NATIONAL RESEARCH UNIVERSITY

NRU Higher School of Economics

PHOTO PRIVACY DETECTION BASED ON TEXT CLASSIFICATION AND FACE CLUSTERING

L.Kopeykina, A.V.Savchenko

VI International Conference of Information Technology and Nanotechnology
ITNT-2020

OUTLINE

- Photo Privacy Detection Problem
- Proposed Approach
- Experimental Results
- Conclusions

PRIVACY DETECTION PROBLEM

EXISTING METHODS & LIMITATIONS

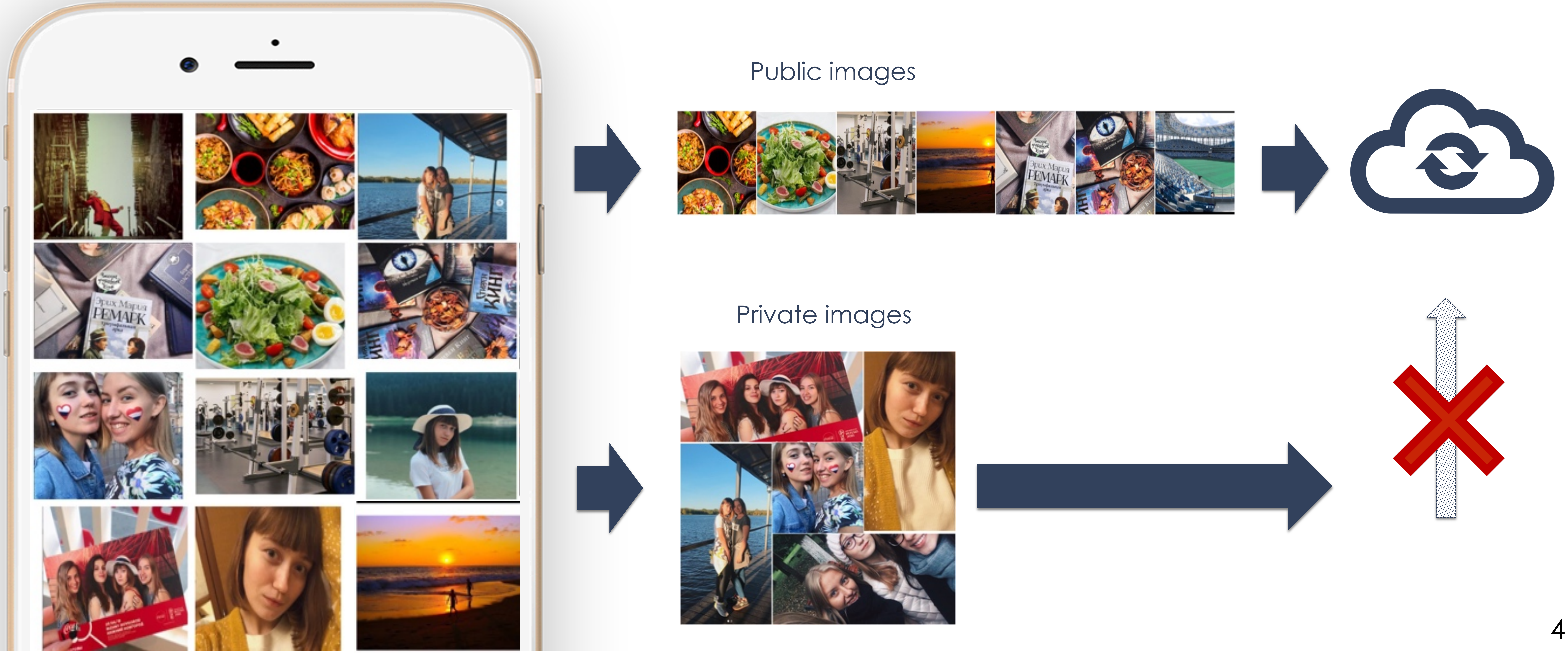
The decision on a particular photo can be made based on its visual appearance.
For example, when a text is detected in an image, the neural network will attribute it to personal data, while this text may not contain any private information at all.
A similar situation is observed with the detection of faces in images: not every image with a face detected on it is confidential.

PROPOSED APPROACH

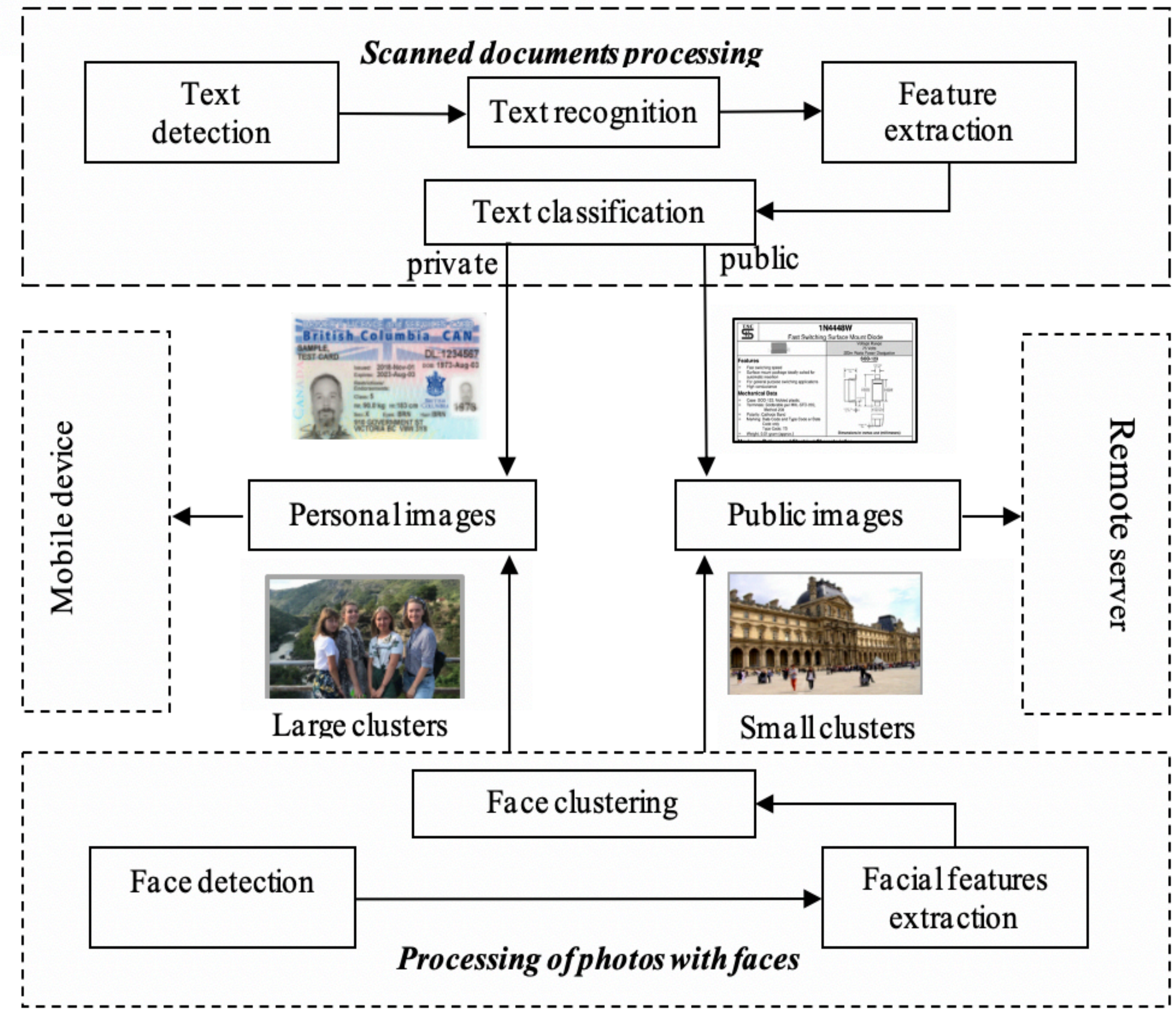
This work proposes a unified approach for personal data detection in photo gallery using well-known methods of face classification and text recognition.

PRIVACY DETECTION PROBLEM

OUR TASK IS:



PROPOSED APPROACH



Step 1:

Detection of scanned documents with EAST text detector, the Tesseract OCR library and the neural network classification of recognized text on images

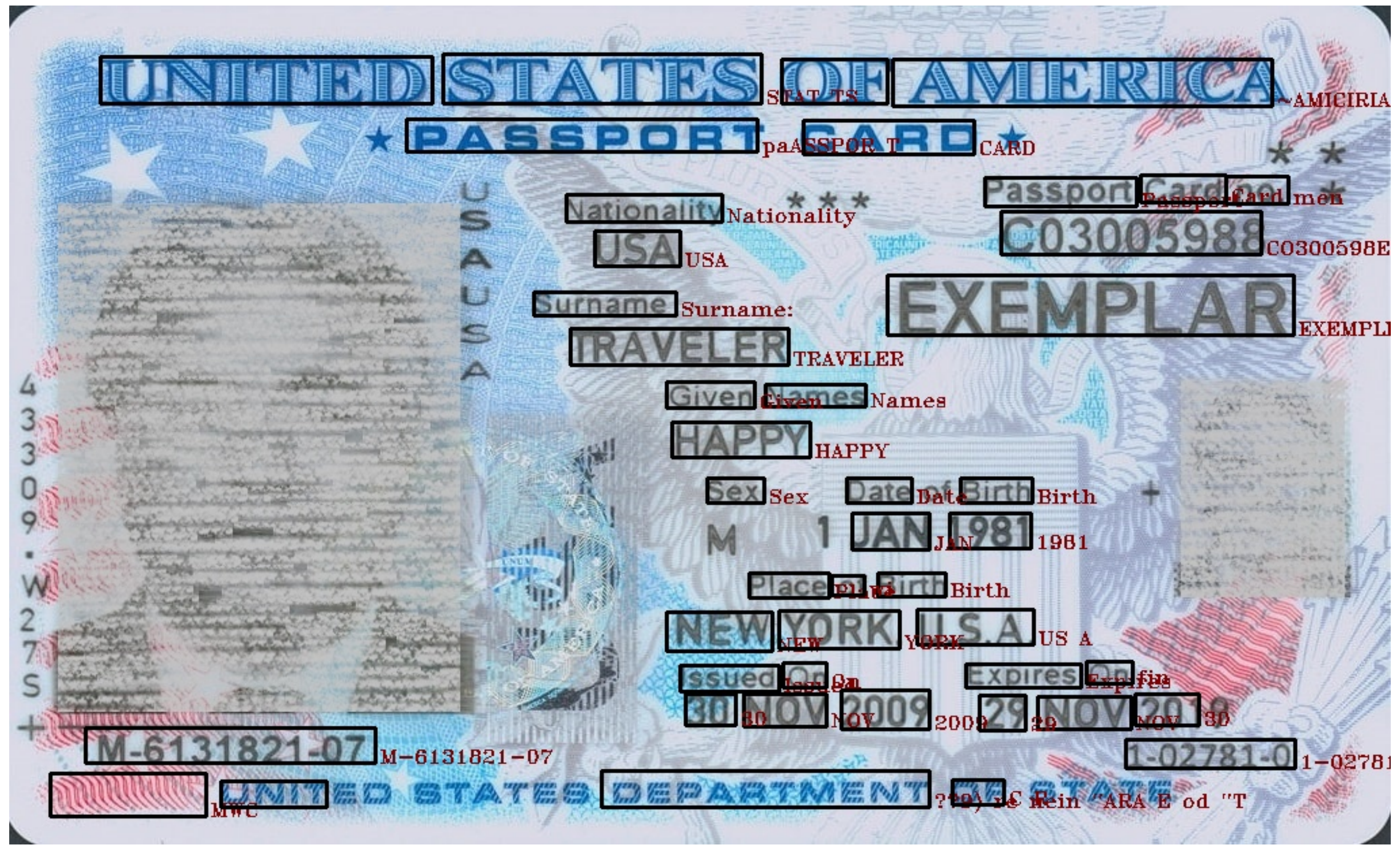
Step 2:

Detection of user’s personal photos based on the well-known methods of face clustering applied to face embeddings

PROPOSED APPROACH

DETECTION OF SCANNED DOCUMENTS

Input image



Extracted text

$$[t_1, t_2, \ldots, t_T]$$

Vocabulary

One-hot encoded vector

$$[0, 1, 0, 1, \ldots, 0]$$

Neural network classifier

$private$

- For each $i \in N$ image $T \geq 0$ text areas are detected using the EAST algorithm.
- Text from each of $t \in T$ detected areas is extracted with Tesseract OCR in *image_to_string* mode
- To classify personal data in the extracted text, a neural network, which is trained based on the input sequence of words recognized in the training set of scanned documents, is used.
- Each text is represented as a V-dimensional binary vector, where the $v-th$ component of the vector is *1* only if the $v-th$ word from the dictionary is presented in the input text

PROPOSED APPROACH

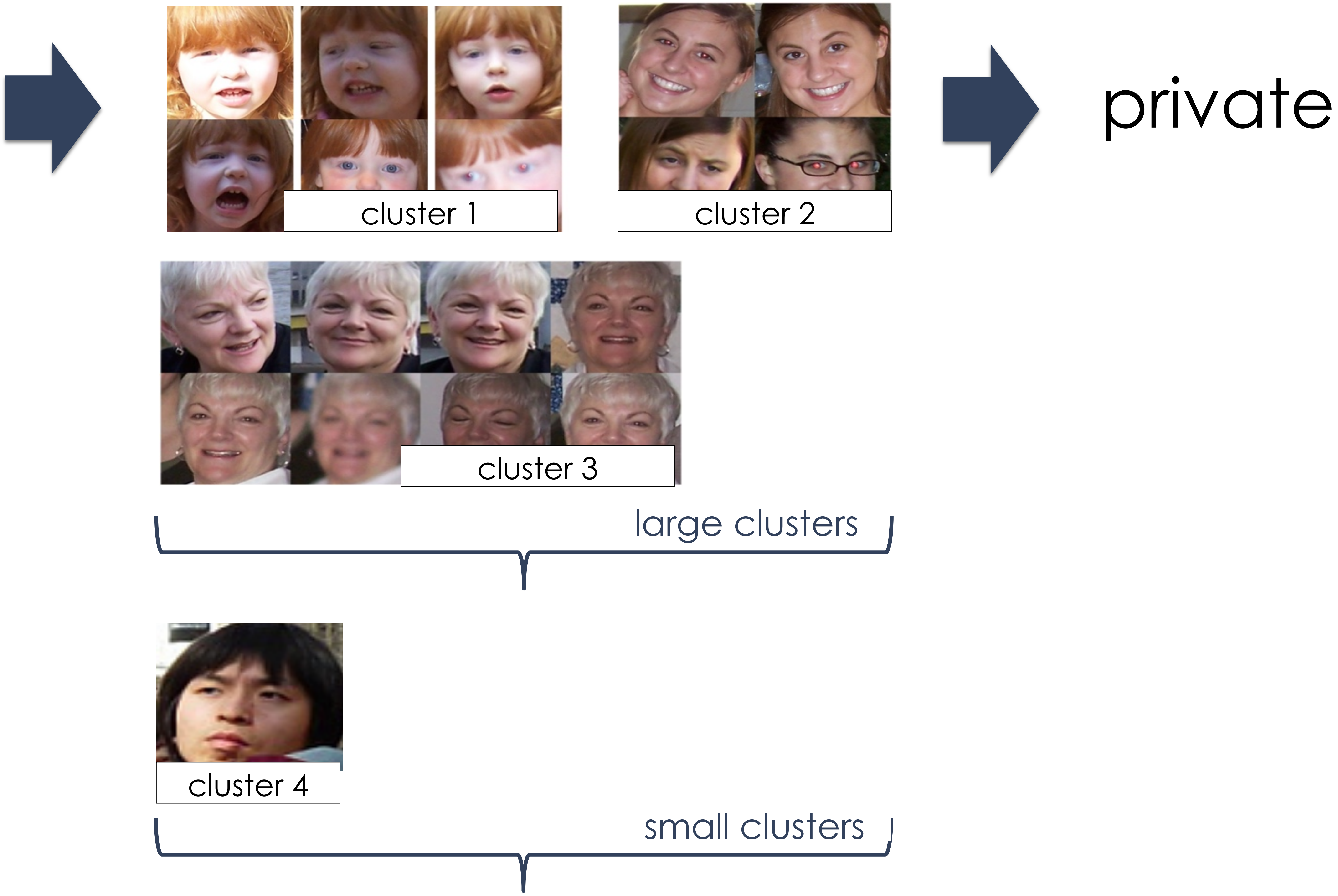
DETECTION OF PERSONAL PHOTOS BASED ON FACE CLUSTERING

- Facial regions are detected in all photographs using MTCNN.
- D-dimensional feature vectors are extracted for each of N > 0 selected facial images by using a CNN
- Each *i-th* facial image *(i = 1, ..., N)* is assigned to one of C ≥ *1* group, where C is usually unknown.
- An image is considered to be private if it contains faces from sufficiently large clusters. In other words, a person presents at least $Kmin$ times on different types of photos, where $Kmin$ is a hyper-parameter of our method

PROPOSED APPROACH

DETECTION OF PERSONAL PHOTOS BASED ON FACE CLUSTERING

Input image



cluster 1

cluster 2

cluster 3

large clusters

cluster 4

small clusters

private

EXPERIMENTAL RESULTS

RESULS FOR CLASSIFICATION OF SCANNED DOCUMENTS

	Model	Precision	Recall	F-score	Error rate
Tesseract	Keyword spotting	0.83	0.62	0.70	0.276
	LSTM	0.97	0.93	0.94	0.043
	CNN	0.88	0.77	0.82	0.161
	Fully-connected	0.98	0.94	0.95	0.028
Proposed (EAST+ Tesseract)	Keyword spotting	0.90	0.75	0.81	0.161
	LSTM	0.93	0.99	0.95	0.038
	CNN	0.89	0.79	0.83	0.144
	Fully-connected	**1.00**	**0.97**	**0.98**	**0.015**

Private class:

350 images of driving license and medical insurance cards, passports and invoices from extension of the MIDV dataset

Public class:

350 photos from publicly available datasets for text classification tasks DIQA and Ghega

EXPERIMENTAL RESULTS

FACE CLUSTERING: DATASETS

Gallagher collection person dataset , which contains 589 images with 931 labeled faces of 32 various people

Subset of labeled faces in the wild (LFW) dataset, which includes photos of those subjects, who has at least 2 images in the original LFW dataset and at least 1 video in the YouTube Faces (YTF) collection.

EXPERIMENTAL RESULTS

FACE CLUSTERING: METHODS AND FEATURE EXTRACTORS

To extract facial features, CNN models were considered:

- VGGFace (VGGNet-16) - 4096-D vectors;
- VGGFace2 (ResNet-50) - 2048-D vectors;
- MobileNet -1024-D vectors;
- InsightFace (ArcFace) - 512-D vectors;
- FaceNet (Inception ResNet v1) - 512-D vectors.

Hierarchical agglomerative clustering with the following types of linkage: single linkage, average linkage, complete linkage, weighted linkage, centroid linkage and median linkage

Rank-order clustering

Approximate rank-order clustering

Graph convolutional neural network

METRICS: *the Rand index (ARI), mutual information index (AMI), homogeneity and completeness, the average number K of selected clusters to the number of groups C and the b-cubed F-measure*

EXPERIMENTAL RESULTS

FACE CLUSTERING: RESULTS FOR GALLAGHER

	CNN	Time, sec	K/C	ARI	AMI	Homogeneity	Completeness	F-score
Rank-order	VGGFace2	32.17	1.25	0.480	0.627	0.794	0.635	0.706
	VGGFace	21.72	1.50	0.439	0.569	0.764	0.585	0.671
	MobileNet	22.71	2.09	0.674	0.678	0.965	0.611	0.725
	InsightFace	27.84	1.59	0.502	0.530	0.729	0.716	0.625
	FaceNet	24.54	1.53	0.674	0.681	0.906	0.633	0.760
Weighted linkage	**VGGFace2**	**0.033**	**1.50**	**0.891**	**0.898**	**0.946**	**0.876**	**0.921**
	VGGFace	0.019	1.03	0.599	0.737	0.704	0.830	0.762
	MobileNet	0.018	0.75	0.751	0.788	0.792	0.818	0.806
	InsightFace	0.018	1.72	0.655	0.697	0.806	0.675	0.734
	FaceNet	0.015	1.47	0.884	0.881	0.934	0.857	0.902
Approximate rank-order	VGGFace2	0.785	3.91	0.515	0.535	0.586	0.641	0.704
	VGGFace	1.312	3.78	0.446	0.485	0.509	0.681	0.653
	MobileNet	1.414	6.68	0.417	0.516	0.522	0.795	0.635
	InsightFace	1.220	5.78	0.324	0.324	0.471	0.656	0.571
	FaceNet	1.092	4.05	0.567	0.621	0.626	0.764	0.724
GCN-D	VGGFace2	5.006	1.67	0.867	0.845	0.954	0.793	0.859
	VGGFace	4.741	0.78	0.641	0.536	0.627	0.539	0.578
	MobileNet	6.290	0.69	0.675	0.748	0.799	0.742	0.728
	InsightFace	6.862	0.65	0.409	0.612	0.603	0.682	0.637
	FaceNet	6.164	0.91	0.636	0.726	0.751	0.749	0.687

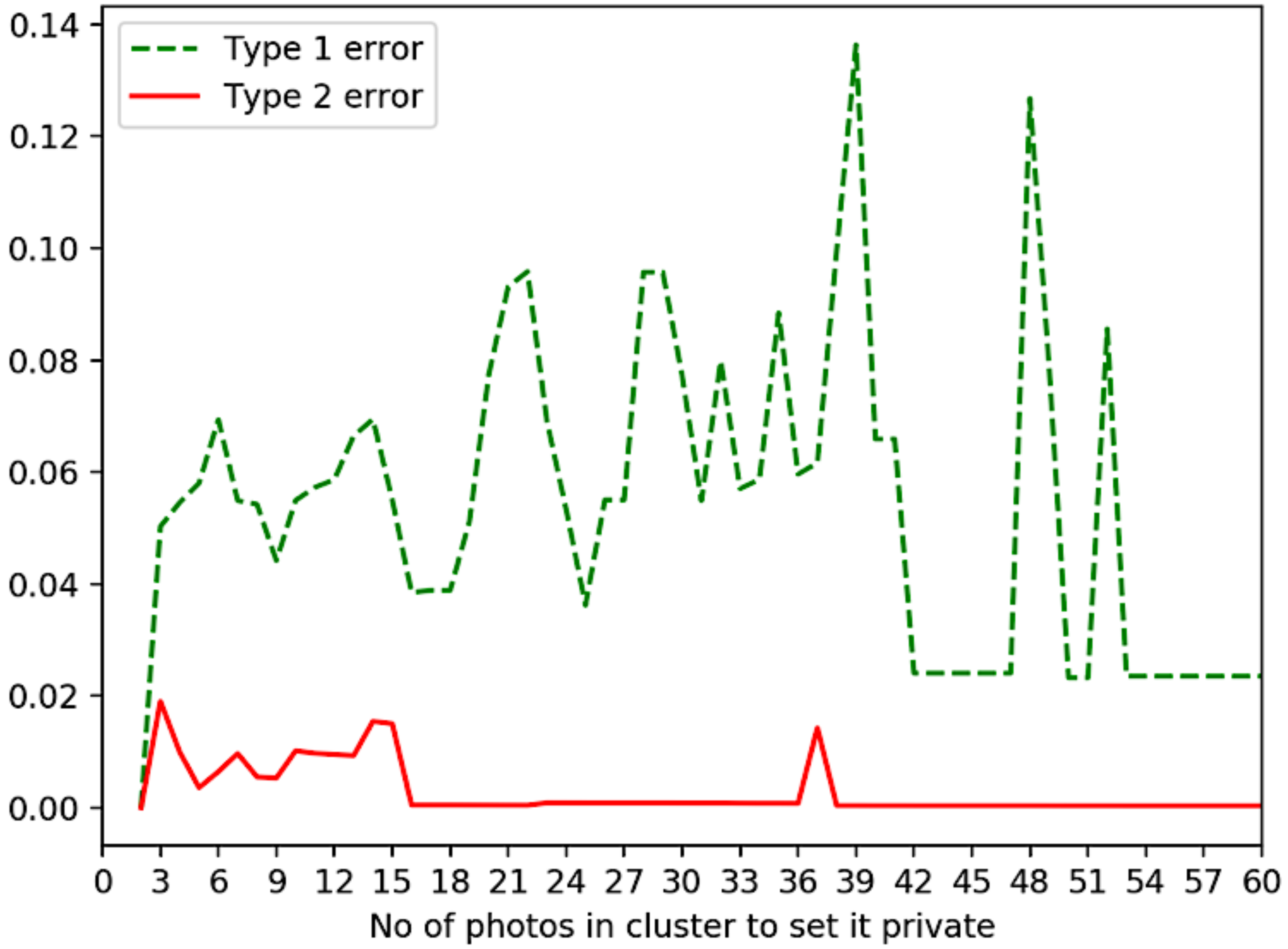
EXPERIMENTAL RESULTS

FACE CLUSTERING: RESULTS FOR LFW

	CNN	Time, sec	K/C	ARI	AMI	Homogeneity	Completeness	F-score
Rank-order	VGGFace2	416.73	0.96	0.719	0.781	0.980	0.911	0.862
	VGGFace	309.44	0.82	0.675	0.748	0.812	0.762	0.746
	MobileNet	305.03	0.77	0.786	0.816	0.944	0.907	0.806
	InsightFace	361.02	1.21	0.673	0.721	0.842	0.912	0.683
	FaceNet	359.62	0.91	0.784	0.832	0.924	0.917	0.812
Weighted linkage	**VGGFace2**	**0.63**	**1.37**	**0.893**	**0.941**	**0.998**	**0.952**	**0.923**
	VGGFace	0.61	1.28	0.925	0.925	0.984	0.950	0.901
	MobileNet	0.59	1.44	0.961	0.940	0.996	0.952	0.919
	InsightFace	0.67	1.42	0.879	0.864	0.972	0.913	0.820
	FaceNet	0.64	1.44	0.935	0.938	0.997	0.950	**0.919**
Approximate rank-order	VGGFace2	9.49	1.42	0.803	0.877	0.924	0.952	0.923
	VGGFace	7.12	1.30	0.621	0.706	0.893	0.816	0.724
	MobileNet	7.06	1.79	0.610	0.741	0.864	0.912	0.740
	InsightFace	12.32	1.57	0.684	0.711	0.849	0.908	0.685
	FaceNet	12.72	1.13	0.782	0.859	0.932	0.937	0.844
GCN-D	VGGFace2	30.33	0.84	0.075	0.395	0.814	0.711	0.512
	VGGFace	28.47	0.69	0.044	0.235	0.866	0.669	0.456
	MobileNet	31.23	0.86	0.332	0.665	0.882	0.825	0.639
	InsightFace	30.18	0.74	0.802	0.732	0.874	0.875	0.666
	FaceNet	31.79	0.92	0.141	0.543	0.828	0.770	0.588

EXPERIMENTAL RESULTS

CLASSIFICATION RESULTS FOR LFW



The dependence between the minimal number Kmin of photos in a personal cluster and type1/type 2 error rates, LFW dataset.

“0” class consists of 3263 private images, whereas public class “1” includes 474. Images from LFW containing faces from clusters that include Kmin=3 or more facial images, were considered personal

Feature extractor	FPR	FNR	Precision	Recall	F1-score	Error rate
VGGFace2	0.051	0.019	0.738	0.978	0.842	0.047
VGGFace	0.055	0.276	0.655	0.723	0.688	0.084
MobileNet	0.054	0.168	0.687	0.831	0.752	0.069
InsightFace	0.115	0.281	0.474	0.719	0.571	0.137
FaceNet	0.056	0.044	0.712	0.952	0.816	0.055

CONCLUSIONS

Novel approach of privacy detection on images was proposed:

- It is proposed to use the EAST text detector and recognize text in the detected areas with Tesseract OCR library to classify scanned documents.
- It has been experimentally shown that a simple fully-connected neural network for text encoded using bag-of-words exceeds more complex network architectures, such as CNN, by more than 10% and achieves high accuracy in detecting personal documents.
- It is proposed to apply face clustering techniques to identify photos of the user himself, his friends and relatives.
- Agglomerative clustering with a weighted linkage performed higher results in extracting groups of user’s faces, friends and relatives

HIGHER SCHOOL OF ECONOMICS

NATIONAL RESEARCH
UNIVERSITY

THANK YOU FOR YOUR ATTENTION

Lyudmila Kopeykina
lnkopeykina@mail.ru