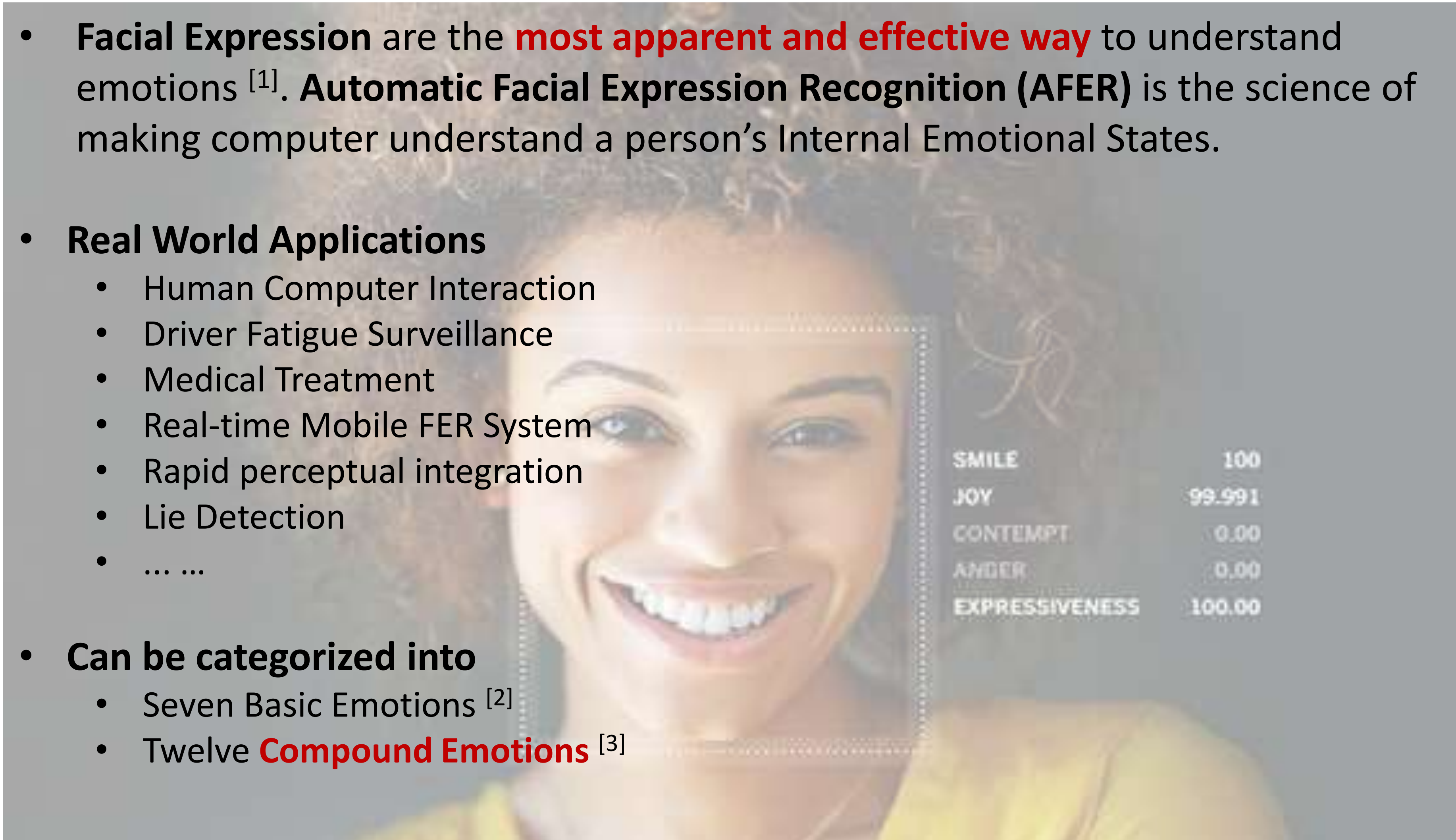
Weihong Deng @ BUPT

Background

- **Facial Expression** are the **most apparent and effective way** to understand emotions [1]. **Automatic Facial Expression Recognition (AFER)** is the science of making computer understand a person's Internal Emotional States.

- **Real World Applications**
  - Human Computer Interaction
  - Driver Fatigue Surveillance
  - Medical Treatment
  - Real-time Mobile FER System
  - Rapid perceptual integration
  - Lie Detection
  - … …

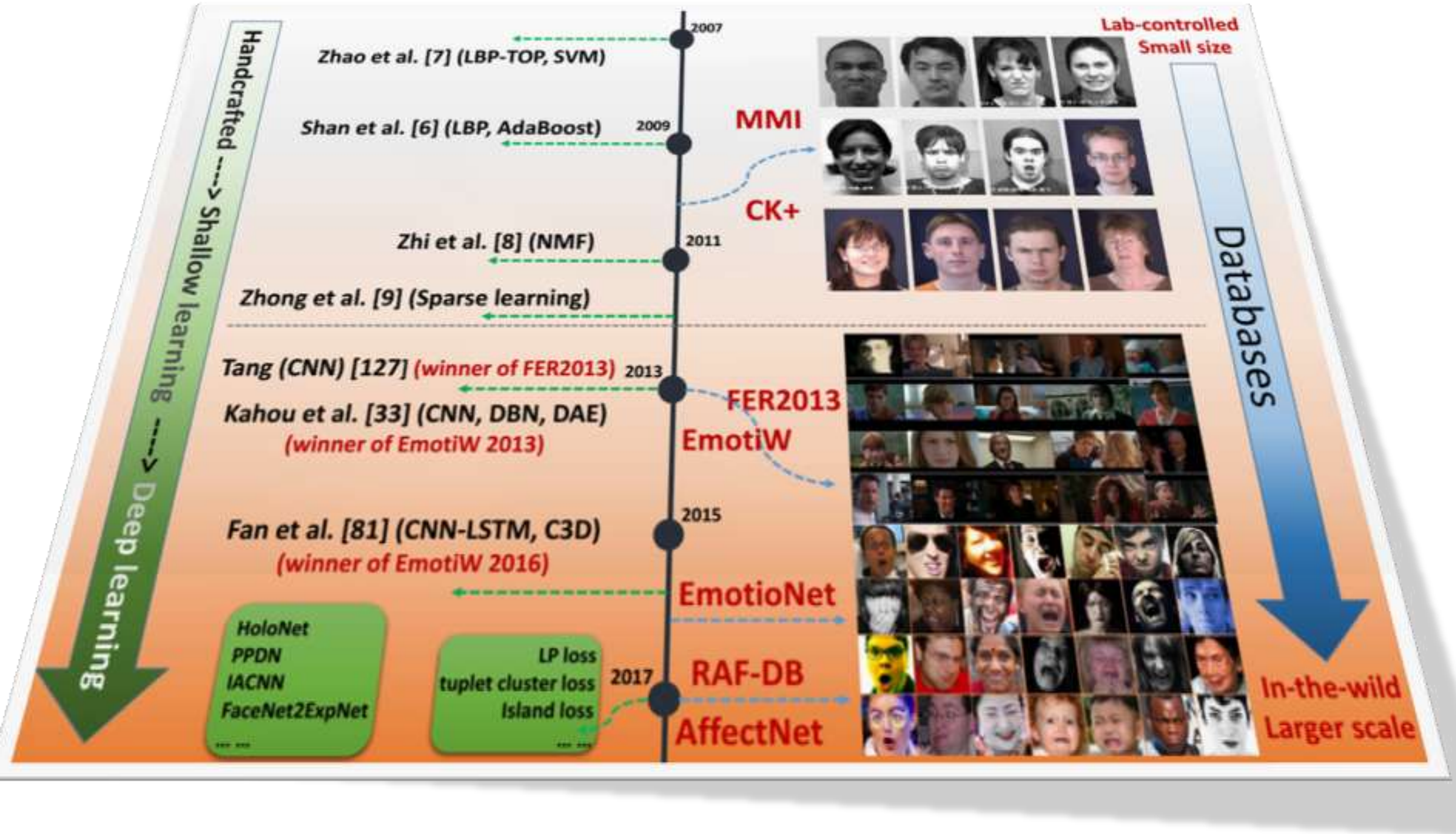
- **Can be categorized into**
  - Seven Basic Emotions [2]
  - Twelve **Compound Emotions** [3]



1. " Nonverbal communication", M. Anderson, 1987.
2. " Facial expression and emotion", P. Ekman, 1993.
3. "Compound facial expressions of emotion", Martinez et al. PNAS 2014.

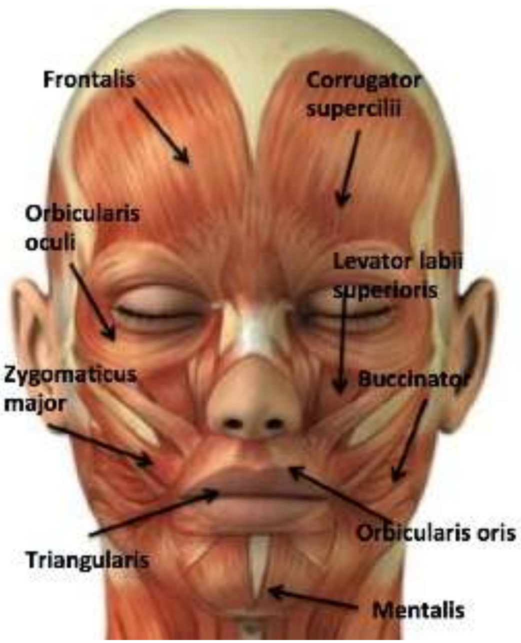
Facial Expression Recognition

Shan Li & W. Deng, Deep Facial Expression Recognition: A Survey
(arXiv:1804.08348)



Data Challenge

- **Opportunities**
  - **Millions of** images are being uploaded every day by users from different events and social gatherings.
- **Challenges**
  - **Annotation of facial expression categories within expertise knowledge is difficult and time-consuming.**



Solution

- **Our Ideas**
  - Expression is perceived by the **public**, rather than experts.
  - **Crowdsourcing** is an efficient tool to collect the judgments of annotation results from a large common population.

Real-world Affective Face Database *(RAF-DB*)*

1. Collection

Keywords
'smile' 'crying' 'OMG'...
60,000 images downloaded

→ flickr's API → XML response → parse → URLs → download → *~ 30,000 images*

- Image Collection
  - **Flickr (Image social network)**
    - https://api.flickr.com/services/rest/?method=flickr.photos.search&api_key={}&text={}&tags={}&per_page={}&page={}&sort=relevance
    - XML response→Interpreted into URLs of the images→Download


```
<collection PhotosOfOneSearch="Search1">
<photos page="1" pages="127" perpage="500" total="63478">
    <photo id="14197338518" owner="74529773@N07" secret="8313e97a1f" server="2909" farm="3" title="null" ispublic="1" isfriend="0" isfamily="0" />
    <photo id="8505470995" owner="69642848@N05" secret="375b0c82bc" server="8111" farm="9" title="null" ispublic="1" isfriend="0" isfamily="0" />
    <photo id="14744669763" owner="96619214@N04" secret="a818044e97" server="5557" farm="6" title="null" ispublic="1" isfriend="0" isfamily="0" />
    <photo id="3568274837" owner="22505098@N04" secret="31f8cd91db" server="3358" farm="4" title="null" ispublic="1" isfriend="0" isfamily="0" />
    <photo id="6109626903" owner="45833131@N03" secret="176b96e284" server="6201" farm="7" title="null" ispublic="1" isfriend="0" isfamily="0" />
    <photo id="8204874510" owner="35456872@N00" secret="61d0c90451" server="8207" farm="9" title="null" ispublic="1" isfriend="0" isfamily="0" />
    <photo id="335131224" owner="85353067@N00" secret="cae5519488" server="151" farm="1" title="null" ispublic="1" isfriend="0" isfamily="0" />
    <photo id="5821864725" owner="55919672@N08" secret="81e246c9fe" server="2603" farm="3" title="null" ispublic="1" isfriend="0" isfamily="0" />
    <photo id="113989596" owner="64705987@N00" secret="601e305e76" server="56" farm="1" title="null" ispublic="1" isfriend="0" isfamily="0" />
    <photo id="858252701" owner="9722602@N05" secret="805210523d" server="1334" farm="2" title="null" ispublic="1" isfriend="0" isfamily="0" />
```

Real-world Affective Face Database *(RAF-DB)*

2. Annotation

Learning from labels →

1,200,000 labels

- Image Annotation
  - **Crowd-sourcing**
    - **315 well-trained annotators** were asked to label facial images with one of the seven basic categories
    - Each image is annotated enough times independently, i.e., around **40 times** in our experiment.

Real-world Affective Face Database *(RAF-DB)*

3. Reliability Estimation → **EM** framework → Filter out unreliable labels → *Optimal Reliability*

- Reliability Estimation
  - **Filter noisy annotators and labels**
    - an Expectation Maximization (EM) framework was used to assess each labeler's reliability.

Algorithm 1 Label reliability estimation algorithm.

Input: Training set $D = \{(x_j, t_j^1, t_j^2, ..., t_j^R)\}_{j=1}^n$

Output: Each annotator's reliability α_i^*

Initialize:

$\forall j = 1, ..., n$, initialize the true label y_j using majority voting

$$\beta_j := -\sum_{i=1}^{R} p(t_j^i) \ln p(t_j^i),\ \alpha_i := 1,$$

The initial value of β_j is image j's entropy. The higher the entropy, the more uncertain the image.

Repeat:

E-step:

$$Q_j(y_j) := \prod_i p(y_j | t_j, \alpha_i, \beta_j)$$

M-step:

$$\alpha_i := \arg\max_{\alpha_i} \sum_j \sum_{y_j} Q_j(y_j) \ln \frac{p(t_j, y_j | \alpha_i, \beta_j)}{Q_j(y_j)}$$

Until convergence

"if the stimulus does contain an **emotion blend**, and the investigator allows only a single choice which does not contain blend terms, low levels of agreement may result, since **some of the observers** may choose a term for one of the blend components, **some** for another."

Paul Ekman. 2013.

Real-world Affective Face Database *(RAF-DB)*

Data collection and Annotation Process

1. Collection

Keywords
'smile' 'crying' 'OMG'…
60,000 images downloaded

flickr → XML → parse → URLs → download → 30K images

2. Annotation

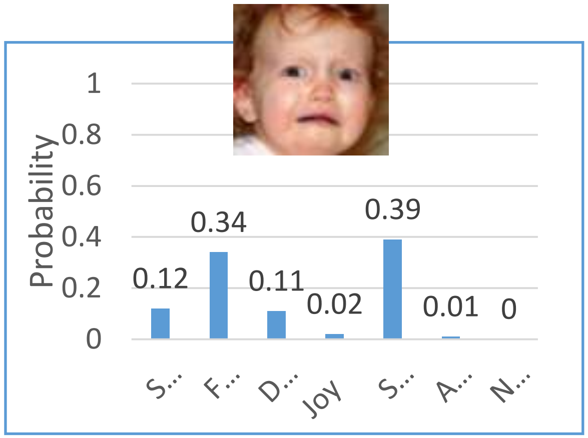
Crowd-sourcing
315 volunteers online
Each image labelled 40 times

Learning from labels → 1.2M labels

3. Reliability Estimation

An **EM** framework → Filter out unreliable labels → *Enhanced Reliability*

Results

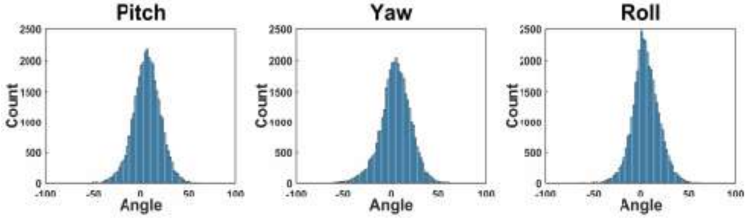


Seven Basic Emotions & Twelve Compound Emotions

RAF-DB

Real-world Affective Face Database *(RAF-DB)*

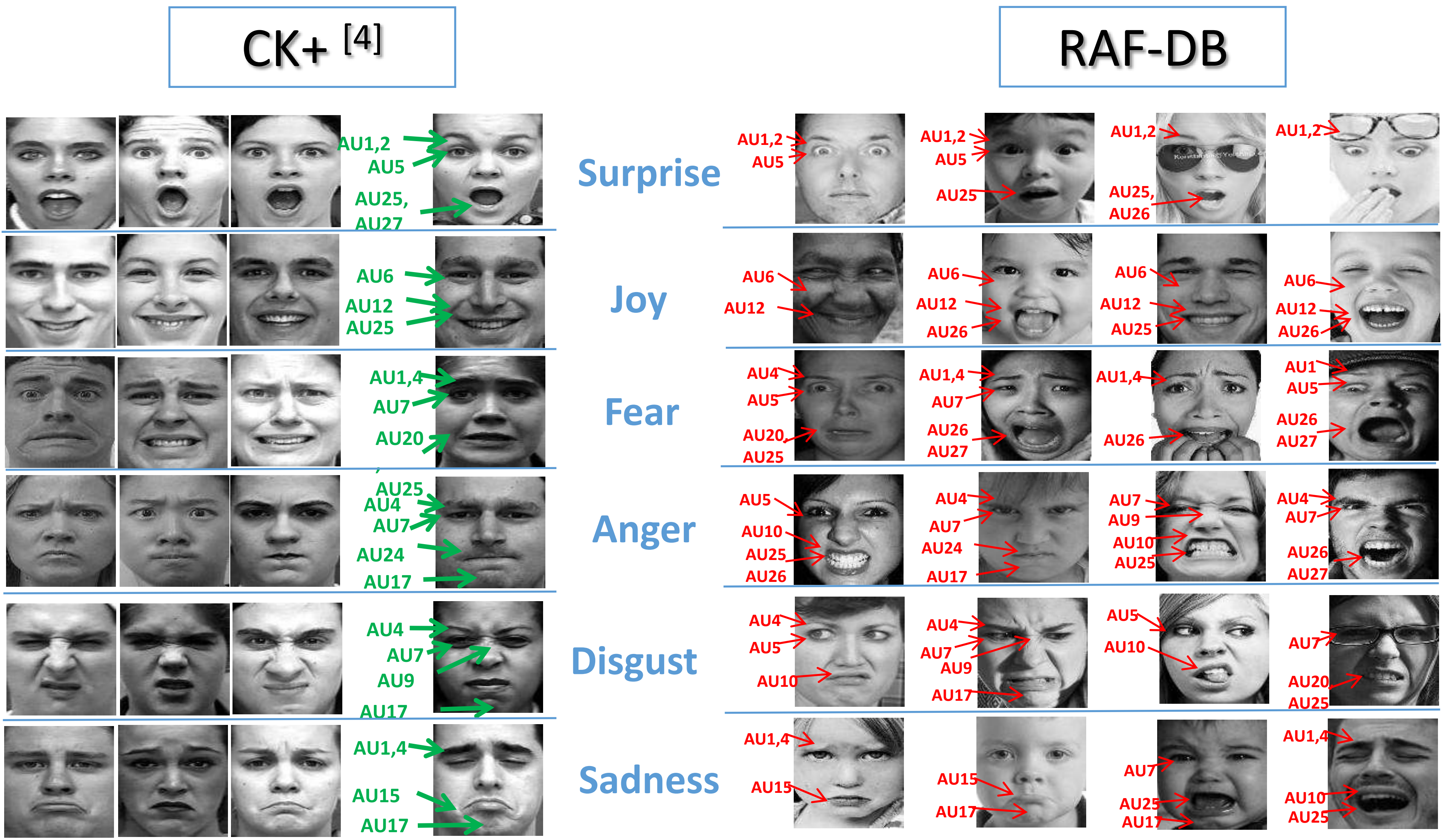
- **Database Statistics**
  - **29672** number of **real-world images**,
  - a 7-dimensional **expression distribution** vector for each image,
  - two different subsets: **single-label subset**, including **7** classes of basic emotions; **two-tab subset,** including **12** classes of compound emotions,
  - 5 **accurate landmark locations**, 37 **automatic landmark locations**, **race**, **age range** and **gender attributes** annotations per image.



Age distribution				
0∼3	**4∼19**	**20∼39**	**40∼69**	**70+**
2696	4731	16460	4696	1089
9.09%	15.94%	55.47%	15.83%	3.67%

Poster 72： Reliable crowdsourcing and deep locality preserving learning for expression recognition in the wild, Shan Li & W. Deng, CVPR 2017

Action Units: RAF-DB is more diverse



4. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression", Lucey, CVPRW 2010

Real-world Affective Face Database *(RAF-DB*)*

7 classes Basic Emotions

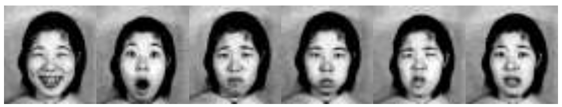
12 classes Compound Emotions

Poster 72： Reliable crowdsourcing and deep locality preserving learning for expression recognition in the wild, Shan Li & W. Deng, CVPR 2017

Comparison: Real-world & lab-controlled

LAB-BASED Datasets

Controlled lab conditions



Prototypical emotions

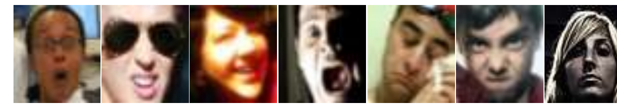
Surprise!

Balanced distribution

≈1:1:1:1:1:1

REAL-WORLD RAF-DB

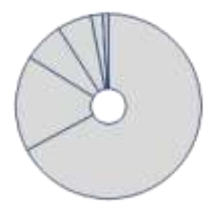
Diverse imaging conditions



Compound emotions

Fear? Sad?

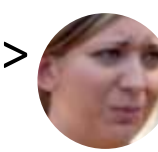
Highly-imbalanced



e.g. 'Joy'

>>

'Disgust'



①

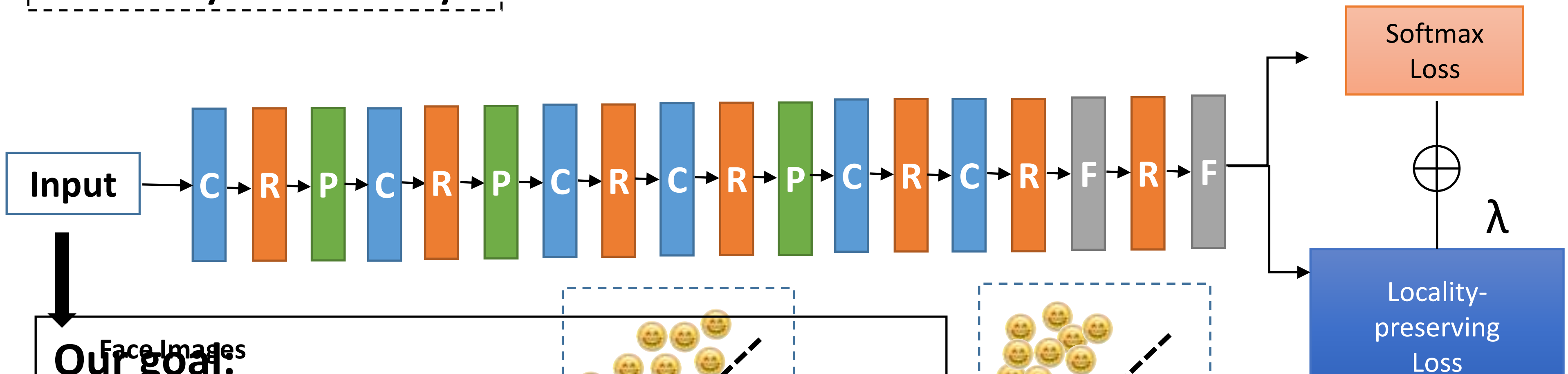
②

③

Poster 72: Reliable crowdsourcing and deep locality preserving learning for expression recognition in the wild, Shan Li & W. Deng, CVPR 2017

DLP-CNN: Deep Locality-preserving CNN

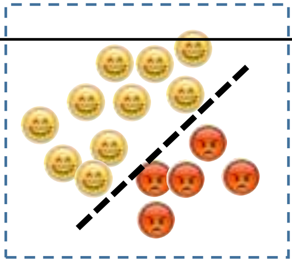
C: The convolution layer
P: The max-pooling layer
R: The ReLU layer
F: The fully connected layer



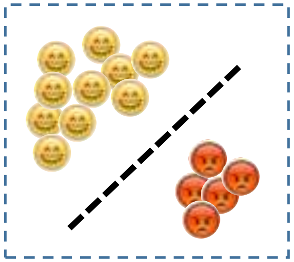
Face Images

Our goal:

$$\min \sum_{i,j} S_{ij} ||x_i - x_j||_2^2$$



Separable Features



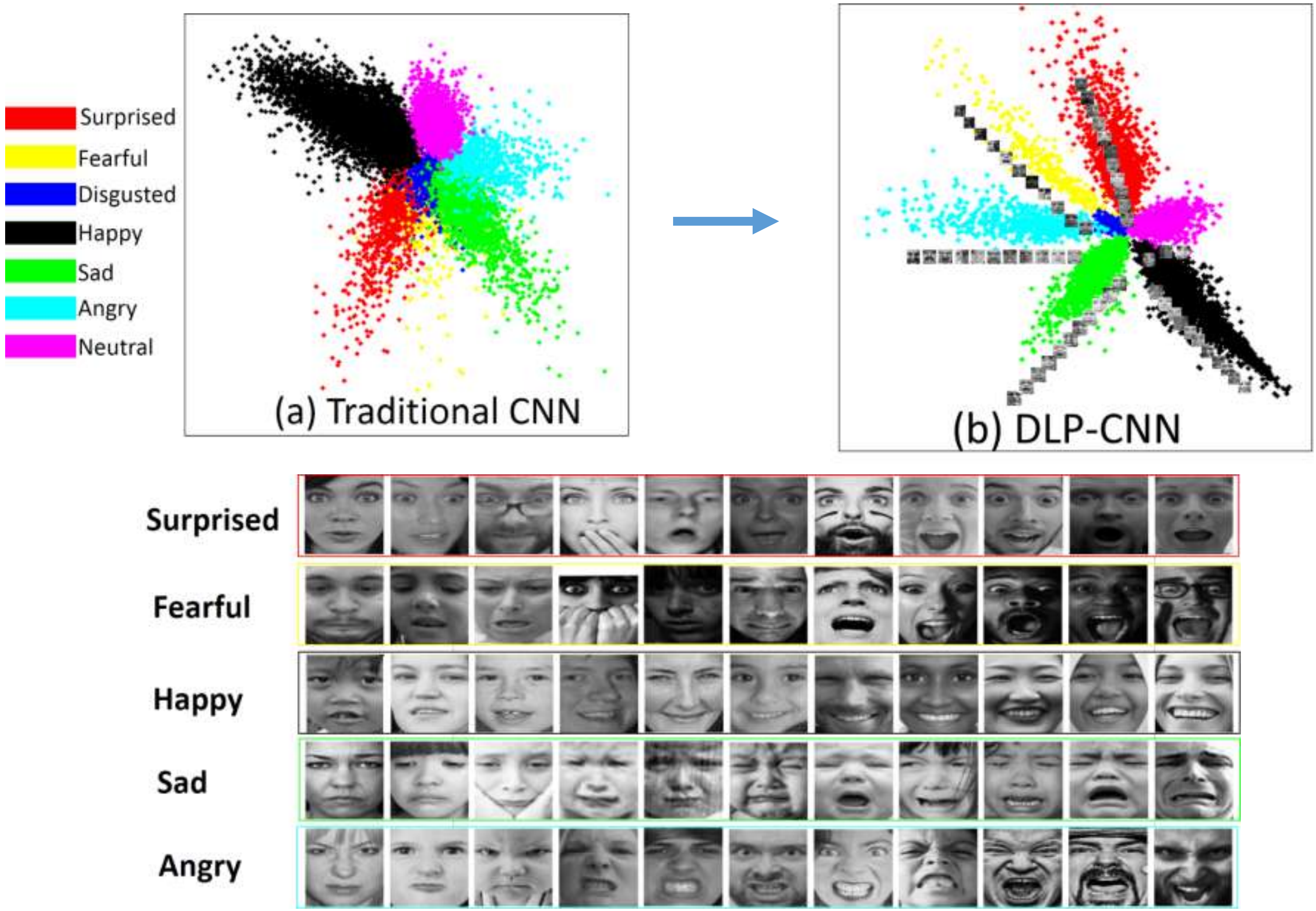
Discriminative Features

$$S_{ij} = \begin{cases} 1, & x_j \text{ is among } k \text{ nearest neighbors of } x_i \textbf{ or} \\ & x_i \text{ is among k nearest neighbors of } x_j \\ 0, & otherwise \end{cases}$$

Locality Preserving Loss:

$$L_{lp} = \frac{1}{2n} ||x_i - \frac{1}{k} \sum_{x \in N_k\{x_i\}} x \,||_2^2$$

DLP-CNN: Deep Locality-preserving CNN



DLP-CNN: Experiment Results

Table 1. Expression recognition performance of different DCNNs on RAF. The metric is the mean diagonal value of the confusion matrix.

		basic								compound
		Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral	**Average**	**Average**
mSVM	VGG [6]	68.52	27.50	35.13	85.32	64.85	66.32	59.88	**58.22**	**31.63**
	AlexNet [7]	58.64	21.87	39.19	86.16	60.88	62.31	60.15	**55.60**	**28.22**
	baseDCNN	70.99	52.50	50.00	92.91	77.82	79.64	83.09	**72.42**	**40.17**
	center loss [8]	68.52	53.13	54.05	93.08	78.45	79.63	83.24	**72.87**	**39.97**
	DLP-CNN	71.60	52.15	62.16	92.83	80.13	81.16	80.29	**74.20**	**44.55**
LDA	VGG [6]	66.05	25.00	37.84	73.08	51.46	53.49	47.21	**50.59**	**16.27**
	AlexNet [7]	43.83	27.50	37.84	75.78	39.33	61.70	48.53	**47.79**	**15.56**
	baseDCNN	66.05	47.50	51.35	89.45	74.27	76.90	77.50	**69.00**	**28.23**
	center loss [8]	64.81	49.38	54.05	92.41	74.90	76.29	77.21	**69.86**	**27.33**
	DLP-CNN	77.51	55.41	52.50	90.21	73.64	74.07	73.53	**70.98**	**32.29**

6. Simonyan & Zisserman, *arXiv:1409.1556* (2014).
7. Krizhevsky et al. NIPS, 1097–1105 (2012).
8. Wen et al. ECCV, 499–515 (2016).

DLP-CNN: Experiment Results

Table 2. Comparison results of DLP-CNN and other state-of-the-art methods on CK+, SFEW and MMI databases. To validate the generalization of our model, the well-trained DLP-CNN has been employed as a feature extraction tool without finetune.

(a) CK+

Methods	Accuracy
CSPL [9]	88.89%
FP+SAE [10]	91.11%
AUDN [11]	92.05 %
AURF [11]	92.22 %
3DCNN-DAP [12]	92.4 %
Inception [13]	93.2%
Dis-ExpLet [14]	95.1%
ESL [15]	95.33%
DLP-CNN (without finetune)	**95.78%**

(b) SFEW 2.0

Methods	Accuracy
DL-GPLVM [16]	24.70%
AUDN [11]	26.14%
STM-ExpLet [17]	31.73%
Inception [13]	47.7%
SFEW third [18]	48.5%
SFEW second [19]	52.29%
SFEW best [20]	52.5%
DLP-CNN (without finetune)	**51.05%**

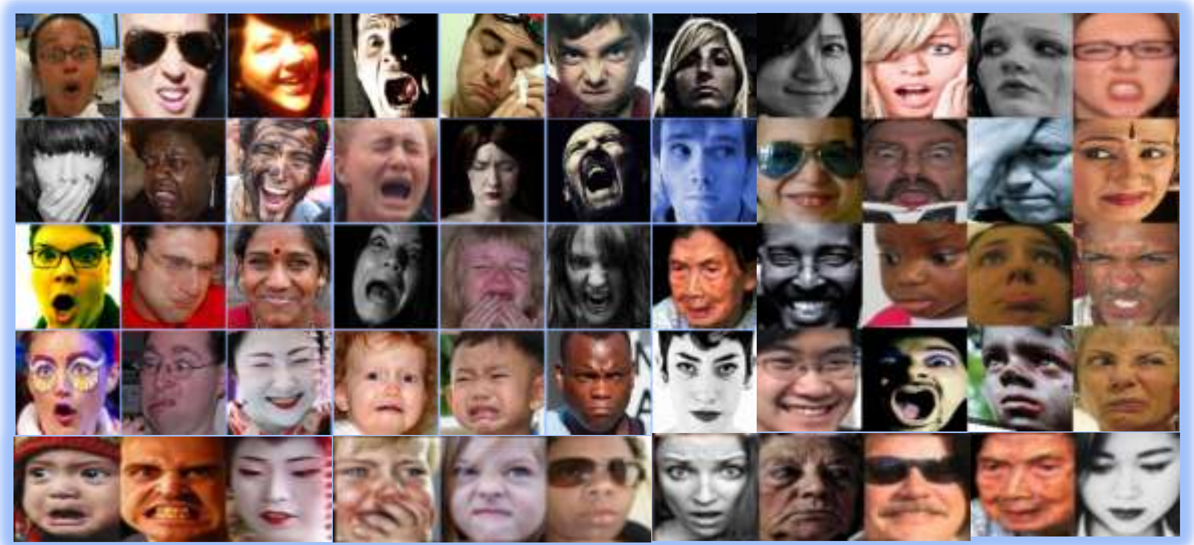
(c) MMI

Methods	Accuracy
3DCNN-DAP [12]	63.4%
DTAGN [21]	70.24%
CSPL [9]	73.53%
AUDN [11]	74.76%
STM-ExpLet [17]	75.12%
F-Bases [22]	75.12%
Inception [13]	77.6%
Dis-ExpLet [14]	77.6%
DLP-CNN (without finetune)	**78.46%**

9. Zhong et al. *CVPR,* 2562–2569 (2012).
10. LV et al. SMARTCOMP, 303–308 (2014).
11. Liu et al. FG, 1–6 (2013).
12. Liu et al. ACCV, 143-157 (2014).
13. Mollahosseini et al. WACV, 1-10 (2016).
14. Liu et al. IEEE TIP, 25(12):5920–5932, (2016).
15. Shojaeilangari et al. IEEE TIP, 24(7):2140–2152, (2015).
16. Eleftheriadis et al. IEEE TIP, 24(1):189–204, (2015).
17. Liu et al. CVPR, 1749–1756 (2014).
18. Ng et al. ICMI, 443-449 (2015).
19. Yu et al. ICMI, 435-442 (2015).
20. Kim et al. ICMI, 427-434 (2015).
21. Jung et al. CVPR, 2983–2991 (2015).
22. Sariyanidi et al. IEEE TIP, 26(4):1965-1978, (2017).

Domain Adaption: From RAF-DB to other datasets

Source:



RAF-DB

Target: JAFFE

Target: CK+

Target: MMI

Domain Adaption: From RAF-DB to other datasets

Table 3. Comparison results of **cross-database** experiments on **CK+**.

	Methods	Source Dataset	Accuracy
Shallow Models	Zhang et al. [23]	MMI	61.20%
	Miao et al. [24]	MMI + JAFFE	65.0%
	Mayer et al. [25]	MMI + JAFFE	66.20%
Deep Models	Mollahosseini et al. [13]	6 Datasets*	64.2%
	Hasani et al. [26]	MMI + JAFFE	67.52%
	Hasani et al. [27]	MMI + JAFFE	73.91%
	Wen et al. [28]	FER2013	76.05%
Our Methods	CNN	RAF-DB	75.49%
	CNN + DA	RAF-DB	78.83%

Table 4. Comparison results of **cross-database** experiments on **MMI**.

	Methods	Source Dataset	Accuracy
Shallow Models	Shan et al. [29]	CK	51.10%
	Mayer et al. [25]	CK	60.30%
	Zhang et al. [23]	CK+	66.90%
Deep Models	Mollahosseini et al. [13]	6 Datasets*	55.6%
	Hasani et al. [26]	CK+	54.76%
Our Methods	CNN	RAF-DB	63.92%
	CNN + DA	RAF-DB	66.05%

Table 5. Comparison results of **cross-database** experiments on **JAFFE**.

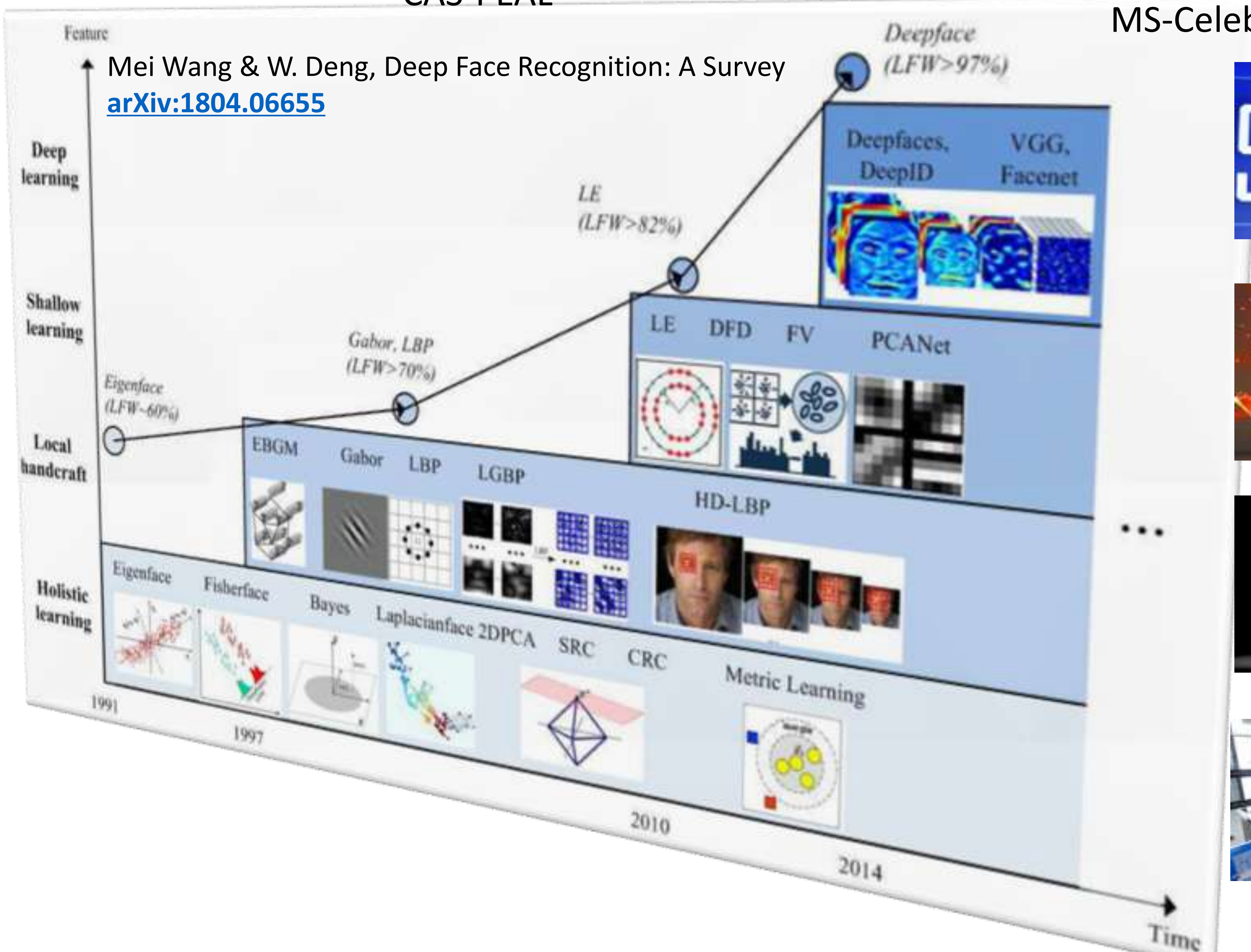
	Methods	Source Dataset	Accuracy
Shallow Models	Shan et al. [29]	CK	41.30%
	El et al. [30]	Bu-3DFE	41.96%
	Zhou et al. [31]	CK	45.71%
Deep Models	Wen et al. [28]	FER2013	50.70%
	Ali et al. [32]	RaFD	48.67%
Our Methods	CNN	RAF-DB	51.17%
	CNN + DA	RAF-DB	57.75%

23. Zhang et al. MVA, 467–483 (2015).
24. Miao et al. ICMLA, 326–332 (2012).
25. Mayer et al. PRIA, 124–132 (2014).
26. Hasani et al. arXiv:1705.07871 (2017).
27. Hasani et al. arXiv:1703.06995 (2017).
28. Wen et al. Cognitive Computation, 1–14 (2017).
29. Shan et al. IVC, 803–816 (2009).
30. El et al. Affective Computing, 141–154 (2014).
31. Zhou et al. 2013
32. Ali et al. PR, 14-27 (2016).

Face Recognition

ORL AR FERET EYB MPIE FRGC CAS-PEAL **LFW** YTF IJB-A/B/C Megaface MS-Celeb-1M

Mei Wang & W. Deng, Deep Face Recognition: A Survey
arXiv:1804.06655



...

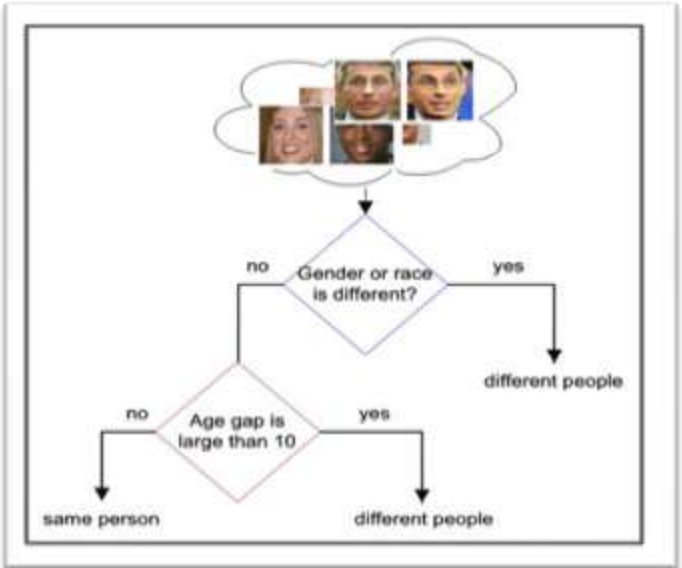
On 100% Accuracy on LFW

Pros: T/N pairwise simplicity, #people invariance, human-machine comparison

Cons: Insufficient challenge due to random selection: inter/intra-personal variations

Negative pairs

Positive pairs



a face blindness decision tree (gender, race & age) yield 86.25% accuracy

On 100% Accuracy on LFW

Pros: T/N pairwise simplicity, #people invariance, human-machine comparison

Cons: Insufficient challenge due to random selection: inter/intra-personal variations

Negative pairs

Positive pairs

Similar-looking

Aging

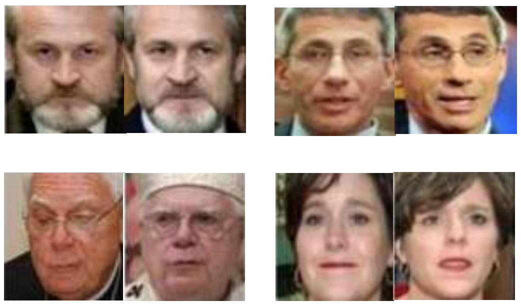
Poses

From LFW to SL/CA/CPLFW

Identical celebrities, scale, and protocols

Similar-Looking

3K positive pairs



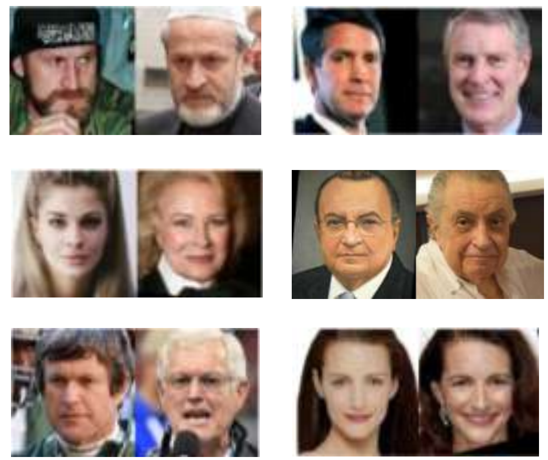
3K negative pairs

Similar-look face pairs
selected by crowd-sourcing

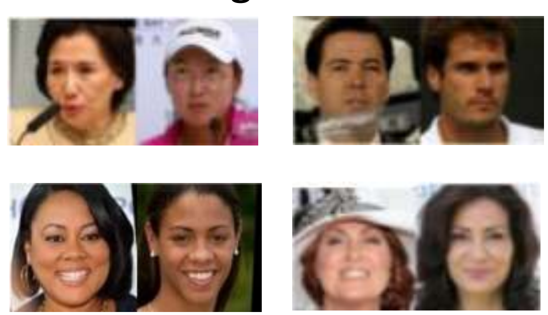
Cross-Age

3K positive pairs

Cross-age face pairs
selected by crowd-sourcing



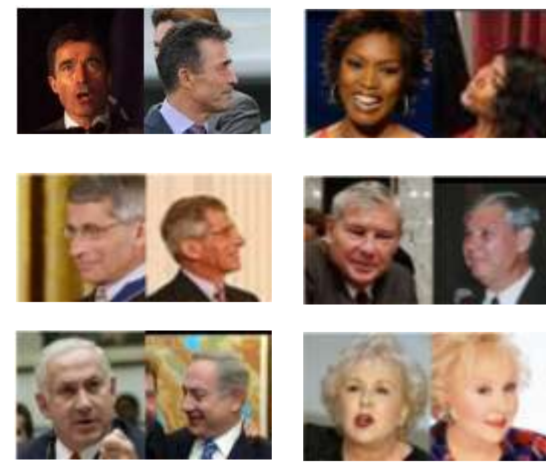
3K negative pairs
with same gender and race



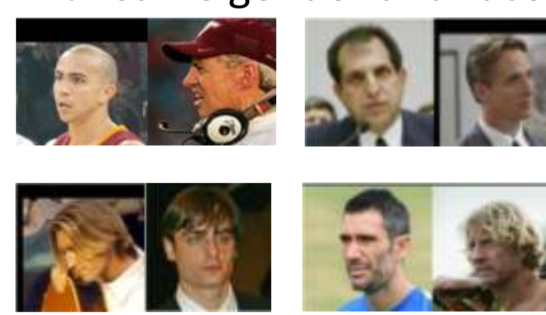
Cross-Pose

3K positive pairs

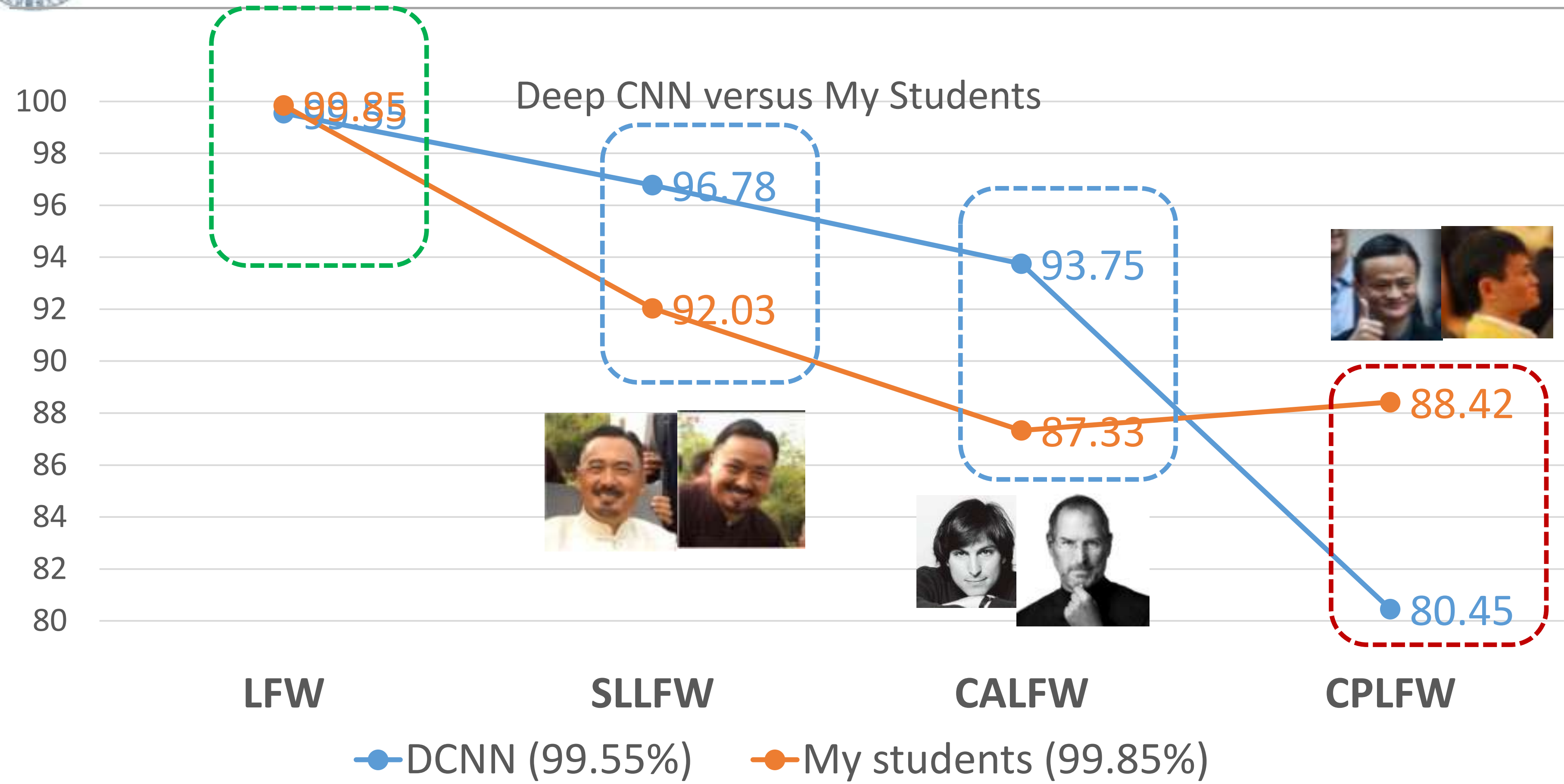
Cross-pose face pairs
selected by crowd-sourcing



3K negative pairs
with same gender and race



Human-Machine Comparison



If serious enough, human (students) do not make any mistake on LFW.

In the mediately difficult cases, DCNN is much better than human

In the extremely challenging cases, human performs more stable than DCNN

Same or Different face?

Angelababy

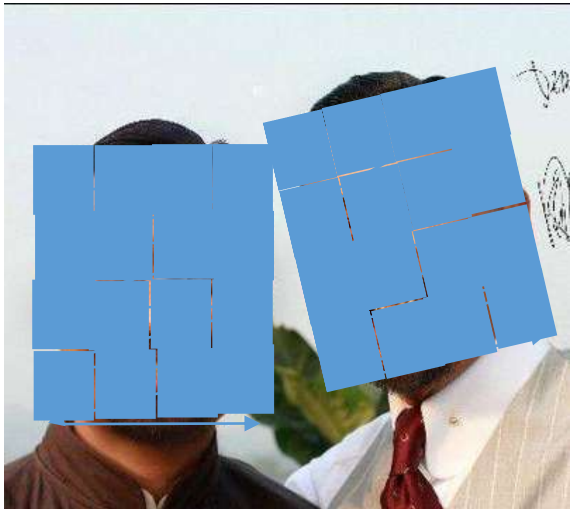
Angelababy

First 4： DCNN correct， Students wrong
The 5th： Students correct, DCNN wrong

The first 4 image pairs are from Similar-Looking LFW database

Human-machine Collaboration

- **DEEP DIFFERENCE ANALYSIS**

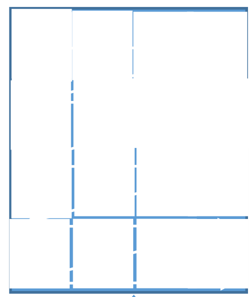


network

feature

network

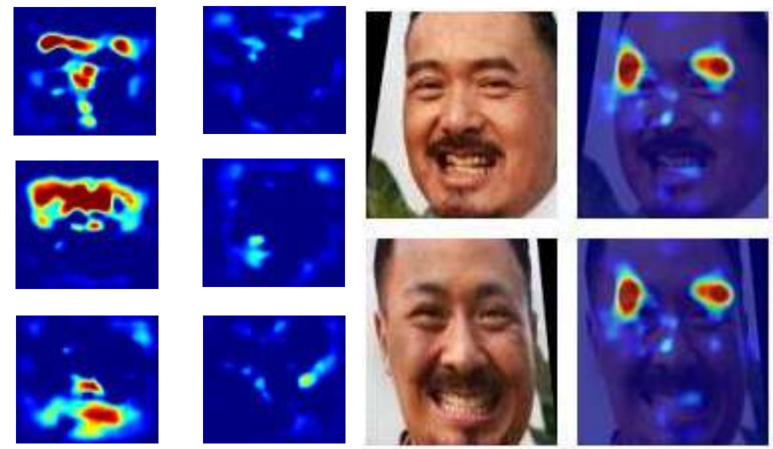
feature



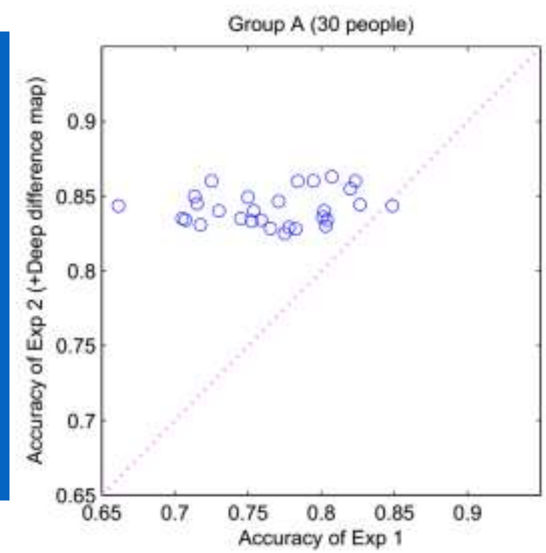
Comparison on local patch

- **Visualization**

Negative Positive Target



Display visualization



Display image pairs

Yaoyao Zhong, W. Deng, Deep Difference Analysis in Similar-looking Face recognition, ICPR, 2018

Summary

- The facial expressions in-the-wild are more complex and diverse than lab-controlled one.
    - RAF-DB is developed to evaluate the facial expression recognition in-the-wild with compound emotions.
    - The DCNN baseline performance is rather low in the RAF-DB recognition task.
- The face verification in-the-wild are more challenging than task in LFW.
    - SL/CA/CPLFW are developed to evaluate the real-world difficulties.
    - Both human and DCNN are insufficient to perform perfect face recognition, and they are complementary.

Advertisement & Acknowledgements

For data, code on

RAF-DB & SL/CA/CPLFW :

http://www.whdeng.cn

Welcome to Poster 72

Collaborators

Shan Li (李珊)

Ph.D student

Yaoyao Zhong (钟瑶瑶)

Ph.D student

Mei Wang (王玫)

Ph.D student