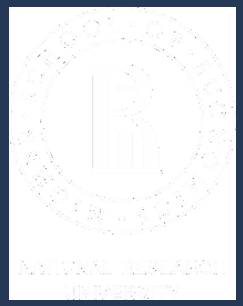


Laboratory of Algorithms and Technologies for Network Analysis (LATNA)

FACIAL ANALYTICS IN MOBILE APPLICATIONS

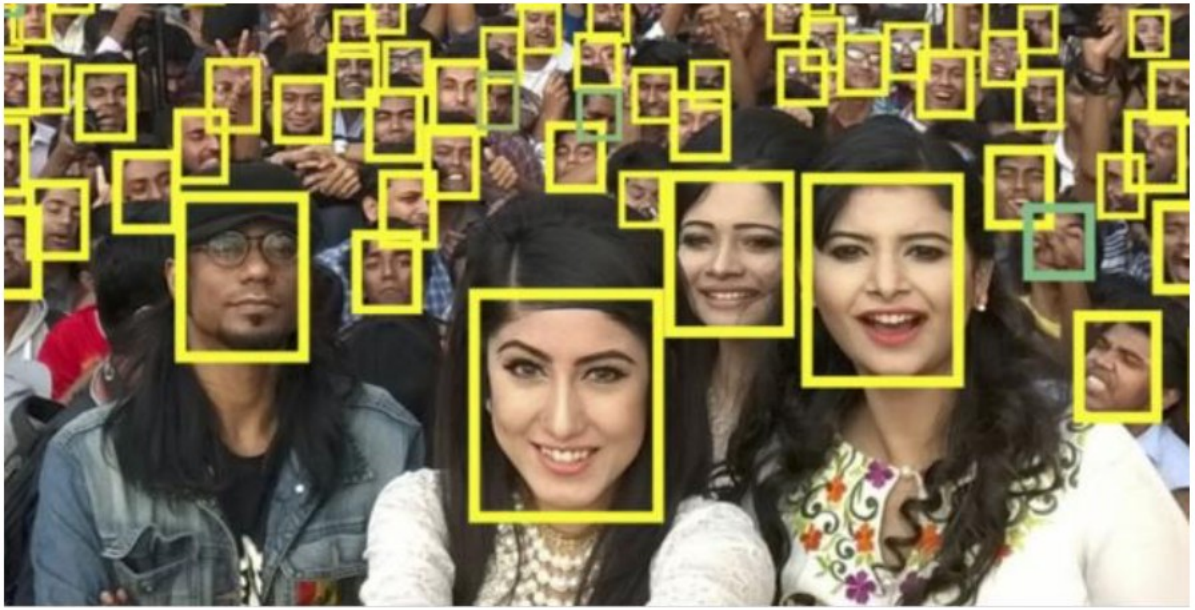
Andrey V. Savchenko
Dr. of Sci., Prof.
Email: avsavchenko@hse.ru
URL: www.hse.ru/en/staff/avsavchenko

3rd Winter School on Data Analytics (DA 2018)

1. Tasks of facial analytics
2. Mobile AI
3. Face identification
4. Video-based age and gender recognition
5. Group-level emotion recognition
6. Organizing Photo and Video Albums

Tasks of facial analytics

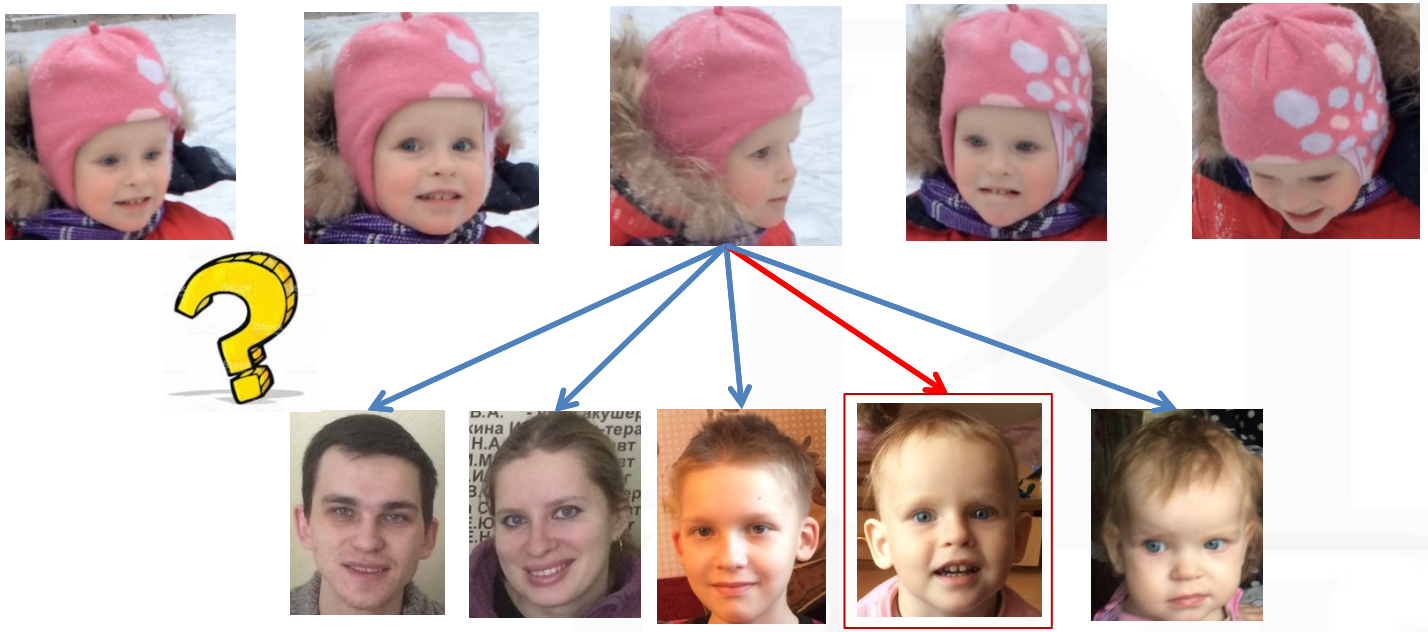
Face detection



Face verification

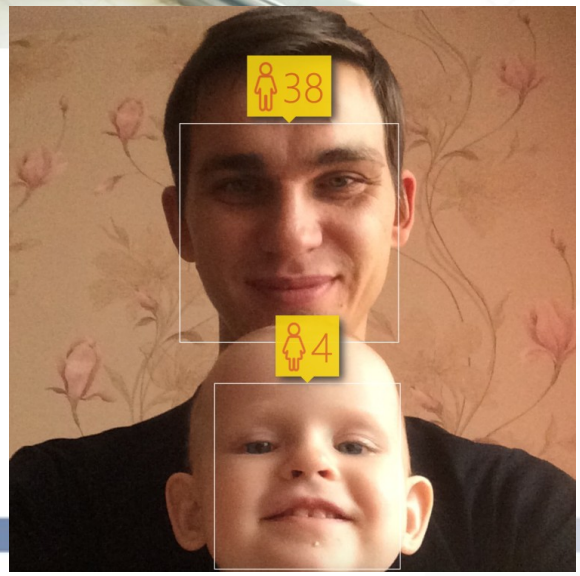
Kumar et al, TPAMI, 2011

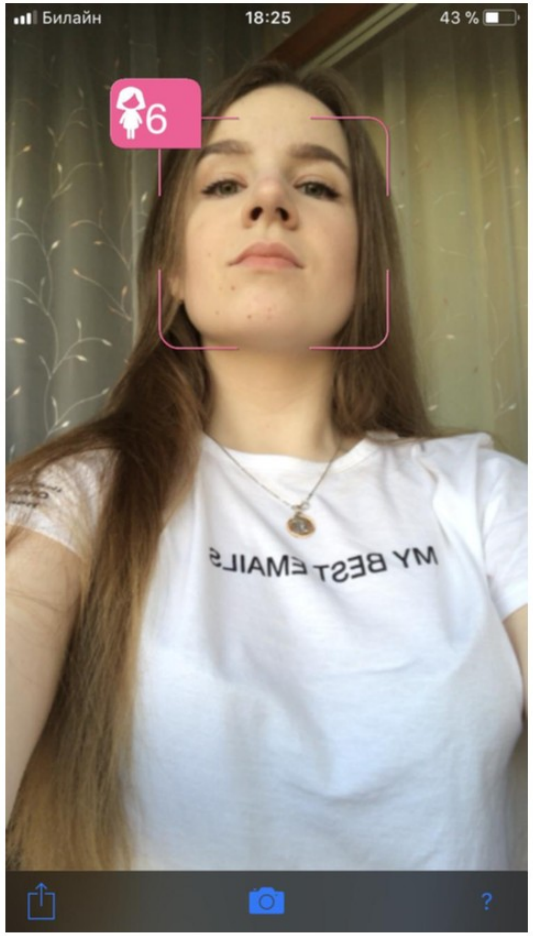
Input video sequence



Reference images

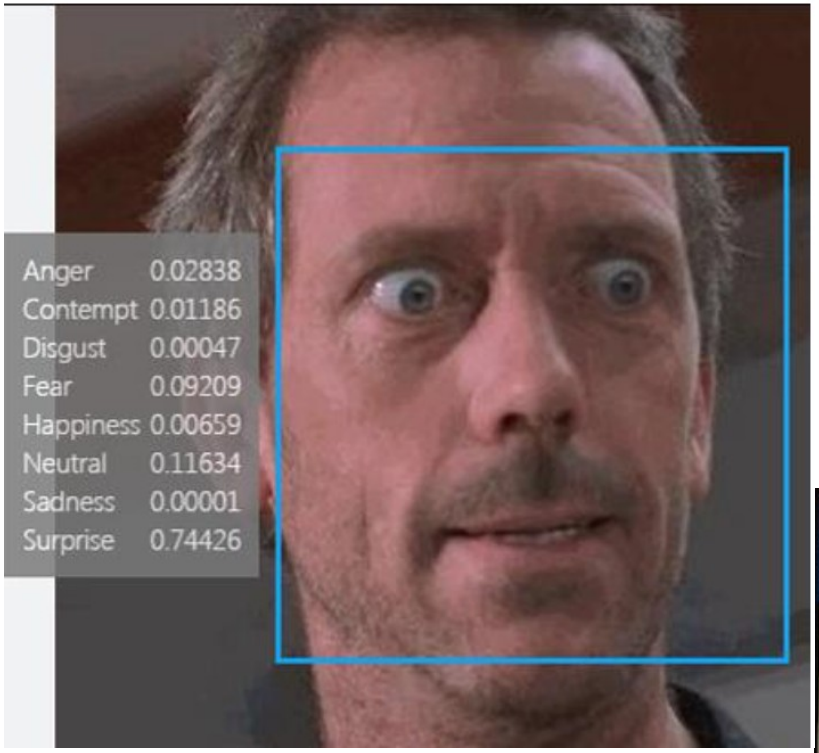
Age, gender, ethnicity recognition



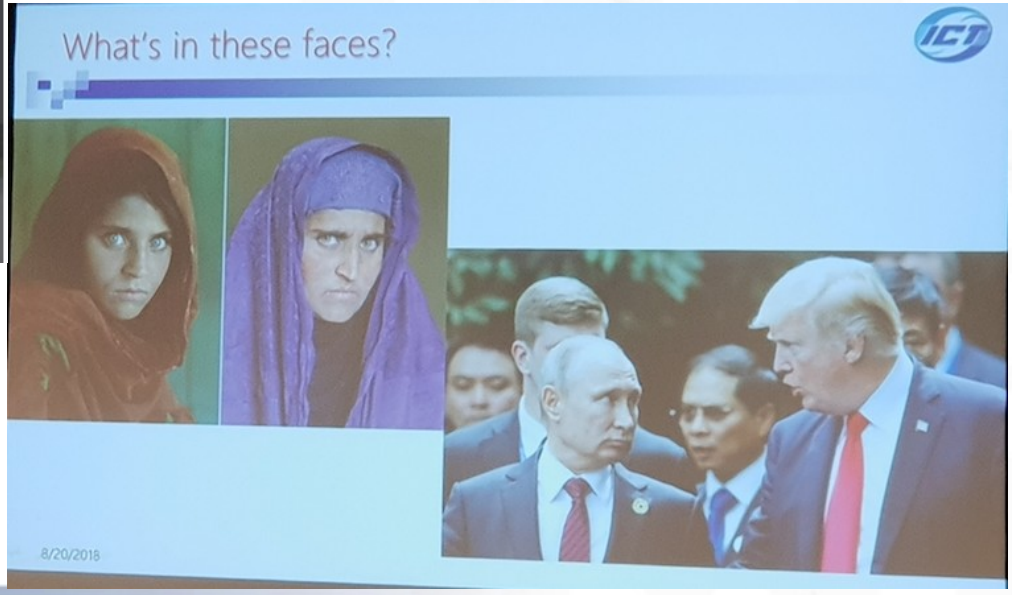


Luxand FaceSDK

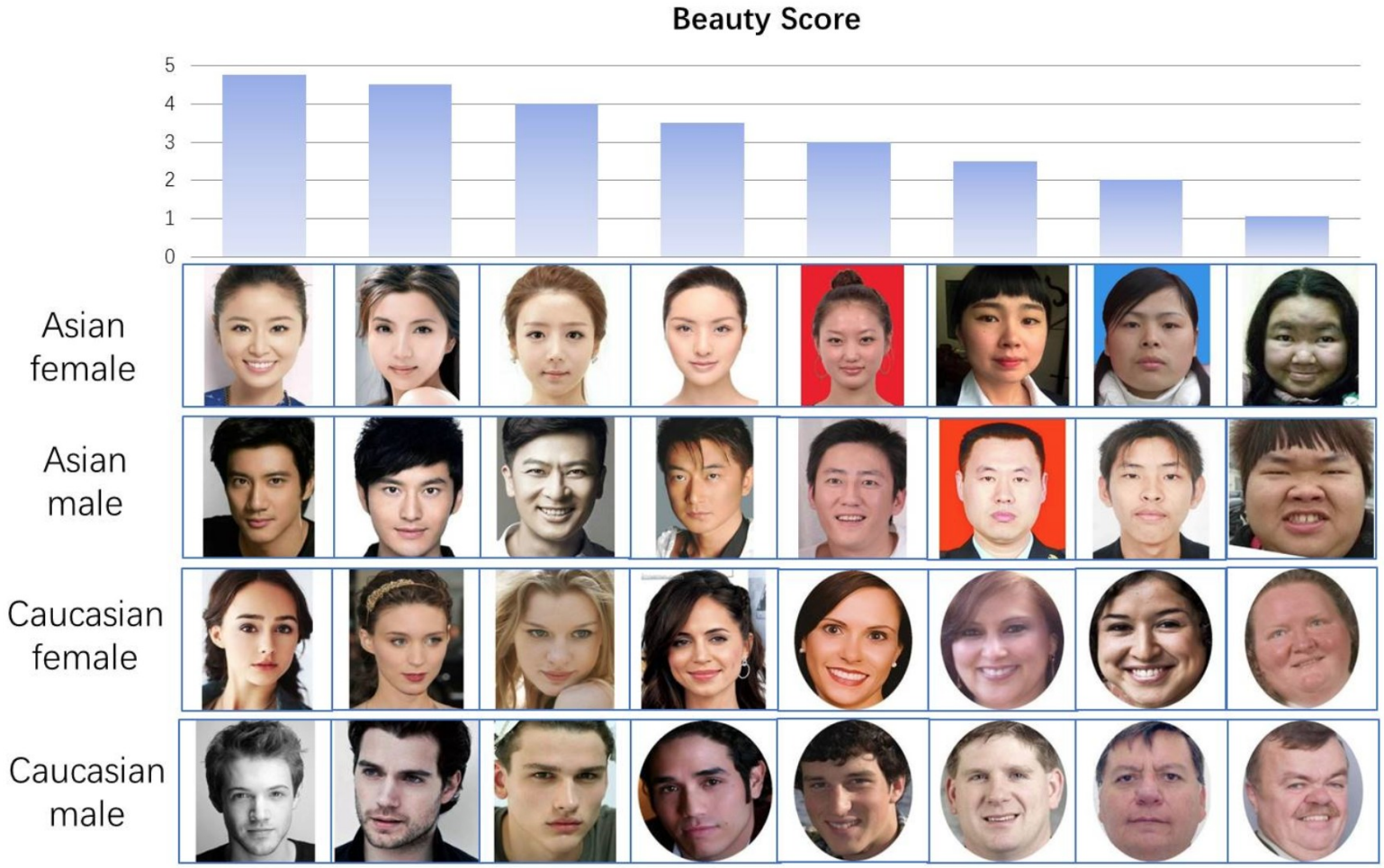
Emotion recognition



Xilin Chen: Challenges in facial expression understanding, ICPR18



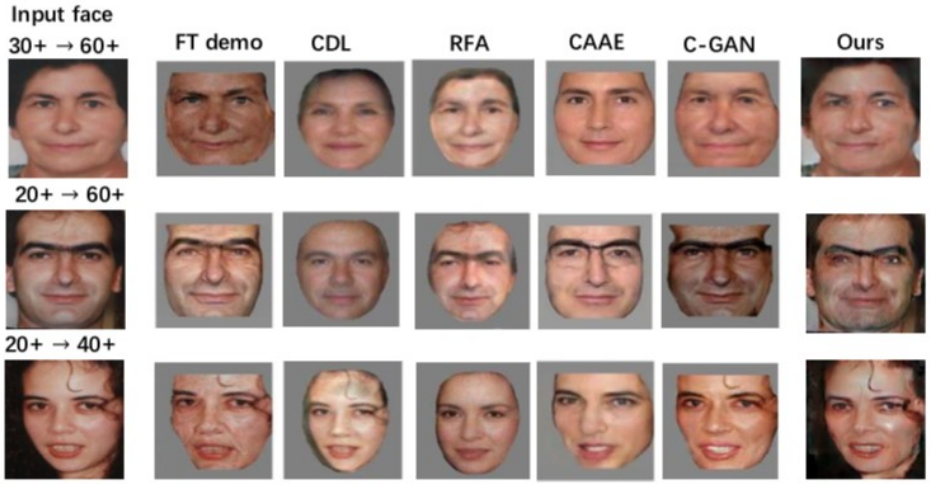
Beauty prediction



https://github.com/HCIILAB/SCUT-FBP5500-Database-Release

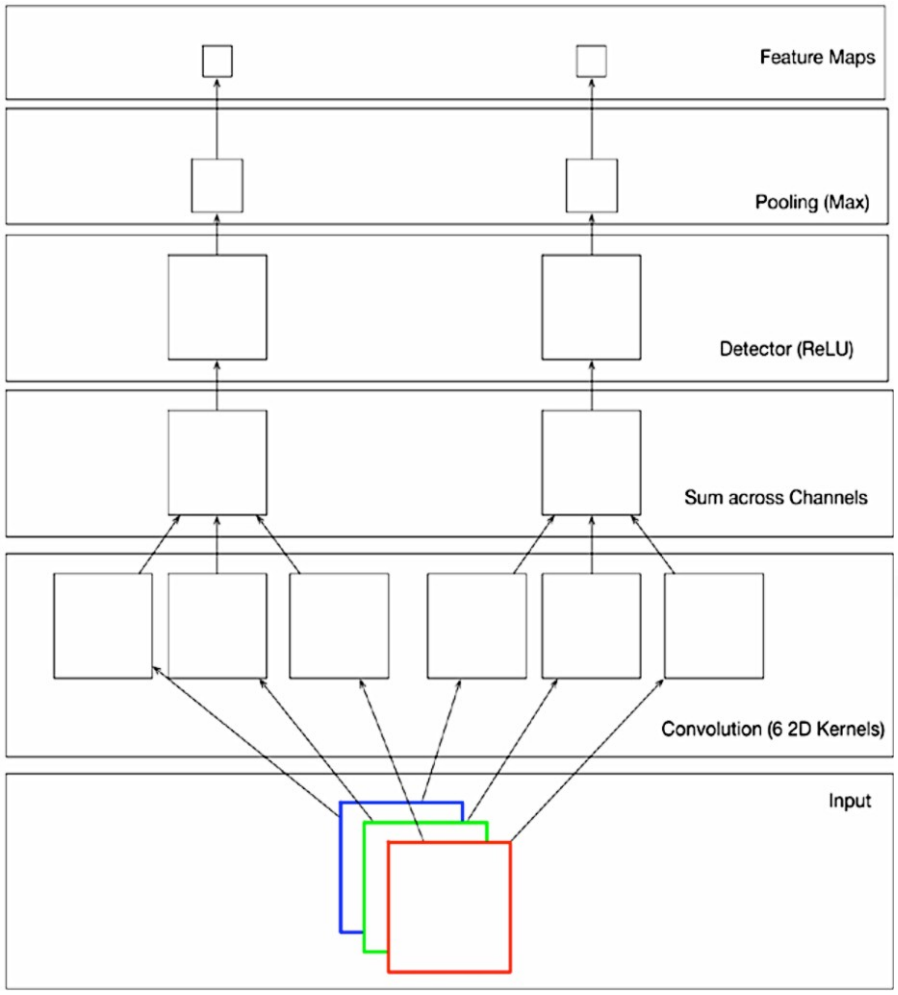
PairedCycleGAN: Asymmetric Style Transfer for Applying and Removing Makeup, CVPR18

Global and Local Consistent Age Generative Adversarial Networks, ICPR18

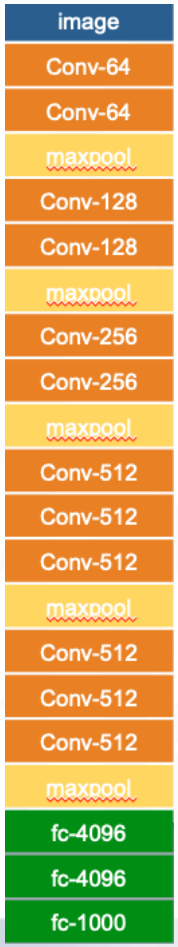


Mobile AI

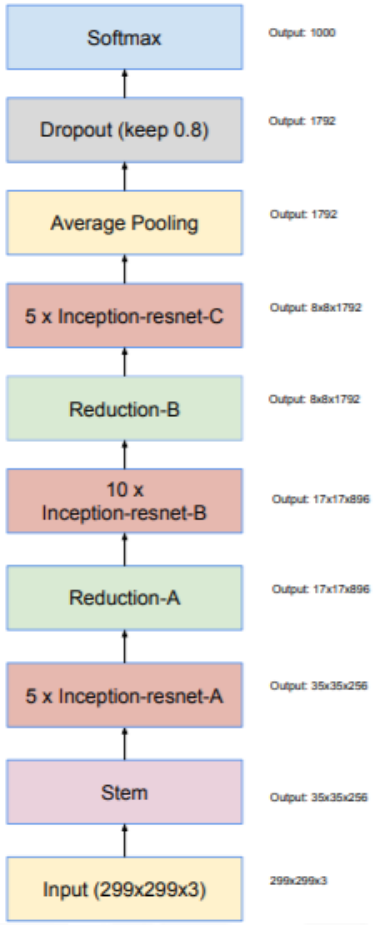
Convolutional Neural Networks (CNN)



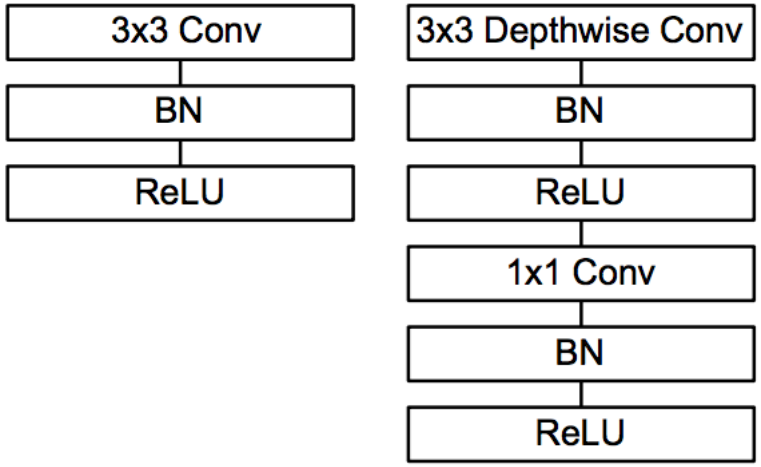
VGGNet (16-19 layers)



InceptionResNet (550 layers)



Depthwise convolution

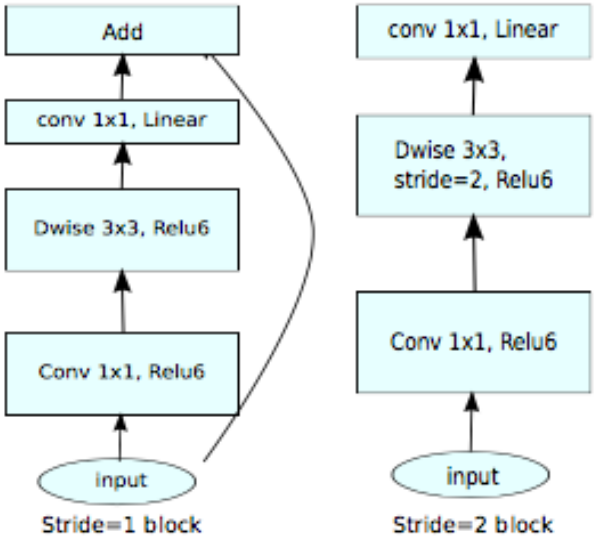


- 28 layers
- ImageNet Top-5 Error rate: 12.81%
- 4.2M parameters

https://arxiv.org/abs/1704.04861

Type / Stride	Filter Shape	Input Size
Conv / s2	3 × 3 × 3 × 32	224 × 224 × 3
Conv dw / s1	3 × 3 × 32 dw	112 × 112 × 32
Conv / s1	1 × 1 × 32 × 64	112 × 112 × 32
Conv dw / s2	3 × 3 × 64 dw	112 × 112 × 64
Conv / s1	1 × 1 × 64 × 128	56 × 56 × 64
Conv dw / s1	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 128	56 × 56 × 128
Conv dw / s2	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 256	28 × 28 × 128
Conv dw / s1	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 256	28 × 28 × 256
Conv dw / s2	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 512	14 × 14 × 256
5× Conv dw / s1	3 × 3 × 512 dw	14 × 14 × 512
5× Conv / s1	1 × 1 × 512 × 512	14 × 14 × 512
Conv dw / s2	3 × 3 × 512 dw	14 × 14 × 512
Conv / s1	1 × 1 × 512 × 1024	7 × 7 × 512
Conv dw / s2	3 × 3 × 1024 dw	7 × 7 × 1024
Conv / s1	1 × 1 × 1024 × 1024	7 × 7 × 1024
Avg Pool / s1	Pool 7 × 7	7 × 7 × 1024
FC / s1	1024 × 1000	1 × 1 × 1024
Softmax / s1	Classifier	1 × 1 × 1000

Bottleneck residual block



Input	Operator	Output
$h \times w \times k$	1x1 conv2d , ReLU6	$h \times w \times (tk)$
$h \times w \times tk$	3x3 dwise s=s, ReLU6	$\frac{h}{s} \times \frac{w}{s} \times (tk)$
$\frac{h}{s} \times \frac{w}{s} \times tk$	linear 1x1 conv2d	$\frac{h}{s} \times \frac{w}{s} \times k'$

Each line describes a sequence of 1 or more identical (modulo stride) layers, repeated n times.

Input	Operator	t	c	n	s
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$28^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1
$7^2 \times 1280$	avgpool 7x7	-	-	1	-
$1 \times 1 \times k$	conv2d 1x1	-	k	-	

- **ImageNet Top-1 Accuracy: 74.7% vs 70.6% for MobileNet v1**
- **6..9M parameters**

http://arxiv.org/abs/1801.04381

[Han et al. 2016] Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding – **ICLR16 Best paper**

- **Pruning**
  [Han et al. 2016], [Molchanov et al. 2016]
- **Distillation The Knowledge (FitNet)**
  [Hinton et al. 2014], [Romero et al. 2014]
- **Weights Hashing / Quantization**
  [Chen et al. 2015], [Han et al. 2016]
- **Tensor Decompositions**
  [Lebedev et al. 2015], [Kim et al. 2015], [Novikov et al. 2015], [Garipov et al. 2016]
- **Binarization**
  [Courbariaux / Hubara et al. 2016], [Rastegari et al. 2016], [Merolla et al. 2016], [Hou et al. 2016]
- **Architectural tricks**
  [Hong et al. 2016], [Iandola et al. 2016], [Teerapittayanon et al. 2016]

- **Memory!?**

64-256 Gb on Samsung S9+ and IPhoneX

- **Computational complexity!?**

MobileNet inference time:

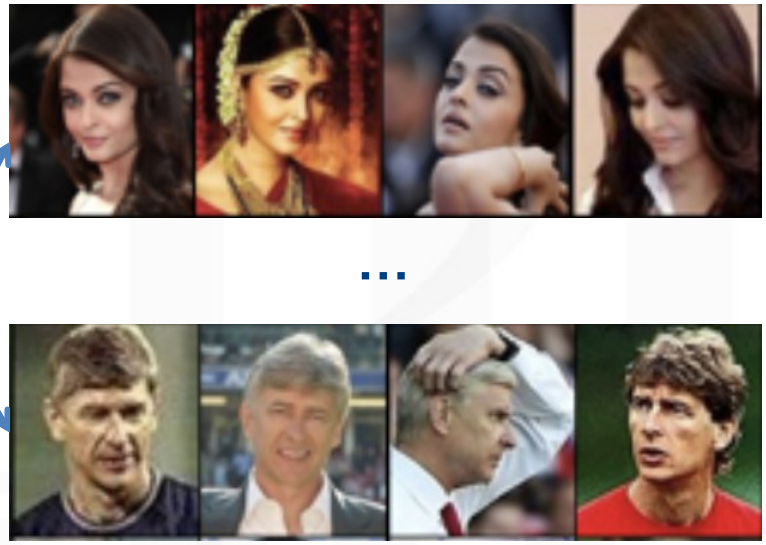
- 50-60 ms on Samsung S9+
- 20 ms on MacBook 2016

- **Power consumption!**

- **Another platform (Android, iOS). Where Python?**

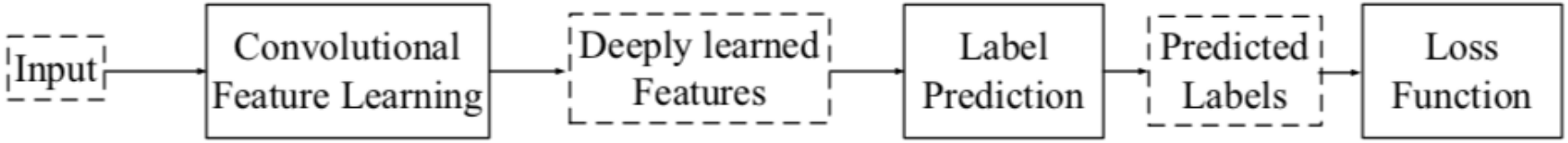
Face identification

Large training set of celebrities (VGGFace2, CASIA-WebFace, MS-Celeb-1M,…)



Representations

Deep CNN → **Fully connected layer**



[Wen et al, ECCV 2016]

Softmax activations

$$\text{softmax}(\boldsymbol{z})_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}.$$

Non-negative log-likelihood

$$L_1 = -\frac{1}{m}\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}$$

- VGGFace (VGG16) /VGGFace2 (ResNet/SeNet-50), https://github.com/rcmalli/keras-vggface

- FaceNet (InceptionResNet v1), https://github.com/davidsandberg/facenet

- LightCNN, https://github.com/AlfredXiangWu/LightCNN

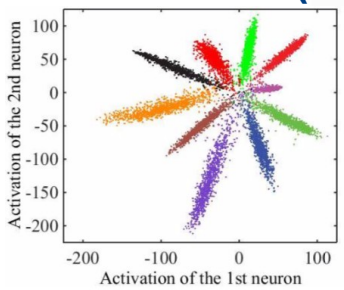
- **MobileNet v1**, http://github.com/HSE-asavchenko/HSE_FaceRec_tf

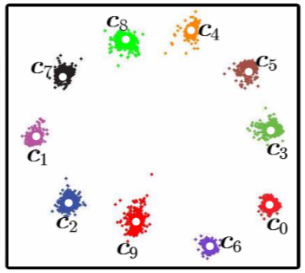
…

- **Center loss** [Wen et al, ECCV 2016] minimizes the intra-class variations

$$-\sum_{i=1}^{m}\log\frac{e^{W_{y_i}^T\boldsymbol{x}_i+b_{y_i}}}{\sum_{j=1}^{n}e^{W_j^T\boldsymbol{x}_i+b_j}}+\frac{\lambda}{2}\sum_{i=1}^{m}\|\boldsymbol{x}_i-\boldsymbol{c}_{y_i}\|_2^2$$

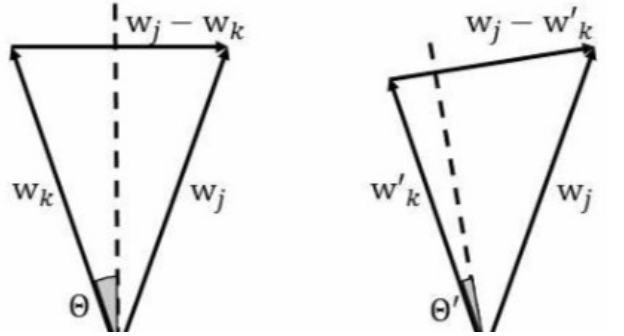
Pure softmax (λ=0) **Center loss (λ=1)**





- **One-shot face recognition** [Guo & Zhang, 2018]: the persons in the novel set *Cn* and the persons in the base set *Cb* should have similar volume sizes

Relationship between the norm of **w***k* and the volume size of the *k*th class



(a) $\|\mathbf{w}_k\|_2 = \|\mathbf{w}_j\|_2$ (b) $\|\mathbf{w}_k\|_2 < \|\mathbf{w}_j\|_2$

$$\mathcal{L}_{up}=\sum_{n}-t_{k,n}\log p_k(x_n)+\frac{1}{|C_n|}\sum_{k\in C_n}\|\|\mathbf{w}_k\|_2^2-\alpha\|_2^2$$

$$\alpha=\frac{1}{|C_b|}\sum_{k\in C_b}\|\mathbf{w}_k\|_2^2$$

- **Contrastive loss in DeepID2** [Sun et al, NIPS 2014]

Identification phase (softmax)

$$-\sum_{i=1}^{n} -p_i \log \hat{p}_i$$

Verification phase

$$\begin{cases} \frac{1}{2}\|f_i - f_j\|_2^2 & \text{if } y_{ij} = 1 \\ \frac{1}{2}\max\left(0, m - \|f_i - f_j\|_2\right)^2 & \text{if } y_{ij} = -1 \end{cases}$$

- **Triplet loss in FaceNet** [Schroff et al, CVPR 2015]: an image (anchor, a) of a specific person should be closer to all other images (positive, p) of the same person than it is to any image (negative, n) of any other person

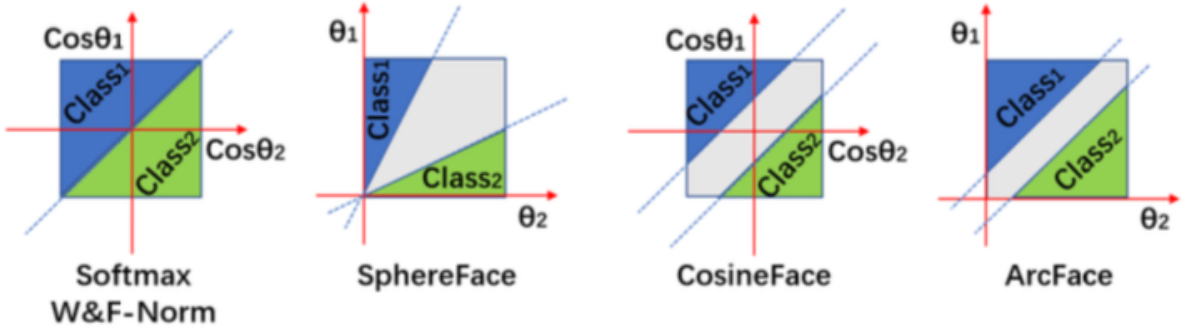
$$\sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

- **Sphere loss** [Liu et al, CVPR 2017]
- **Cosine loss in FaceCNN** [Wang et al, CVPR 2018]
- **ArcFace** [Deng et al, 2018], https://github.com/AIInAi/tf-insightface

Example: binary classification

Loss Functions	Decision Boundaries
Softmax	$(W_1 - W_2)\,x + b_1 - b_2 = 0$
W-Norm Softmax	$\|x\|\,(\cos\theta_1 - \cos\theta_2) = 0$
SphereFace [23]	$\|x\|\,(\cos m\theta_1 - \cos\theta_2) = 0$
F-Norm SphereFace	$s(\cos m\theta_1 - \cos\theta_2) = 0$
CosineFace [44, 43]	$s(\cos\theta_1 - m - \cos\theta_2) = 0$
ArcFace	$s(\cos(\theta_1 + m) - \cos\theta_2) = 0$

- θ_i is the angle between $\mathbf{W}_i$ and $\boldsymbol{x}$.
- Parameter $m>1$ requires $\cos(m\,\theta_1) > \cos(\theta_2)$



$$f(\rho_1(\mathbf{x}), \ldots, \rho_C(\mathbf{x})|W_c) = f(\rho_c(\mathbf{x})|W_c) \cdot \prod_{i=1, i\neq c}^{C} f(\rho_i(\mathbf{x})|W_c)$$

KL divergence is asymptotically distributed as non-central chi-squared, normal approximation

$$\mathcal{N}\left(\rho_{KL}(\mathbf{x}_r, \mathbf{x}_i) + \frac{D-1}{WH}, \frac{4WH\rho_{KL}(\mathbf{x}_r, \mathbf{x}_i) + (D-1)}{2(WH)^2}\right)$$

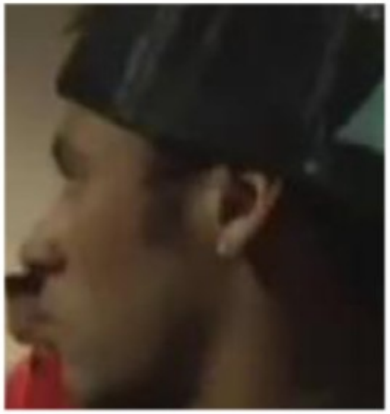
Final criterion

$$\min_{c\in\{c_1,\ldots,c_M\}} \left(\rho_c(\mathbf{x}) + \frac{\lambda}{C} \sum_{i=1, i\neq c}^{C} \frac{(\rho_i(\mathbf{x}) - \rho_{c;i} - \frac{D-1}{WH})^2}{\rho_{c;i} + \frac{D-1}{4WH}}\right)$$

It can be considered as adding regularization term to the nearest neighbor (NN) rule

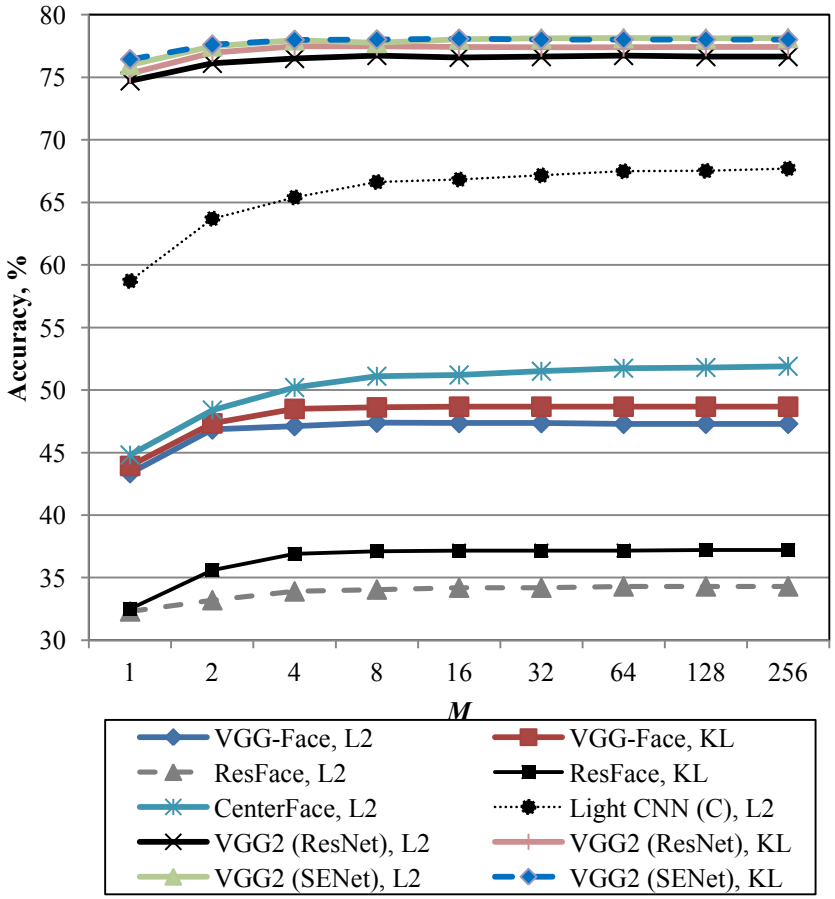
[Savchenko, Belova, Expert Systems with Applications 2018]

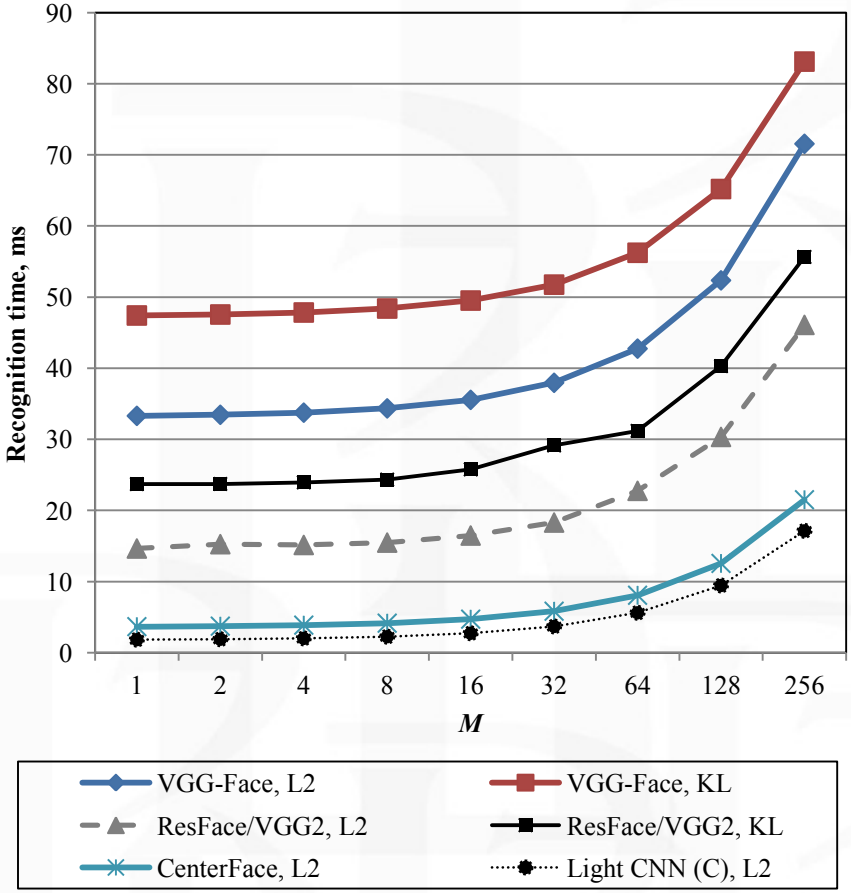
Probe (SubjectId 942)	first NN (SubjectId 1001)	second NN (SubjectId 1303)	Our result (seventh NN, SubjectId 942)



SubjectId	L_2 distance	Regularization term	Total objective
1001	0.0236	0.0152	0.0247
1303	0.0238	0.0152	0.0249
314	0.0239	0.0113	0.0247
1980	0.0240	0.0166	0.0252
387	0.0241	0.0300	0.0262
1289	0.0241	0.0140	0.0251
942	**0.0242**	**0.0059**	**0.0246**
1287	0.0242	0.0174	0.0258

Experimental study: YTF/LFW





Rank-1 accuracy (%) in 1:N identification

Features	Distance	IJB-A bounding boxes			MTCNN face detection		
		SVM	NN	Proposed	SVM	NN	Proposed
CenterFace	L_2	52.6 ± 1.8	49.5 ± 1.9	54.6 ± 1.7	64.6 ± 1.0	64.9 ± 1.1	66.8 ± 1.1
Light CNN	L_2	57.2 ± 0.5	58.3 ± 0.7	61.8 ± 0.6	79.9 ± 0.4	81.3 ± 0.4	83.3 ± 0.3
ResFace	L_2	79.8 ± 0.7	77.9 ± 0.5	80.7 ± 0.7	86.4 ± 0.5	85.6 ± 0.6	87.6 ± 0.6
ResFace	KL	–	79.2 ± 0.6	81.3 ± 0.7	–	86.0 ± 0.5	88.3 ± 0.6
VGG-Face	L_2	87.7 ± 0.6	87.0 ± 0.5	88.5 ± 0.6	89.2 ± 0.4	89.0 ± 0.4	90.3 ± 0.4
VGG-Face	KL	–	87.5 ± 0.6	88.9 ± 0.6	–	90.0 ± 0.3	91.5 ± 0.4
VGG2 (ResNet)	L_2	95.7 ± 0.3	95.7 ± 0.5	96.1 ± 0.4	96.8 ± 0.4	97.7 ± 0.4	98.0 ± 0.3
VGG2 (ResNet)	KL	–	96.4 ± 0.4	96.7 ± 0.3	–	97.5 ± 0.4	98.0 ± 0.5
VGG2 (SENet)	L_2	95.3 ± 0.3	95.9 ± 0.5	96.8 ± 0.5	97.3 ± 0.5	97.8 ± 0.5	98.1 ± 0.4
VGG2 (SENet)	KL	–	96.0 ± 0.5	96.7 ± 0.3	–	98.0 ± 0.4	98.2 ± 0.6
LBP histograms	L_2	25.8 ± 1.2	24.2 ± 1.1	26.2 ± 1.3	29.0 ± 1.0	28.1 ± 1.2	29.2 ± 1.2
LBP histograms (synthetic images)	L_2	28.2 ± 1.5	27.3 ± 1.2	28.9 ± 1.4	32.3 ± 1.6	31.0 ± 1.3	32.8 ± 1.4
LBP histograms	KL	–	31.0 ± 1.3	33.7 ± 1.5	–	33.6 ± 1.1	34.7 ± 1.4
LBP histograms (synthetic images)	KL	–	33.7 ± 1.4	35.2 ± 1.6	–	35.4 ± 1.2	36.3 ± 1.4
HOG	L_2	11.7 ± 3.4	19.7 ± 2.0	27.9 ± 2.1	12.1 ± 3.3	23.0 ± 1.8	29.1 ± 2.0
HOG (synthetic images)	L_2	12.3 ± 3.5	23.9 ± 3.1	31.9 ± 3.2	12.9 ± 3.1	26.6 ± 2.6	32.3 ± 2.8
HOG	KL	–	29.4 ± 1.9	32.3 ± 2.0	–	31.5 ± 1.6	33.3 ± 1.8
HOG (synthetic images)	KL	–	32.9 ± 1.4	34.4 ± 1.4	–	35.1 ± 1.6	37.2 ± 1.2

Video-based age and gender recognition

97803_1991-04-04_2007.jpg

167310_1943-02-06_1959.jpg

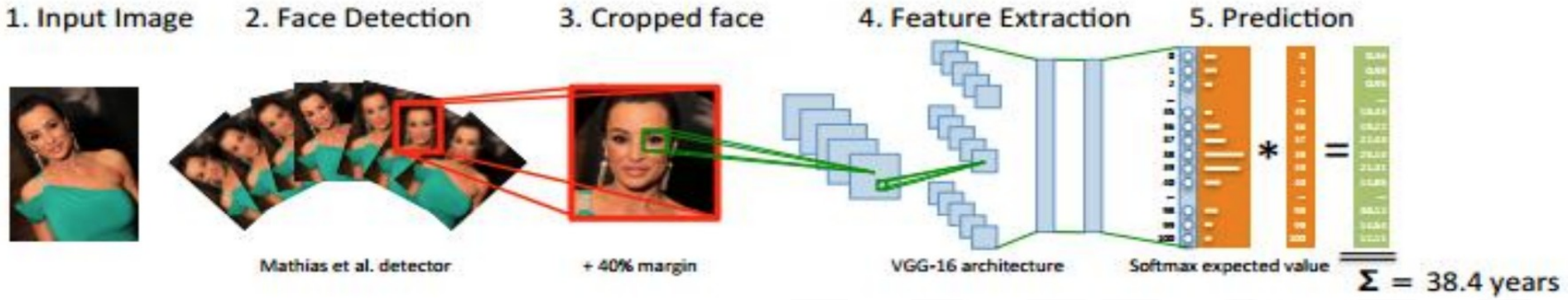
205707_1982-06-10_1998.jpg

610819_1966-02-07_1982.jpg

701129_1934-03-02_1950.jpg

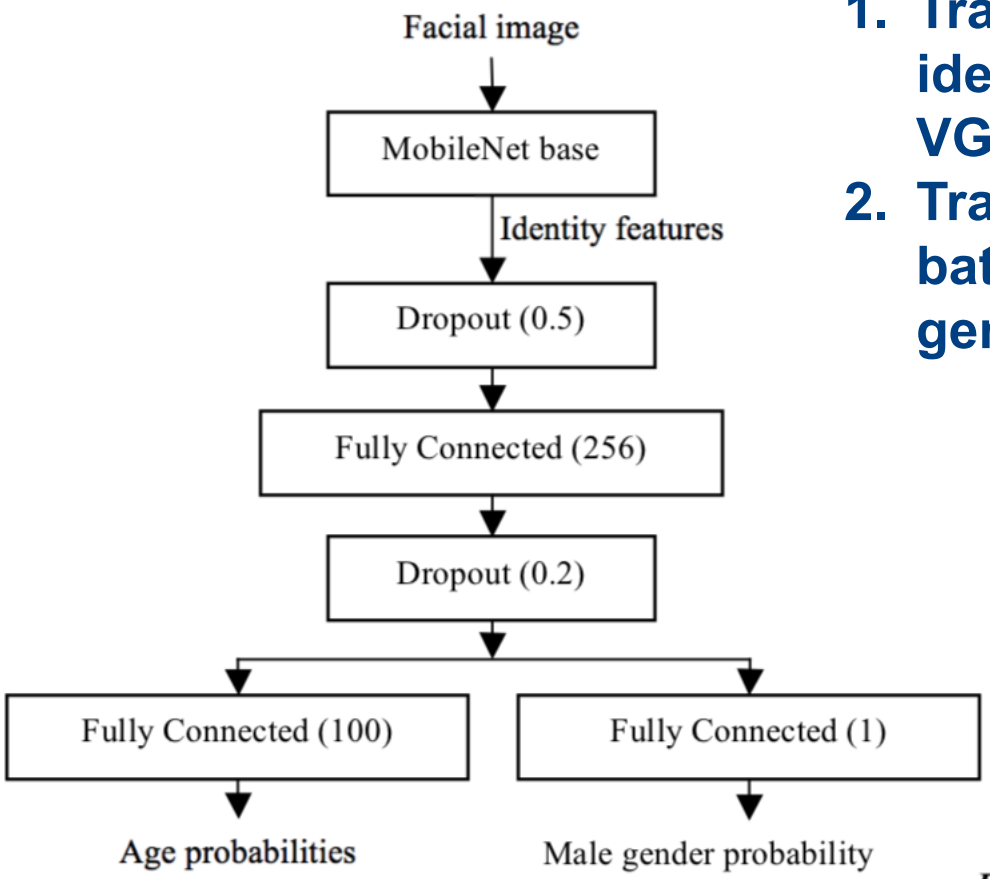
804602_1928-04-23_1944.jpg

VGG16-Net



DEX: Deep EXpectation of apparent age from a single image, 2015

Our simple multi-output model



1. **Train identity features for face identification (softmax loss, VGGFace2 dataset)**
2. **Train outputs by alternating batches of data with age and gender labels**

Age prediction

$$l^* = \frac{\sum_{k=1}^{K} l_{(k)} \cdot P\left(l_{(k)} \middle| X(t)\right)}{\sum_{k=1}^{K} P\left(l_{(k)} \middle| X(t)\right)}.$$

$$P\left(l_{(1)} \middle| X(t)\right) \geq P\left(l_{(2)} \middle| X(t)\right) \geq \ldots \geq P\left(l_{(L)} \middle| X(t)\right)$$

https://github.com/HSE-asavchenko/HSE_FaceRec_tf/tree/master/age_gender_identity

1. Training. Estimate decision templates for each class label *j*

$$DT_j^l = \frac{1}{N_j} \sum_{X \in \omega_j} P(l|X)$$

2. Inference.

a) Estimate the closeness of the frame to each class *j*

$$\Phi_j(t) = \frac{\left(1 + \left(\sum_{l=1}^{L} \left(DT_j^l - P(l|X(t))\right)\right)^{1/2}\right)^{-1}}{\sum_{i=1}^{L} \left(1 + \left(\sum_{l=1}^{L} \left(DT_i^l - P(l|X(t))\right)\right)^{1/2}\right)^{-1}}$$

b) Compute Dempster belief

$$b_j(t) = \frac{\Phi_j(t) \prod_{l \neq j} (1 - \Phi_l(t))}{1 - \Phi_j(t) \left[1 - \prod_{l \neq j} (1 - \Phi_l(t))\right]}.$$

c) Final aggregation for all frames

$$l^* = \underset{j=\overline{1,L}}{\operatorname{argmax}} \prod_{t=1}^{T} b_j(t)$$

Experimental results: Gender recognition accuracy

CNN	Aggregation	Eurecom Kinect	IMFDB	AFEW	IJB-A
gender_net	Simple Voting	0.73	0.71	0.75	0.60
	Product rule	0.77	0.75	0.75	0.59
	Dempster-Shafer	0.83	0.80	0.78	0.63
DEX	Simple Voting	0.84	0.81	0.80	0.81
	Product rule	0.84	0.88	0.81	0.82
	Dempster-Shafer	0.92	0.90	0.90	0.81
Our MobileNet	Simple Voting	0.94	0.98	**0.93**	0.95
	Product rule	0.93	**0.99**	**0.93**	**0.96**
	Dempster-Shafer	0.95	0.98	**0.93**	**0.96**
Our MobileNet, quantized	Simple Voting	0.88	0.96	0.92	0.93
	Product rule	0.86	0.96	0.93	0.94
	Dempster-Shafer	0.89	0.97	0.93	0.94
Our MobileNet, fine-tuned	Simple Voting	0.93	0.95	0.91	0.94
	Product rule	**0.95**	0.97	0.92	0.95
	Dempster-Shafer	**0.96**	0.96	0.92	0.95

Experimental results: Age recognition accuracy

CNN	Aggregation	Eurecom Kinect	IMFDB	AFEW
age_net	Simple Voting	0.41	0.68	0.27
	Product Rule	0.45	0.48	0.27
	Expected Value	0.69	0.32	0.30
	Dempster-Shafer	0.73	0.29	0.52
DEX	Simple Voting	0.60	0.29	0.47
	Product Rule	0.71	0.29	0.48
	Expected Value	0.71	0.54	0.52
	Dempster-Shafer	0.59	0.68	0.30
Our MobileNet	Simple Voting	0.92	0.32	0.46
	Product Rule	**0.94**	0.36	0.46
	Expected Value	**0.94**	**0.77**	**0.54**
	Dempster-Shafer	0.64	0.73	0.48
Our MobileNet, quantized	Simple Voting	0.86	0.34	0.44
	Product Rule	0.88	0.36	0.46
	Expected Value	0.85	0.58	0.50
	Dempster-Shafer	0.56	0.69	0.47
Our MobileNet, fine-tuned	Simple Voting	0.74	0.33	0.45
	Product Rule	0.77	0.35	0.45
	Expected Value	0.92	0.72	0.51
	Dempster-Shafer	0.58	0.79	0.45

Group-level emotion recognition

"positive" "neutral" "negative"

EmotiW2017 sub-challenge:

- 3 classes (positive, neutral, negative)
- Training set: 3630 photos
- Validation set: 2065 photos

We got the 4th place (75.4% accuracy on validation set - 23+% better than the baseline, 78.53% on test set - 25+% than the baseline)

Our ensemble based on Dlib face detector and VGGFace was too complicated

[Rassadin, A., Gruzdev, A., Savchenko, Proc. ICMI, 2017]

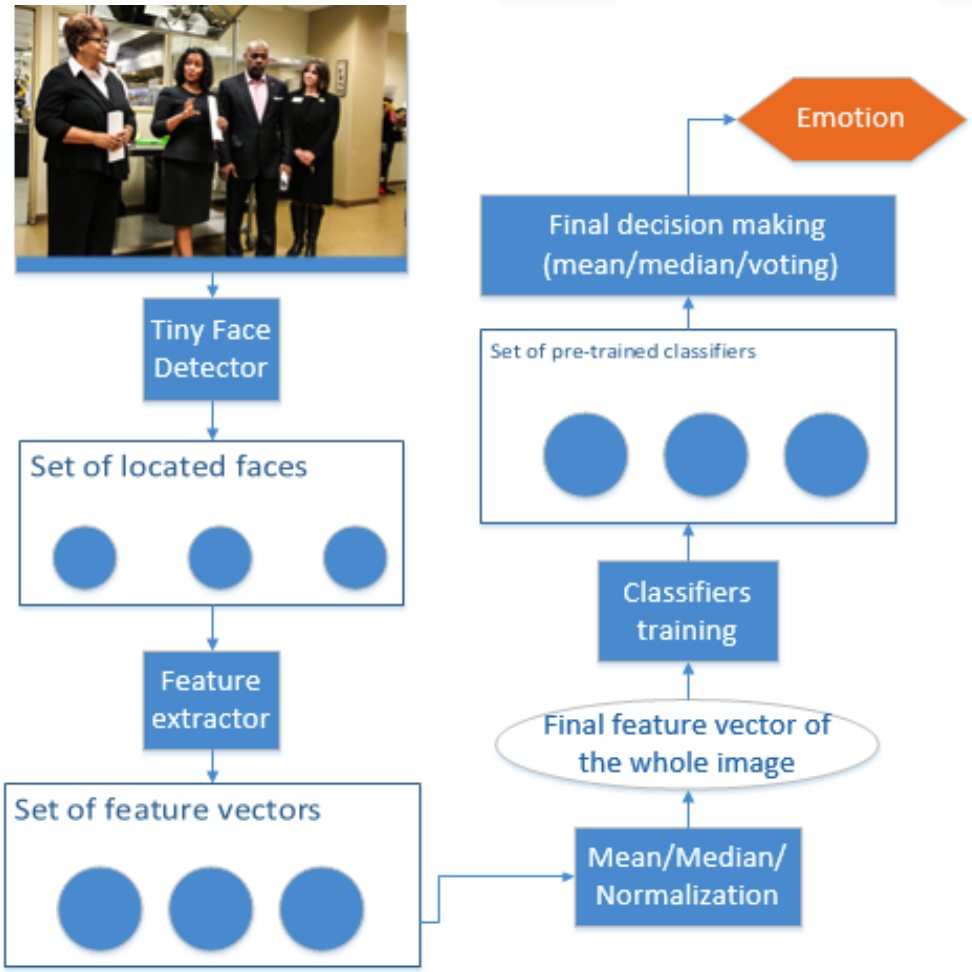
Leaderboard:

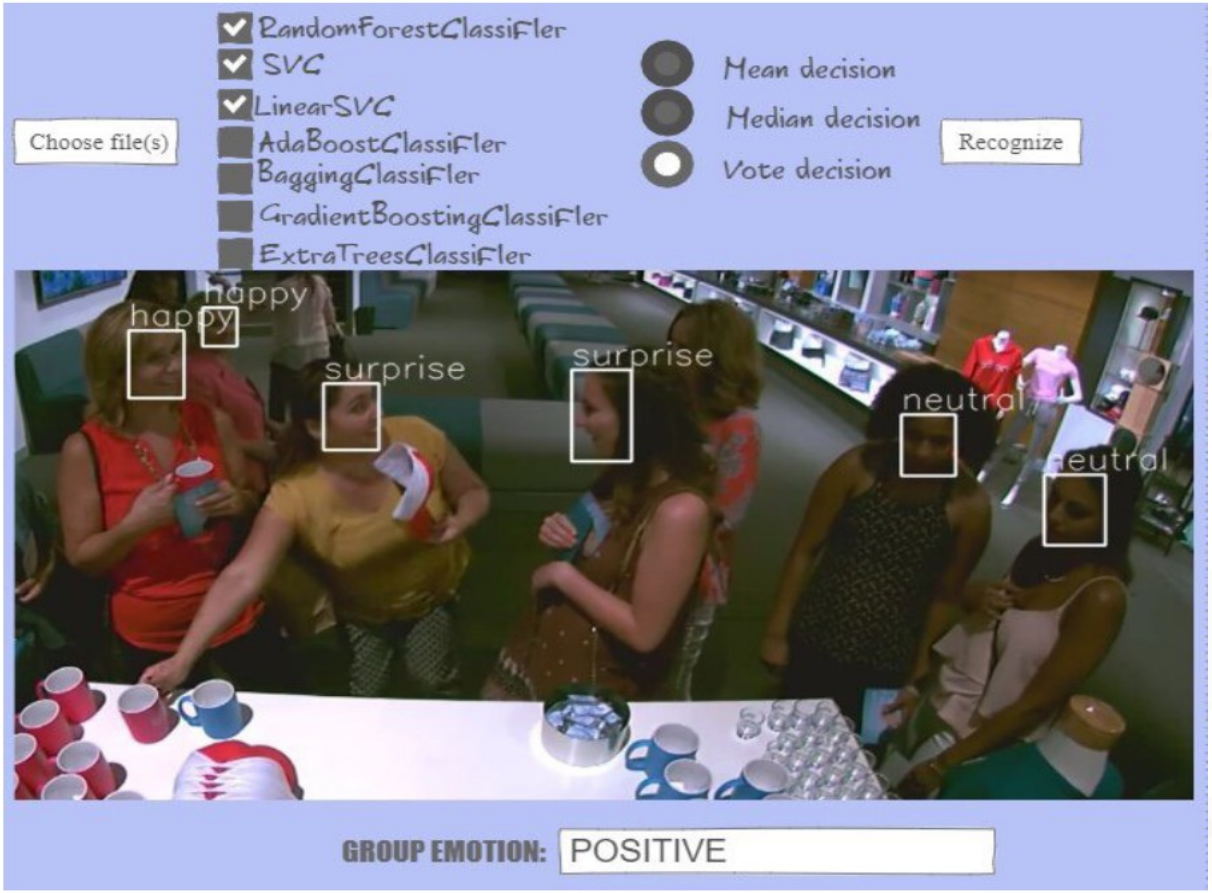
Group-level Emotion Recognition	
SIAR	80.89%
UDI-GPB	80.61%
BNU	79.79%
AMD	**78.53%**
AmritaEEE	75.07%

Small convolutional neural networks (CNNs) pre-trained on emotion recognition with FER-2013 (64x64 facial images)

[Tarasov, Savchenko, LNCS, 2018]

Pipeline





The most accurate (70.9% val acc) Random Forest returns a wrong Neutral emotion. The neutral label gets a score of 0.709. When Linear SVM (70.3% val acc) and RBF SVM (69.3% val acc) are in an ensemble, correct positive label is chosen by both SVMs and gets maximal score 1.396 (0.703+0.693).

Experimental results on EmotiW 2017 data

	CNN-5		FCN	
Classifier	**7 softmax scores**	**128 embeddings**	**7 softmax scores**	**112 embeddings**
RBF SVM	57.0	60.5	68.5	69.3
Gradient Boosting	61.3	61.9	67.2	70.1
Extra Trees	59.1	61.9	69.6	70.2
Linear SVM	59.6	61.5	69.8	70.3
Bagging	59.0	61.5	68.8	70.7
Random Forest (RF)	58.0	59.9	69.0	70.9
AdaBoost	57.4	58.1	69.5	70.9
Proposed approach (all classifiers)	61.5	62.2	70.7	72.3
Proposed approach (RF, RBF SVM and Linear SVM)	**61.9**	**64.2**	**73.2**	**75.5**

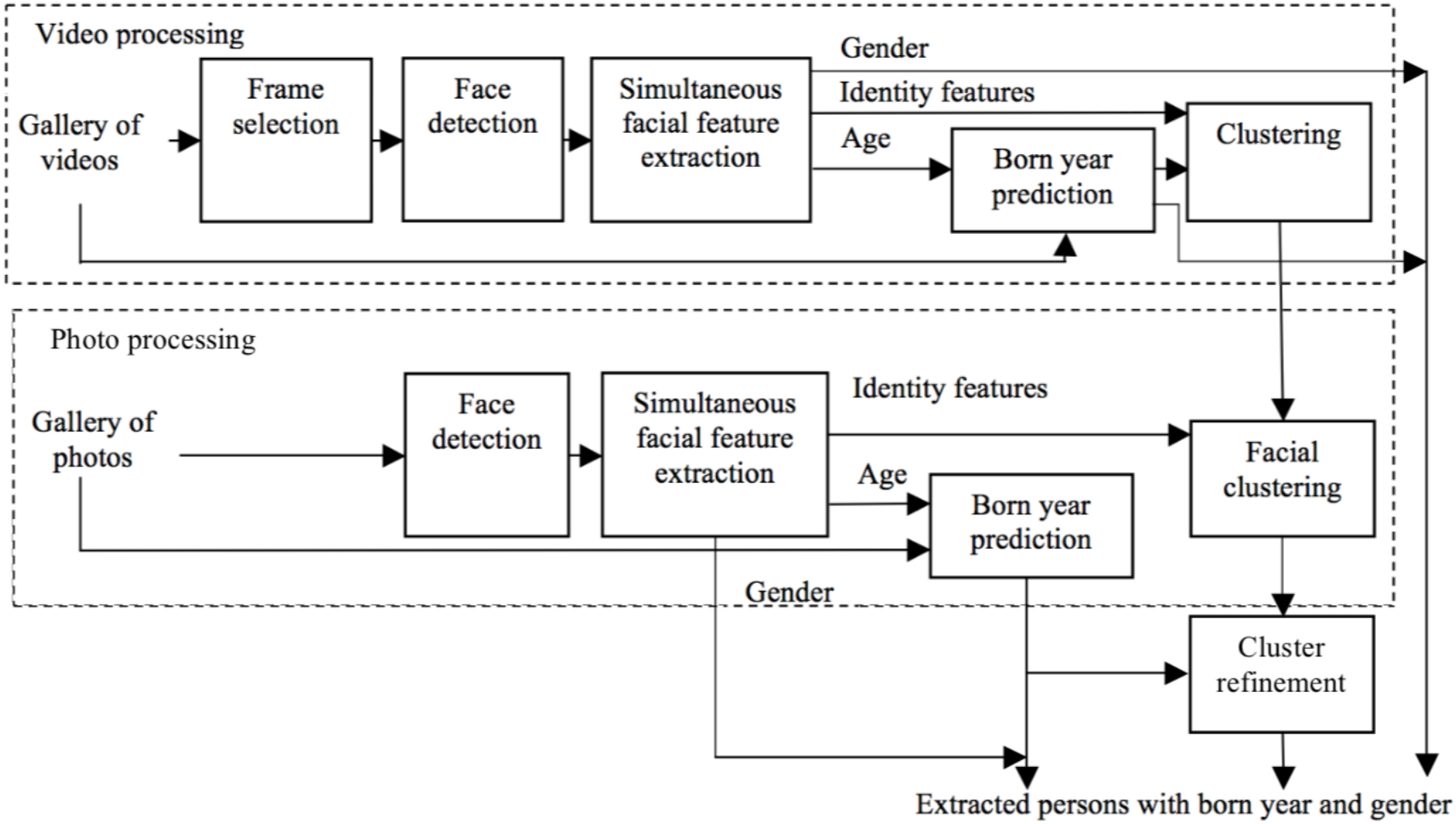
CNN compression. Visual emotion recognition (Radboud Faces Database)

	Training time per epoch, ms	Inference time, ms	Model size, MB	Accuracy, %
VGG-S (baseline)	43.7	33.4	372.2	97.13
SqueezeNet-1.1 (baseline)	**22.94**	**4.94**	2.8	89.14
SqueezeNet-1.1, CP-Decomposition	**22.94**	7.74	**2.1**	87.5
HashedNets	294.8	158.2	68.3	96.31
Binary-Weight-Network (BWN)	83.8	33.5	11.6	**98.57**
XNOR-Net	84.3	34.2	11.6	58.81
XNOR-Net w/o weights activation	43.4	34.1	11.6	88.32

[Rassadin, Savchenko, CEUR-WS, 2017]

Organizing Photo and Video Albums

Organizing Photo and Video Albums



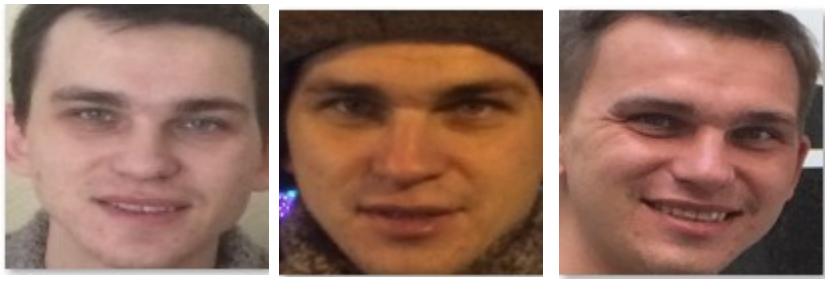
Funded by SAMSUNG

Experimental results: GFW (Grouping Faces in the Wild) dataset

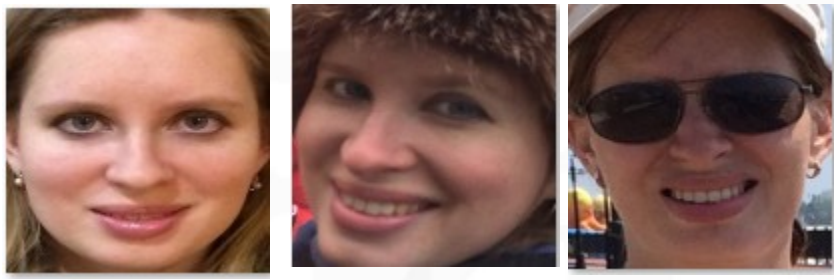
		K/C	ARI	AMI	Homogeneity	Completeness	F-measure
Single	VGGFace	4.10	0.440	0.419	0.912	0.647	0.616
	VGGFace2	3.21	0.580	0.544	0.942	0.709	0.707
	Ours	4.19	0.492	0.441	0.961	0.655	0.636
Average	VGGFace	1.42	0.565	0.632	0.860	0.751	0.713
	VGGFace2	1.59	0.603	**0.663**	0.934	0.761	0.746
	Ours	1.59	**0.609**	0.658	0.917	**0.762**	**0.751**
Complete	VGGFace	**0.95**	0.376	0.553	0.811	0.690	0.595
	VGGFace2	1.44	0.392	0.570	0.916	0.696	0.641
	Ours	1.28	0.381	0.564	0.886	0.693	0.626
Weighted	VGGFace	1.20	0.464	0.597	0.839	0.726	0.662
	VGGFace2	1.05	0.536	0.656	0.867	**0.762**	0.710
	Ours	1.57	0.487	0.612	0.915	0.727	0.697
Median	VGGFace	5.30	0.309	0.307	0.929	0.587	0.516
	VGGFace2	4.20	0.412	0.422	0.929	0.639	0.742
	Ours	6.86	0.220	0.222	**0.994**	0.552	0.411
Rank-Order	VGGFace	0.82	0.319	0.430	0.650	0.694	0.630
	VGGFace2	1.53	0.367	0.471	0.937	0.649	0.641
	Ours	1.26	0.379	0.483	0.914	0.658	0.652

In average, $C = 46$ subjects

male 1987

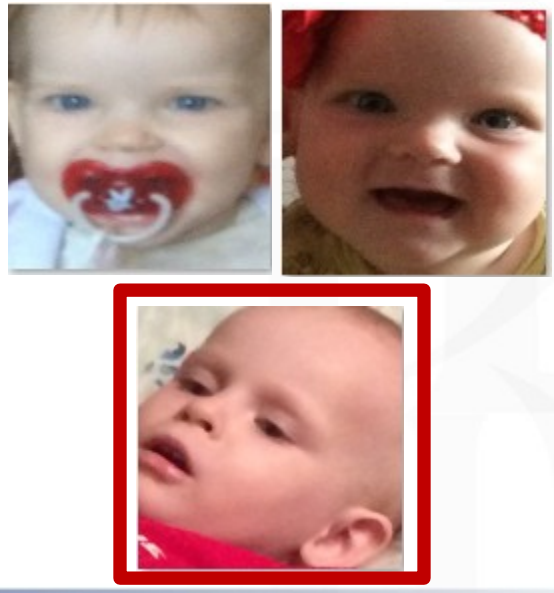


female 1982

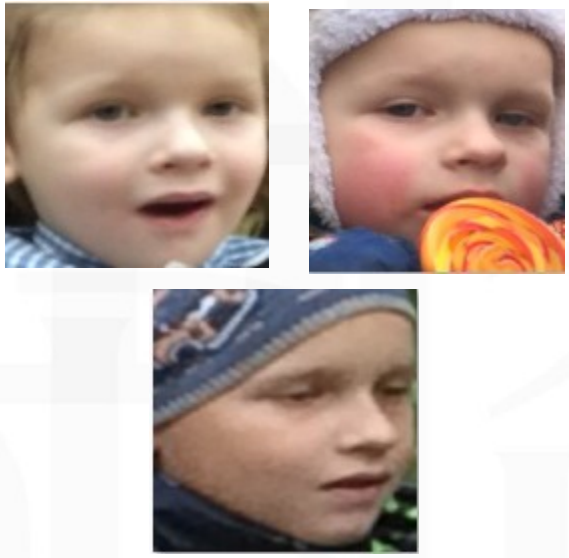


female 2015

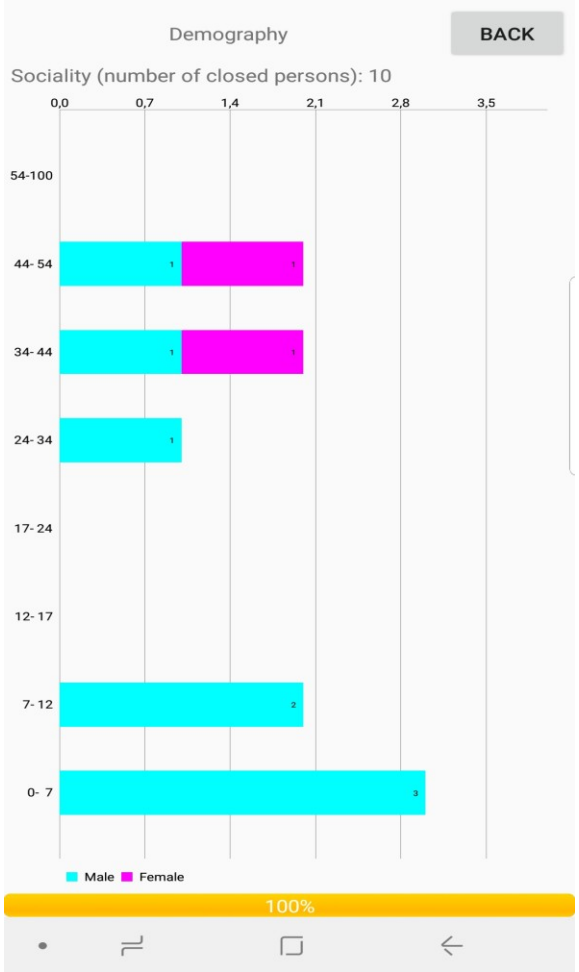
female 2016

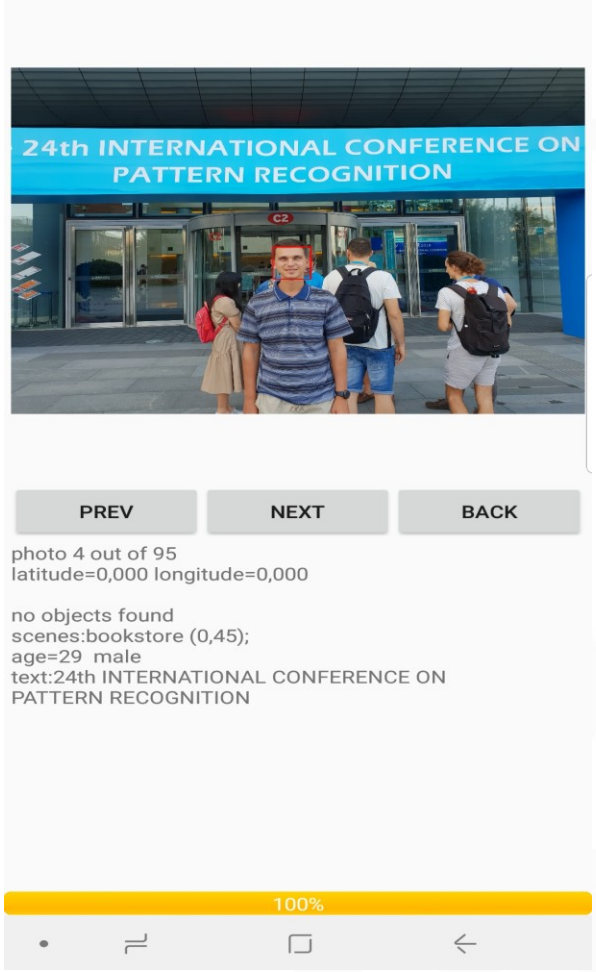


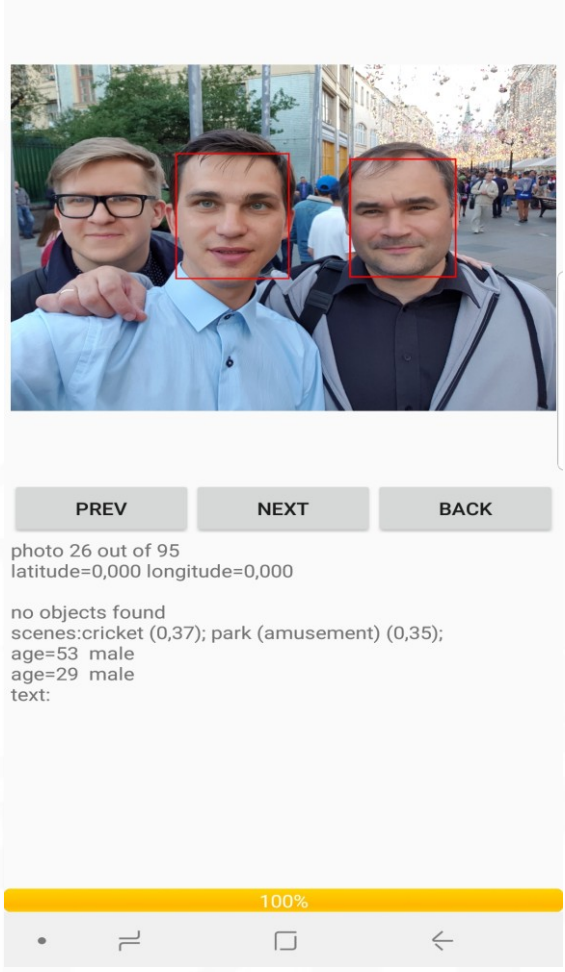
male 2009



Example (2)







Funded by SAMSUNG

1. Though facial analytics is well established industrial field, there are still plenty of research questions

2. AI in mobile systems is still under active research

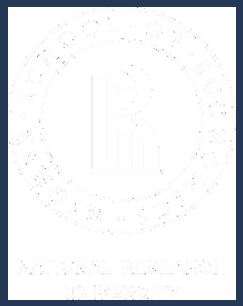
3. Combination of conventional machine learning techniques with face descriptors extracted by contemporary CNNs has prooved its potential

Kharchevnikova Angelina

Sokolova Anastasiya

Rassadin Alexander

Tarasov Alexander



Thank you!