

CVPRW2017

<https://github.com/MarekKowalski/DeepAlignmentNetwork>

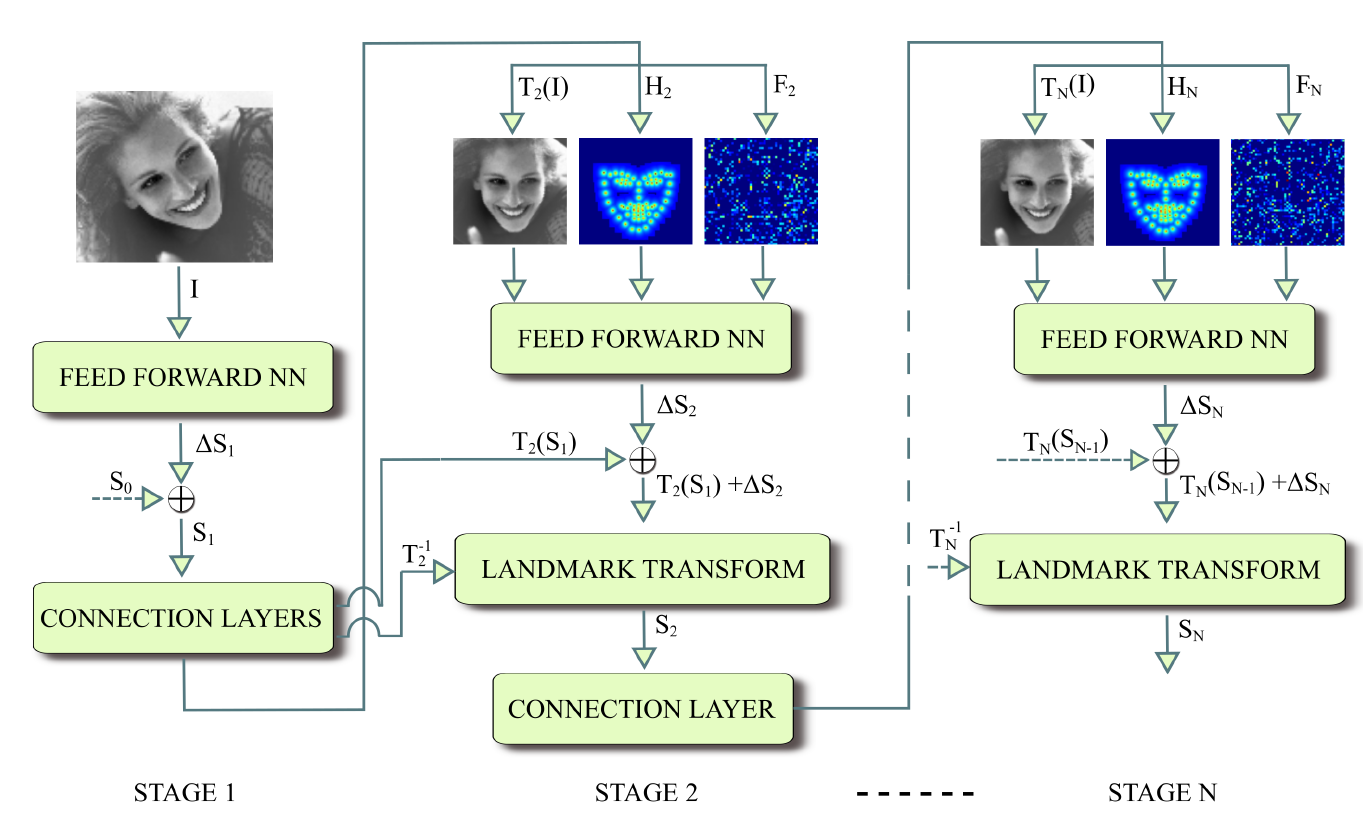
人脸关键点检测（**facial feature point detection**）也称为人脸关键点检测、定位或者人脸对齐（**face alignment**），是指给定人脸图像，定位出人脸面部的关键区域位置，包括眉毛、眼睛、鼻子、嘴巴、脸部轮廓等；我们把关键点的集合称作**形状(shape)**，形状包含了关键点的位置信息。



DAN是一个**级联**思想的关键点检测方法，通过引入关键点热图作为补充，DAN可以从整张图片进行提取特征，从而获得更为精确的定位。

Feature work:

1. **Multiple stages** where each stage improves the locations of the facial landmarks estimated by the previous stage.
2. Using entire face images at all stages by introducing a **landmark heatmap**, which is a key element of the system.
3. 输入模型过多，需要进行复杂函数运算。



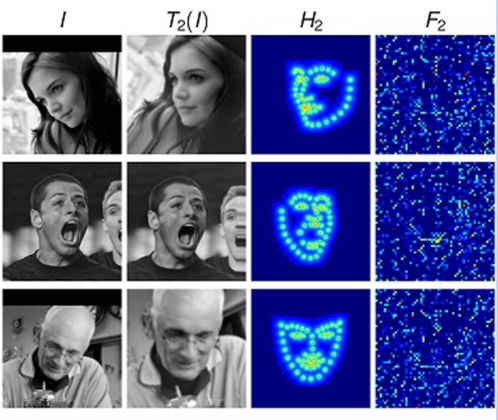
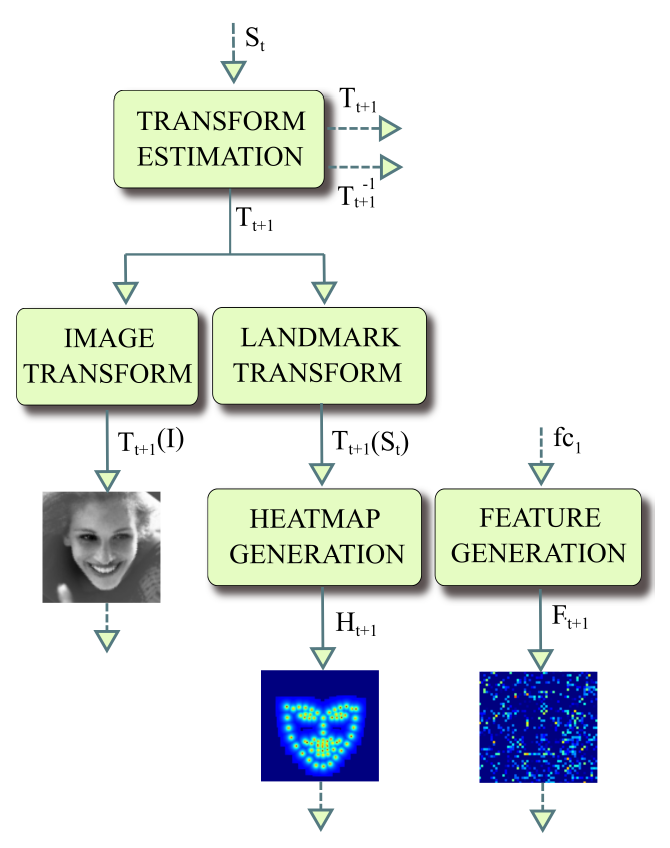
Each DAN stage consists of the **Feed Forward Neural Network**, which predicts an update **ΔSt** to the landmark location, and **the connection layers**.

**The** **connection layers** serve several purposes:

1. Generation of the transform **Tt+1**which normalizes the landmarks and the image to the canonical pose.
2. Generation of the landmark heatmap **Ht+1** based on **Tt+1(St)**.
3. Generation of the feature image **Ft+1** which is learned based on the penultimate layer of the previous stage.

* 第一阶段的输入仅有原始灰度图片和S0。面部关键点的初始化即为S0，S0是由所有关键点取平均得到，第一阶段输出S1。
* 对于第二阶段，首先，S1经第一阶段的CONNECTION LAYERS进行转换，分别得到转换后图片T2（I）、S1所对应的热图H2和第一阶段fc1层输出，这三个正是第二阶段的输入。
* 如此周而复始，直到最后一个阶段输出SN。

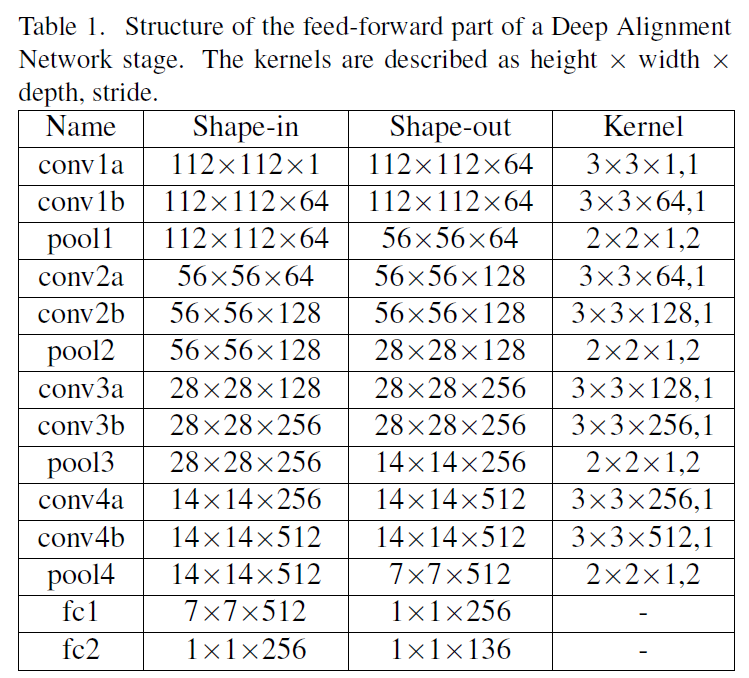
新的关键点位置会送入 “Connection Layers”，该网络示意图如下：





St is the output of the last layer of stage t and Tt-1 is the inverse of transform Tt

Feed-forward network outputs the update ΔSt to the current estimate of the landmark positions. 输出136用于预测68个关键点

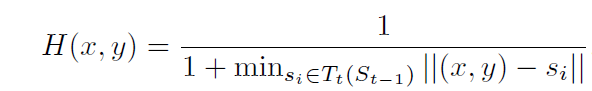


Inspired by VGG16.

Batch Normalization, ReLU activation.

A dropout layer is added before fc1.

**关键点热度图 landmark heatmap**的计算就是一个中心衰减，关键点处值最大，越远则值越小，公式如下：



**特征图feature map** 如下计算的：

The feature image layer is a dense layer which has 3136 units with ReLU activations.

输出特征为1x3136，reshape为 56x56, 然后上采样到 112x112，和输入图像一样大。

为什么需要从fc1层生成一张特征图？文中提到“Such a connection allows any information learned by the preceding stage to be transferred to the consecutive stage.”其实就是人为给CNN增加上一阶段信息。

**Implementation**:

**Data augmentation** is performed by mirroring around the Y axis as well as random translation, rotation and scaling, all sampled from normal distributions. During data augmentation a total of 10 images are created from each input image in the training set.

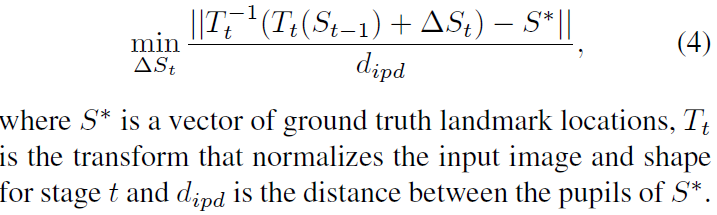
Both models (DAN and DAN-Menpo) consist of **two stages**. For optimization we use Adam stochastic optimization with an initial step size of 0.001 and mini batch size of 64. For validation we use a random subset of 100 images from the training set.

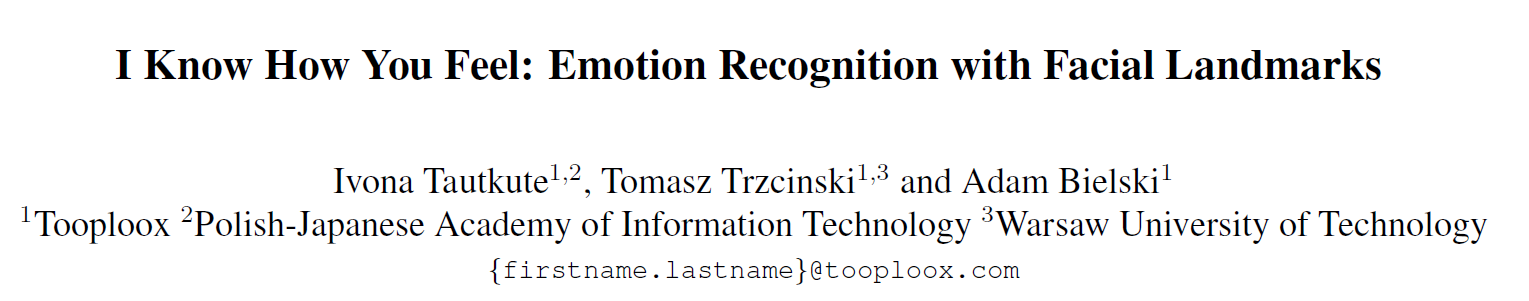
**Training Procedure**

The stages of DAN are trained **sequentially.**

The first stage is trained by itself until the validation error stops improving.

Subsequently the connection layers and the second stage are added and trained. This procedure is repeated until further stages stop reducing the validation error.

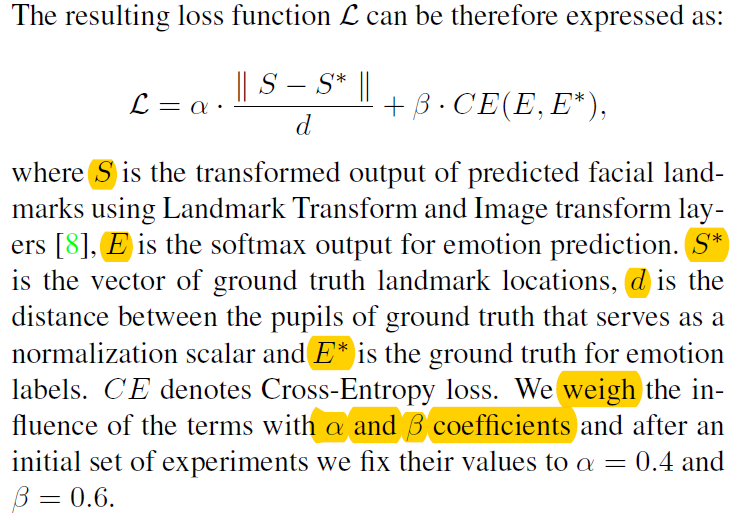




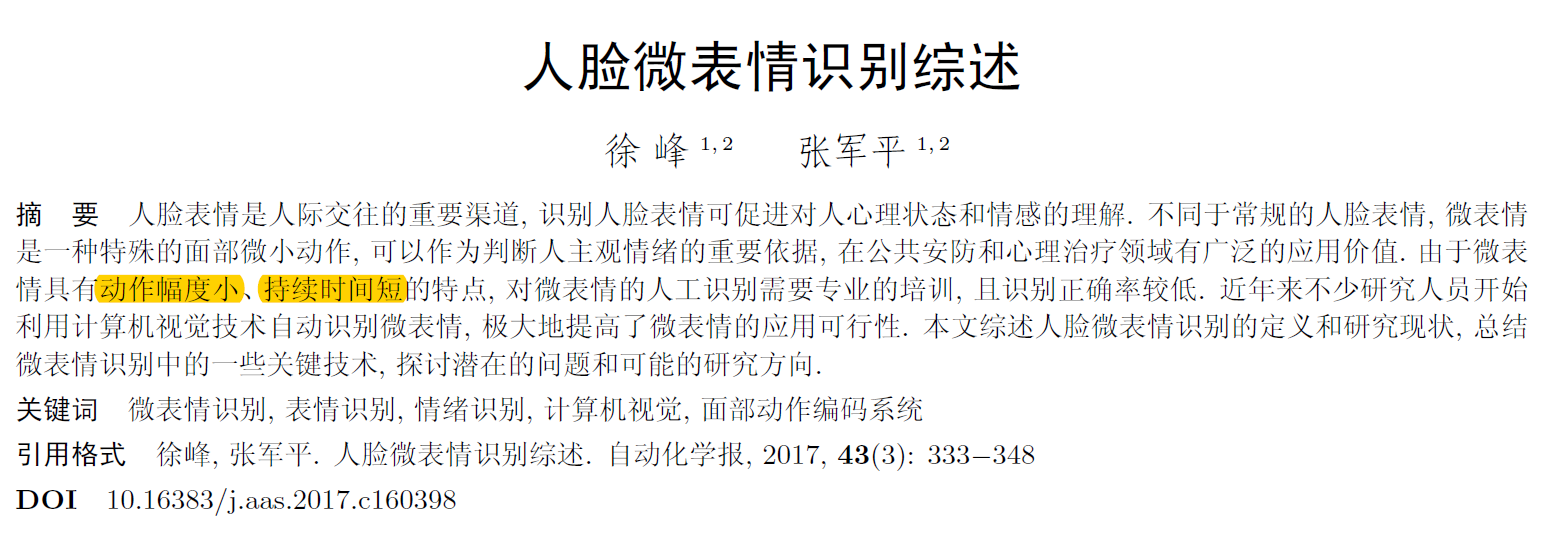
CVPR WiCV workshop 2018

Based on DAN, extend the network learning task with an additional goal of estimating expressed facial emotions, which is called **EmotionalDAN**.

Minimize both landmark location and emotion recognition terms **jointly**.



语焉不详



Affective estimation

Short duration and subtle movement

Micro expression, spontaneous expression

表明微表情仅持续1/25 s~1/3 s, 且动作幅度非常小, 不会同时在上半脸和下半脸出现

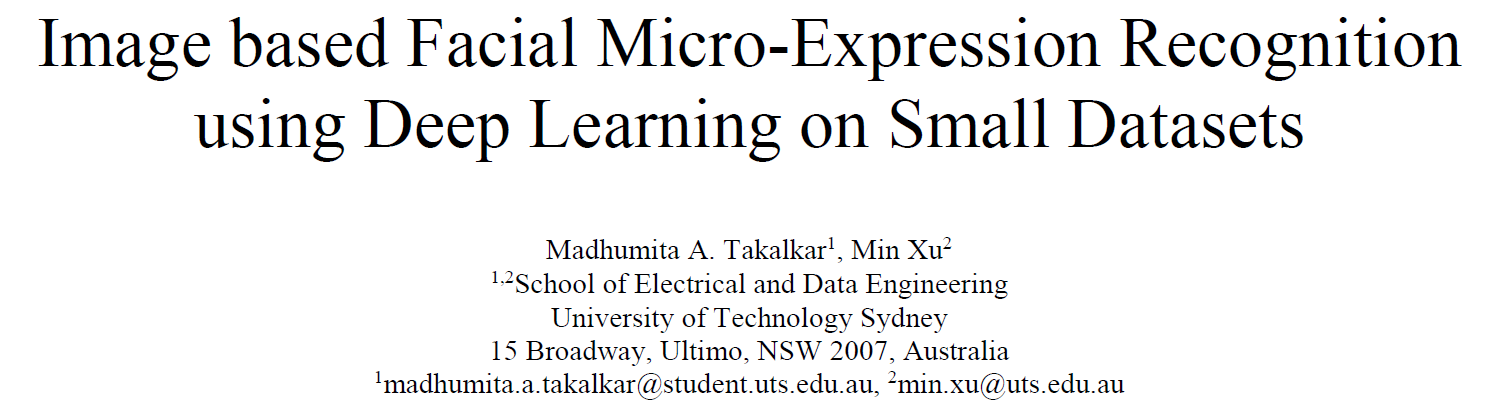
高速摄像机用于微表情的捕捉

问题分类：

1. 微表情检测
2. 微表情分类

人在试图掩盖自己情绪时的微小面部动作。严格地说，人主观模拟的微小表情不能称为微表情。因此诱导方法决定了微表情数据集的可靠程度。

现有方法其共性在于从时空纹理的角度挖掘面部表情的变化，具有很强的描述能力, 但是计算得到的特征的可解释性欠佳。

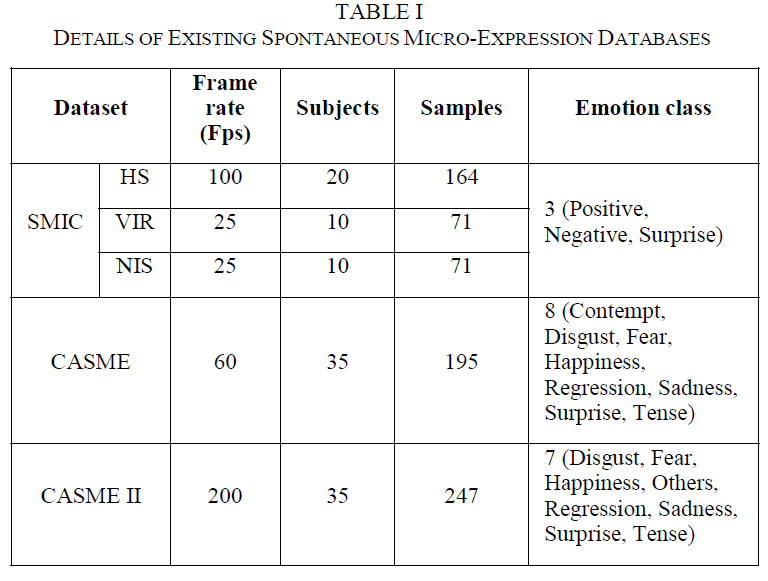


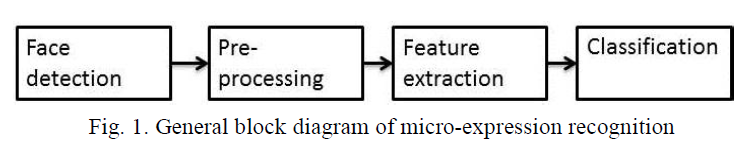
978-1-5386-2839-3/17/$31.00 ©2017 IEEE

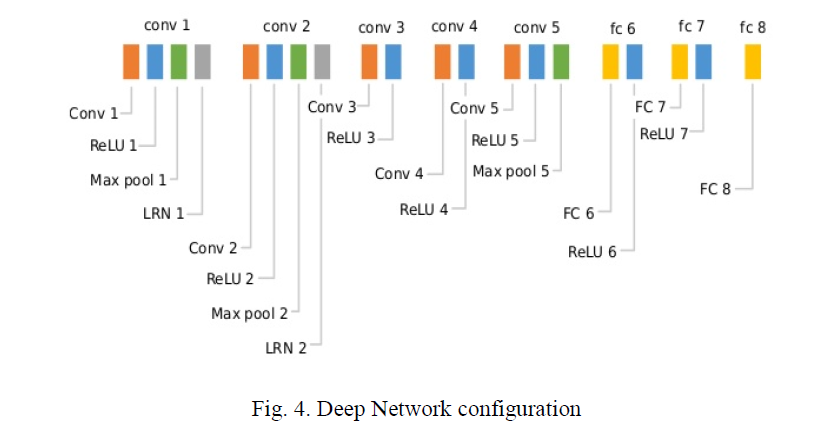
To convert the labelled micro-expression **videos into frames** and use these labelled frames for training the CNN model for image based micro-expression recognition. The frames are extracted every 0.2 seconds (or 200 milliseconds) from the video.

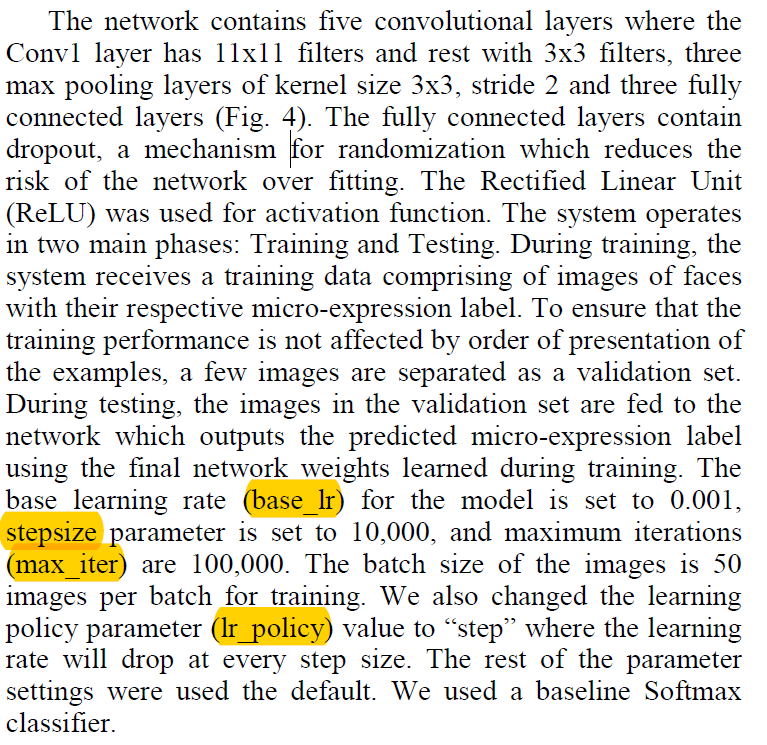
Databases:

To combine the widely used micro-expression databases CASME and CASME II to increase the number of samples for training CNN model. The Chinese Academy of Sciences Micro-Expression (CASME)









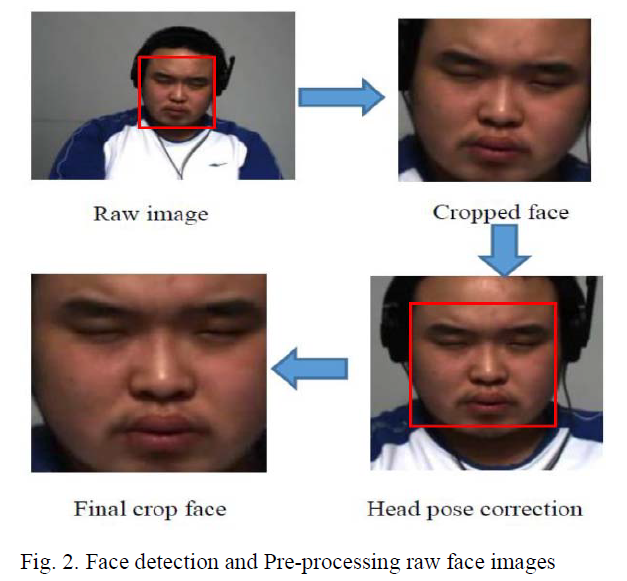
**Data Augmentation:**

both original and vertically mirrored images are used for training

**Face detection and Pre-processing:**

implemented the DLib face detector in **OpenCV** to detect and crop the face region from the raw image. The cropped face is then processed for head pose correction by computing the angle between the eye centroids and later applying **the affine transformation**. The transformed image is again passed to DLib face detector to crop and save the more accurate face region.

The head pose is aligned with the use of affine transformation.



**Finetuning** VGGFace for micro-expressions

VGGFace, a network trained on a very large-scale face images dataset (2.6M images, 2.6k people) for the task of face recognition.

**Difficulties with certain expressions:**

using a small **unbalanced** dataset with classes that are not visually distinctive

**Features and classifiers**

a good feature should be informative, invariant to noise and a set of transformations (e.g., rotation and translation), and fast to compute

The task of the classifier is to use the feature vector to assign an image or region of interest (RoI) to a category



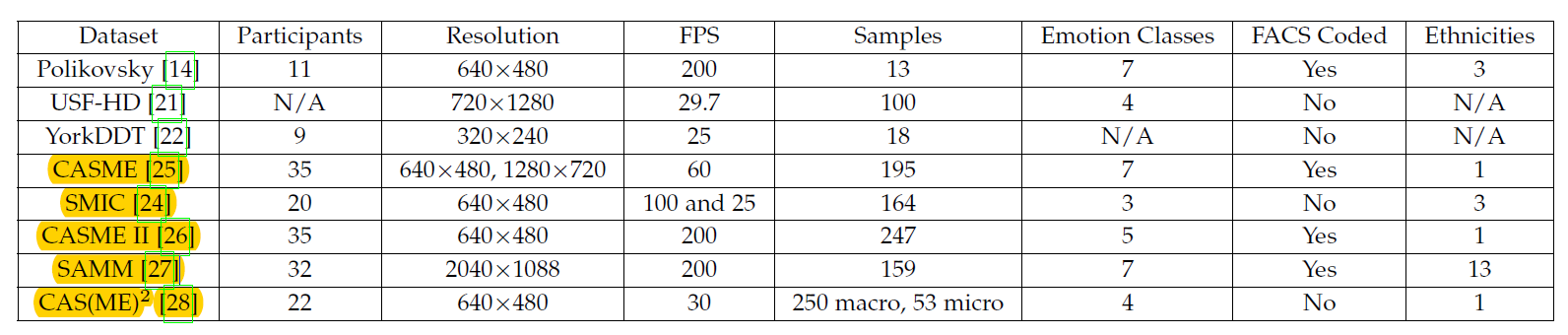
arXiv:1805.02397v1 [cs.CV] 7 May 2018

Micro-expression occur when a person attempts to conceal their true emotion.

Duration is the main feature that distinguish **micro-expressions** from **macro-facial expressions**. These two topics should be looked upon as **different** research problems.

Automated facial micro-expression recognition is still in its **infancy** when compared to the facial macro-expression.

**Datasets:**



Non-spontaneous datasets and Spontaneous datasets (yellow background)

Mainstream of research: CASME II and SAMM.

**Features:**

3D Histogram of Oriented Gradients

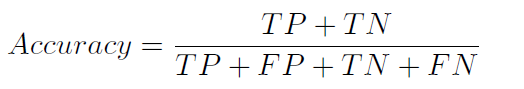
Local Binary Pattern-Three Orthogonal Planes

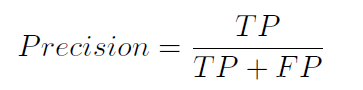
Histogram of Oriented Optical Flow

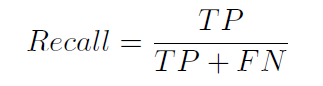
Deep learning Approaches

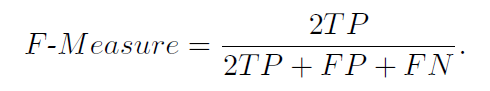
Other Feature Extraction Methods

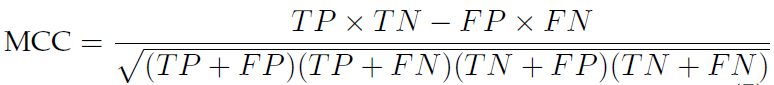
**Metrics：**











**Validation Techniques：**

n-fold cross validation - n折交叉验证

leave-one-subject-out (LOSO) - 留一交叉验证（ n折交叉验证的特殊情形是n=N，这里N是给定数据集的容量 ）

**Challenges:**

Datasets: resolution (critical) and frame rates.

The conventional methods are relies on **feature descriptors** and varies from one to another.

Emotional classes versus objective classes (FACS) in data labelling.

Face regions in data analysis: block-based face regions, FACS-based region, and Delaunay triangulation.

**The main factors for conventional approaches:**

Spatial temporal settings during data collection. Preprocessing stage of dataset including face alignment and face regions split, feature extraction methods and the type of classifiers.

**Deep learning method difficulty:**

Minimal data available to train from,

GPU memory (12GB) cannot load such large amounts of data (long video), leading to the minimization of the batch size and reduction of resolution to allow for training to proceed.

**3D ConvNets**, capable of extracting and learning from both motion and appearance.

Micro-facial movements describe the facial muscle activations.

**Future directions:**

How the dataset is captured (special temporal settings)

Labeling of the dataset based on Action Unit based objective classes, FACS-based face regions for better localization, end-to-end solution using deep learning

Fair evaluation using standardized metrics (ideally F1-score and MCC)

LOSO as the validation technique