ODANet: Online Deep Appearance Network for Identity-Consistent Multi-Person Tracking

Guillaume Delorme[1], Yutong Ban[2], Guillaume Sarrazin[1], Xavier Alameda-Pineda[1]
[1]Inria, LJK, Univ. Grenoble Alpes, France [2] MIT CSAIL Distributed Robotics Lab

10/01/2021, MPRSS 2020, Milano, Italy

1. Introduction
2. Bayesian Model
3. Appearance modeling
4. Results

Introduction

A robotic context

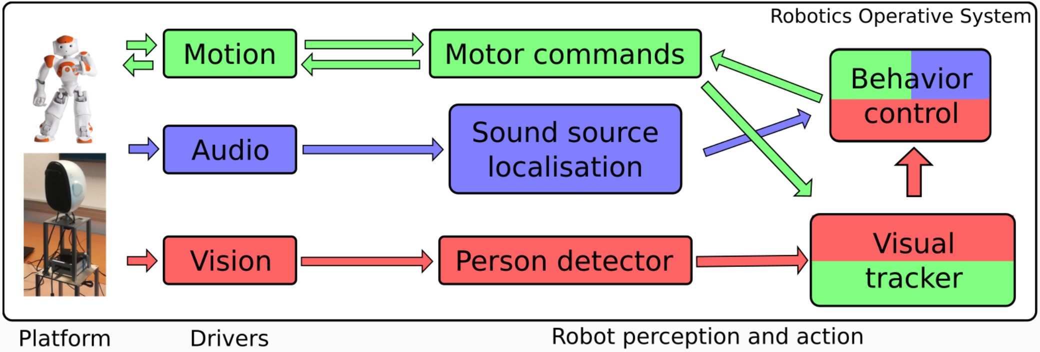


Figure 1: Software architecture of our robotic platform.

A robotic context

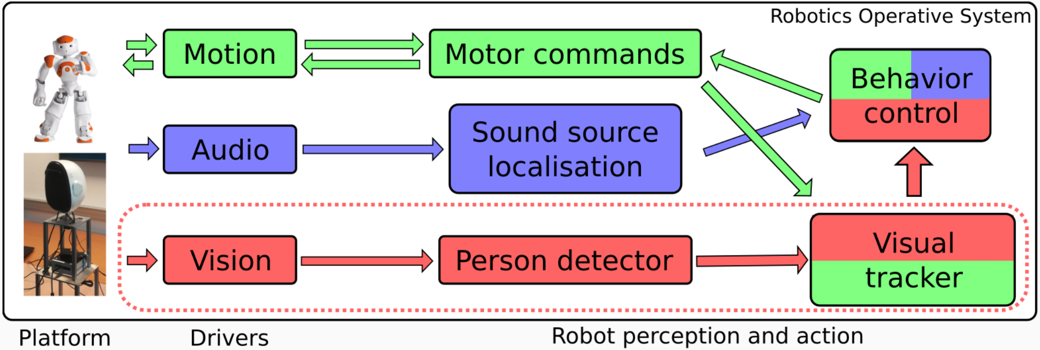
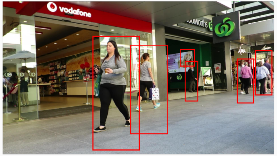


Figure 1: Software architecture of our robotic platform.

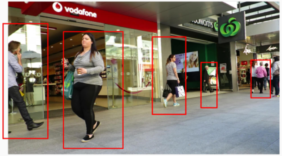
In today's presentation we'll focus on the visual tracker part of our implementation.

Tracking by detection



t

t + 1

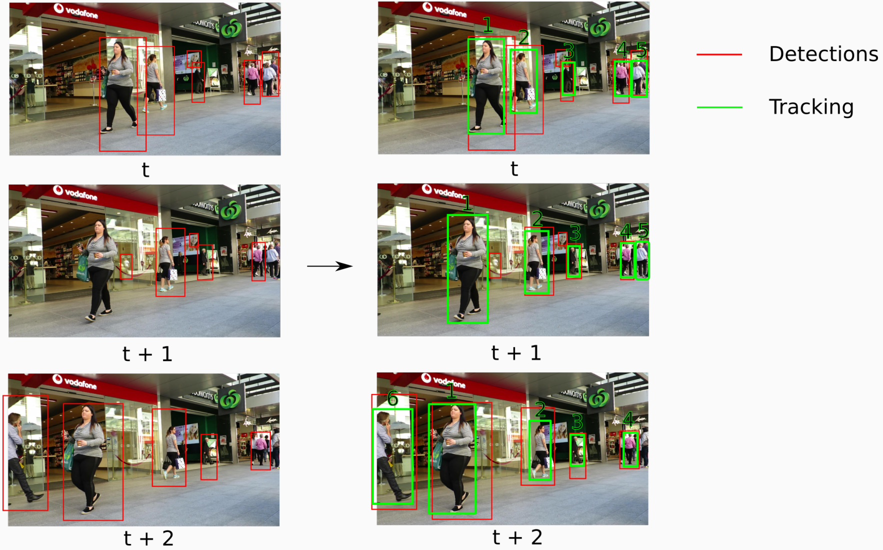


t + 2

Detections

Tracking

Tracking by detection



Multiple Object Tracking

- Our tracker can be seen as a generalization of a **Kalman Filter**, dealing with **multiple** objects at once.

Multiple Object Tracking

- Our tracker can be seen as a generalization of a **Kalman Filter**, dealing with **multiple** objects at once.
- Dealing with multiple objects/targets introduces a detection-to-target **assignment problem**, which makes probabilistic formulations **intractable** and motivates the use of a **variational approximation**.

Multiple Object Tracking

- Our tracker can be seen as a generalization of a **Kalman Filter**, dealing with **multiple** objects at once.
- Dealing with multiple objects/targets introduces a detection-to-target **assignment problem**, which makes probabilistic formulations **intractable** and motivates the use of a **variational approximation**.
- To fully **disambiguate** the assignment problem, we need a **discriminative** appearance model, which **adapts** to the situation at hand.

Bayesian Model

Model

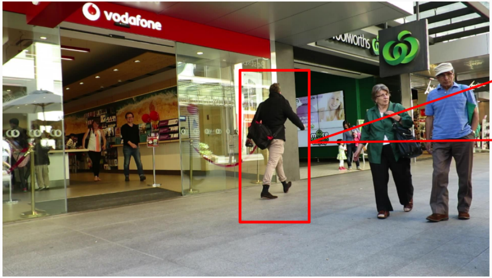
Notations:

- $\mathbf{X}_t \in (\mathbb{R}^6)^N$: State Variables (tracks).

Model

Notations:

- $\mathbf{X}_t \in (\mathbb{R}^6)^N$: State Variables (tracks).
- $\mathbf{O}_t = (\mathbf{Y}_t, \mathbf{U}_t) \in (\mathbb{R}^4 \times \mathcal{I})^{K_t}$ observations at t (detections), $\mathcal{I}$ the image space.



$y_{t,k}$: Position, and size of $o_{t,k}$

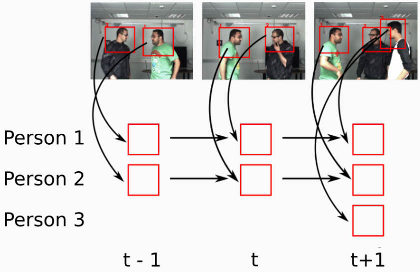
u_{tk}: photometric/appearance information:

Model

Notations:

- $\mathbf{X}_t \in (\mathbb{R}^6)^N$: State Variables (tracks).
- $\mathbf{O}_t = (\mathbf{Y}_t, \mathbf{U}_t) \in (\mathbb{R}^4 \times \mathcal{I})^{K_t}$ observations at t (detections), $\mathcal{I}$ the image space.
- $\mathbf{Z}_t \in [1; N]^{K_t}$: Observation-to-person assignment variables.

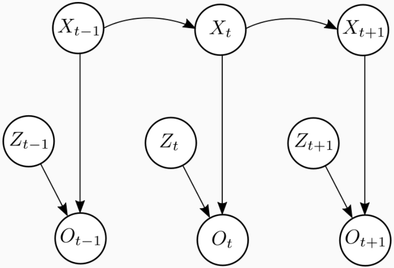
  k-th observation is assigned to track n iff $Z_{tk} = n$



Model

Notations:

- $\mathbf{X}_t \in (\mathbb{R}^6)^N$: State Variables (tracks).
- $\mathbf{O}_t = (\mathbf{Y}_t, \mathbf{U}_t) \in (\mathbb{R}^4 \times \mathcal{I})^{K_t}$ observations at t (detections), $\mathcal{I}$ the image space.
- $\mathbf{Z}_t \in [1; N]^{K_t}$: Observation-to-person assignment variables.
  k-th observation is assigned to speaker n iff $Z_{tk} = n$



Posterior Intractability

At each time t, our goal is to solve

$$\hat{\mathbf{x}}_t = \arg\max_{\mathbf{x}_t} p(\mathbf{x}_t | \mathbf{o}_{1:t})$$

Posterior Intractability

At each time t, our goal is to solve

$$\hat{\mathbf{x}}_t = \arg\max_{\mathbf{x}_t} p(\mathbf{x}_t|\mathbf{o}_{1:t})$$

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{o}_{1:t}) &= \sum_{\tau=1}^{t} \sum_{n=1}^{N} \sum_{k=1}^{K} \int_{\mathbf{x}_1,\ldots\mathbf{x}_{t-1}} p(\mathbf{x}_1,\ldots\mathbf{x}_t, Z_{\tau k} = n|\mathbf{o}_{1:t};) d\mathbf{x}_1 \ldots d\mathbf{x}_{t-1} \\ &= \underbrace{\sum_{c=1}^{C} \pi_c p(\mathbf{x}_t;\Theta_c)}_{\text{mixture}}, \text{with} \quad C = (N+1)^{tK} \text{mixture components} \end{aligned}$$

Its direct estimation is intractable.

Variational Approximation

To solve this, we make use of the following variational approximation

$$p(\mathbf{x}_t, \mathbf{z}_t | \mathbf{o}_{1:t}) \approx q(\mathbf{x}_t) q(\mathbf{z}_t),$$

Variational Approximation

To solve this, we make use of the following variational approximation

$$p(\mathbf{x}_t, \mathbf{z}_t | \mathbf{o}_{1:t}) \approx q(\mathbf{x}_t) q(\mathbf{z}_t),$$

Using this approximation we derive a **variational EM** which minimizes the **Kullback-Leibler divergence** between the approximation and the true posterior.

Variational Approximation

To solve this, we make use of the following variational approximation

$$p(\mathbf{x}_t, \mathbf{z}_t | \mathbf{o}_{1:t}) \approx q(\mathbf{x}_t)q(\mathbf{z}_t),$$

Using this approximation we derive a **variational EM** which minimizes the **Kullback-Leibler divergence** between the approximation and the true posterior.

It boils down to alternate between updating $q(\mathbf{z}_t)$ with a **GMM responsibility** computation, and updating $q(\mathbf{x}_t)$ with a weighted **Kallman forward pass**.

Appearance modeling

The observation model

- A key component of our model is the definition of the Observation Model:

$$\overbrace{p(\mathbf{o}_{t,k}|\mathbf{x}_{t,n}, Z_{t,k} = n)}^{\text{Which track to associate } \mathbf{o}_{t,k} \text{ with?}} = \underbrace{p(\mathbf{y}_{t,k}|\mathbf{x}_{t,n}, Z_{t,k} = n)}_{\text{Geometric Model, the closest one}} \times \underbrace{p(\mathbf{u}_{t,k}|Z_{t,k} = n)}_{\text{Appearance Model, the most similar one}} \quad (1)$$

$t-1$

The observation model

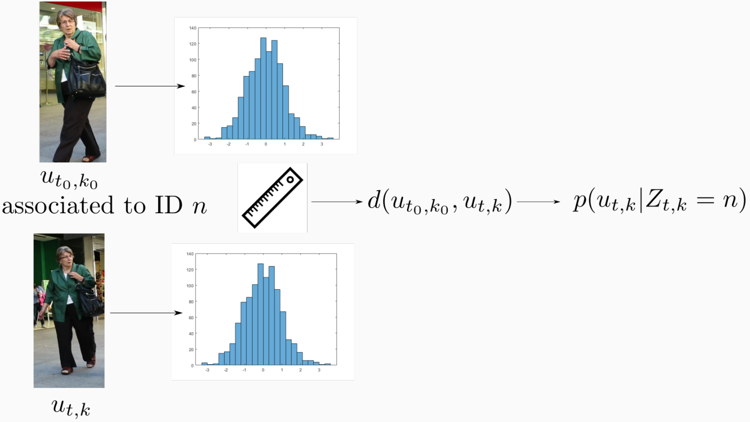
- A key component of our model is the definition of the Observation Model:

$$\overbrace{p(\mathbf{o}_{t,k}|\mathbf{x}_{t,n}, Z_{t,k}=n)}^{\text{Which track to associate } \mathbf{o}_{t,k} \text{ with?}} = \underbrace{p(\mathbf{y}_{t,k}|\mathbf{x}_{t,n}, Z_{t,k}=n)}_{\text{Geometric Model, the closest one}} \times \underbrace{p(\mathbf{u}_{t,k}|Z_{t,k}=n)}_{\text{Appearance Model, the most similar one}} \tag{1}$$

$t-1$ $\qquad$ t

Histogram-based appearance model

- Previous strategy: use of **hand-crafted** descriptors and metrics to compute a distance interpreted as a density function.



Histogram-based appearance model

- Previous strategy: use of **hand-crafted** descriptors and metrics to compute a distance interpreted as a density function.
- Lack **discriminative** power and **robustness**, due to appearance variations (Illumination, pose, background, occlusions...)

Deep Appearance Model

Inspired by model-based tracking methods[1], our strategy is to **learn** a descriptor, using a neural network, with appearances extracted from the **tracker's history**.

- Deep Appearance Models are generally trained **offline**, on large manually annotated datasets[2].

[1] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. "Deep metric learning for visual tracking". In: *IEEE Transactions on Circuits and Systems for Video Technology* 26.11 (2016), pp. 2056–2068.

[2] Siyu Tang et al. "Multiple People Tracking by Lifted Multicut and Person Re-identification". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington, DC, USA: IEEE Computer Society, July 2017, pp. 3701–3710.

Deep Appearance Model

Inspired by model-based tracking methods[1], our strategy is to **learn** a descriptor, using a neural network, with appearances extracted from the **tracker's history**.

- Deep Appearance Models are generally trained **offline**, on large manually annotated datasets[2].
- While they seek generality, they **lack discriminative power**.

[1] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. "Deep metric learning for visual tracking". In: *IEEE Transactions on Circuits and Systems for Video Technology* 26.11 (2016), pp. 2056–2068.

[2] Siyu Tang et al. "Multiple People Tracking by Lifted Multicut and Person Re-identification". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington, DC, USA: IEEE Computer Society, July 2017, pp. 3701–3710.

Deep Appearance Model

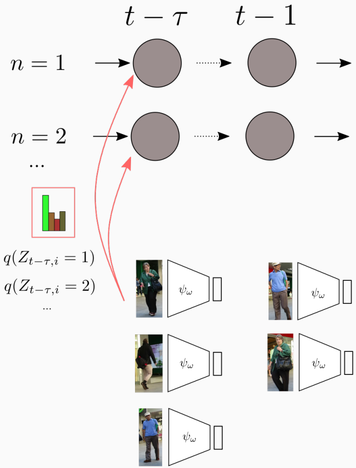
Inspired by model-based tracking methods[1], our strategy is to **learn** a descriptor, using a neural network, with appearances extracted from the **tracker's history**.

- Deep Appearance Models are generally trained **offline**, on large manually annotated datasets[2].
- While they seek generality, they **lack discriminative power**.
- We want to train a shallow NN, in an online fashion, updated **every few frames**.

[1] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. "Deep metric learning for visual tracking". In: *IEEE Transactions on Circuits and Systems for Video Technology* 26.11 (2016), pp. 2056–2068.

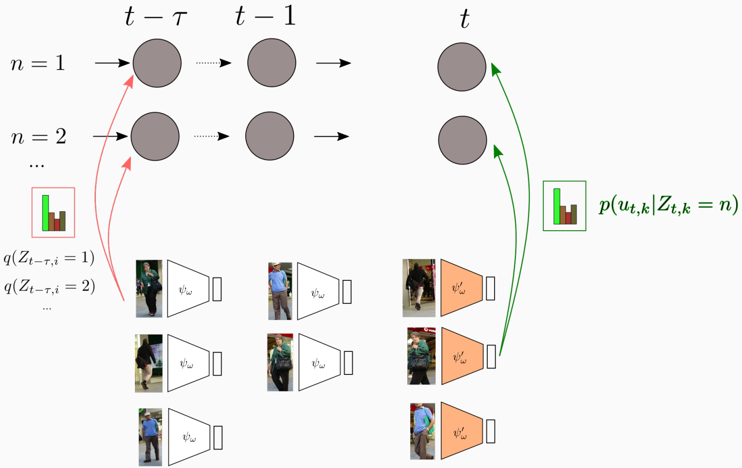
[2] Siyu Tang et al. "Multiple People Tracking by Lifted Multicut and Person Re-identification". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington, DC, USA: IEEE Computer Society, July 2017, pp. 3701–3710.

Online Appearance Model Update



Training set labeled with past assignment posterior by the tracker

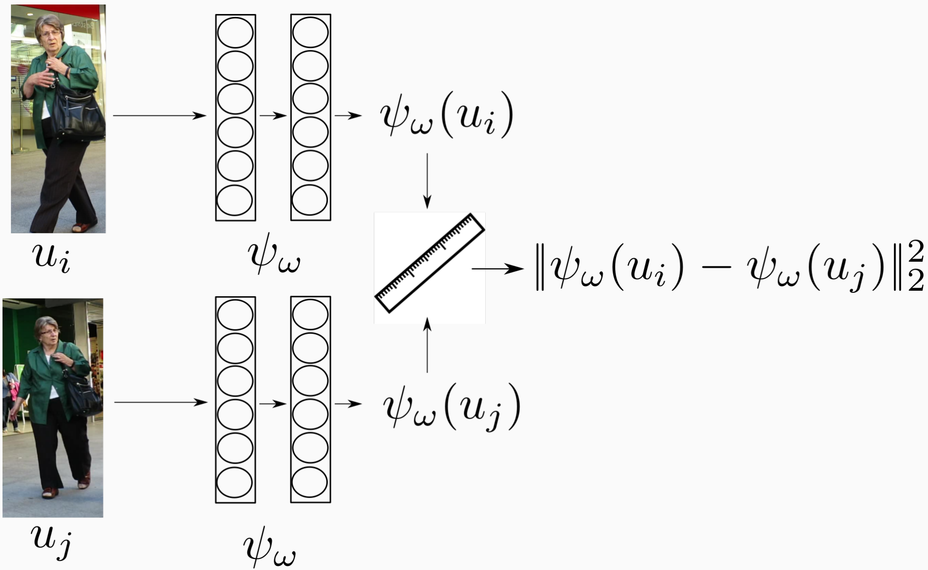
Online Appearance Model Update



Training set labeled with past assignment posterior by the tracker

Predictions

Siamese Neural Network



Appearance model update

We update the top layers of ψ_ω in a siamese setting every few frames using the contrastive loss

$$\mathcal{J}(\omega) = \frac{1}{2} \sum_{i,j=1} \max(0, 1 - l_{ij}(\tau - \|\psi_\omega(\mathbf{u}_i) - \psi_\omega(\mathbf{u}_j)\|^2)),$$

Appearance model update

We update the top layers of ψ_ω in a siamese setting every few frames using the contrastive loss

$$\mathcal{J}(\omega) = \frac{1}{2}\sum_{i,j=1} \max(0, 1 - l_{ij}(\tau - \|\psi_\omega(\mathbf{u}_i) - \psi_\omega(\mathbf{u}_j)\|^2)),$$

This loss needs to be supervised with **binary** (positive or negative pairs) labels. With the tracker's supervision we have access to past **posterior** estimation $q(\mathbf{z}_t)$,thus we propose to a **soft labelisation** instead:

Appearance model update

We update the top layers of ψ_ω in a siamese setting every few frames using the contrastive loss

$$\mathcal{J}(\omega) = \frac{1}{2}\sum_{i,j=1} \max(0, 1 - l_{ij}(\tau - \|\psi_\omega(\mathbf{u}_i) - \psi_\omega(\mathbf{u}_j)\|^2)),$$

This loss needs to be supervised with **binary** (positive or negative pairs) labels. With the tracker's supervision we have access to past **posterior** estimation $q(\mathbf{z}_t)$,thus we propose to a **soft labelisation** instead:

$$\gamma_{ij} = p(Z_{t_i k_i} = Z_{t_j k_j} | \mathbf{o}_{1:t-1}) \approx \sum_{n=1}^{N} q(Z_{t_i k_i} = n) q(Z_{t_j k_j} = n).$$

We label positive pairs with $l_{ij} = \gamma_{ij}$ and negative pairs with $l_{ij} = -(1 - \gamma_{ij})$.

Implementation details

- To make it work smoothly, we use 2 models in parallel, one for training and the other for inference. Our implementation reaches 10 FPS in our framework.

[3] Ergys Ristani et al. "Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking". In: *ECCV Workshops*. 2016.

Implementation details

- To make it work smoothly, we use 2 models in parallel, one for training and the other for inference. Our implementation reaches 10 FPS in our framework.
- The convolutional layers of ψ are pretrained using external Re-ID dataset, using a standard training framework[3].

[3] Ergys Ristani et al. "Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking". In: *ECCV Workshops*. 2016.

Results

Quantitative results: evaluation settings

We evaluate our tracker in a **multi-party conversation**: the robot has a **low fov** and often change position, increasing **identity switches**.

[4] A. Milan et al. "MOT16: A Benchmark for Multi-Object Tracking". In: *arXiv:1603.00831 [cs]* (Mar. 2016). arXiv: 1603.00831. URL: http://arxiv.org/abs/1603.00831.

[5] Keni Bernardin and Rainer Stiefelhagen. "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics". In: *EURASIP Journal on Image and Video Processing* (2008).

Quantitative results: evaluation settings

We evaluate our tracker in a **multi-party conversation**: the robot has a **low fov** and often change position, increasing **identity switches**.
We also want to use a **standard** dataset, thus we use of the MOT16 dataset[4] that we divide in 2 evaluation settings:

- *moving surveillance camera* for the sequences with camera **fixed**: we **simulate** the **camera movement** to increase identity switches.

[4] A. Milan et al. "MOT16: A Benchmark for Multi-Object Tracking". In: *arXiv:1603.00831 [cs]* (Mar. 2016). arXiv: 1603.00831. URL: http://arxiv.org/abs/1603.00831.
[5] Keni Bernardin and Rainer Stiefelhagen. "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics". In: *EURASIP Journal on Image and Video Processing* (2008).

Quantitative results: evaluation settings

We evaluate our tracker in a **multi-party conversation**: the robot has a **low fov** and often change position, increasing **identity switches**.
We also want to use a **standard** dataset, thus we use of the MOT16 dataset[4] that we divide in 2 evaluation settings:

- *moving surveillance camera* for the sequences with camera **fixed**: we **simulate** the **camera movement** to increase identity switches.
- *robot navigating in the crowd* for the sequences where the camera is **moving**.

We use the CLEAR metrics[5] to evaluate the quality of the tracker results.

[4] A. Milan et al. "MOT16: A Benchmark for Multi-Object Tracking". In: *arXiv:1603.00831 [cs]* (Mar. 2016). arXiv: 1603.00831. URL: `http://arxiv.org/abs/1603.00831`.
[5] Keni Bernardin and Rainer Stiefelhagen. "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics". In: *EURASIP Journal on Image and Video Processing* (2008).

Quantitative results: moving surveillance setting

Model	Detection		Tracking		Identities		
	Rcll	Prcn	MOTA	MOTP	IDP	IDR	IDF1
CH[6]	49.4	88.2	42.5	84.5	70.3	39.4	50.5
ODA-FR	49.5	**88.7**	43.0	**84.8**	66.7	37.2	47.8
ODA-UP	**54.7**	86.7	**45.6**	84.0	**75.4**	**45.7**	**56.0**

Table 1: Results on the *moving surveillance camera* setting.

- ODA-UP stands for our *online deep appearance update*.
- ODA-FR refers to the same appearance model architecture, but frozen (FR).
- CH stands for Color Histogram based appearance model.

[6] Yutong Ban et al. "Variational bayesian inference for audio-visual tracking of multiple speakers". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

Quantitative results: robot navigating in the crowd

Model	Detection		Tracking		Identities		
	Rcll	Prcn	MOTA	MOTP	IDP	IDR	IDF1
CH[7]	45.8	91.8	41.2	80.7	74.1	37.0	49.3
ODA-FR	45.8	**93.1**	42.0	81.0	73.8	36.3	48.6
ODA-UP	**52.3**	90.5	**46.2**	**81.5**	**79.0**	**45.7**	**57.9**

Table 2: Results on the *robot navigating in the crowd* setting.

- ODA-UP stands for our *online deep appearance update*.
- ODA-FR refers to the same appearance model architecture, but frozen (FR).
- CH stands for Color Histogram based appearance model.

[7] Yutong Ban et al. "Variational bayesian inference for audio-visual tracking of multiple speakers". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

Conclusion

Thank you for your attention.