Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images

Tribhuvanesh Orekondy, Mario Fritz, Bernt Schiele

Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany

max planck institut informatik

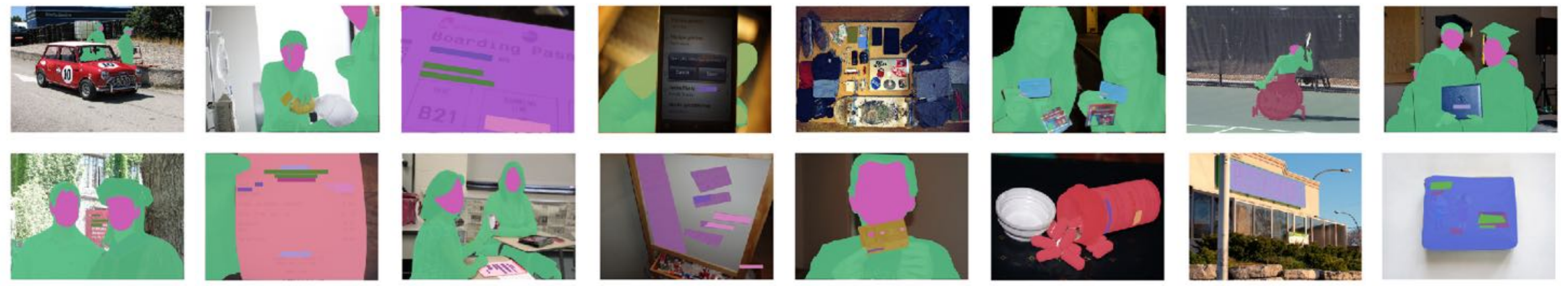
UNIVERSITÄT DES SAARLANDES

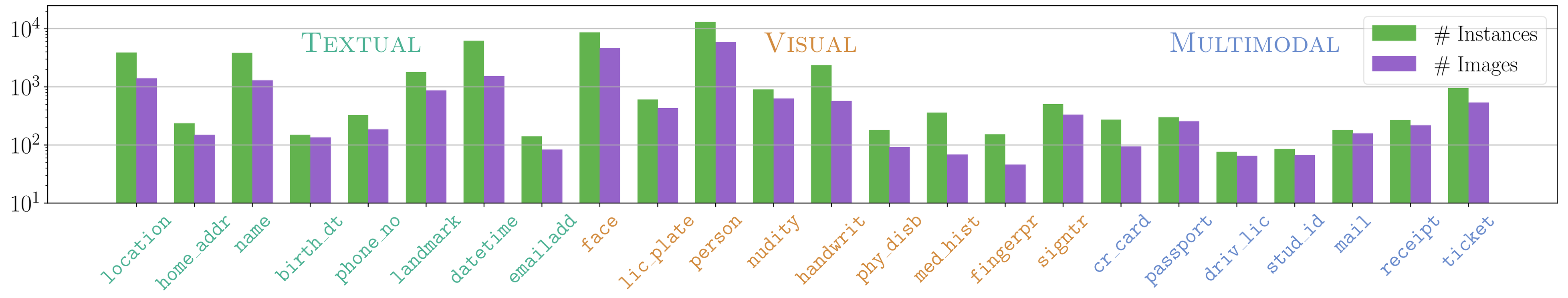
CVPR 2018 SALT LAKE CITY • JUNE 18-22

Motivation

- Numerous personal photos containing a broad range of private information are shared on the Internet everyday
- Previous works: Image classification or redact one/narrow range of privacy classes
- Ours: How can we sanitise a wide spectrum of private content in images?

The Visual Redactions Dataset

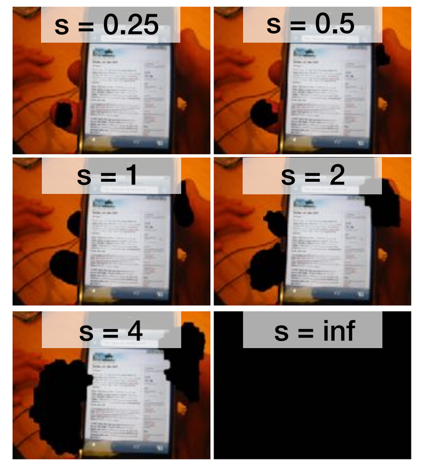


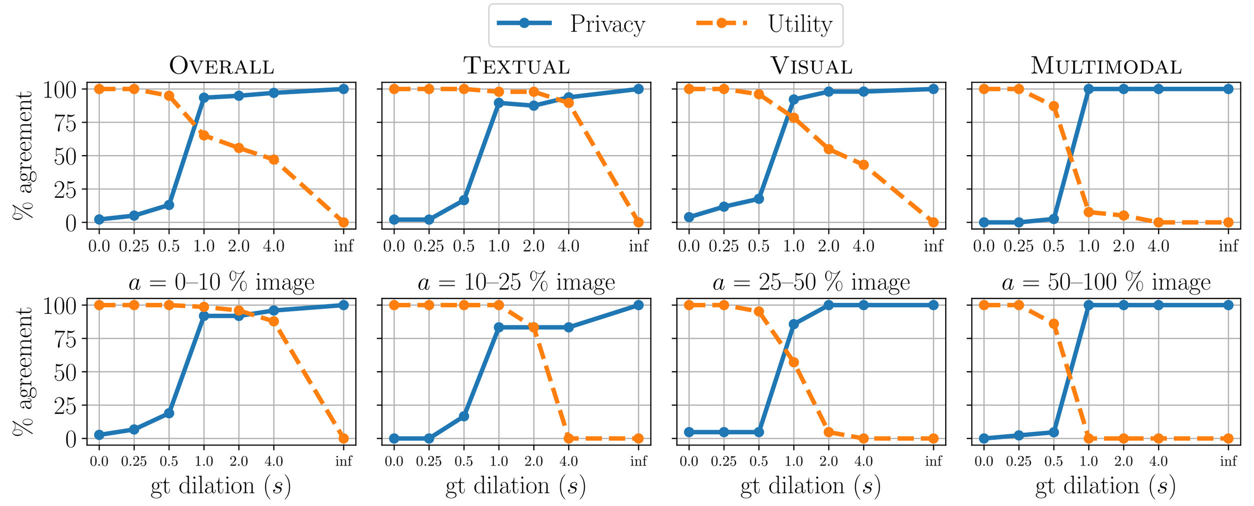


- 8.4k images, 47.6k high-quality instances, 24 privacy attributes, 3 modalities
- Helpful for other tasks too: 9k face, 13k person instances
- Other goodies: Text detections, OCR, etc. using Google Cloud Vision API
- Dataset and Code: resources.mpi-inf.mpg.de/d2/orekondy/redactions

Influence of Redacted Pixels on Privacy and Utility

- User study on AMT over various dilations (s) of GT redactions: 24 privacy attributes x 6 images x 7 scales x 5 yes/no responses
- Privacy Question: "Is X visible in the image?" (e.g. X: fingerprint)
- Utility Question: "Is the image intelligible, so that it can be shared on social networking websites?"
- Measuring privacy/utility of a redacted image: Majority agreement (y-axis)



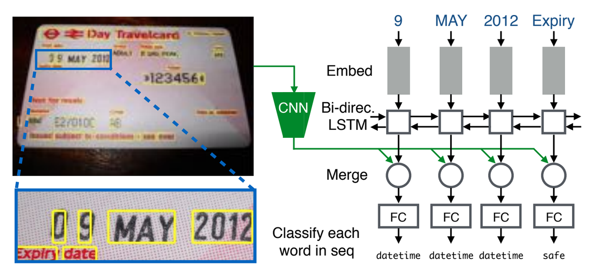


- Privacy is a step-like function
- Utility gradually decreases
- Different operating points for different modalities/attributes
- GT segmentation = great proxy

Segmentation of Private Regions

Textual

SEQ



NER

Named Entity Recognition on image word sequence using Stanford NER CRF [2] classifier

RULES

Hand-crafted text rules (e.g., look-ups, pattern matching) to recognise privacy attributes

Visual

FCIS [3]

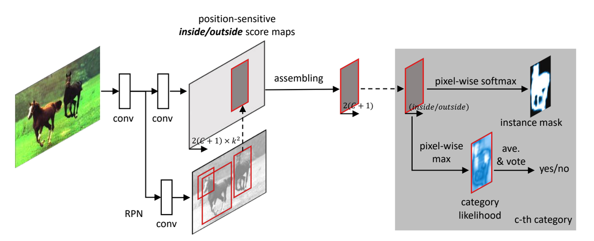


Image source: [3]

WSL:I

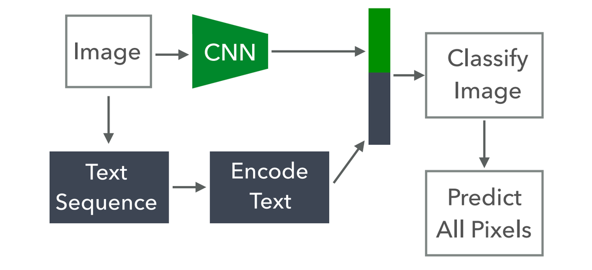
Weakly supervised pixel-labelling using ResNet-50 [4]. Predicts all pixels of image.

PTM

Popular pretained models to predict well-studied attributes: face, person, license plates

Multimodal

WSL:I+T



SAL

Using WSL:I+T as a base, approximate attribute's location with salient region

IR

Refine convex hull of text regions using DenseCRF [5] to "spill into" text-heavy document regions

NN

Predict mask of nearest-neighbour from train+val set

References

[1] *Towards a visual privacy advisor: Understanding and predicting privacy risks in images*, Orekondy et al., ICCV '17

[2] *Incorporating non-local information into information extraction systems by gibbs sampling*, Finkel et. al, ACL '05

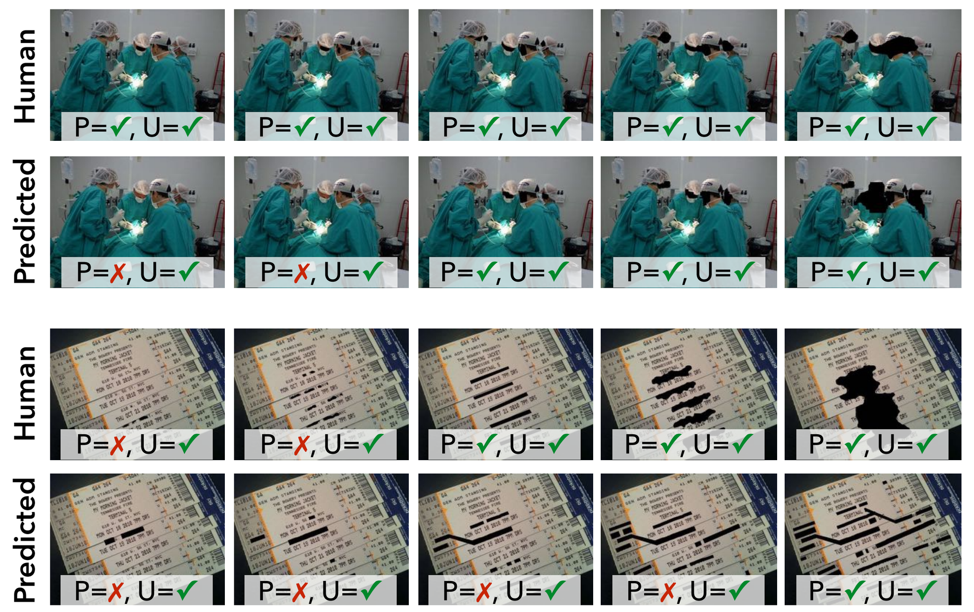
[3] *Fully Convolutional Instance-aware Semantic Segmentation*, Li et. al, CVPR '17

[4] *Deep residual learning for image recognition*, He et al., CVPR '16

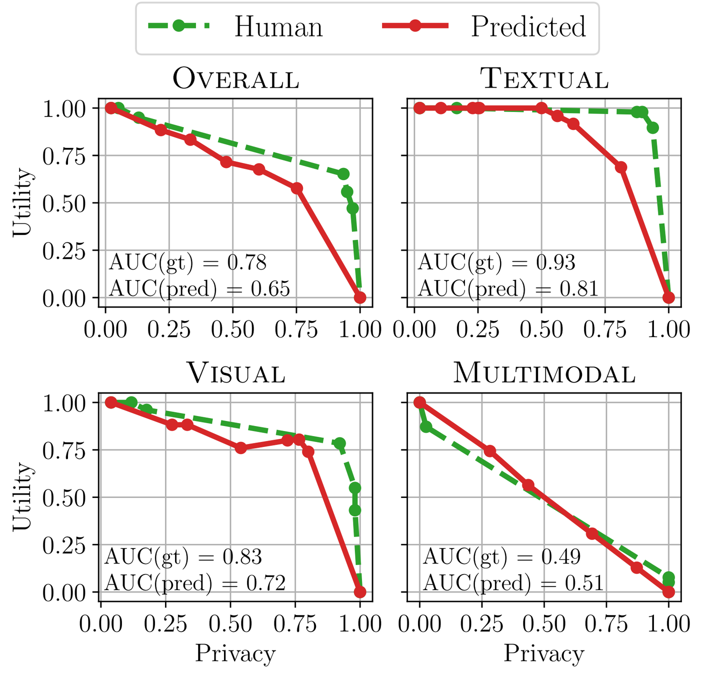
[5] *Efficient inference in fully connected crfs with gaussian edge potentials*, Krähenbühl et al., NIPS '11

Acknowledgment This research was partially supported by the German Research Foundation (DFG CRC 1223)

Privacy vs. Utility Trade-off



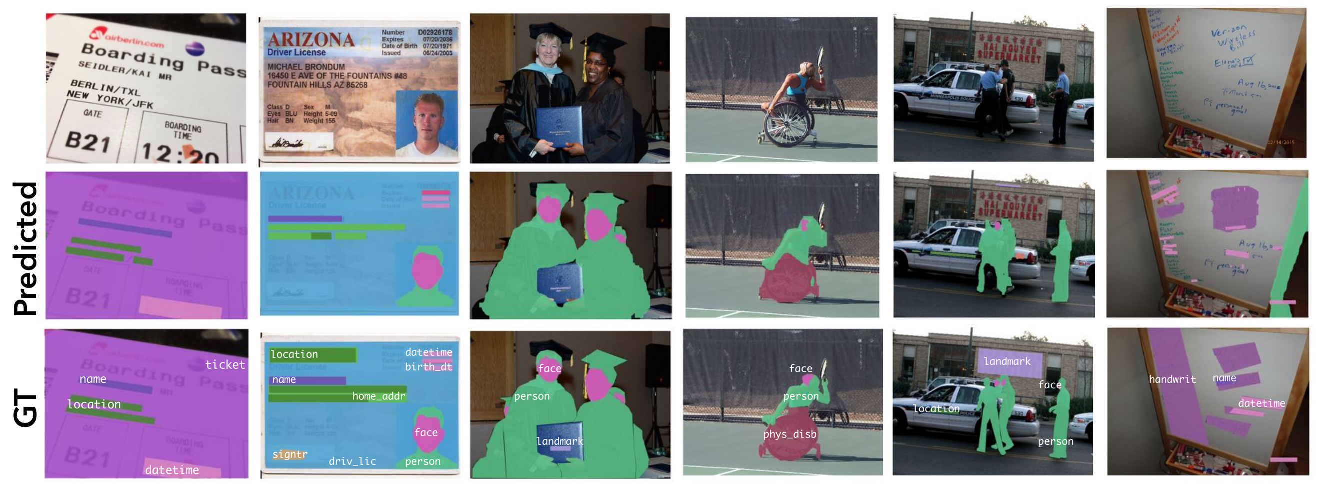
- Segmentation of privacy attributes across modalities is performed as an intermediate step
- Unlike segmentation which requires pixel-perfect prediction, redaction allows for leeway
- Metric: Area under Privacy-Utility curve (AUC)
- User-study to evaluate redactions. We achieve 83% performance of human-based redactions!
- Can predict more pixels "for free" e.g. Textual attributes (low 26.8 mAP for segmentation, but high 81% privacy-utility AUC)



Take-home messages

- Task: Visual redaction across broad range of private content
- Large pixel-annotated dataset for task
- Privacy vs. Utility trade-off in redactions
- Methods to pixel-label private content across multiple modalities
- We approach human-based performance for redactions

Segmentation Evaluation



TEXTUAL

Method	mAP	location	home addr	name	birth dt	phone no	land mark	date time	email add
PROXY	45.0	31.7	37.8	48.7	52.5	52.6	33.6	52.4	50.8
NN	0.9	0.3	1.9	0.4	*0.7*	0.0	3.1	0.6	0.0
NER	3.0	*6.0*	1.7	4.4	0.5	0.0	0.5	10.9	0.0
RULES	4.2	3.1	0.5	2.8	0.6	1.4	1.2	6.4	*17.5*
FCIS	*7.2*	4.3	0.2	*9.8*	0.1	*2.5*	**27.6**	*12.9*	0.0
SEQ	**26.8**	**18.4**	**19.4**	**19.1**	**25.1**	**45.8**	*13.9*	**33.4**	**38.9**

VISUAL

Method	mAP	face	licp late	per son	nud ity	hand writ	phy disb	med hist	fing erpr	sig ntr
NN	16.6	9.0	16.0	33.6	6.2	37.5	11.4	18.9	16.9	0.1
WSL:I	*20.8*	5.0	4.3	30.3	*16.4*	*49.9*	*13.7*	*37.7*	*28.8*	*1.3*
PTM	20.0	*47.6*	*44.5*	**88.3**	0.0	0.0	0.0	0.0	0.0	0.0
FCIS	**68.3**	**83.8**	**77.9**	*87.0*	**69.7**	**80.7**	**59.0**	**45.8**	**68.1**	**42.6**

MULTIMODAL

Method	mAP	cr card	pass port	driv lic	stud id	mail	rece ipt	tic ket
NN	24.1	10.5	49.5	19.9	14.5	20.6	17.1	36.7
WSL:I+T	*55.6*	27.7	*68.8*	**83.3**	**56.1**	**41.4**	54.2	58.0
SAL	36.2	**55.9**	37.2	23.8	30.4	8.1	42.5	55.1
IR	53.6	41.7	51.2	67.8	48.1	36.9	*57.2*	*72.5*
FCIS	**59.2**	*53.2*	**76.3**	66.5	*50.3*	33.1	**59.4**	**75.4**

- Metric: Mean Average Precision (à la Pascal VOC)
- Textual: (+) Patterns in text help (-) Bottlenecked by challenging text detections/OCR
- Visual: (+) FCIS is highly effective across many visual attributes
- Multimodal: (+) Text-understanding helps disambiguation (-) Large object bias
- Redactions performed using ENSEMBLE (SEQ, FCIS, WSL:I+T) at calibrated thresholds