NATIONAL RESEARCH UNIVERSITY

National Research University Higher School of Economics (HSE) – N. Novgorod

OFFLINE ANALYSIS AND RECOGNITION OF PHOTOS IN A GALLERY ON MOBILE DEVICE

Andrey V. Savchenko
Dr. of Sci., Prof.,
Lead Researcher in HSE's international laboratory LATNA
Email: avsavchenko@hse.ru
URL: www.hse.ru/en/staff/avsavchenko

Huawei Workshop on Fundamental and Applied Problems of Machine Learning
December 19, 2019

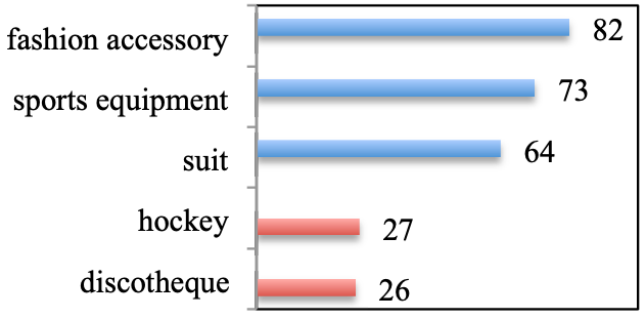
Deep understanding of user characteristics by analyzing user images and videos in a mobile device

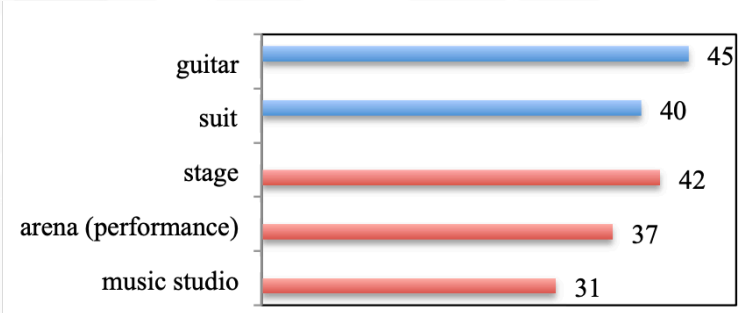
User images

…

…

Profile of interests





Outline

1. Event recognition in still images
2. Event recognition in a gallery of images
3. Sequential analysis of high-dimensional features
4. PNN with complex exponential activation functions
5. Organizing photo and video albums on mobile device

Event recognition in still images

Event recognition

"An event captures the complex behavior of a group of people, interacting with multiple objects, and taking place in a specific environment. Images from the same event category may vary even more in visual appearance and structure" (Wang et al, IJCV 2018)

WIDER (Web Image Dataset for Event Recognition)

PEC (Photo Event Collection)

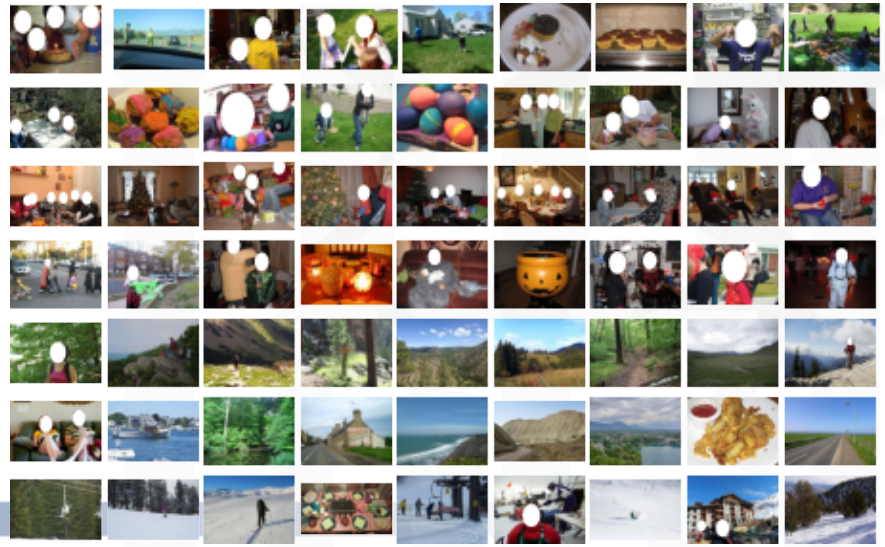
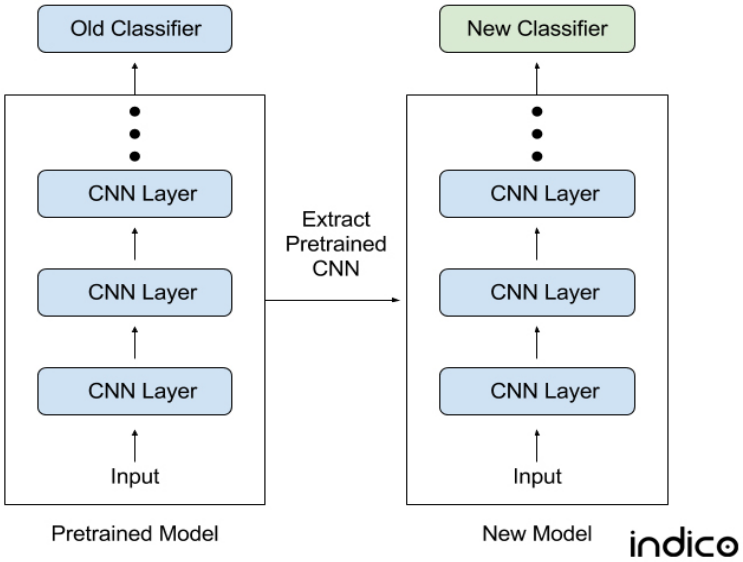


Image recognition: it is required to assign an observed image X to one of C classes. Training set contains N reference images (examples) $\{X_n\}$, $n \in \{1,\dots N\}$, with known class label $c_n \in \{1,\dots C\}$

Conventional approach

1 Fine-tune convolutional neural network (CNN) pre-trained on ImageNet, Places, etc.

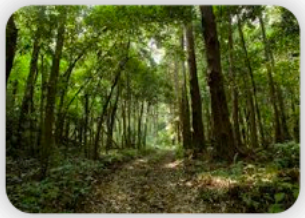


2 Classify *embeddings (features)* from one of the last CNN's layers: D-dimensional feature vector $\mathbf{x}=[x_1,\dots, x_D]$

Training set is associated with embeddings $\{\mathbf{x}_n\}$, $\mathbf{x}_n=[x_{n,1},\dots, x_{n;D}]$.

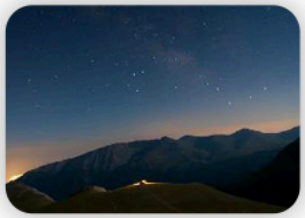
Generate textual descriptions of images

Google's Conceptual Captions



by Joi Ito

the trail climbs steadily uphill most of the way.



by Danail Nachev

the stars in the night sky.

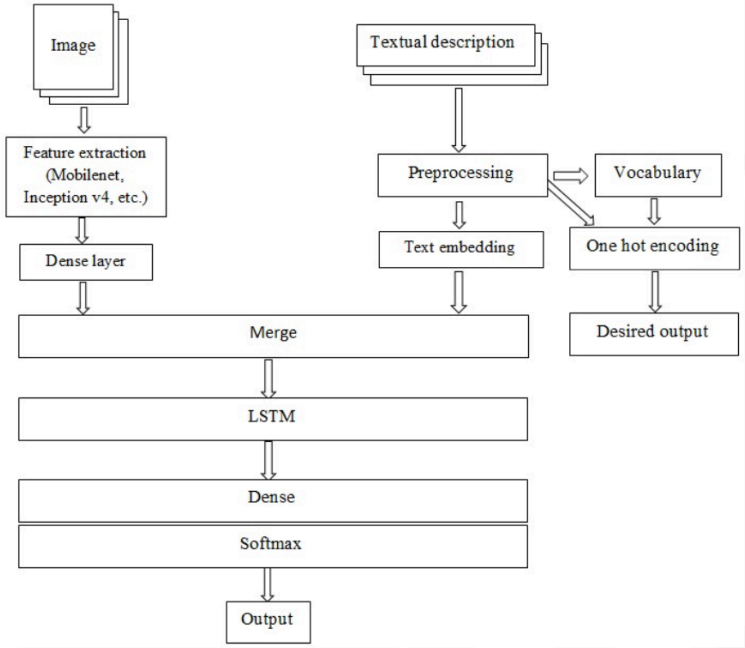
by Justin Higuchi

musical artist performs on stage during festival.

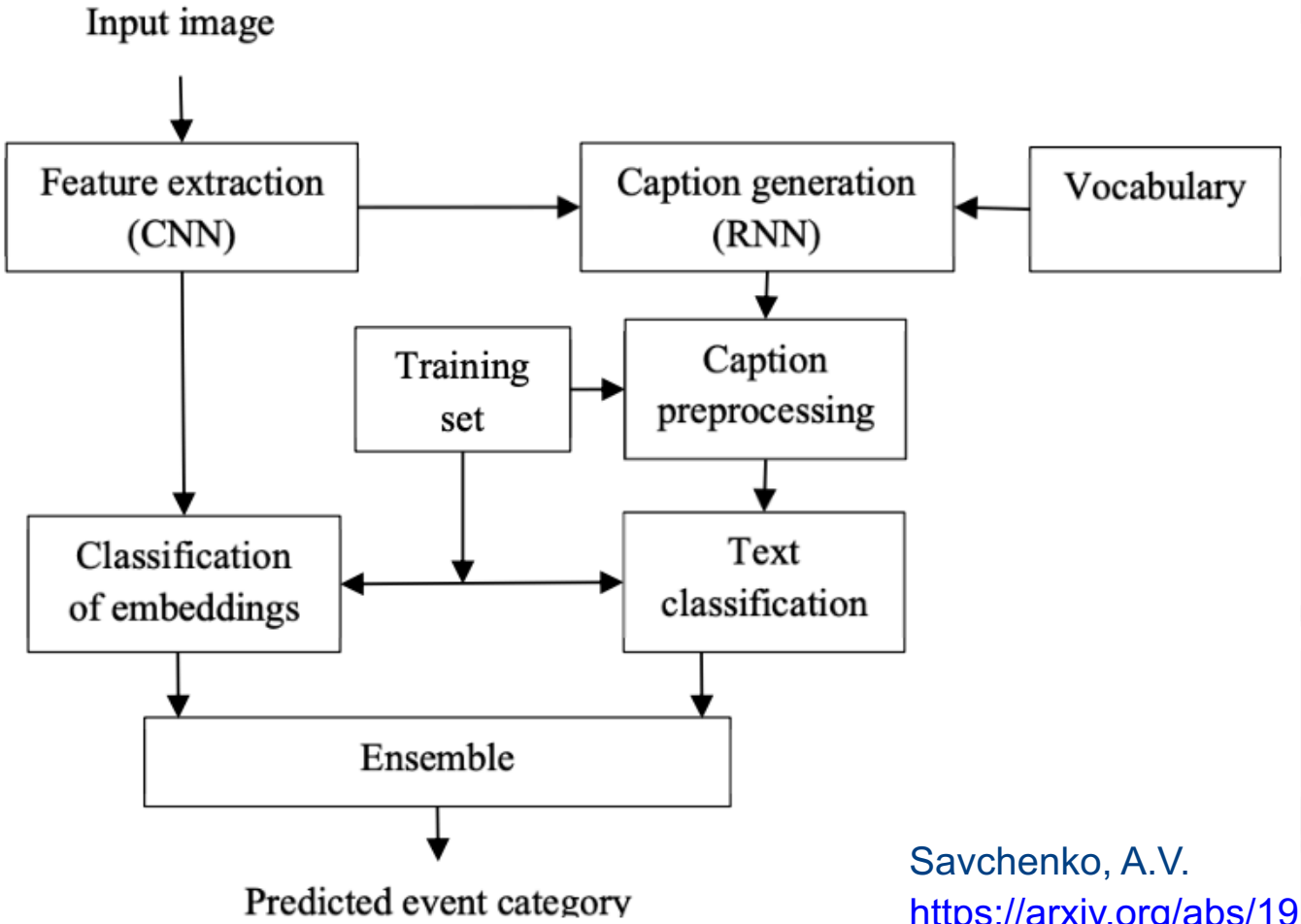
by Viaggio Routard

popular food market showing the traditional foods from the country.

- Show and Tell
- Show, Attend and Tell
- Neural Baby Talk
- Multimodal RNN
- **Auto-Reconstructor Network (ARNet)**



Proposed pipeline for event recognition in single images



Savchenko, A.V. https://arxiv.org/abs/1911.11010, 2019

Qualitative results

the statue of liberty and the moon
ThemePark (texts)
Christmas (embeddings)
ThemePark (ensemble)

a woman is doing a handstand at a local fair
PersonalSports (texts)
ReligiousActivity (embeddings)
PersonalArtActivity (ensemble)

person , a painting by person
Museum (texts)
UrbanTrip (embeddings)
PersonalArtActivity (ensemble)

the tower of the city
ThemePark (texts)
Architecture (embeddings)
ThemePark (ensemble)

Experimental results

WIDER

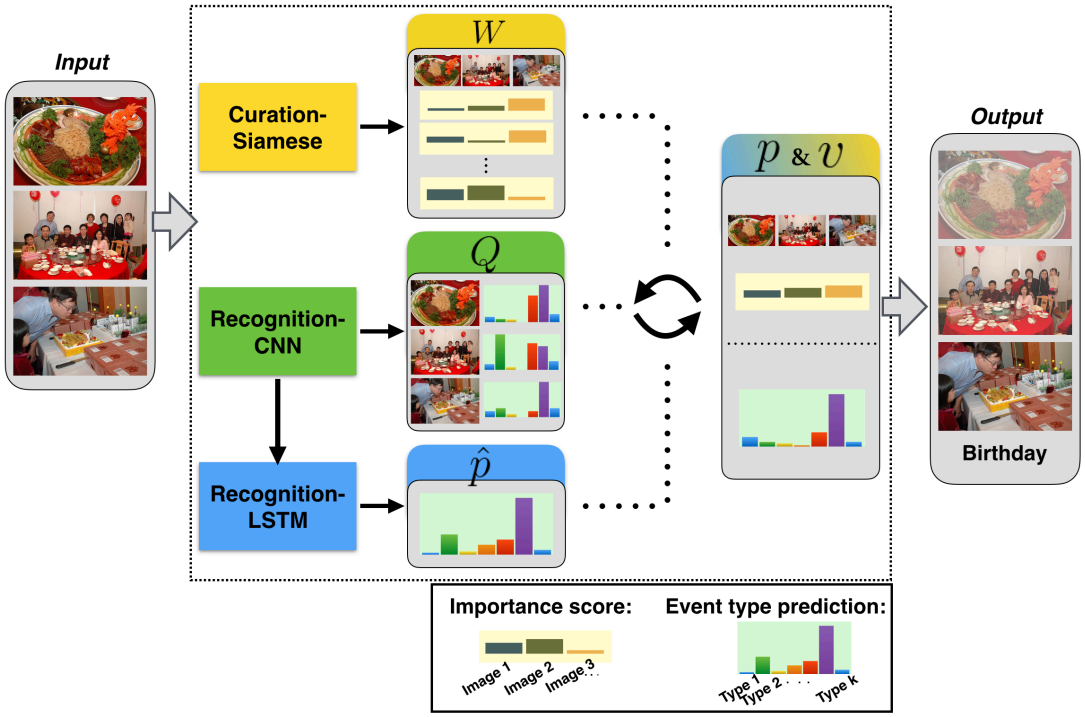
Classifier	Features	Lightweight models	Deep models
SVM	Embeddings	48.31	50.48
	Objects	19.91	28.66
	Texts	26.38	31.89
	Proposed ensemble (4), (5)	48.91	51.59
Fine-tuned CNN	Embeddings	49.11	50.97
	Objects	12.91	21.27
	Texts	25.93	30.91
	Proposed ensemble (4), (5)	49.80	51.84
Baseline CNN [35]			39.7
Deep channel fusion [35]			42.4
Initialization-based transfer learning [32]			50.8
Transfer learning of data and knowledge [32]			53.0

ML-CUFED (Multi-Label Curation of Flickr Events Dataset)

Classifier	Features	Lightweight models	Deep models
SVM	Embeddings	53.54	57.27
	Objects	34.21	40.94
	Texts	37.24	41.52
	Proposed ensemble (4), (5)	55.26	58.86
Fine-tuned CNN	Embeddings	56.01	57.12
	Objects	32.05	40.12
	Texts	36.74	41.35
	Proposed ensemble (4), (5)	57.94	60.01

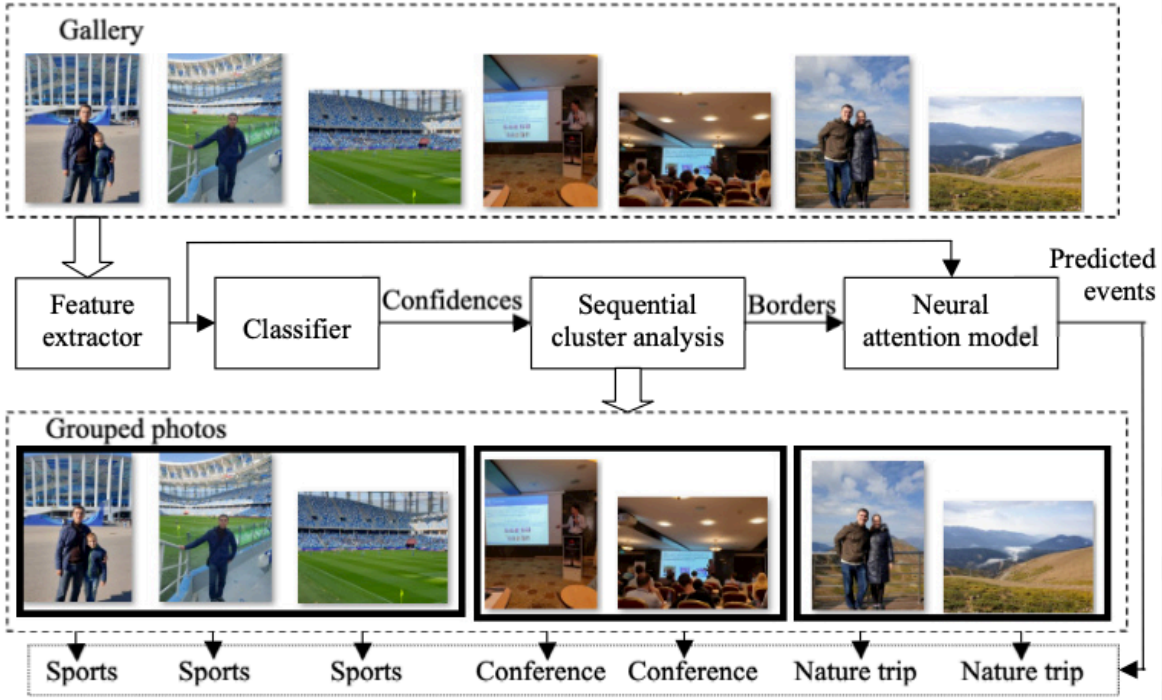
Event recognition in a gallery of images

Image-set recognition: it is required to an album of images X_t, $t \in \{1,\dots T\}$ to one of C event classes. Training set of N albums is given: n-th reference album with known class label $c_n \in \{1,\dots C\}$ is defined by a set of images $\{X_n(1),\dots X_n(L_n)\}$



Wang, Y. et al. Recognizing and Curating Photo Albums via Event-Specific Image Importance, BMVC17

In practice: it is required to assign **each photo** X_t, $t \in \{1, \dots T\}$ from a **gallery** to one of C event classes. Training set is the same as above



1. Combine sequential photos using distance between L_2-normed **classifier's scores** for CNN embeddings

2. Learn distance threshold by random permutation of albums from the training set

3. **Attention mechanism**

$$\mathbf{x}(k) = \sum_{t=t_{k-1}+1}^{t_k} w(\mathbf{x}_t)\mathbf{x}_t,$$

$$w(\mathbf{x}_t) = \frac{\exp(\mathbf{q}^T\mathbf{x}_t)}{\sum_{j=t_{k-1}+1}^{t_k} \exp(\mathbf{q}^T\mathbf{x}_j)}$$

Savchenko, A.V.
https://arxiv.org/abs/1911.11010, 2019

Experimental results (1). Conventional tasks

Accuracy (%) of event recognition in a set of images (album)

CNN	Aggregation	PEC	ML-CUFED
MobileNet2, $\alpha = 1.0$	AvgPool	86.42	81.38
	Attention	89.29	84.04
MobileNet2, $\alpha = 1.4$	AvgPool	87.14	81.91
	Attention	87.36	84.31
Inception v3	AvgPool	86.43	82.45
	Attention	87.86	84.84
AlexNet	CNN-LSTM-Iterative [33]	84.5	79.3
	Aggregation of representative features [34]	87.9	84.5
ResNet-101	CNN-LSTM-Iterative [33]	84.5	71.7
	Aggregation of representative features [34]	89.1	83.4

Accuracy (%) of event recognition in a single image

Dataset	CNN	Baseline	Embeddings		Scores
			L_2	χ^2	L_2
PEC	MobileNet2, $\alpha = 1.0$	58.32	60.42	60.69	58.44
	MobileNet2, $\alpha = 1.4$	60.34	61.25	61.92	60.58
	Inception v3	61.82	64.19	64.22	61.97
ML-CUFED	MobileNet2, $\alpha = 1.0$	54.41	57.03	57.45	54.56
	MobileNet2, $\alpha = 1.4$	53.54	54.97	55.98	54.03
	Inception v3	57.26	59.19	60.12	57.87

Experimental results (2). Event recognition is a gallery

Accuracy (%) for PEC

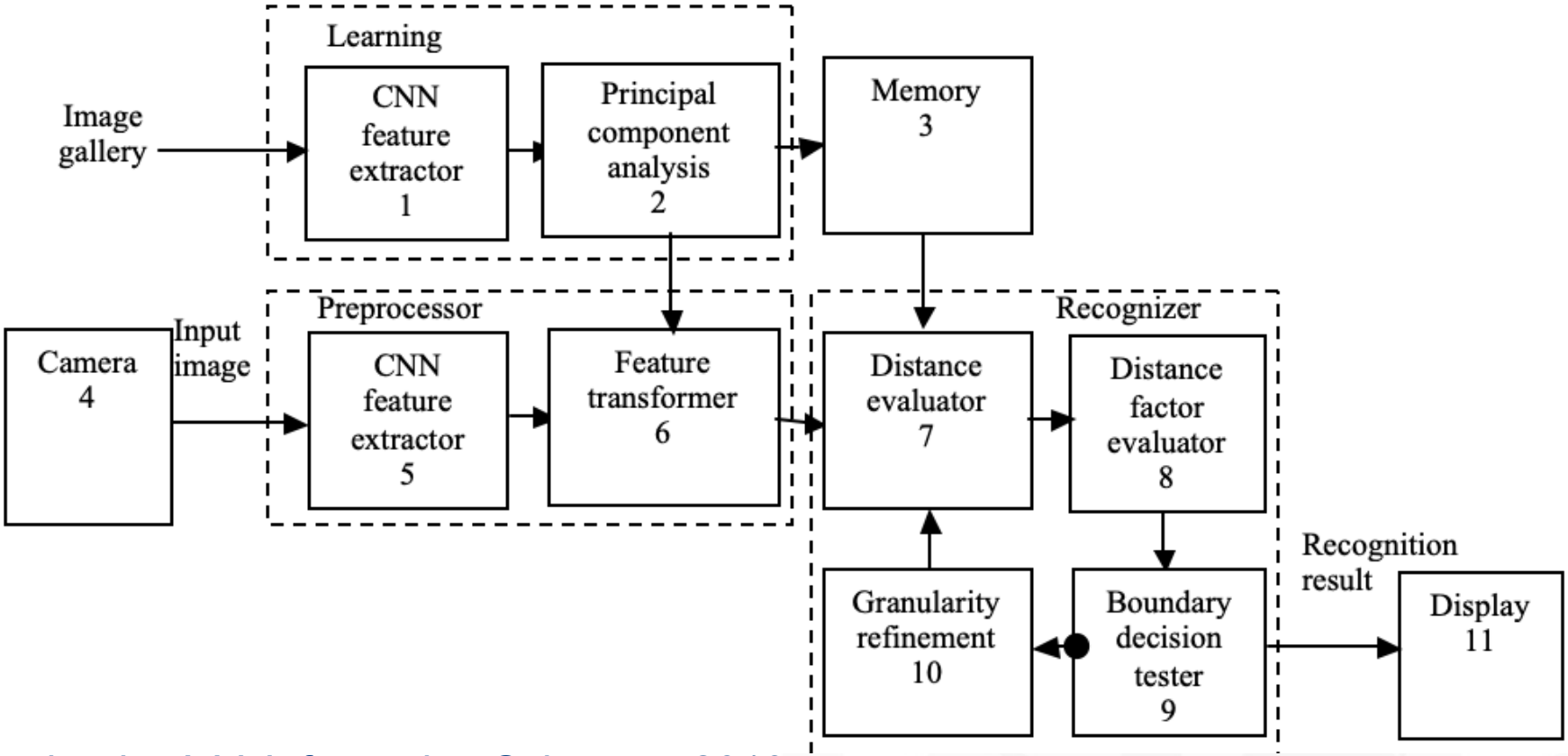
CNN	Aggregation	Baseline	Embeddings		Scores		Scores (L_2-normed)
			L_2	χ^2	L_2	χ^2	L_2
MobileNet2, $\alpha = 1.0$	AvgPool	58.32	66.85 ± 0.59	68.52 ± 0.89	71.08 ± 0.59	-	72.68 ± 0.56
(pre-trained), embeddings	Attention	54.43	68.51 ± 0.41	70.65 ± 1.20	74.49 ± 0.70	-	80.48 ± 1.01
MobileNet2, $\alpha = 1.4$	AvgPool	60.34	68.85 ± 0.59	69.57 ± 0.57	72.59 ± 1.49	-	73.49 ± 0.86
(pre-trained), embeddings	Attention	55.36	70.53 ± 0.79	71.16 ± 0.72	78.20 ± 1.47	-	81.27 ± 0.81
MobileNet2, $\alpha = 1.4$	AvgPool	61.89	-	-	75.66 ± 0.55	76.96 ± 0.97	-
(fine-tuned), scores	Attention	61.55	-	-	78.77 ± 0.49	81.33 ± 0.69	-
Inception v3	AvgPool	61.82	72.29 ± 1.28	72.32 ± 1.54	74.54 ± 1.04	-	76.48 ± 0.47
(pre-trained), embeddings	Attention	56.94	72.38 ± 1.13	71.96 ± 0.67	76.76 ± 0.70	-	80.17 ± 1.14
Inception v3	AvgPool	63.56	-	-	78.87 ± 0.67	79.92 ± 0.65	-
(fine-tuned), scores	Attention	62.91	-	-	81.03 ± 0.77	81.95 ± 1.11	-

Accuracy (%) for ML-CUFED

CNN	Aggregation	Baseline	Embeddings		Scores		Scores (L_2-normed)
			L_2	χ^2	L_2	χ^2	L_2
MobileNet2, $\alpha = 1.0$	AvgPool	54.41	67.54 ± 0.76	67.42 ± 0.93	69.83 ± 0.74	-	70.42 ± 0.41
(pre-trained), embeddings	Attention	51.05	68.71 ± 0.71	68.55 ± 0.61	71.44 ± 0.82	-	71.61 ± 0.69
MobileNet2, $\alpha = 1.4$	AvgPool	53.54	66.93 ± 0.60	67.21 ± 0.55	68.56 ± 0.73	-	69.47 ± 0.36
(pre-trained), embeddings	Attention	51.12	68.34 ± 0.68	68.62 ± 0.50	70.79 ± 0.75	-	71.78 ± 0.74
MobileNet2, $\alpha = 1.4$	AvgPool	56.01	-	-	70.57 ± 0.48	71.61 ± 0.28	-
(fine-tuned), scores	Attention	56.09	-	-	72.90 ± 0.59	73.46 ± 0.58	-
Inception v3	AvgPool	57.26	69.91 ± 0.58	70.01 ± 0.62	72.25 ± 0.61	-	72.78 ± 0.71
(pre-trained), embeddings	Attention	50.89	69.30 ± 0.47	68.52 ± 0.89	72.73 ± 0.72	-	73.00 ± 0.65
Inception v3	AvgPool	57.12	-	-	72.18 ± 0.63	73.20 ± 0.74	-
(fine-tuned), scores	Attention	57.29	-	-	73.06 ± 0.74	73.92 ± 0.81	-

Sequential analysis of high-dimensional features

Three-way decisions to choose robust representation of the input image



- Savchenko A.V. Information Sciences, 2019
- Savchenko A.V. Knowledge-Based Systems, 2016
- Patent RU 2706960 (22.11.2019) / Author: Savchenko A.V. Assignee: Samsung

[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

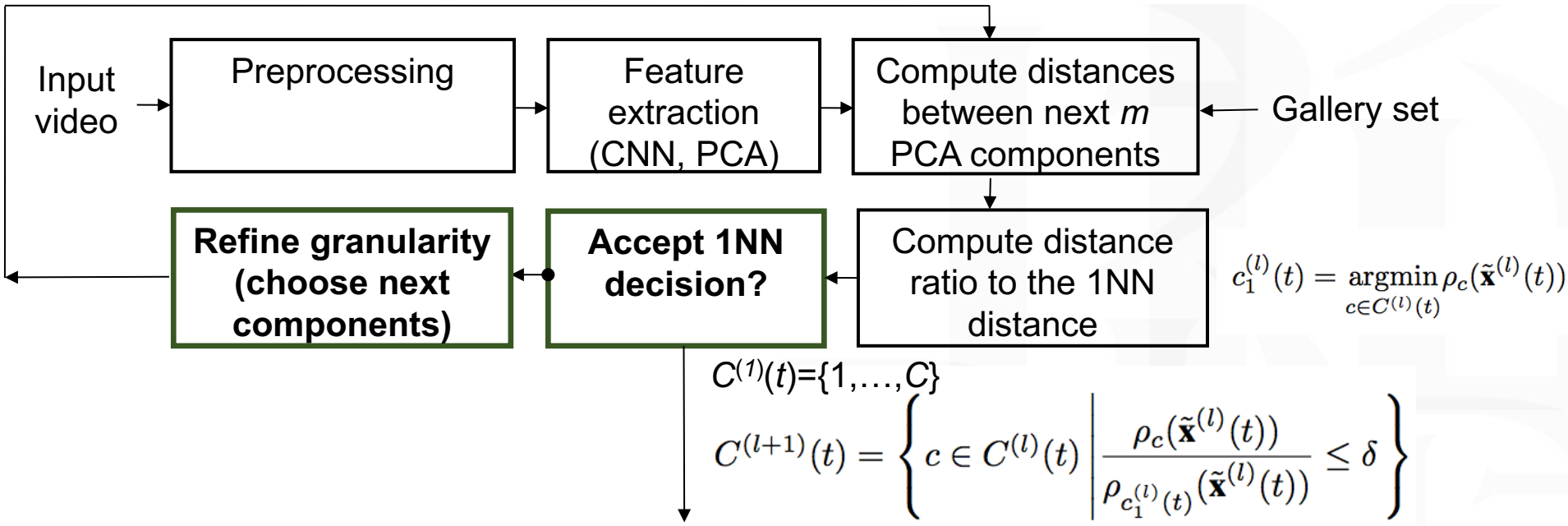
$$\tilde{\mathbf{x}}(t) = [\tilde{x}_1(t), ..., \tilde{x}_D(t)]$$

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

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

Distance ratio

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



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

Probe photo Closest gallery photos

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

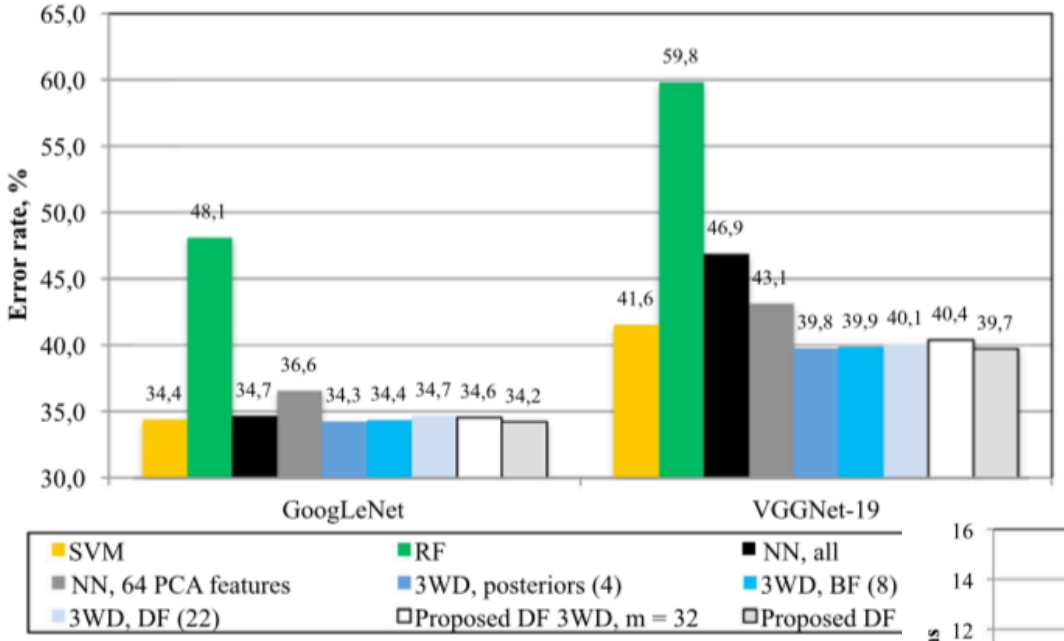
Recognition results, distance factor threshold 0.7

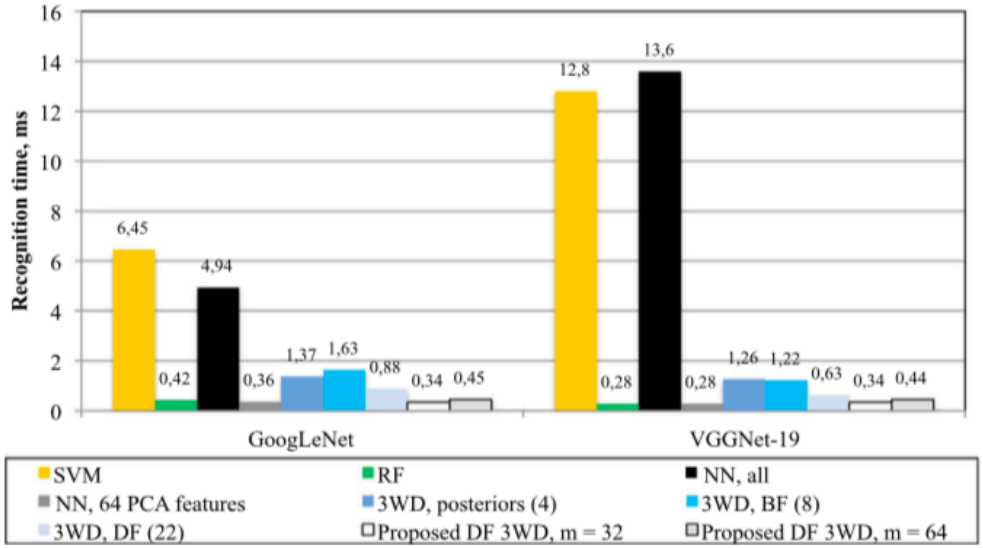
Subject	$l=1$		$l=2$		$l=3$	
	$\rho[r]$	$\rho_{\min}/\rho[r]$	$\rho[r]$	$\rho_{\min}/\rho[r]$	$\rho[r]$	$\rho_{\min}/\rho[r]$
Armstrong (d)	0.0086	0.87	0.0122	1	0.0129	1
Auriemma (b)	0.0074	1.00	0.0170	0.71	0.0195	0.66
McEwen (c)	0.0104	0.72	0.0188	0.65	-	-
Williams	0.0100	0.75	0.0300	0.41	-	-
Wirayuda	0.0103	0.73	0.0217	0.56	-	-
LeBron	0.0105	0.71	0.0200	0.61	-	-

Better recognition performance (no need to process all features)

Higher recognition accuracy

Experimental results (Caltech-256 dataset)





Probabilistic Neural Network (PNN) with complex exponential activation functions

Statistical approach: empirical Bayesian classifier with naïve assumption about independent features

$$c^* = \arg\max_{c\in\{1,\dots,C\}} \frac{R(c)}{R} f(\mathbf{x}|W_c) \qquad \hat{f}(\mathbf{x}|W_c) = \prod_{d=1}^{D} \hat{f}_d(x_d|W_c)$$

1 We propose to estimate the individual likelihood as the average of the first J partial sums. Here the right-hand side is the non-negative Fejér kernel

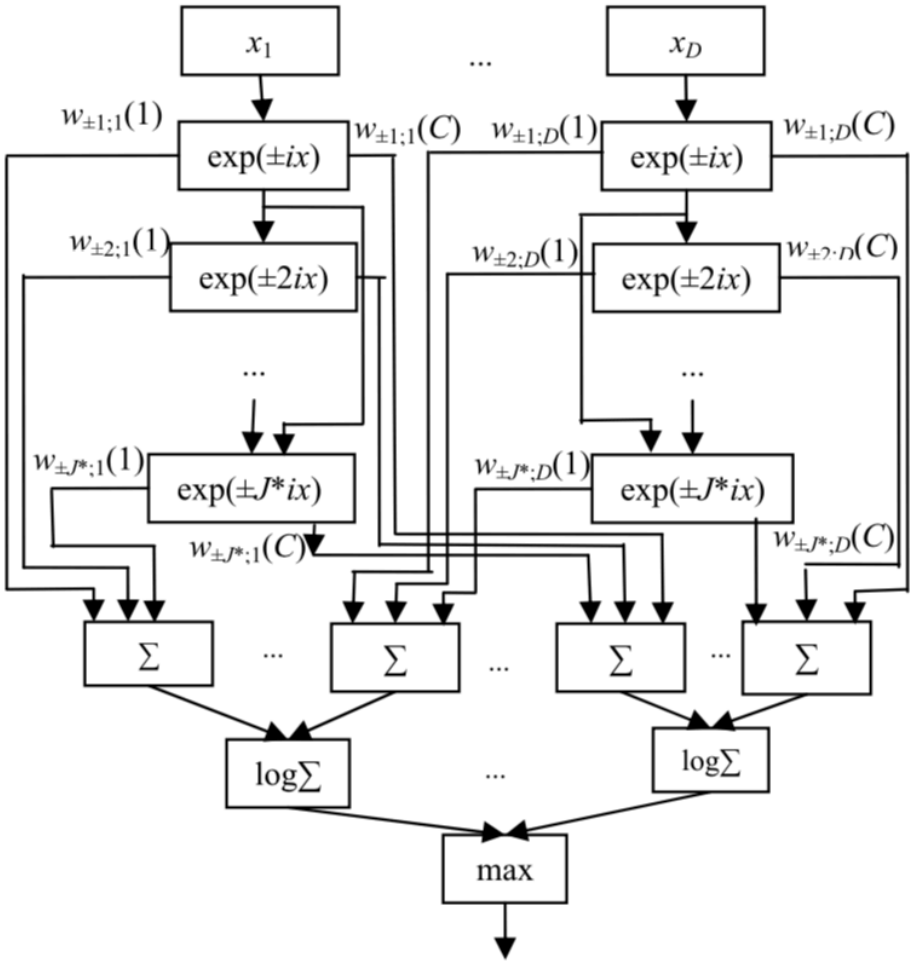
$$\hat{f}_d(x_d|W_c) = \frac{1}{J+1}\sum_{j=0}^{J} \hat{f}_{j;d}(x_d|W_c) \qquad F_{J+1} = \frac{1}{J+1}\left(\frac{\sin((J+1)\pi(x_d - x_{r;d}(c))/2)}{\sin(\pi(x_d - x_{r;d}(c))/2)}\right)^2$$

2 We replace canonical form of density estimate to the equivalent form, **which does not implement the brute force**

$$\hat{f}_{J;d}(x_d|W_c) = \sum_{j=-J}^{J} a_{j;d}(c)\exp(ij\pi x_d) \qquad \psi_j(x) = \exp(ij\pi x)$$

- Savchenko, A.V. IEEE Transactions on Neural Networks and Learning Systems, 2019
- Savchenko A.V., IEEE ICPR 2018

PNN based on trigonometric series



$$c^* = \underset{c \in \{1,\ldots,C\}}{\operatorname{argmax}} \sum_{d=1}^{D} \log \sum_{j=-J}^{J} w_{j;d}(c) \cdot \psi_j(x_d)$$

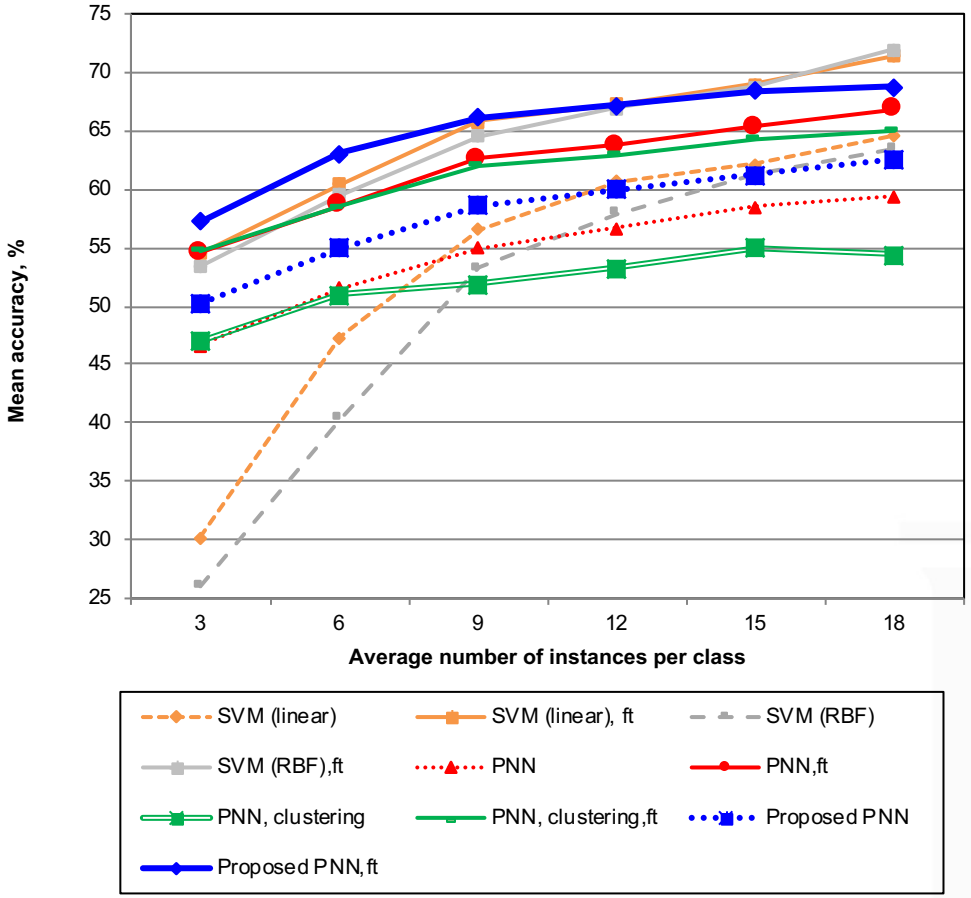
$$\psi_{j+1}(x_d) = \psi_j(x_d) \cdot \psi_1(x_d),$$
$$\psi_{-j}(x_d) = \overline{\psi_j(x_d)}.$$

- **Converges to Bayesian solution**
- **Very high training speed**
- **Runtime complexity and memory space complexity:** $O(DR^{1/3}C^{1/3})$.

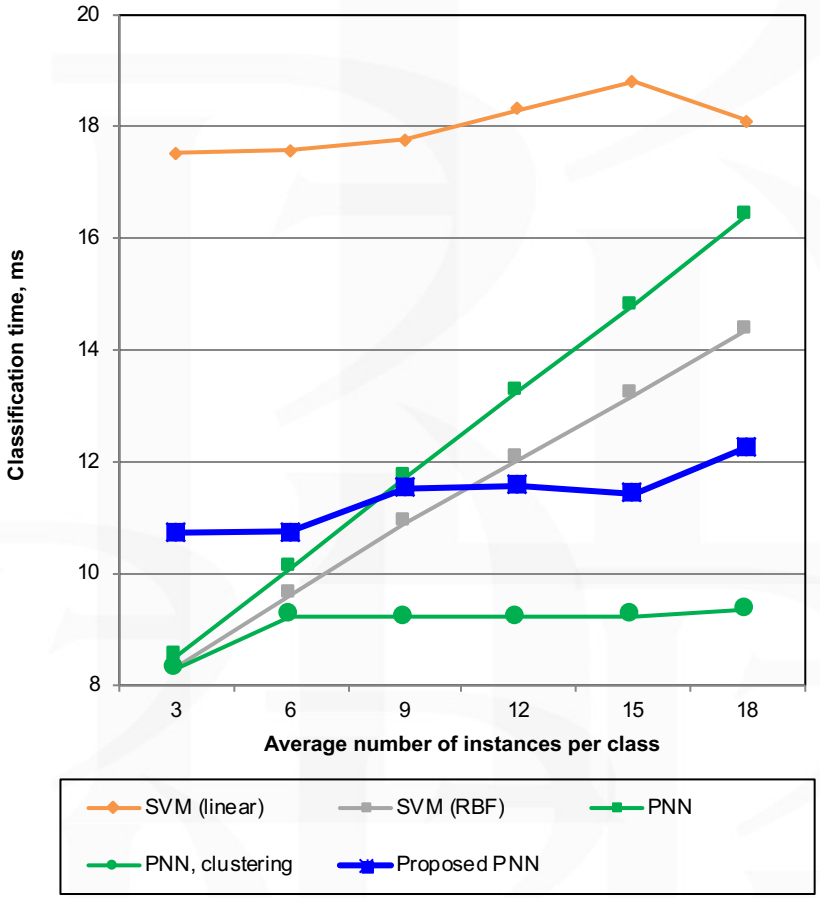
Online classification is approximately $R^{2/3}$–times faster than instance-based learning (PNN, k-NN) if at least 5 photos per subject are available

Experimental results (Caltech-256 dataset)

Accuracy (%)

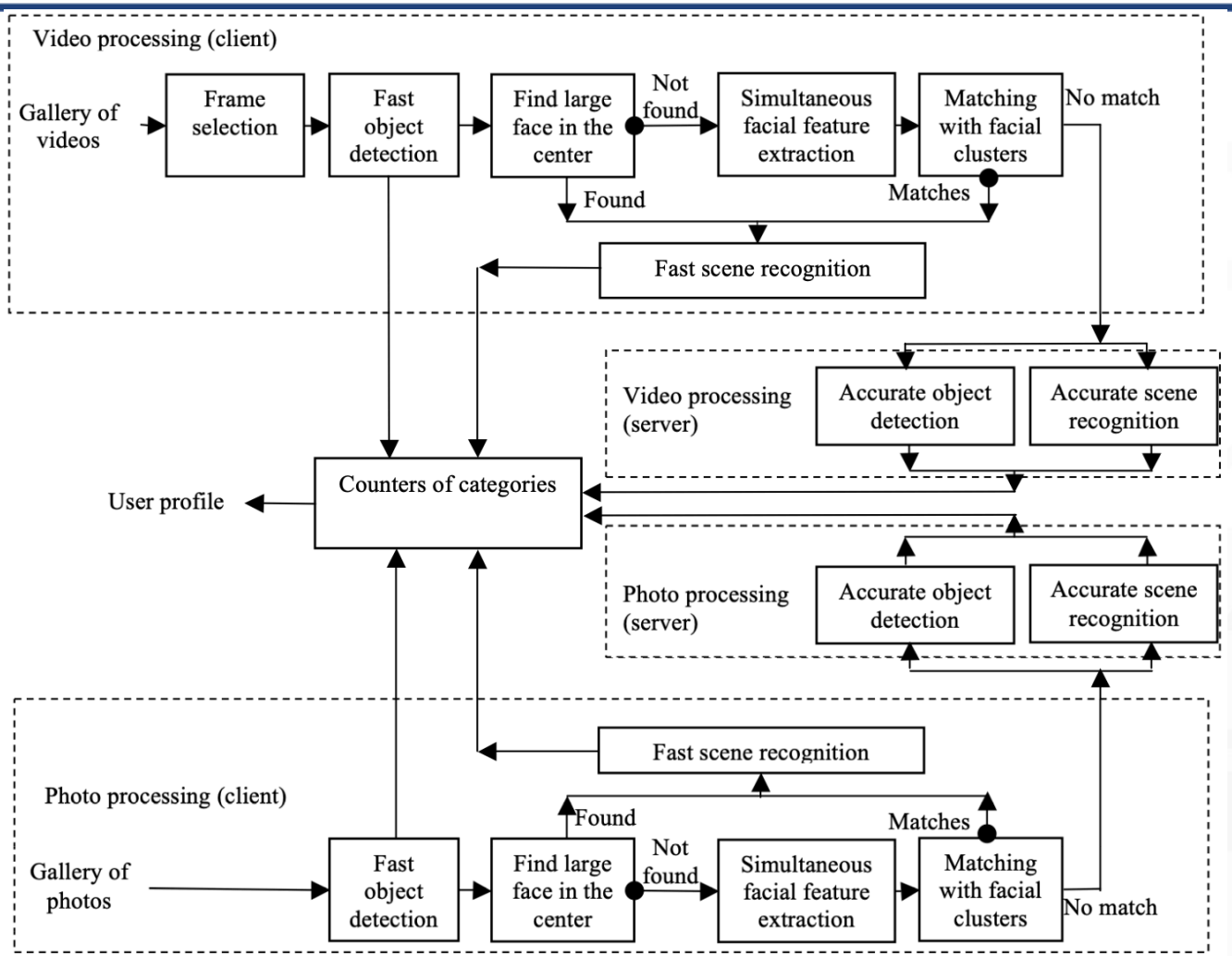


Classification time (ms.)



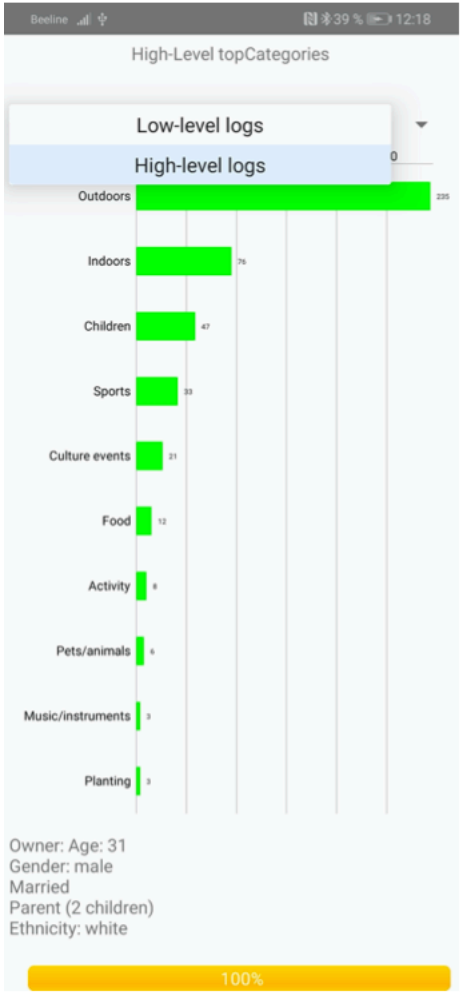
Organizing photo and video albums on mobile device

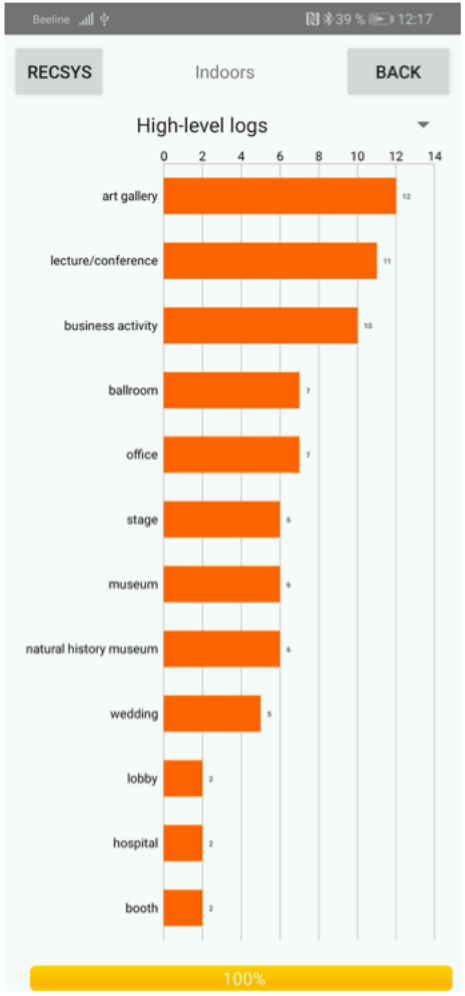
Organizing Photo and Video Albums



Savchenko A.V. et al., https://arxiv.org/abs/1907.04519, 2019

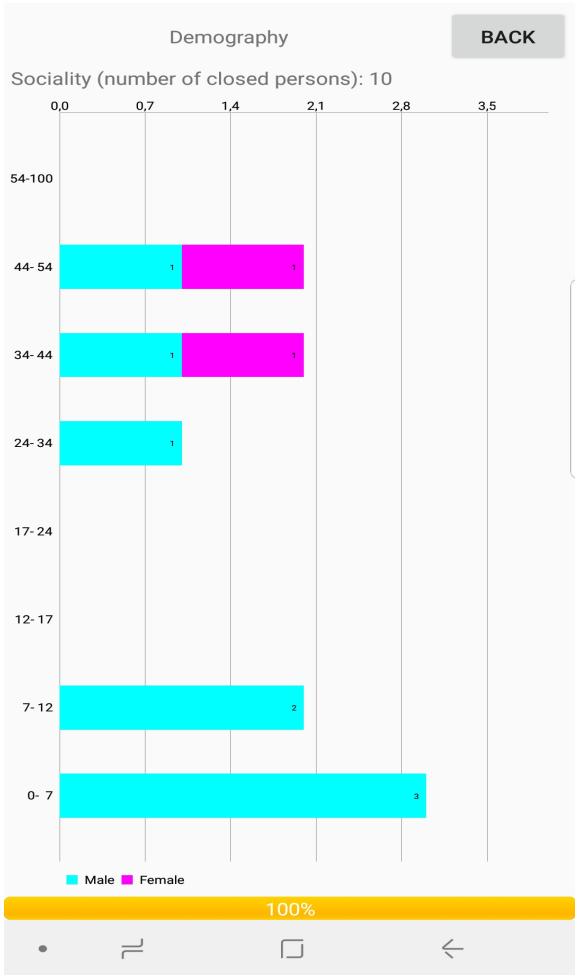
Example (1)

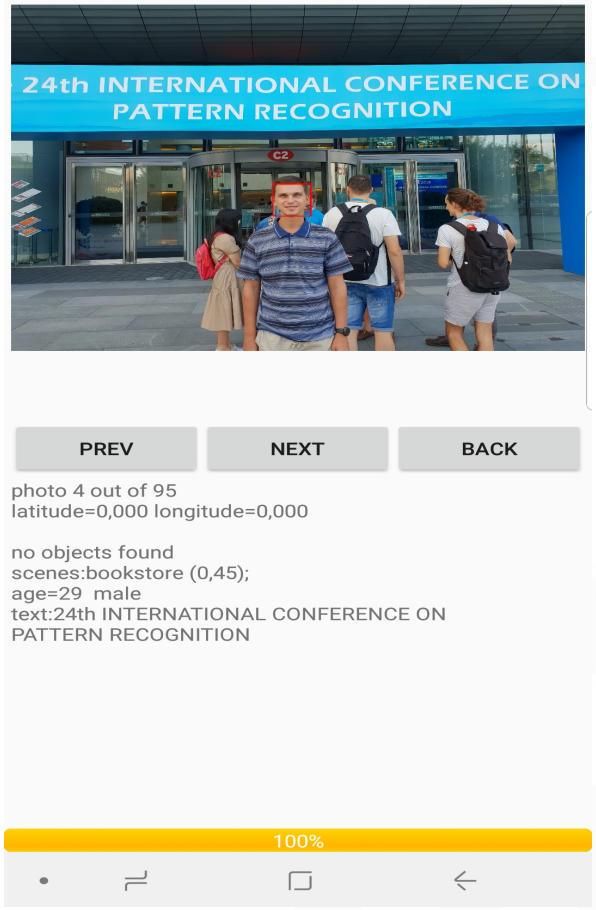

Beeline
39 %
12:18
High-Level topCategories
Low-level logs
High-level logs
Outdoors
Indoors
Children
Sports
Culture events
Food
Activity
Pets/animals
Music/instruments
Planting
Owner: Age: 31
Gender: male
Married
Parent (2 children)
Ethnicity: white
100%

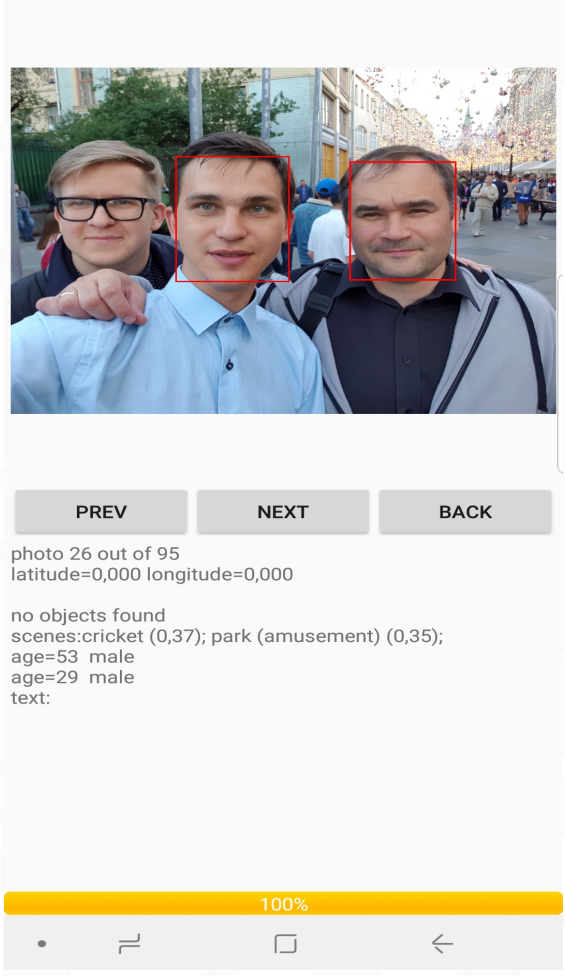

Beeline
39 %
12:17
RECSYS
Indoors
BACK
High-level logs
0 2 4 6 8 10 12 14
art gallery
lecture/conference
business activity
ballroom
office
stage
museum
natural history museum
wedding
lobby
hospital
booth
100%


Beeline
39 %
12:15
Sep 8, 2018
Sep 25, 2019
Sep 11, 2018
Jun 27, 2018
Aug 31, 2019
PREV
NEXT
BACK
photo 1 out of 2
Public photo
latitude=0,000 longitude=0,000
no objects found
scenes: football (0,98); events: sports (0,17);
No faces found
text:OAMMn JAKO
Instat
OOAMH OL
BAdSIRRU
1AAIH
OLMP.BET
100%

Example (2)

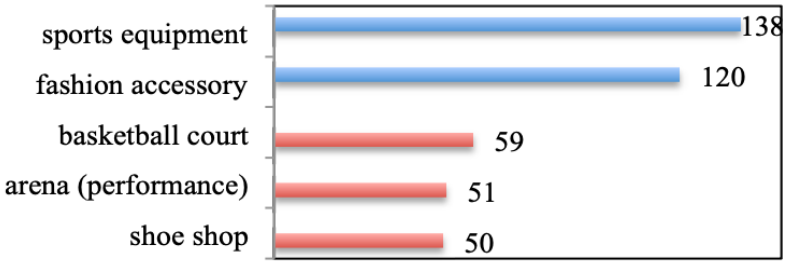

Demography
BACK
Sociality (number of closed persons): 10
0,0
0,7
1,4
2,1
2,8
3,5
54-100
44- 54
34- 44
24- 34
17- 24
12- 17
7- 12
0- 7
Male
Female
100%


24th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
PREV
NEXT
BACK
photo 4 out of 95
latitude=0,000 longitude=0,000
no objects found
scenes:bookstore (0,45);
age=29 male
text:24th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
100%

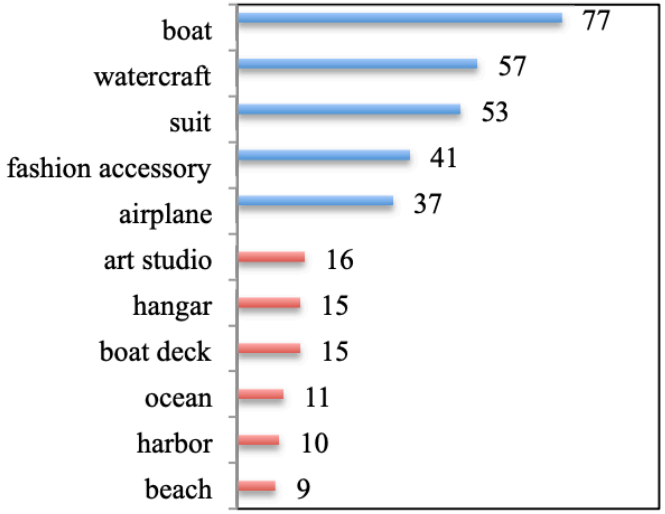

PREV
NEXT
BACK
photo 26 out of 95
latitude=0,000 longitude=0,000
no objects found
scenes:cricket (0,37); park (amusement) (0,35);
age=53 male
age=29 male
text:
100%

Example (3). Profiles from Instagram accounts

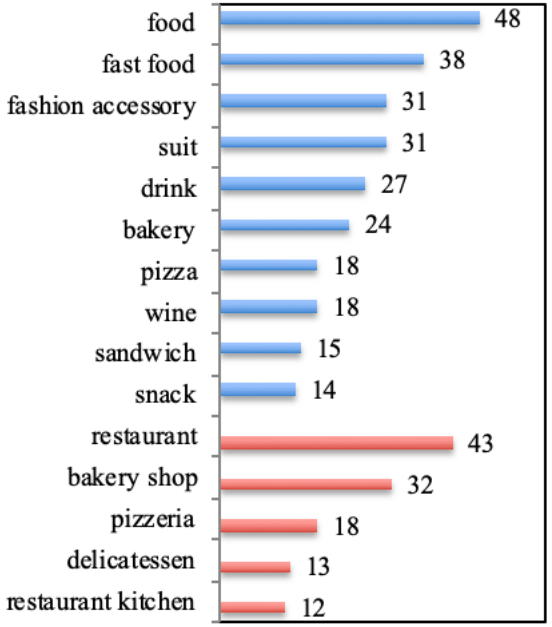
LeBron James



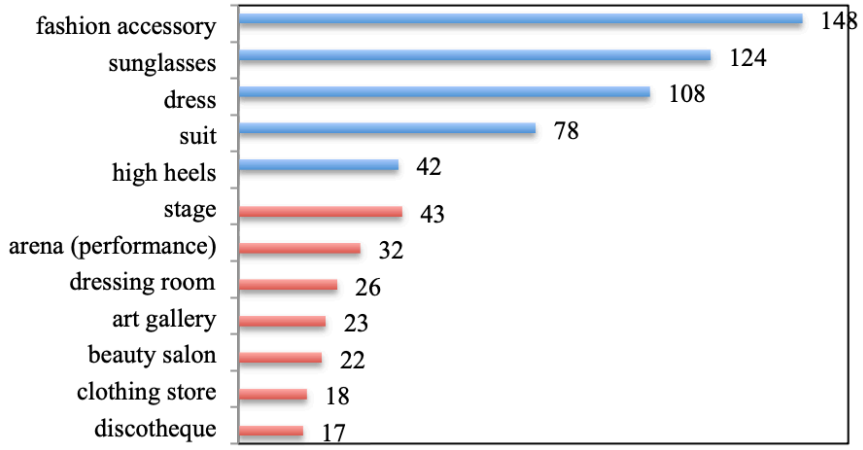
Fedor Konyukhov



Gordon Ramsay



Beyonce



Conclusion

1. Event recognition in a gallery is a practical task that still needs further study: accuracy is much higher if the albums are known
2. Sequential analysis of CNN features/layers can potentially provide high performance without losses in accuracy
3. PNN with complex exponential activations proves the possibility to create fast and accurate classifier (when compared to k-NN and PNN)
4. User modeling based on visual data from mobile phones can be used to deal with cold start problem in recommender systems

annual International Conference **AIST (Analysis of Images, Social networks and Texts)**

- Main proceedings – Springer LNCS (Lecture Notes in Computer Science);
- Companion volume – Springer CCIS (Communications in Computer and Information Science)

http://aistconf.org/

NATIONAL RESEARCH
UNIVERSITY

Thank you!