Semi-supervised Learning with Constraints for Person Identification in Multimedia Data

Martin Bäuml, Makarand Tapaswi, Rainer Stiefelhagen

Computer Vision for Human Computer Interaction, Karlsruhe Institute of Technology, Germany

Contributions

- A unified multi-class learning framework taking into account (weakly-)supervised and unsupervised data and constraints
- Its application to the task of character naming in Multimedia Data, achieving state-of-the-art results
- An extensive data set of more than 9200 face tracks from a total of 12 episodes over two TV series

Motivation

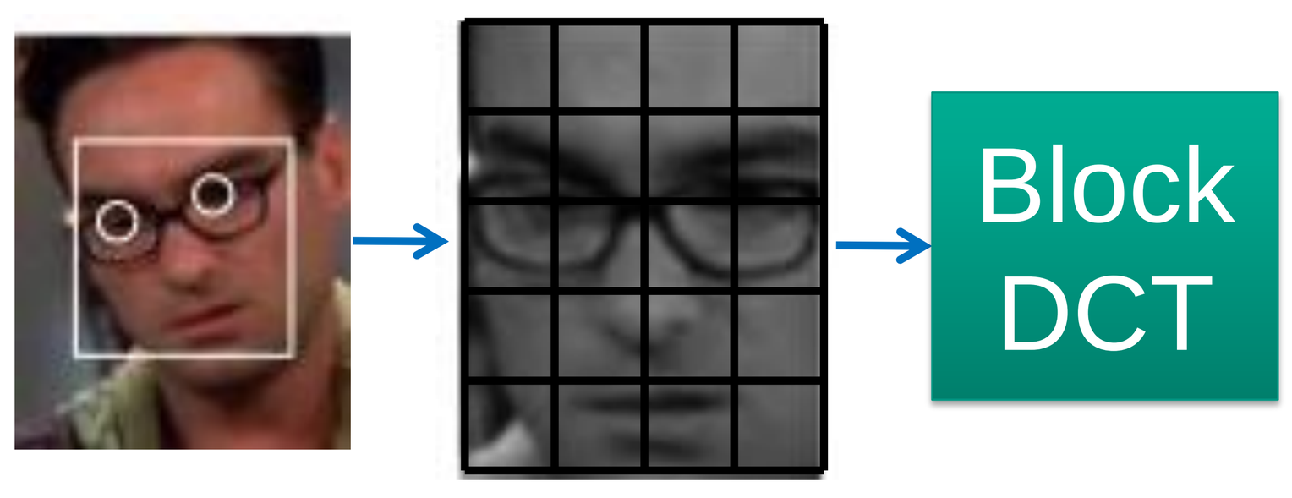
Person identification in multimedia data has many applications such as smart video browsing or as basis for higher-level applications. **Weak supervision** provides labels to a small portion of samples. Most of the face tracks remain **unlabeled**, yet provide information of the underlying distribution of the data. Furthermore, **constraints** between face tracks help restrict possible identities. We integrate all this information in a **common learning framework**.

Face Tracking

- MCT-based multi-pose face detector
- Particle filter tracker

Features

- Eye detection / warp to canonical pose
- Block DCT features



Weak labels from Transcripts/Subtitles

1. align (fan) transcripts to subtitles

who speaks what? ←→ what is spoken **when**?

Leonard
At least I didn't have to invent twenty-six dimensions just to make the math come out.

Sheldon
I didn't invent them, they're there.

Leonard
In what universe?

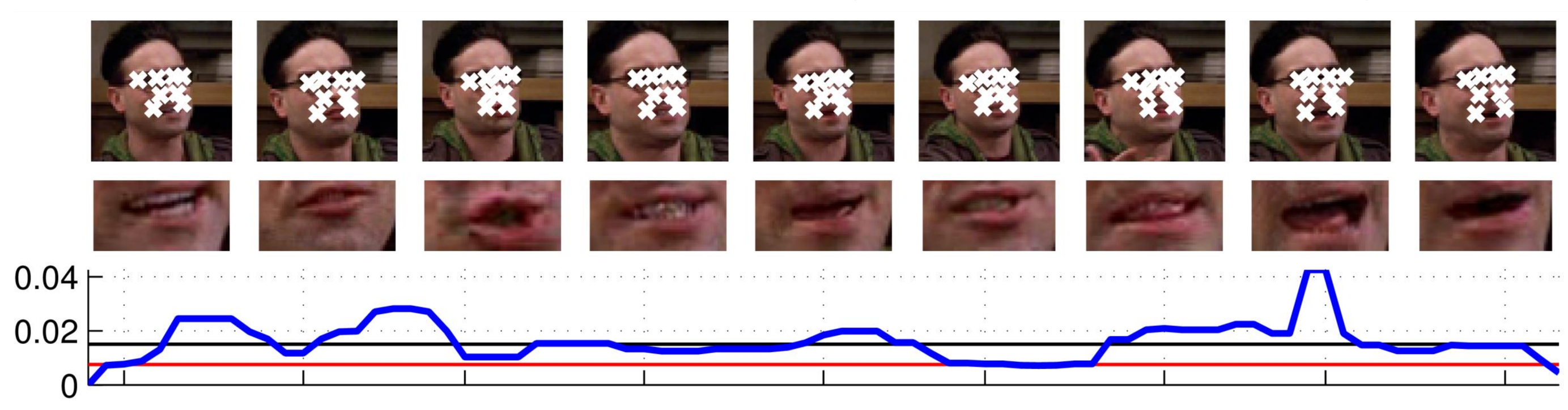
Sheldon
In all of them, that is the point.

00:07:00,733 --> 00:07:04,612
At least I didn't have to invent 26 dimensions just to make the math come out.

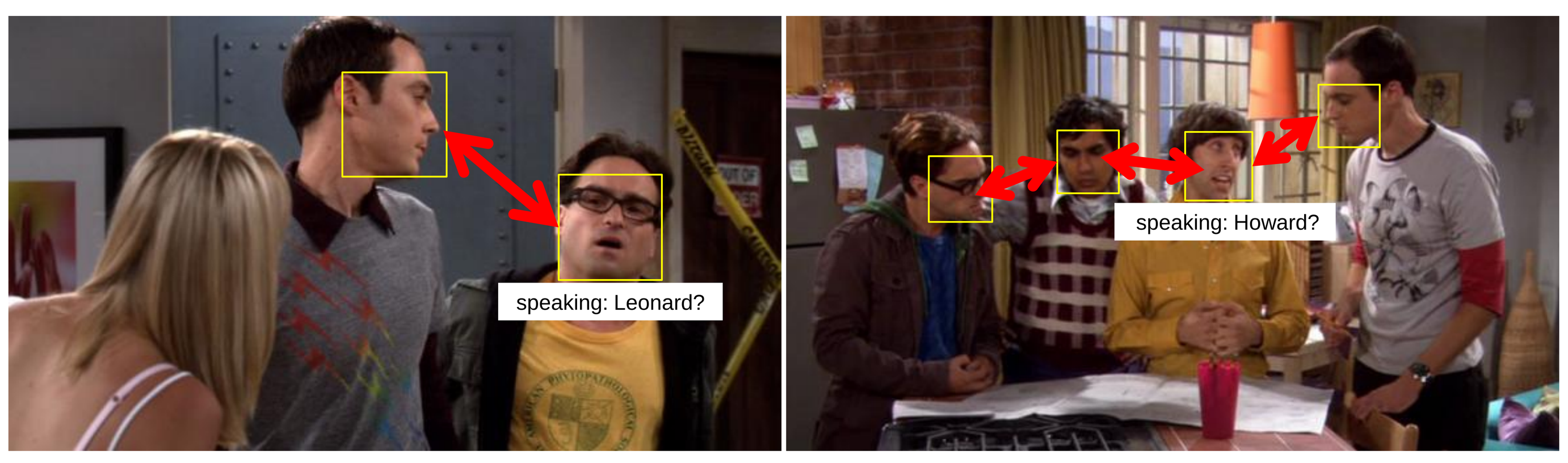
00:07:04,773 --> 00:07:07,412
- I didn't invent them. They're there.
- In what universe?

00:07:07,573 --> 00:07:09,291
In all of them, that is the point.

2. assign names to speaking faces
   - facial feature detection + lip movement analysis



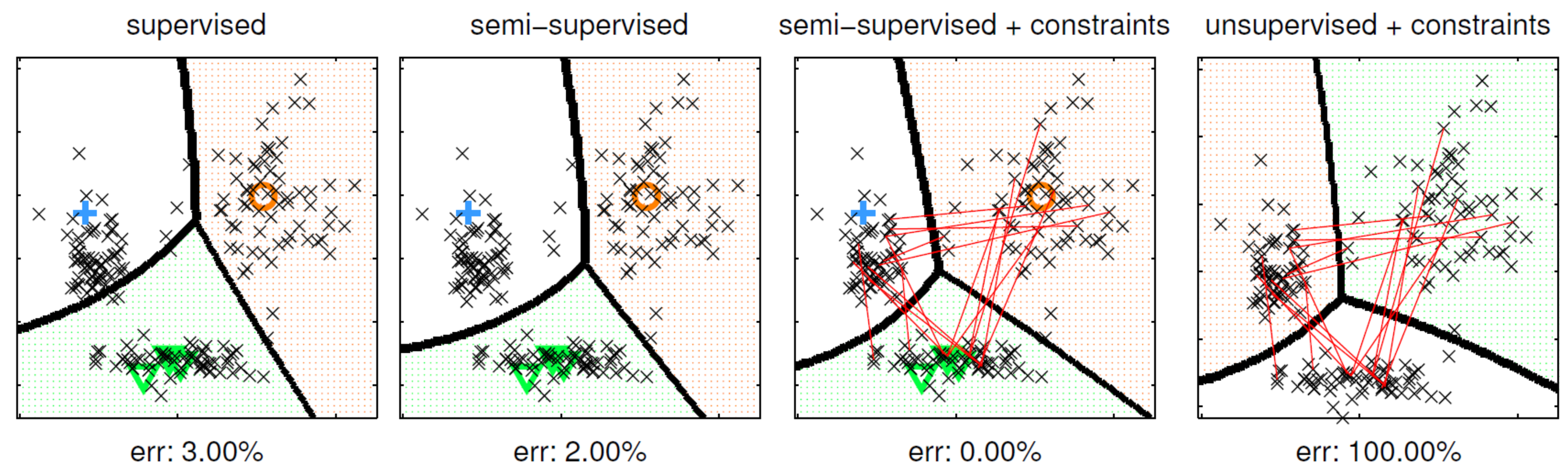
Constraints between co-occurring Face Tracks



Contact
{baeuml, tapaswi}@kit.edu

Project page (tracks, ground truth, etc.)
http://cvhci.anthropomatik.kit.edu/projects/mma

Joint Learning Framework



- combined loss function

$$\mathcal{L}(\mathcal{X};\theta) = \mathcal{L}(y_l, y_c; \mathcal{X}_l, \mathcal{X}_u, \mathcal{C}, \theta)$$
$$= \mathcal{L}_l(y_l; \mathcal{X}_l, \theta) + \mathcal{L}_u(\mathcal{X}_u, \theta) + \mathcal{L}_c(y_c; \mathcal{C}, \theta)$$

- model: multinomial logistic regression + kernelization

$$P(y=k|\mathbf{x};\theta) = \frac{e^{\theta_k^T \mathbf{x}}}{\sum_z e^{\theta_z^T \mathbf{x}}} \qquad f(\mathbf{x}) = \sum_{i=1}^{n} \theta_{ki} K(\mathbf{x}, \mathbf{x}_i)$$

- *supervised loss*

$$\mathcal{L}_l(y_l; \mathcal{X}_l, \theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{1}[y_i{=}k] \ln(P_\theta^k(\mathbf{x}_i)) + \lambda||\theta||^2$$

- *unsupervised loss*

$$\mathcal{L}_u(\mathcal{X}_u;\theta) = -\frac{\mu}{M}\sum_{i=1}^{M}\sum_{k} P_\theta^k(\mathbf{x}_i)\ln(P_\theta^k(\mathbf{x}_i))$$

- *constraint loss*

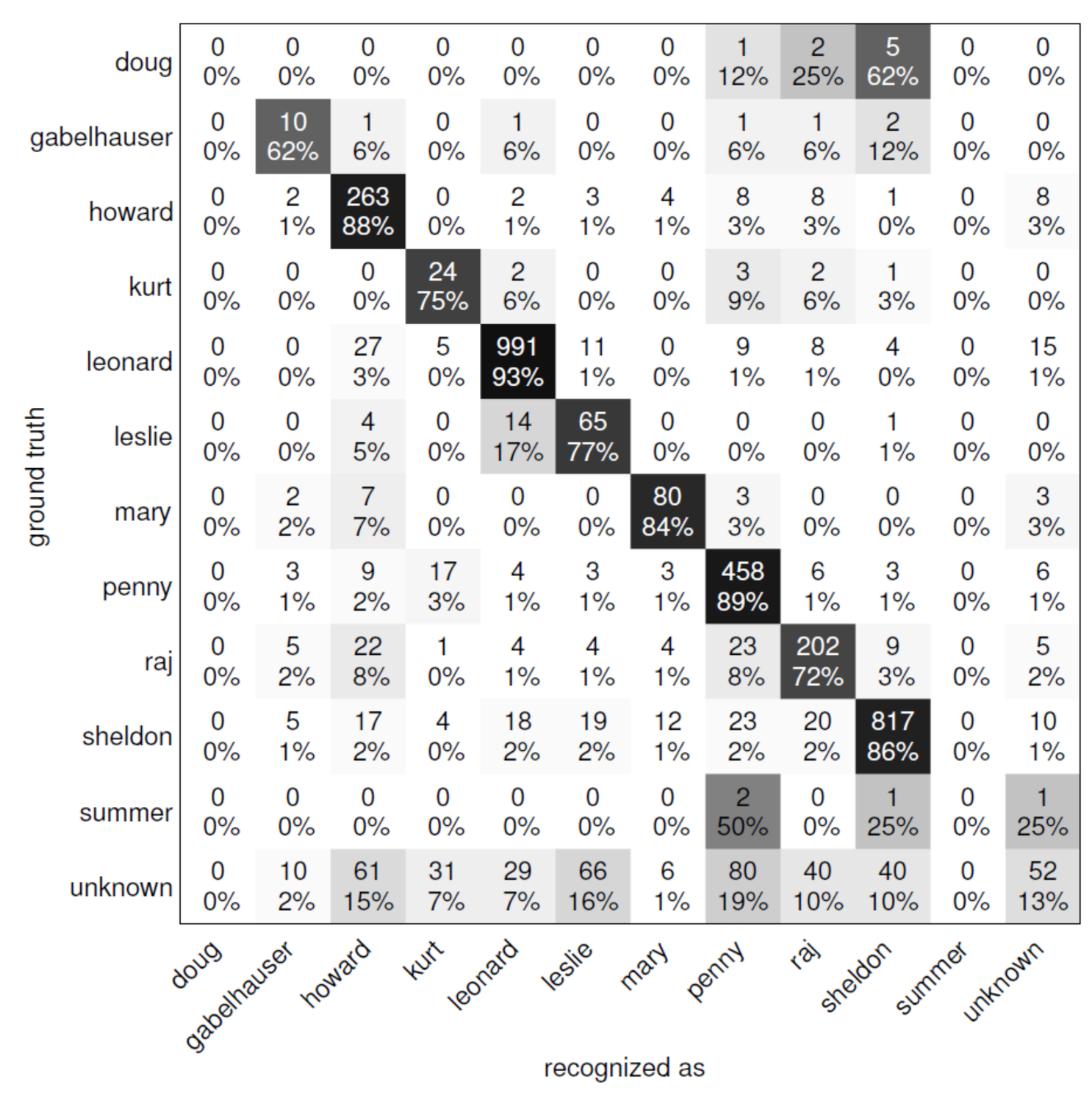
$$\mathcal{L}_c(c_i;\mathcal{C},\theta) = -\frac{\gamma}{L}\sum_{i=1}^{L}\ln(P(y_{i1} \neq y_{i2})) = -\frac{\gamma}{L}\ln\left(1 - \sum_{k=1}^{K} P_\theta^k(\mathbf{x}_{i1}) P_\theta^k(\mathbf{x}_{i2})\right)$$

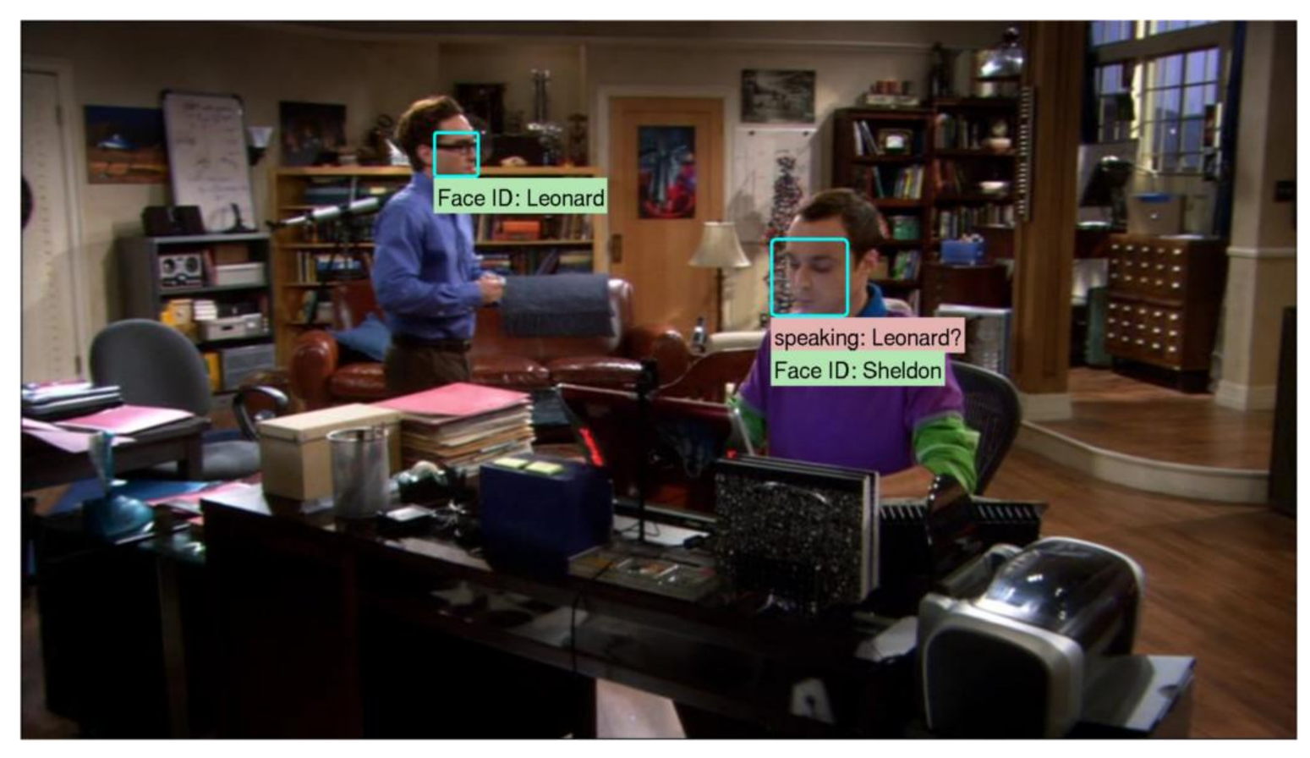
Results

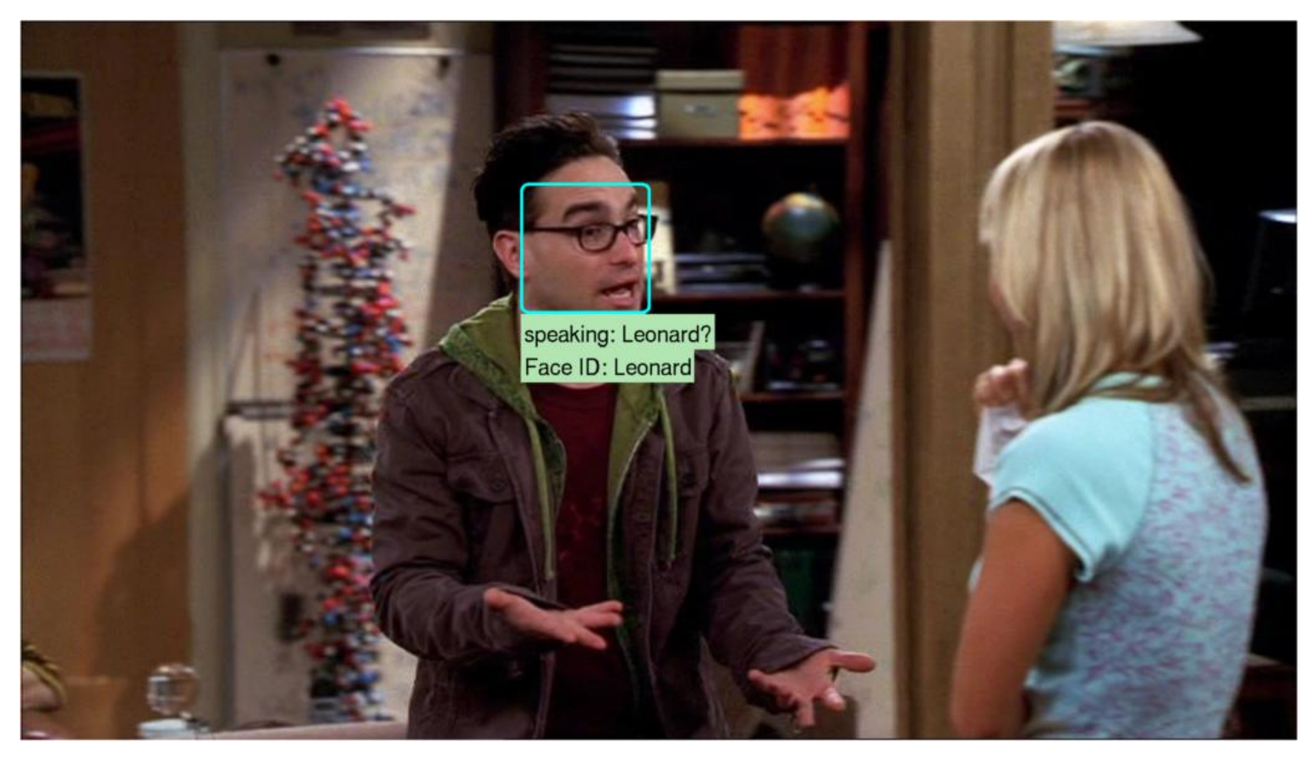
- Evaluation on 12 episodes of two TV series (~9200 face tracks)
- Goal: identify ALL characters (including minor characters & unknowns)
- Evaluation criterion: Track Identification Accuracy

	BBT-1	BBT-2	BBT-3	BBT-4	BBT-5	BBT-6	**BBT Avg.**	BF-1	BF-2	BF-3	BF-4	BF-5	BF-6	**BF Avg.**
Max Prior	37.94	33.98	34.42	17.56	24.19	23.66	28.63	29.97	19.31	18.69	25.75	35.24	14.58	23.92
baseline: NN [5]	72.19	71.86	66.88	59.04	59.50	55.98	64.24	60.34	51.92	55.13	58.92	61.96	50.74	56.50
baseline: one-vs-all LR	88.42	84.60	73.57	73.84	70.97	65.73	76.19	69.50	59.29	66.05	65.87	67.56	60.33	64.77
baseline: one-vs-all SVM [15]	87.46	84.96	74.06	74.87	70.25	66.46	**76.34**	69.90	59.71	66.23	66.47	68.07	61.44	**65.30**
baseline: one-vs-all SVM + MRF [15]	94.05	92.21	76.18	79.00	75.63	74.51	**81.93**	–	–	–	–	–	–	–
ours: MLR	88.59	87.61	76.18	74.01	72.76	65.24	77.40	68.85	61.37	65.96	67.19	69.85	61.72	65.82
ours: MLR + $\mathcal{L}_u$	88.59	87.61	76.35	74.01	72.94	65.24	77.46	71.60	60.54	66.42	67.78	70.10	61.44	66.31
ours: MLR + $\mathcal{L}_u$ + $\mathcal{L}_c$	89.23	89.20	78.47	76.59	75.09	68.05	**79.44**	71.99	61.27	66.60	67.07	69.59	61.72	**66.37**
ours: MLR + $\mathcal{L}_u$ + $\mathcal{L}_c$ + MRF [15]	95.18	94.16	77.81	79.35	79.93	75.85	**83.71**	–	–	–	–	–	–	–

- Confusion Matrix for BBT-1..6:







Acknowledgment: This work was supported by the Quaero Programme, funded by OSEO; and BMBF contract no. 01ISO9052E and 13N12063.