SACI 2018

May 17-19, 2018

Timisoara, Romania

Granular Computing and Sequential Analysis of Deep Embeddings in Fast Still-to-Video Face Recognition

Andrey V. Savchenko
Dr. of Sci., PhD,
- Professor in Department of Information Systems and Technology
National Research University Higher School of Economics, Nizhny Novgorod, Russia
URL: http://www.hse.ru/en/staff/avsavchenko

-Senior Researcher in PDMI-Samsung AI Joint Laboratory
URL: https://samsung.pdmi.ras.ru/staff/
E-mail: avsavchenko@hse.ru

May 19, 2018

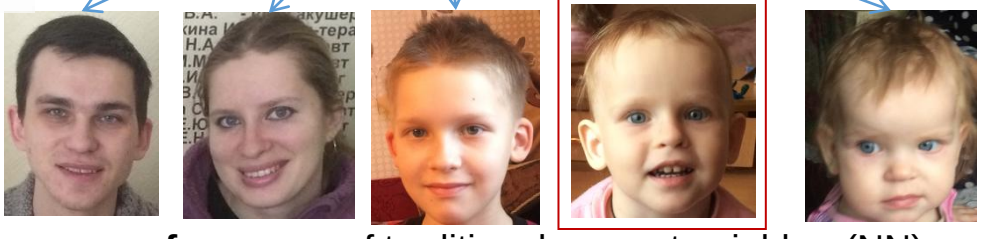
Still-to-video face identification

What for?

Let the input sequence $\{X(t)\}$ of $T>1$ frames be specified. The problem is to **assign** this sequence to one of $R>1$ **identities** specified by R **reference** images $\{X_r\}$ from the gallery set.
We consider the **small-sample-size** case: number of reference photos per class is small (including one sample per person)

Input video sequence

Gallery set



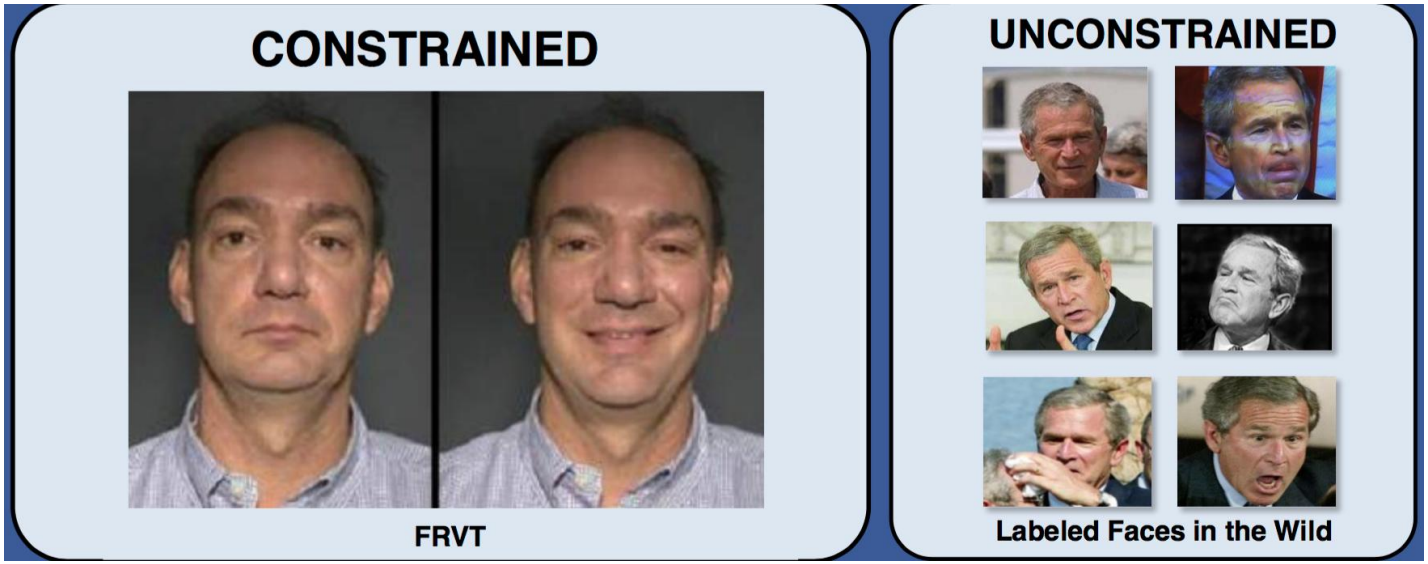
Key idea

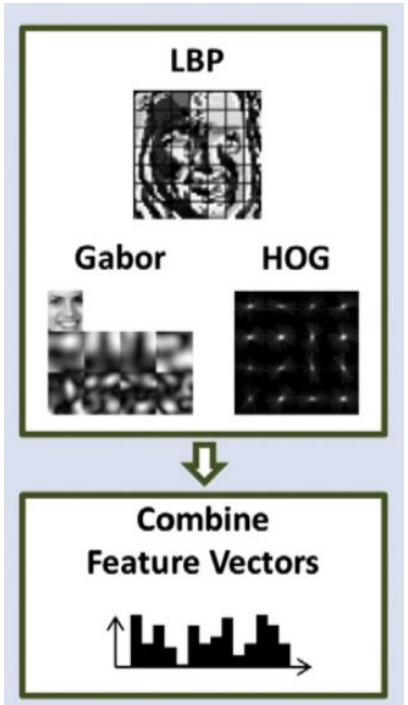
We improve **performance** of traditional nearest neighbor (NN) methods by using sequential three-way decisions (TWD) and granular computing

And now we introduce the agenda of our talk

1. State-of-the-art: Deep Embeddings
2. Proposed Method
3. Experimental results
4. Conclusion and future work

Unconstrained face recognition

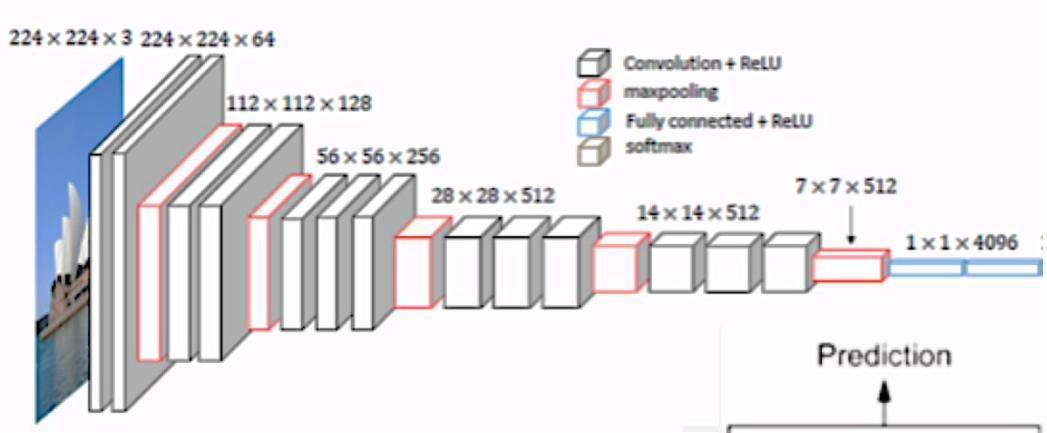




Conventional features (HOG+LBP+Gabor).
Ortiz E.G., Becker B.C. Face recognition for web-scale datasets. CVIU 2014
Accuracy 50-80%

Deep Embeddings

Deep Convolutional Neural Networks (CNN)



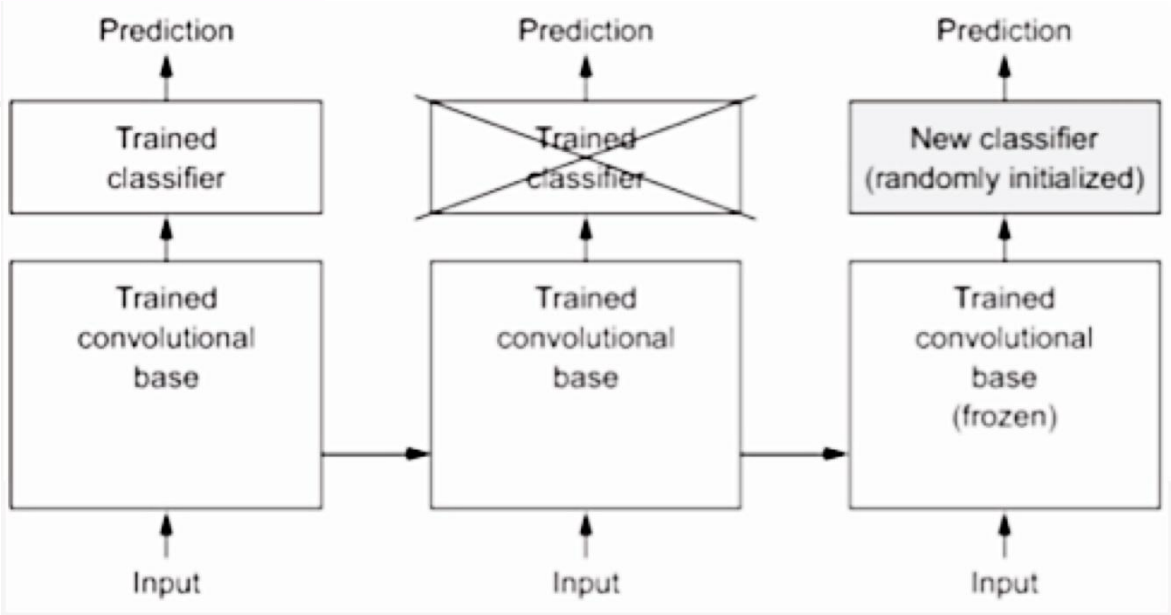
Deep embeddings - CNN features

Facial images are described by *D*-dimensional feature vectors

State-of-the-art CNNs

VGGFace, FaceNet, CenterFace, DeepID (2/3) LightCNN, VGGFace2, SphereFace, ArcFace

Accuracy: 85-99%



Still-to-video recognition

Nearest neighbor (NN) classifier for small sample size problem

Recognition of each frame

$$\min_{r\in\{1,\dots,R\}} \sum_{t=1}^{T} \rho(\mathbf{x}(t), \mathbf{x}_r)$$

Run-time complexity per frame: O(*RD*).
Brute force of the whole database

[Learnable] pooling (aggregation)

$$\bar{\mathbf{x}} = \sum_{t=1}^{T} w_t \mathbf{x}(t)$$

Deep embeddings are **high-dimensional**

Performance is usually **insufficient**, especially for implementation at autonomous mobile platforms

Multi-objective optimization:

$$\bar{\alpha} \to \min \qquad \bar{t} \le t_0$$

Sequential three-way decisions and granular computing

Yao Y., Information Sciences, 2010:
“A **positive** rule makes a **decision of acceptance**, a **negative** rule makes a **decision of rejection**, and a **boundary** rule makes a **decision of abstaining**”

Key question: how to make a decision if the boundary region was chosen?
Yao Y. Proc. of RSKT, LNCS, 2013: "Objects with a non-commitment decision may be further investigated by using fine-grained granules"

Sequential three-way decisions and granular computing

Yao Y., Information Sciences, 2010:
"A **positive** rule makes a **decision of acceptance**, a **negative** rule makes a **decision of rejection**, and a **boundary** rule makes a **decision of abstaining**"

Key question: how to make a decision if the boundary region was chosen?
Yao Y. Proc. of RSKT, LNCS, 2013: "Objects with a non-commitment decision may be further investigated by using fine-grained granules"

PCA (principal component analysis), scores are ordered by corresponding eigenvalues

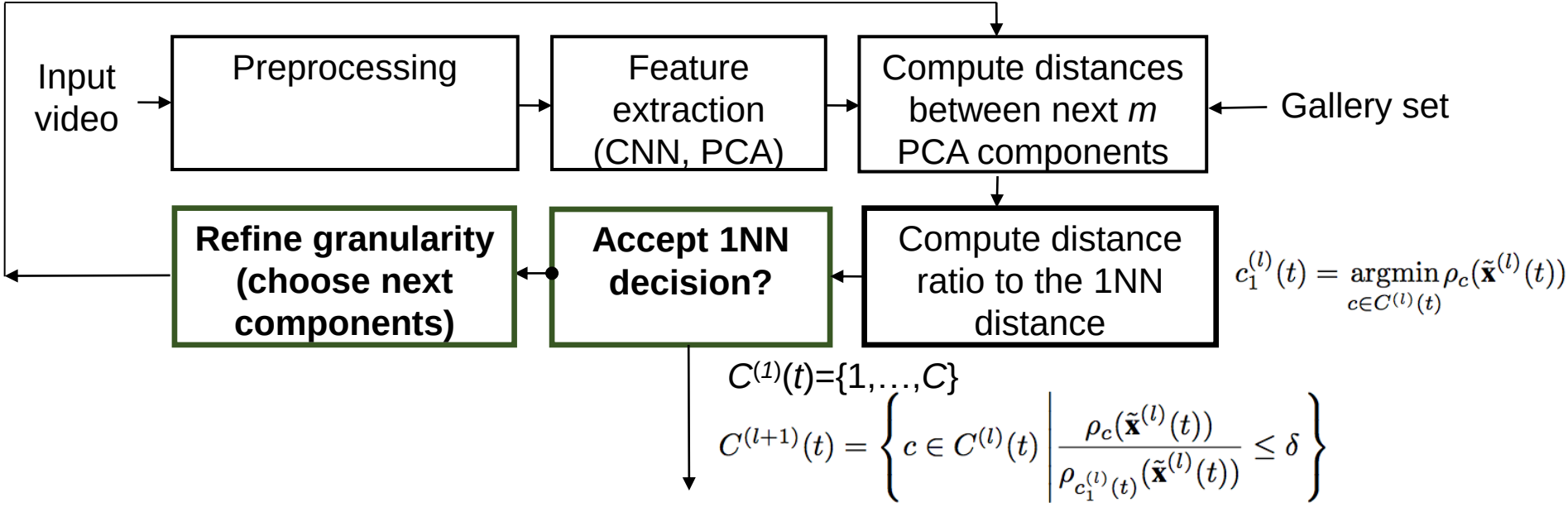
$$\tilde{\mathbf{x}}(t) = [\tilde{x}_1(t), ..., \tilde{x}_D(t)]$$

Proposed: representation of frame at the *l*-th granularity level includes first $d^{(l)}=lm$ principal components. This representation is **computationally cheap** for additive distances

$$\rho\left(\tilde{\mathbf{x}}^{(l+1)}(t), \tilde{\mathbf{x}}_r^{(l+1)}\right) = \rho\left(\tilde{\mathbf{x}}^{(l)}(t), \tilde{\mathbf{x}}_r^{(l)}\right) + \sum_{d=d^{(l)}+1}^{d^{(l+1)}} \rho(\tilde{x}_d(t), \tilde{x}_{r;d})$$

Proposed approach based on sequential TWD

$$\rho_c(\tilde{\mathbf{x}}^{(l)}(t)) = \min_{r\in\{1,\dots,R\},c(r)=c} \rho(\tilde{\mathbf{x}}^{(l)}(t), \tilde{\mathbf{x}}_r^{(l)})$$



Final Maximum a-posterior (MAP) decision

$$\max_{c\in C^{(L)}} \sum_{t=1}^{T} \frac{\exp(-n\rho_c(\tilde{\mathbf{x}}^{(l)}(t)))}{\sum_{i\in C^{(L)}} \exp(-n\rho_c(\tilde{\mathbf{x}}^{(l)}(t))}$$

Strong theoretical foundations for the Jensen-Snannon and Kullback-Leibler divergences

Here is exactly how our method works in practice. Granularity level l=1

Probe photo Closest gallery photos

(a) (b) (c) (d)

Distance factor threshold 0.7
32 principal components

Subject	$l = 1$ $\rho[r]$	$\rho_{\min}/\rho[r]$
Armstrong (d)	0.0086	0.87
Auriemma (b)	0.0074	1.00
McEwen (c)	0.0104	0.72
Williams	0.0100	0.75
Wirayuda	0.0103	0.73
LeBron	0.0105	0.71

Here is exactly how our method works in practice. Granularity level l=2

Probe photo Closest gallery photos

(a) (b) (c) (d)

Distance factor threshold 0.7
64 principal components

Subject	$l = 2$ $\rho[r]$	$\rho_{\min}/\rho[r]$
Armstrong (d)	0.0122	1
Auriemma (b)	0.0170	0.71
McEwen (c)	0.0188	0.65
Williams	0.0300	0.41
Wirayuda	0.0217	0.56
LeBron	0.0200	0.61

Here is exactly how our method works in practice. Granularity level *l*=3

Probe photo Closest gallery photos

(a) (b) (c) (d)

Distance factor threshold 0.7
96 principal components

Subject	$l = 3$	
	$\rho[r]$	$\rho_{\min}/\rho[r]$
Armstrong (d)	0.0129	1
Auriemma (b)	0.0195	0.66

Better recognition performance (no need to process all features)

Higher recognition accuracy

Experiments.

OpenCV Caffe

1 **Pre-trained deep CNNs**

1.1 VGG-Face (VGG-16 Network): *D* = 4096 embeddings at “fc8” layer

2.2 ResFace-101(ResNet-101 for face recognition): *D* = 2048 features from “pool5” layer

2.3 VGGFace2_ft (ResNet-50 trained on VGGFace2): *D* = 2048 embeddings at “pool5/7x7_s1” layer

2 **Datasets**

2.1 **Testing set**: YTF (YouTube Faces). **Training set**: subset of LFW (Labeled Faces in the Wild)

Number of subjects *C* 1589 Number of training photos *R* 4732

Number of testing videos 3353 183 frames per clip in average

2.2 **Testing set**: videos from IJB-A (IARPA Janus Benchmark). **Training set**: still images from IJB-A

Number of subjects *C* 500 Number of training photos *R* 5712

Number of testing videos 2043 8 frames per clip in average

10-times random sub-sampling cross-validation: 80% of training set

YTF/LFW

	ResFace		VGGFace		VGGFace2_ft	
Classifier	**Accuracy (%)**	**Running time (ms)**	**Accuracy (%)**	**Running time (ms)**	**Accuracy (%)**	**Running time (ms)**
SVM	18.53±0.5	10213±33	22.49±0.6	19840±40	52.85±1.2	10137±35
SVM, AvgPooling	22.15±1.2	10269±39	26.90±0.7	20070±37	55.91±1.3	10252±42
k-NN	30.92±0.1	12.3±0.0	43.44±0.1	23.3±0.1	74.48±0.1	12.7±0.0
MAP	31.49±0.0	12.3±0.0	43.50±0.1	23.7±0.1	74.41±0.1	12.6±0.1
k-NN/32 PCA	26.28±0.3	1.7±0.0	37.69±0.1	2.2±0.1	58.81±0.0	19±0.1
k-NN/256 PCA	31.21±0.1	2.3±0.0	44.64±0.2	2.9±0.1	75.02±0.0	2.5±0.1
Proposed, k-NN	**31.21±0.1**	**0.8±0.0**	**44.64±0.1**	**1.1±0.1**	**74.77±0.1**	**0.7±0.0**
Proposed, MAP	**31.52±0.2**	**0.9±0.1**	**45.46±0.1**	**1.2±0.1**	**74.87±0.2**	**0.7±0.0**

IJB-A still-to-video recognition

	ResFace		VGGFace		VGGFace2_ft	
Classifier	**Accuracy (%)**	**Running time (ms)**	**Accuracy (%)**	**Running time (ms)**	**Accuracy (%)**	**Running time (ms)**
SVM	55.98±0.3	117.3±0.9	70.88±0.3	230.3±1.6	84.12±0.4	117.8±1.1
SVM, AvgPooling	57.17±0.5	114.7±0.7	73.62±0.0	233.0±4.7	87.06±0.1	124.4±1.4
k-NN	54.41±0.0	12.9±0.0	68.88±0.3	24.9±0.2	85.67±0.2	12.8±0.1
MAP	55.16±0.2	13.3±0.3	70.50±0.8	25.0±0.0	87.09±0.1	12.8±0.1
k-NN/32 PCA	48.49±0.4	1.2±0.1	66.26±0.3	1.9±0.2	77.91±0.2	1.3±0.1
k-NN/256 PCA	53.91±0.3	1.9±0.1	70.94±0.1	2.5±0.1	85.48±0.1	2.1±0.0
Proposed, k-NN	**54.12±0.2**	**0.8±0.1**	**70.94±0.1**	**1.2±0.0**	**85.65±0.2**	**0.8±0.0**
Proposed, MAP	**56.01±0.1**	**0.8±0.1**	**73.06±0.5**	**1.2±0.1**	**86.89±0.1**	**0.8±0.1**

And summarizing our results we have the following conclusions

Proposed implementation of sequential TWD has a list of **advantages**

1. The usage of the distance ratio made it possible to dramatically reduce the search space without losses in accuracy.
2. Our method may be successfully applied with any additive dissimilarity measures.
3. Our approach is 2-10 times faster than conventional methods (k-NN, SVM) and leads to the most accurate decisions in practically all cases
4. C++ implementation is freely available: https://github.com/HSE-asavchenko/HSE_FaceRec/tree/master/windows

and **disadvantages**

1. It is important to tune the threshold for the distance ratio. Thresholds potentially differs for various granulation levels
2. Current implementation only works with 1-NN methods. State-of-the-art classifiers are more accurate if the training sample is not small

What we are going to do in the future

Further research directions

1. From recognition of a set of objects to the state-of-the-art video processing
   - 1.1 Take into account temporal coherence of sequential frames
   - 1.2 End-to-end learning for frames weighting to make distinction between frames with different quality
2. Application with various classifiers
   - 2.1 Point-to-set metric learning
   - 2.2 More complex classifiers at the finest granularity level for reduced number of classes, e.g., one-vs-one SVM

Related papers

1. Savchenko, A. V. (2016). *Search techniques in intelligent classification systems*. Springer.
2. Savchenko, A. V., & Belova, N. S. (2018). Unconstrained Face Identification Using Maximum Likelihood of Distances Between Deep Off-the-shelf Features. *Expert Systems with Applications*.
3. Savchenko, A. V. (2017). Maximum-likelihood approximate nearest neighbor method in real-time image recognition. *Pattern Recognition*, 61, 459-469.
4. Savchenko, A. V. (2017). Clustering and maximum likelihood search for efficient statistical classification with medium-sized databases. *Optimization Letters*.
5. Savchenko, A. V. (2016). Fast multi-class recognition of piecewise regular objects based on sequential three-way decisions and granular computing. *Knowledge-Based Systems*

Thank you for your attention

Any Questions?