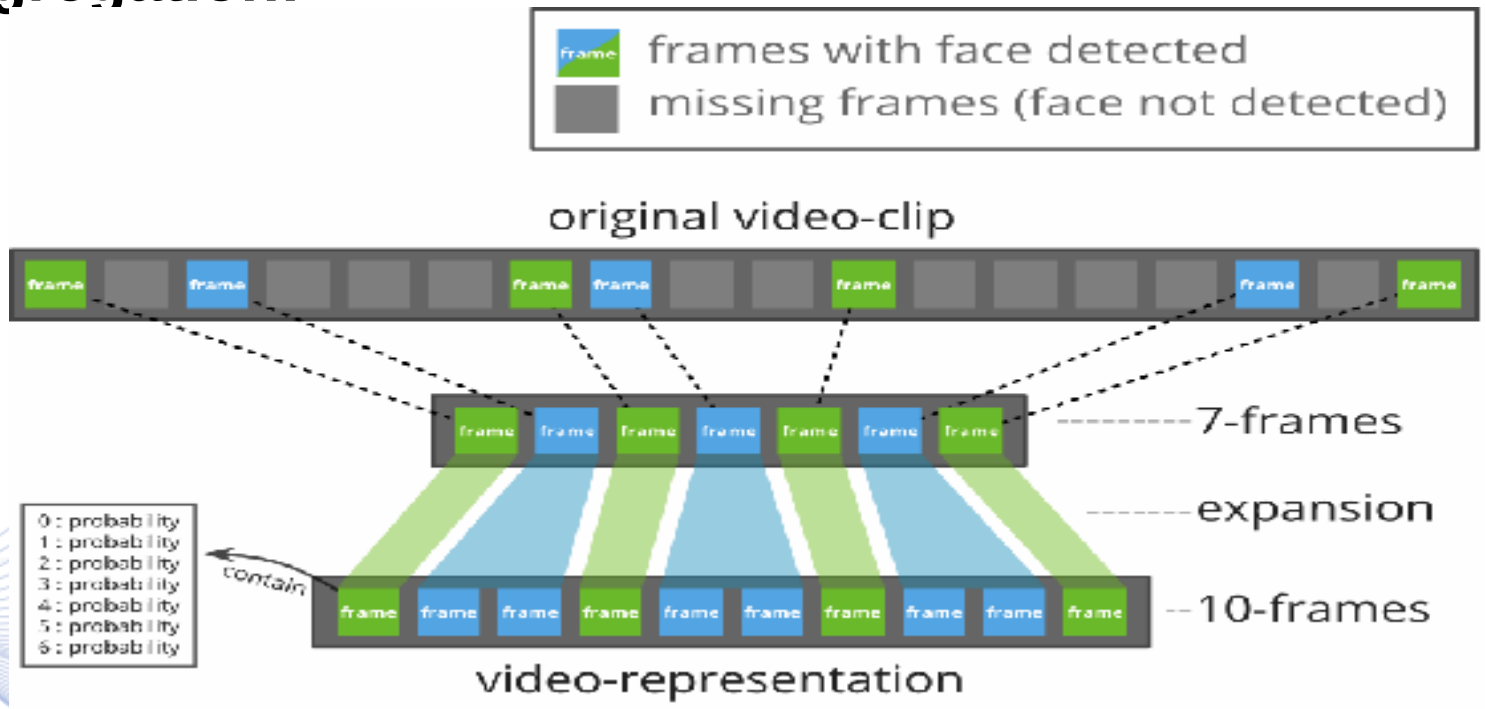
Frame aggregation:



(a) Frame averaging
(Image from SemanticScholar)

Deep FER networks for videos

Frame aggregation:



(b) Frame expansion
(Image from SemanticScholar)

Deep FER networks for videos

RNN and C3D:

RNN can robustly derive information from sequences by exploiting the fact that feature vectors for successive data are connected semantically and are therefore interdependent.
Compared with RNN, CNN is more suitable for computer vision applications.

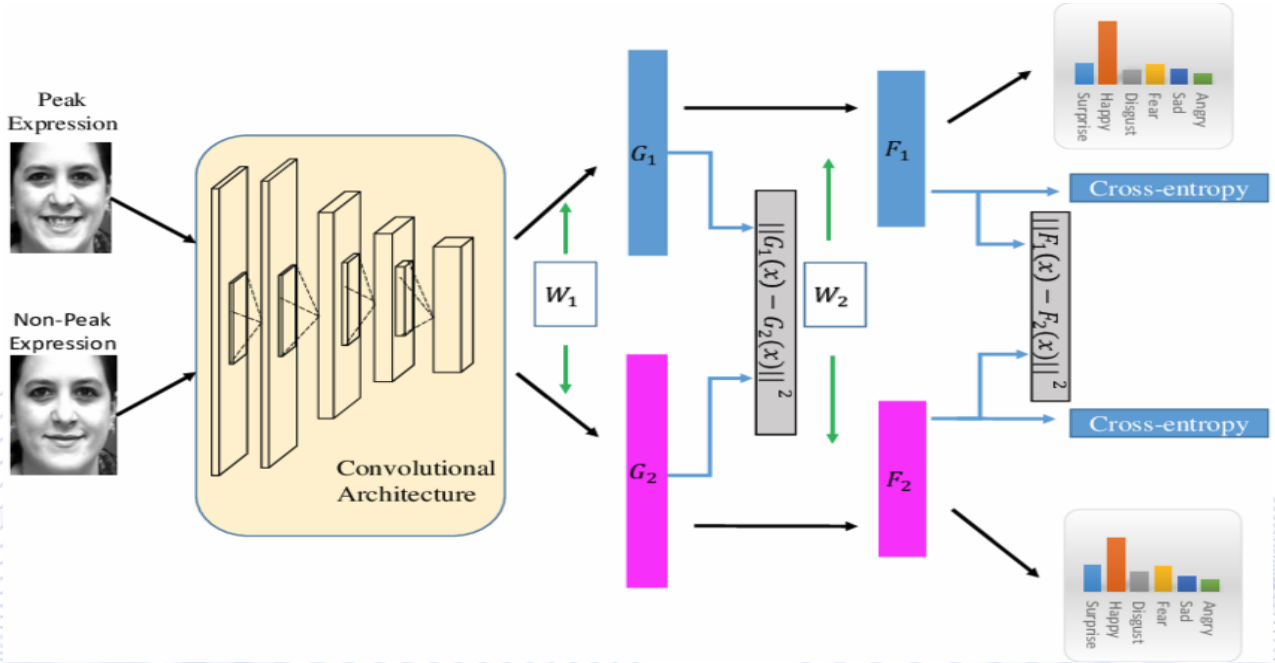
Deep FER networks for videos

RNN and C3D:

During training, PPDN is trained by jointly optimizing the L2-norm loss and the cross-entropy losses of two expression images. During testing, the PPDN takes one still image as input for probability prediction.

Deep FER networks for videos

RNN and C3D:



(Image from SemanticScholar)

Deep FER networks for videos

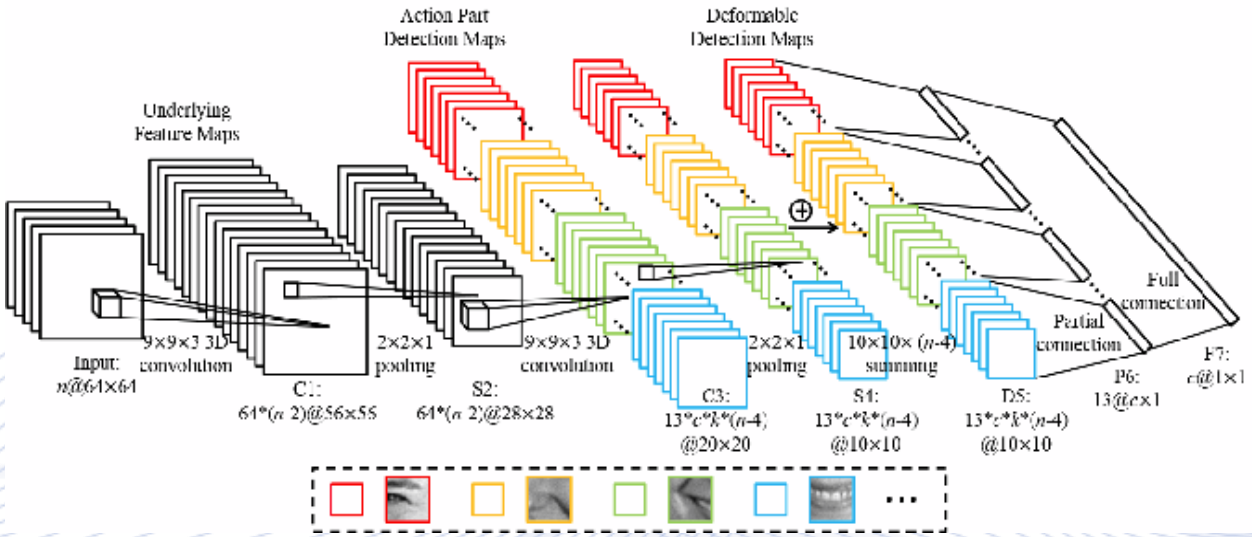
Facial landmark trajectory:

Related psychological studies have shown that expressions are invoked by dynamic motions of certain facial parts (e.g., eyes, nose and mouth) that contain the most descriptive information for representing expressions.

To obtain more accurate facial actions for FER, facial landmark trajectory models have been proposed to capture the dynamic variations of facial components from consecutive frames.

Deep FER networks for videos

Facial landmark trajectory :



(Image from SemanticScholar)

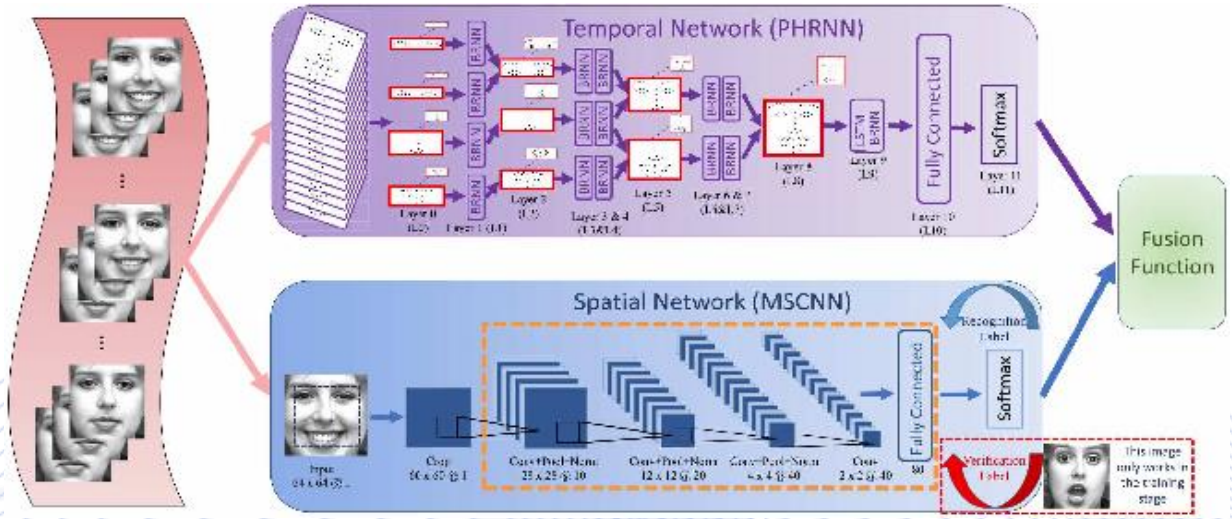
Deep FER networks for videos

The spatio-temporal network:

The temporal network PHRNN for landmark trajectory and the spatial network MSCNN for identity-invariant features are trained separately. Then, the predicted probabilities from the two networks are fused together for spatio-temporal FER.

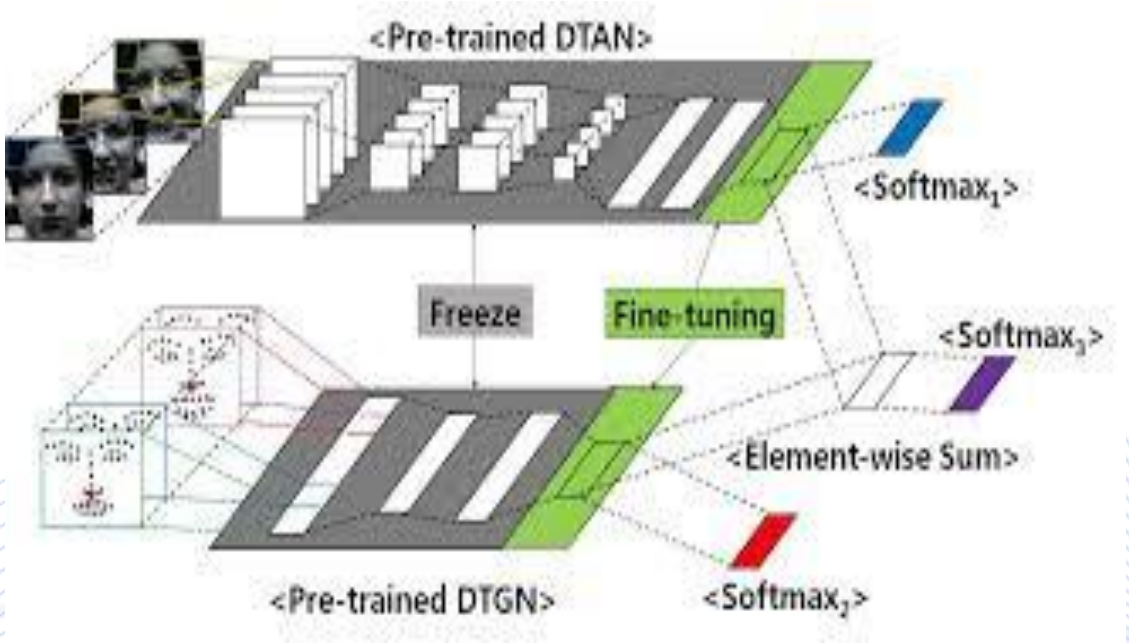
Deep FER networks for videos

The spatio-temporal network:



(Image from SemanticScholar)

Deep FER networks for videos



(Image from The Computer Vision Foundation)

Deep FER networks for videos

Network type		data	spatial	temporal	frame length	accuracy	efficiency
Frame aggregation		low	good	no	depends	fair	high
Expression intensity		fair	good	low	fixed	fair	varies
Spatio-temporal network	RNN	low	low	good	variable	low	fair
	C3D	high	good	fair	fixed	low	fair
	$\mathcal{FLT}$	fair	fair	fair	fixed	low	high
	$\mathcal{CN}$	high	good	good	variable	good	fair
	$\mathcal{NE}$	low	good	good	fixed	good	low

FLT = Facial Landmark Trajectory, CN = Cascaded Network, NE = Network Ensemble (Image from Semantic Scholar)

Deep FER networks for videos

ADDITIONAL RELATED ISSUES:

- **Occlusion and non-frontal head pose**
- **FER on infrared data**
- **FER on 3D static and dynamic data**
- **Facial expression synthesis**

Deep FER networks for videos

- **Occlusion and non-frontal head pose**

Occlusion and non-frontal head pose, which may change the visual appearance of the original facial expression, are two major obstacles for automatic FER, especially in real-world scenarios.

Deep FER networks for videos

- **FER on infrared data**

Although RBG or gray data are the current standard in deep FER, these data are vulnerable to ambient lighting conditions.
While, infrared images that record the skin temporal distribution produced by emotions are not sensitive to illumination variations, which may be a promising alternative for investigation of facial expression.

Deep FER networks for videos

- **FER on 3D static and dynamic data**

Despite significant advances have achieved in 2D FER, it fails to solve the two main problems: illumination changes and pose Variations.

3D FER that uses 3D face shape models with depth information can capture subtle facial deformations, which are naturally robust to pose and lighting variations.

Depth images and videos record the intensity of facial pixels based on distance from a depth camera, which contain critical information of facial geometric relations.

Deep FER networks for videos

CHALLENGES AND OPPORTUNITIES

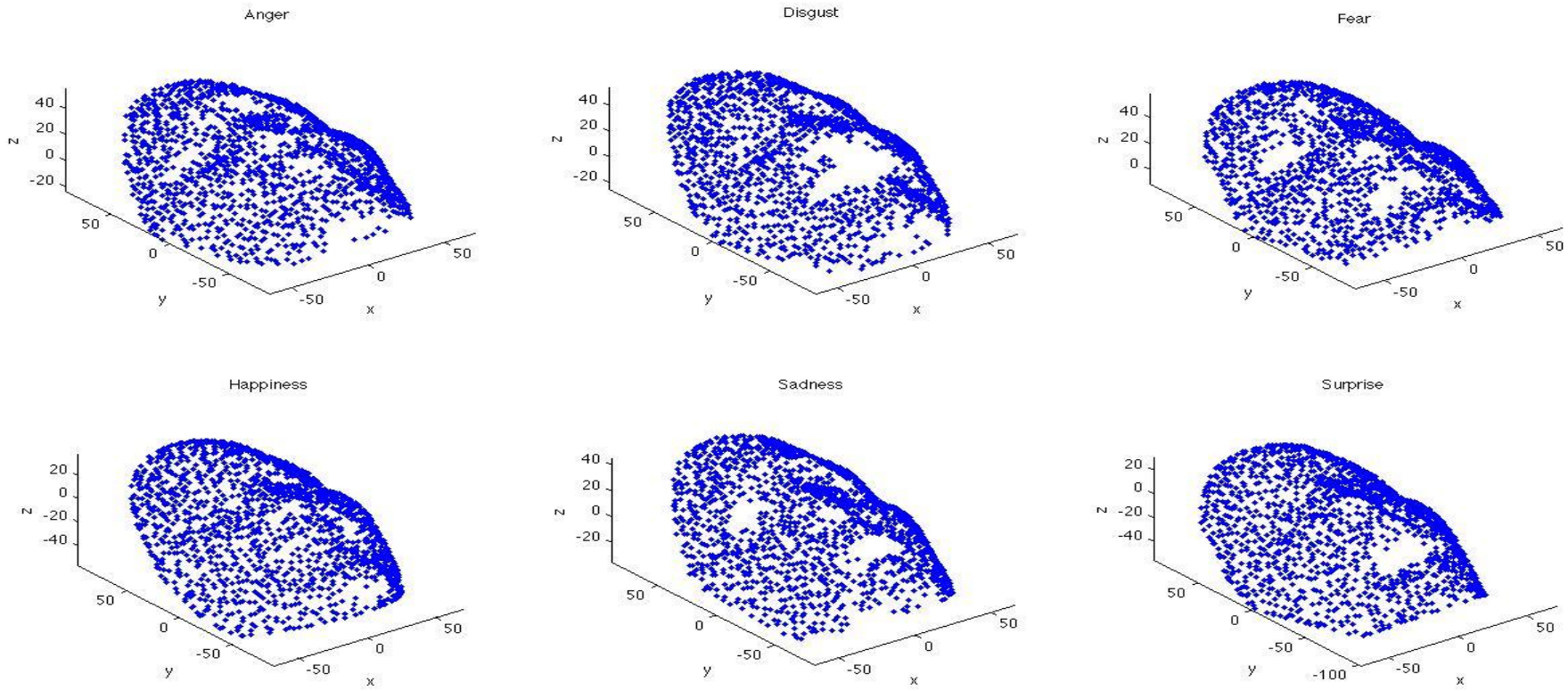
- Facial expression datasets
- Incorporating other affective models
- Dataset bias and imbalanced distribution
- Multimodal affect recognition

Facial Expression Recognition

- Classical Facial Expression Recognition
  - Grid-Based Methods
  - Subspace methods
- DNN Facial Expression Recognition
  - DNN Facial Expression Recognition on static images
  - DNN Facial Expression Recognition on videos
- **3D Facial Expression Recognition**
- Facial Expression Recognition datasets

3D Facial Expression Recognition

- Use of 3D facial point clouds.



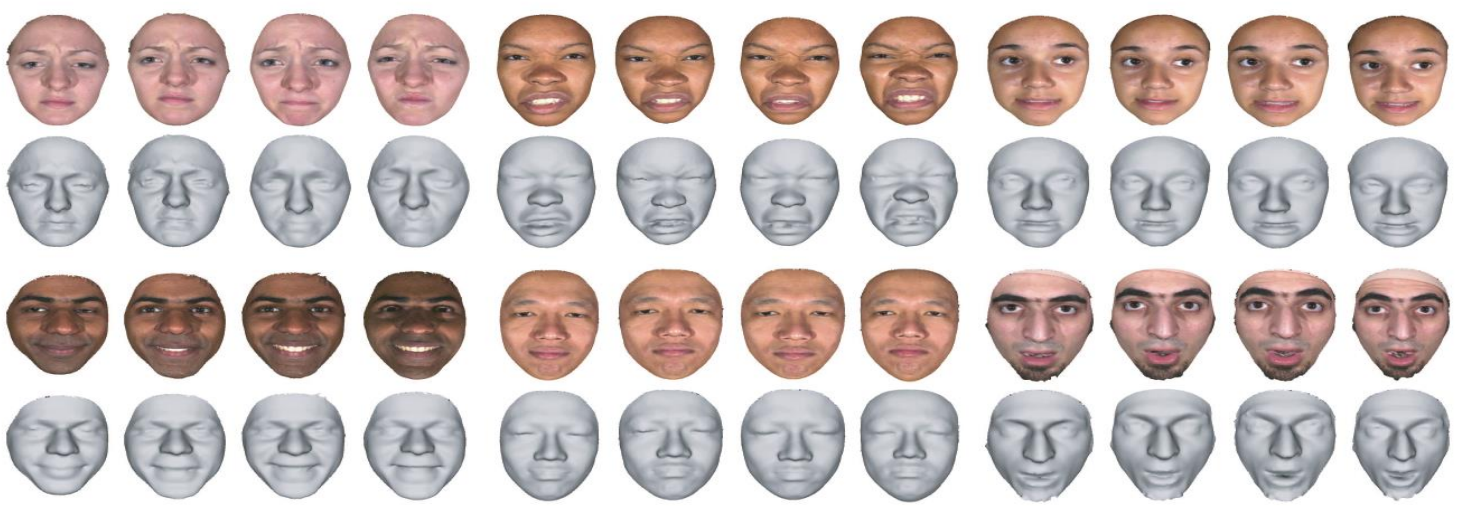
3D Facial Expression Recognition

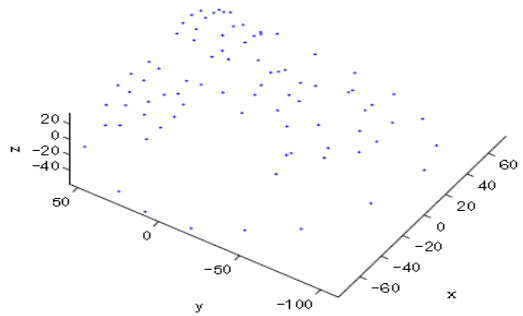
- Experiments on the BU-3DFE database

  100 subjects.

  6 facial expressions x 4 different intensities +neutral per subject.

  83 landmark points (used in our method).





Facial Expression Recognition

- Classical Facial Expression Recognition
  - Grid-Based Methods
  - Subspace methods
- DNN Facial Expression Recognition
  - DNN Facial Expression Recognition on static images
  - DNN Facial Expression Recognition on videos
- 3D Facial Expression Recognition
- **Facial Expression Recognition datasets**

Facial expression databases

These Data Bases are the followings:

- CK+
- MMI
- JAFFE
- TFD
- FER2013
- AFEW
- SFEW
- Multi-PIE
- BU-3DFE
- Oulu-CASIA
- RaFD
- KDEF
- EmotioNet
- RAF-DB
- AffectNet
- ExpW

FER databases

CK+:

CK+ Dataset (Image from portointeractivecenter.org).

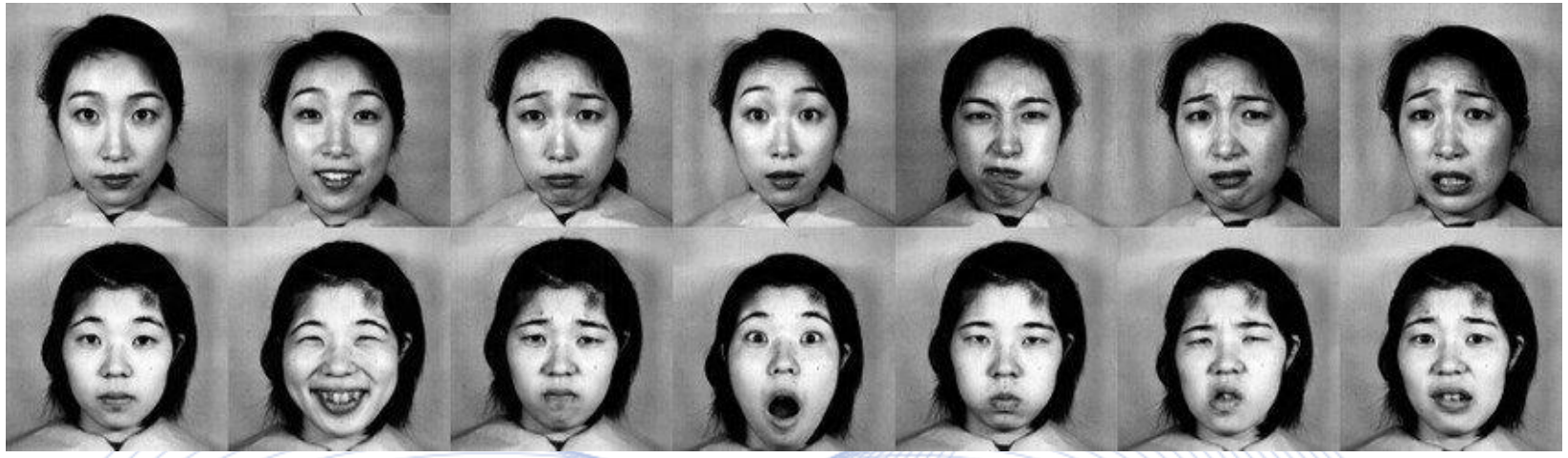
FER databases

MMI:

MMI Dataset (Image from mmiface.eu).

FER databases

JAFFE:



JAFFE Dataset (Image from ResearchGate).

FER databases

TFD:

TFD Dataset (Image from www.cs.toronto.edu).

FER databases

FER2013:

FER2013 Dataset (Image from ResearchGate).

FER databases

AFEW:

AFEW Dataset (Image from ResearchGate).

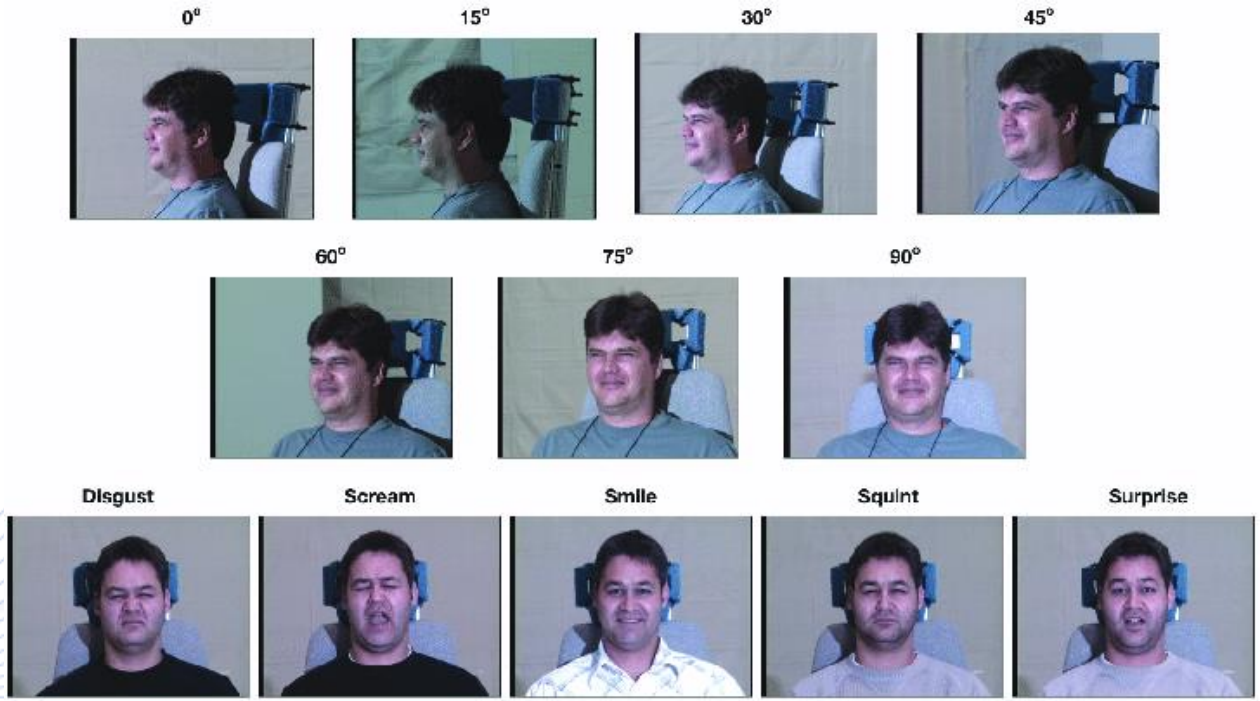
FER databases

SFEW:

SFEW Dataset (Image from ResearchGate).

FER databases

Multi-PIE:



Multi-PIE Dataset (Image from ResearchGate).

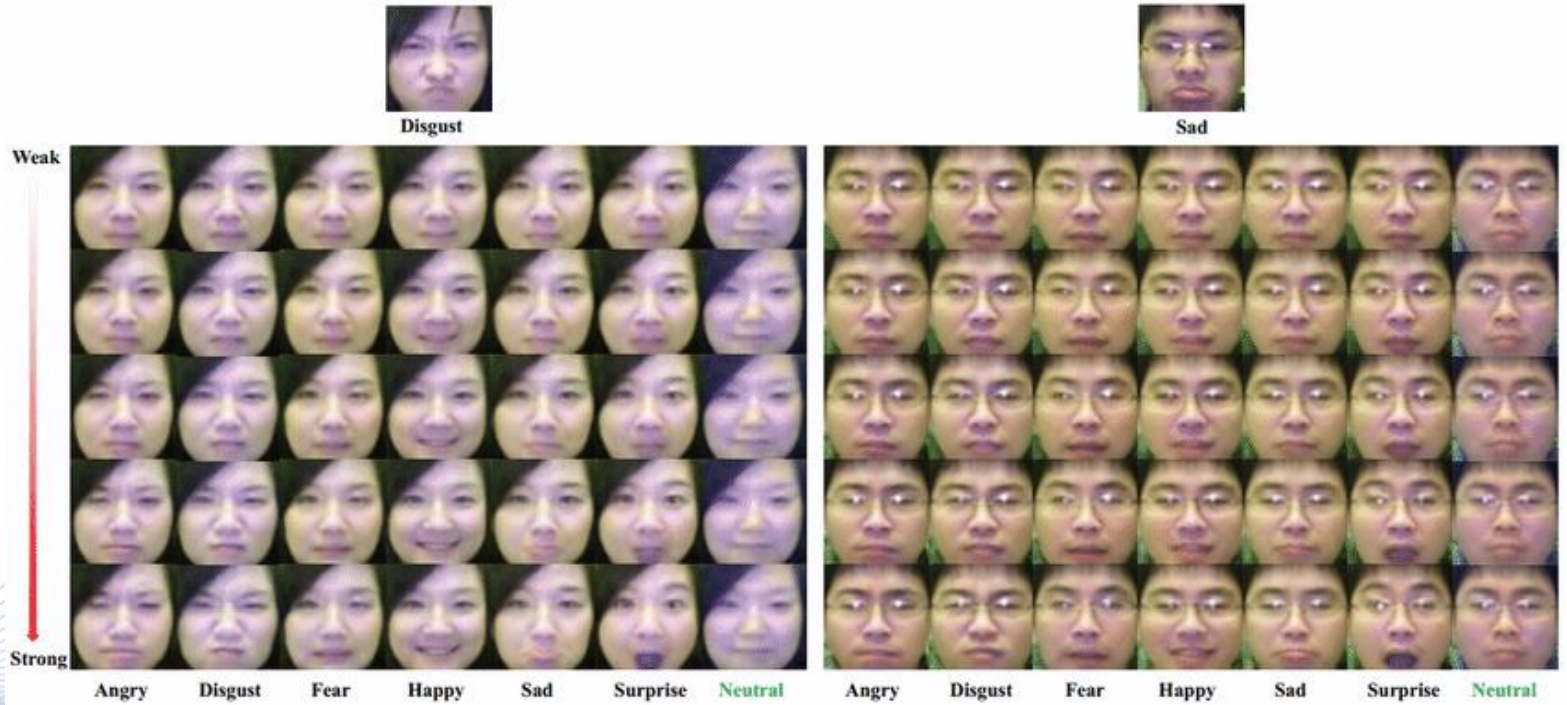
FER databases

BU-3DFE:

BU-3DFE Dataset (Image from ResearchGate).

FER databases

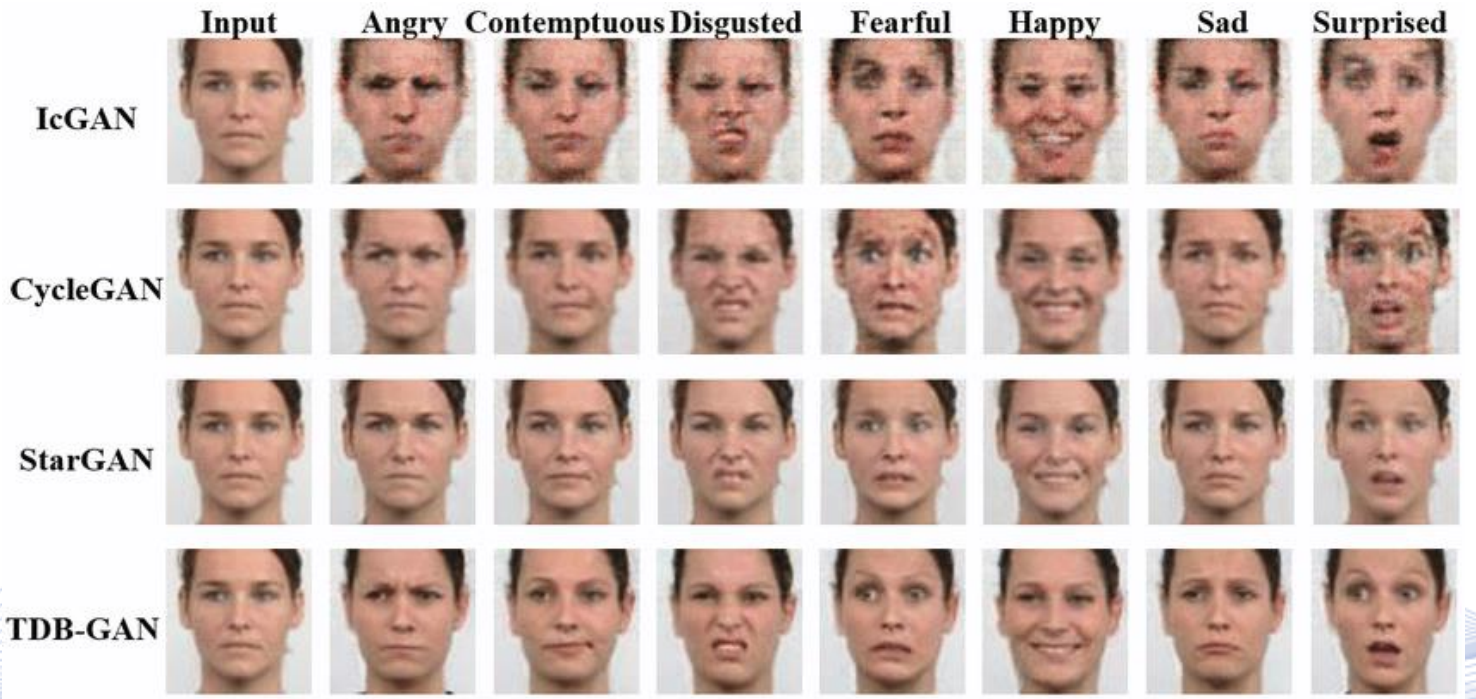
Oulu-CASIA:



Oulu-CASIA Dataset (Image from ResearchGate).

FER databases

RaFD:



RaFD Dataset (Image from ResearchGate).

FER databases

KDEF:

KDEF Dataset (Image from ResearchGate).

FER databases

EmotioNet:

EmotioNet Dataset
(Image from Computational Biology and Cognitive Science Lab).

FER databases

RAF-DB:

RAF-DB Dataset
(Image from Semantic Scholar).

FER databases

AffectNet:

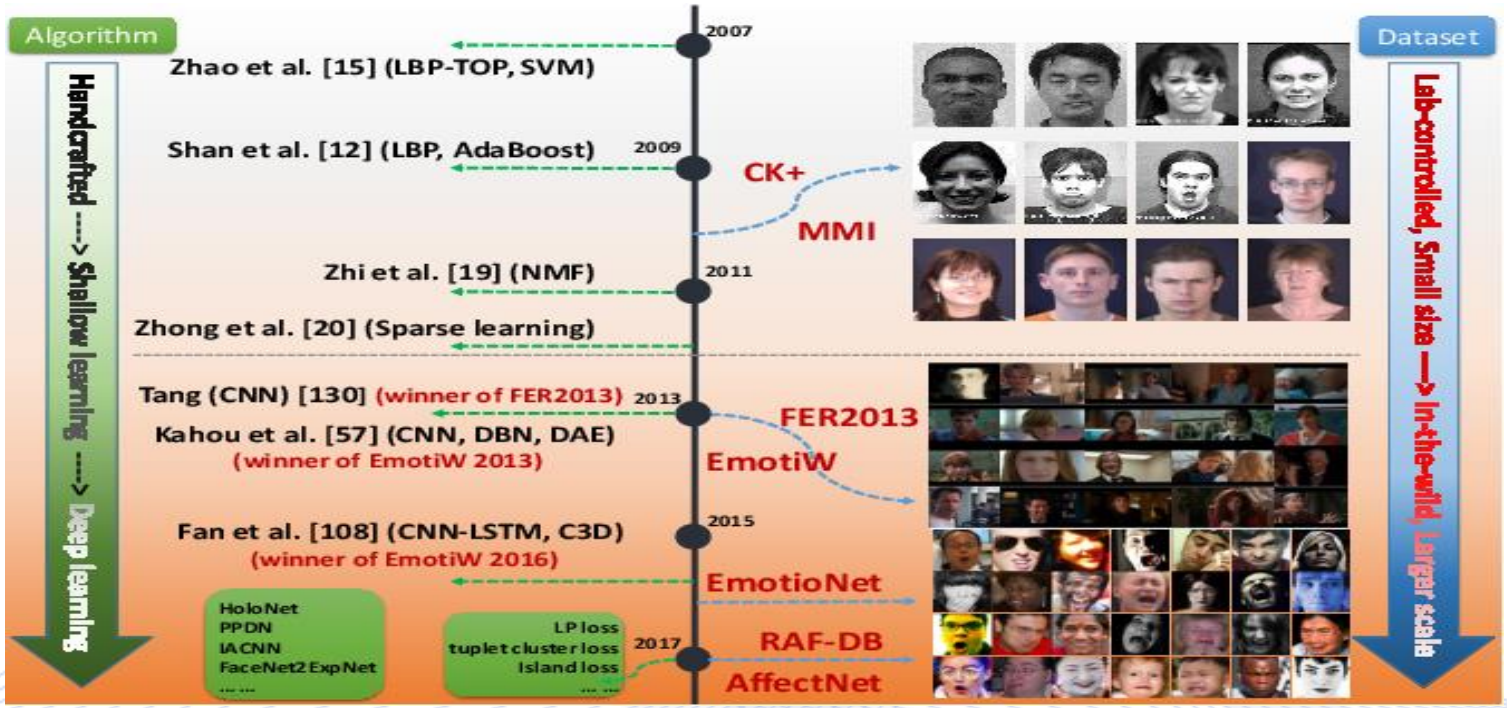
AffectNet Dataset
(Image from Scinapse).

FER databases

ExpW:

ExpW Dataset
(Image from ResearchGate).

FER databases



The evolution of facial expression recognition in terms of datasets and methods (Image from Semantic Scholar).

Bibliography

[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.
[PIT2017] I. Pitas, “Digital video processing and analysis” , China Machine Press, 2017 (in Chinese).
[PIT2013] I. Pitas, “Digital Video and Television” , Createspace/Amazon, 2013.
[NIK2000] N. Nikolaidis and I. Pitas, “3D Image Processing Algorithms”, J. Wiley, 2000.
[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr