Master thesis:

Gaze Estimation and Gazeable Objects

Presented by:	Vadim Sushko	
Thesis Supervisor:	Dr. Xavier Alameda-Pineda	INRIA Grenoble
Co-supervisor:	Dr. Pablo Mesejo	University of Granada
	Dr. Radu Horaud	INRIA Grenoble

25/06/2019

Perception team, INRIA Grenoble Rhône-Alpes, France

Gaze-following.

Gaze-following aims to infer the direction of gaze of a person.

The gaze of a person:

- Highlights the focus of attention
- Indicates involvement and interest in the talk
- Reveals intentions and future actions
- Shows relations with other people

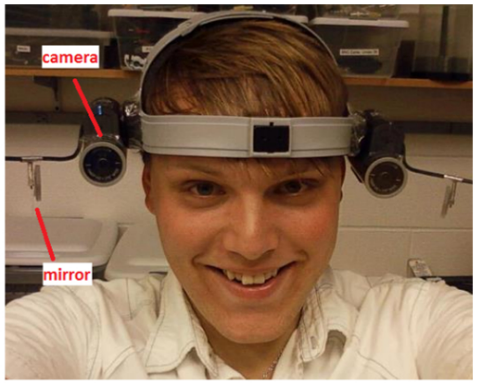
Gaze-following becomes useful in the context of human-robot interaction.

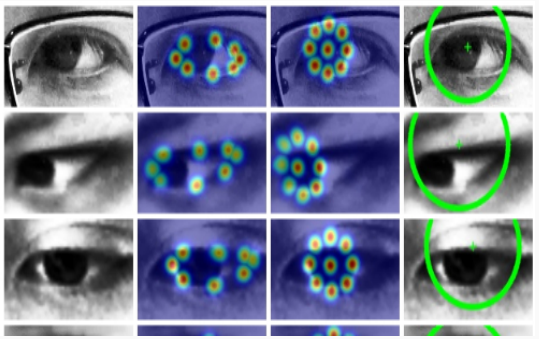
Gaze-following. Early studies.

The early studies concentrated on limited settings:

- Head-mounted systems (massive setups)
- Eye-tracking (high resolution images of faces)

- Cannot be applied in a natural scenario!





Upper image: [Hong 2012], center image: [Fisheret al. 2011], lower image: [Parket al. 2011]

Gaze-following. Recent approaches.

Less limited settings:

- Plane images, only RGB data
- Occluded heads, face and eyes not visible
- Difficult scenes

Solved with recent deep learning approaches as a combination of:

- Head pose estimation
- Saliency detection

Upper image: `https://www.shutterstock.com/video/clip-29933698`,
center image: [Jugget al. 2014], lower image: [Zhanget al. 2015]

Gaze-following. State-of-the-art in the unrestricted setting.

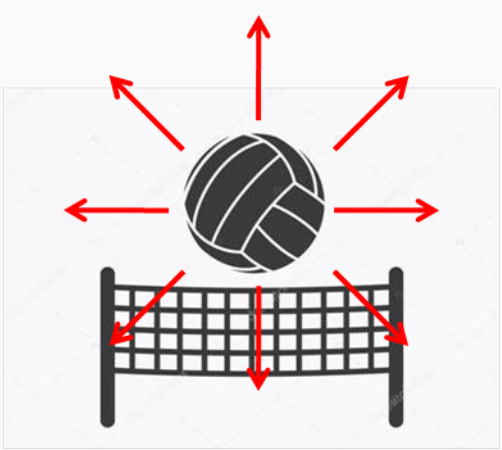
- [Recasens, Khosla, 2016] – The pioneering work where the setting was proposed. Provides a dataset and a deep learning solution. The solution is a combination of **saliency detection** and **head pose estimation**

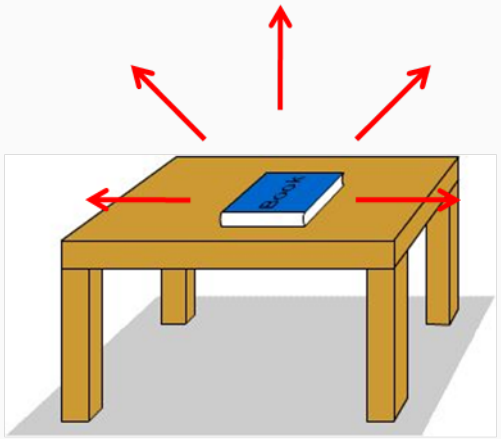
- [Chonget et al., 2018] – Different head pose estimation
- [Gorji, Clark, 2017] – Different saliency detection models

- Other recent works study different settings (Gadget screens, conversations between people, laboratory conditions)

Object Gazeability

- Saliency detection has been studied from the point of view of an external person *looking at the image*, not for an internal observer *within the image*.
- **Idea**: Objects can seem salient from outside an image, but be visually unreachable for an observer inside it
- We denote **Object Gazeability** as the property of objects to have different probabilities of being looked at from different positions
- **Objective**: study gazeability of objects and find its application to gaze-following





Outline of the presentation

- Introduction

- You are here -

- Study of Object Gazeability
- 3D analysis of scenes for gaze prediction
- Conclusion

Gazeability of objects

Gazeability of objects [1/6] – Objective

In this section, we understand object gazeability as a map $p(x, y)$, where p is the function of coordinates of the image representing the probability that the object may be looked from this position.

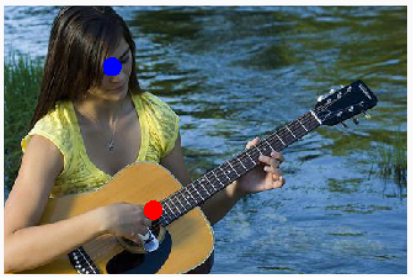
Gazeability of objects [1/6] – GazeFollow dataset

- Large-scale collection of images (125000 in the train set and 5000 in the test set)
- Collected with Amazon Mechanical Turk (AMT)
- Pixel coordinates of observers and gazed objects for each image
- People perform diverse activities in various scenarios



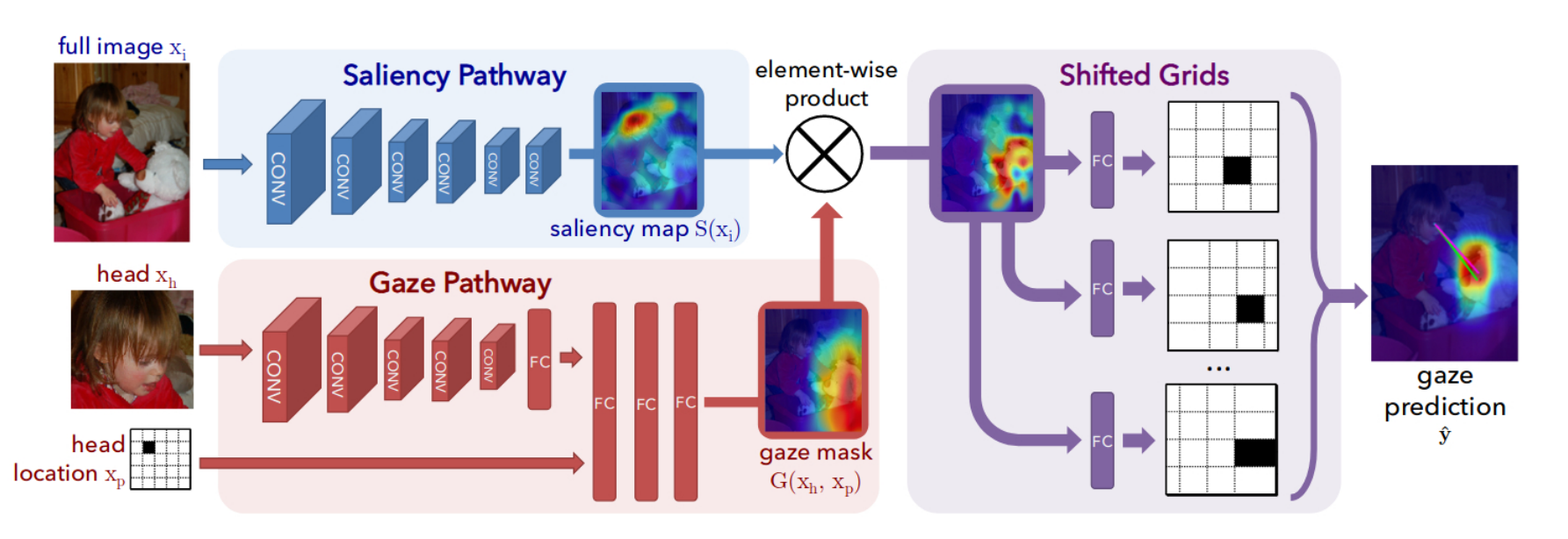




GazeFollow dataset: [Recasens*, Khosla* 2016]

Gazeability of objects [1/6] – GazeFollow network



The GazeFollow neural network was proposed for the purpose of gaze estimation. It requires three inputs:

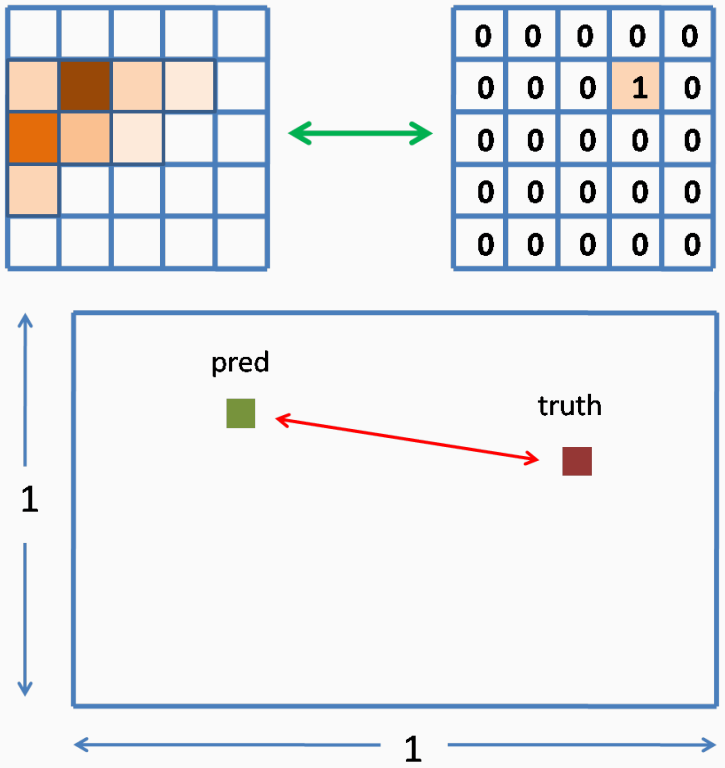
- The processed image
- The close-up image of the observer
- The location of the face with respect to the 13×13 uniform grid

GazeFollow model: [Recasens*, Khosla* 2016]

Gazeability of objects [1/6] – Metrics

The quality of a solution for gaze estimation can be assessed by the two metrics:

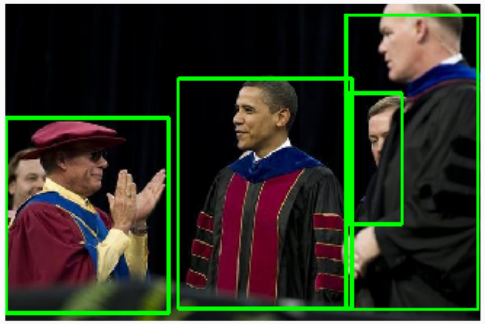
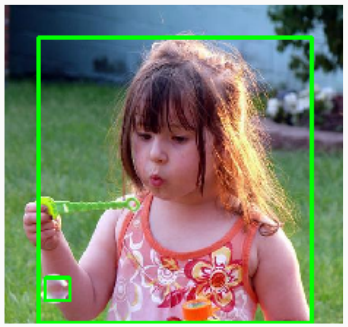
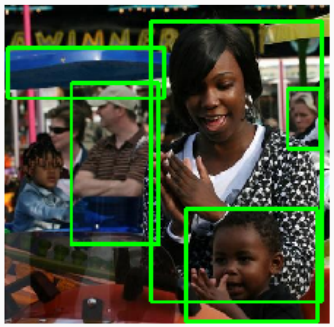
- **ROC AUC** of the constructed binary classification task
- **Average distance** between the predictions and annotations of gazed objects



The results of the GazeFollow network on the GazeFollow dataset:

ROC AUC	Average distance
0.878	0.240

Gazeability of objects [2/6] – Step 1. Motivation for a fully-convolutional network

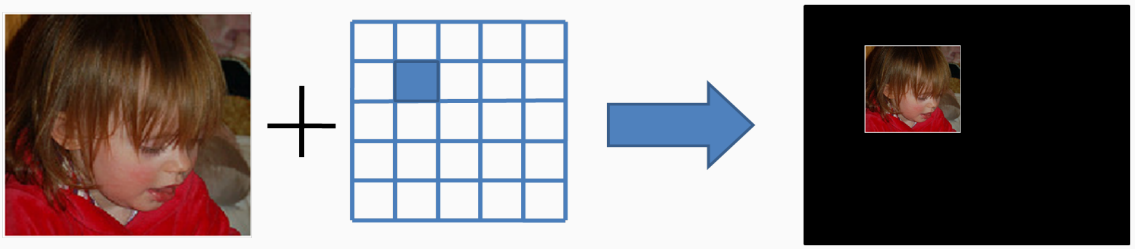





- The GazeFollow network expects the inputs (full image and face image) to have fixed size (227×227)
- Object images are not necessarily squares
- It might be reasonable not to downsize or crop images

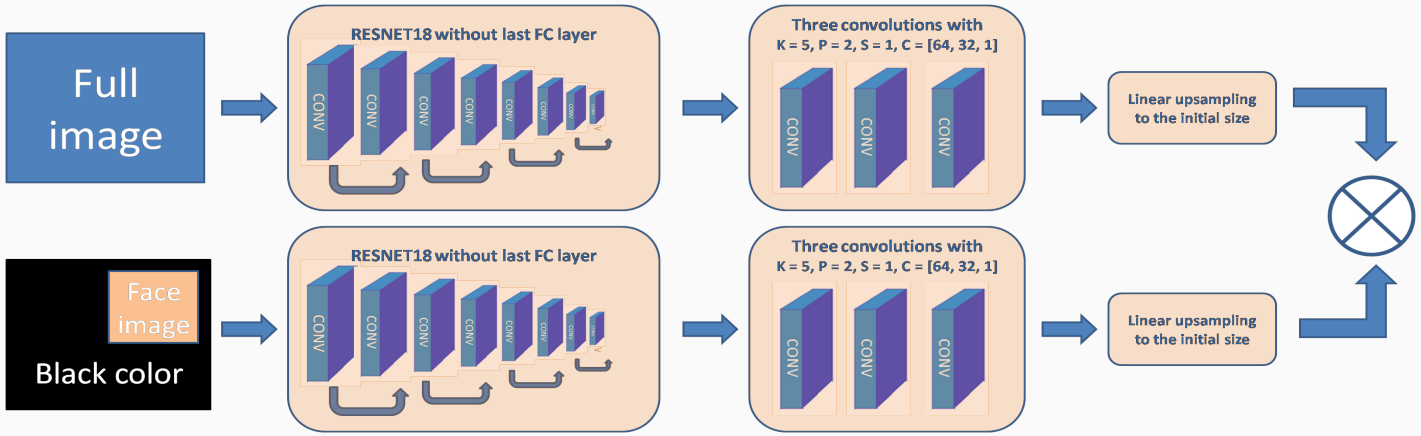
A fully-convolutional architecture would be a good choice for our purposes

Gazeability of objects [2/6] – Step 1. Fully-convolutional neural network

- To avoid fully-connected layers, the coordinates of the face are provided together with the face image



- The proposed network consists of ResNet18 feature extractor, mixing convolutions and linear upsampling. The input can be of arbitrary size



Gazeability of objects [2/6] – Step 1. Training

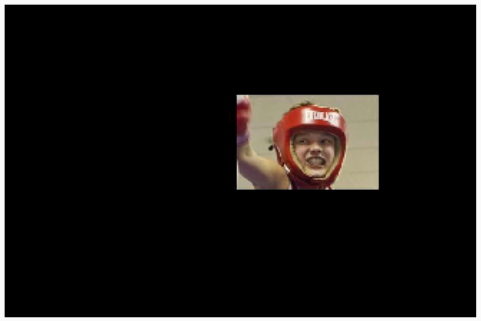
- As ground truth, synthesised Gaussian distributions centered at gazed location were used
- Binary cross entropy loss was used for each pixel:

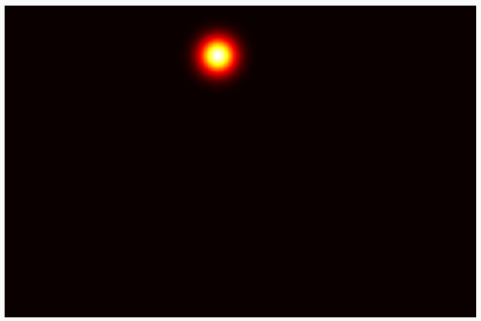
$$L(a, b) = -[b \cdot \ln a + (1 - a) \cdot \ln(1 - b)],$$

- The optimization problem (heatmap regression):

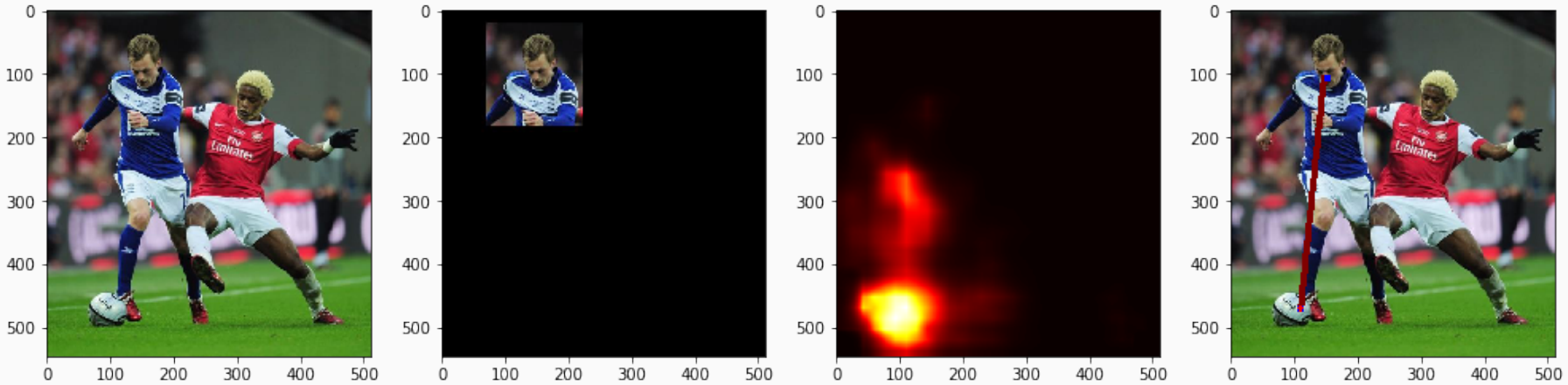
$$w = \arg\min_{w} \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} L(N(w, x, y), Y(x, y))$$

- Adam optimizer was used [Kingma, Ba, 2014], the training took 13 hours on Titan XP GPU





Gazeability of objects [2/6] – Step 1. Results of the training

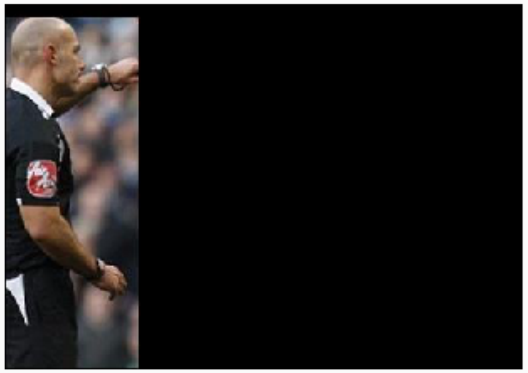


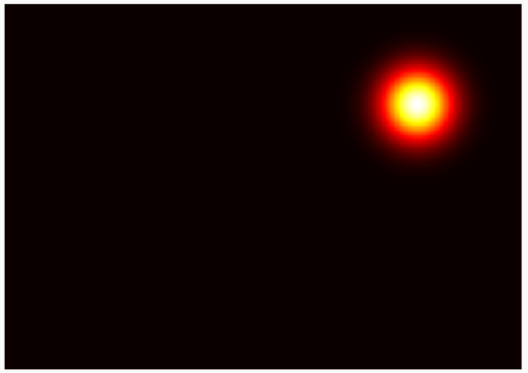
- The proposed architecture is suitable for gaze estimation
- It outperforms the GazeFollow network on the GazeFollow test set:

	ROC AUC	Average distance
Original GazeFollow model	0.878	0.240
Our fully-convolutional network	0.893	0.232

Gazeability of objects [3/6] – Step 2. Observer localisation

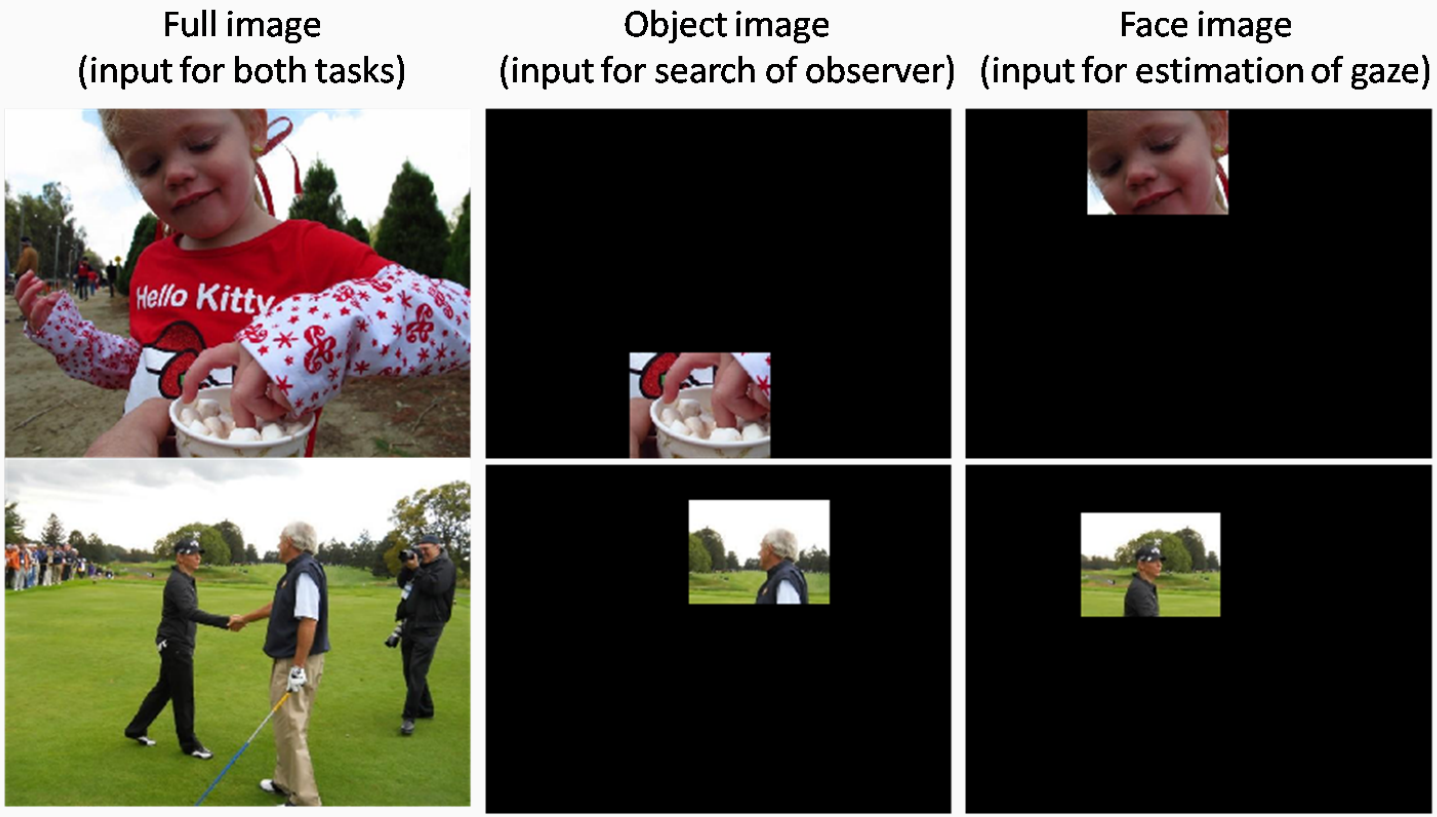
- Object gazeability aims at **observer localisation**
- In this setting faces and objects exchange their roles
- Objects are used as input, location of observer becomes ground truth
- All the details of the training could be passed directly





Gazeability of objects [3/6] – Step 2. Training without object detectors

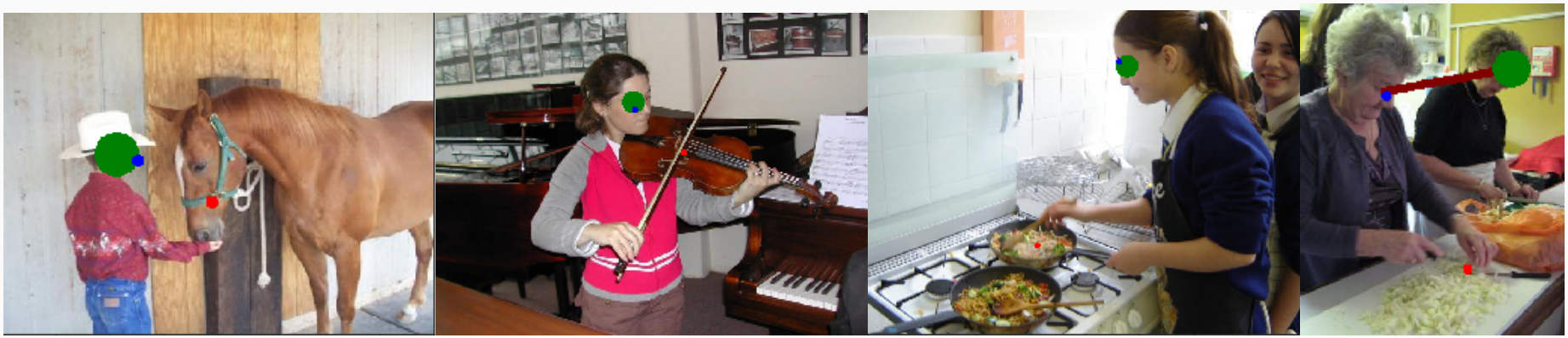
Idea: Bounding boxes of the same shape as face images:



- Our experiments showed that this type of bounding boxes increases the generalization ability of the network

Gazeability of objects [3/6] – Step 2. Results of the training

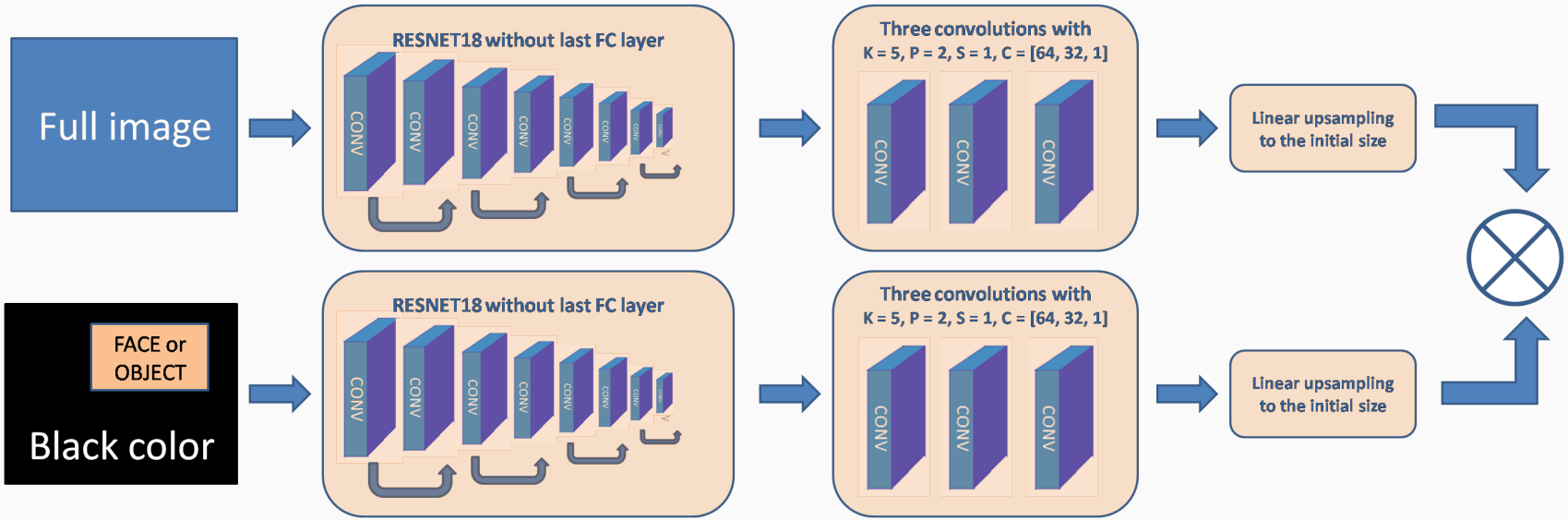
The network learnt to search for heads of observers.



- **Face accuracy**. The proportion of predictions separated from the ground truth not further than 0.15 part of the image. (Dist < 0.15)

ROC AUC	Distance	Face Accuracy
0.969	**0.212**	**0.53**

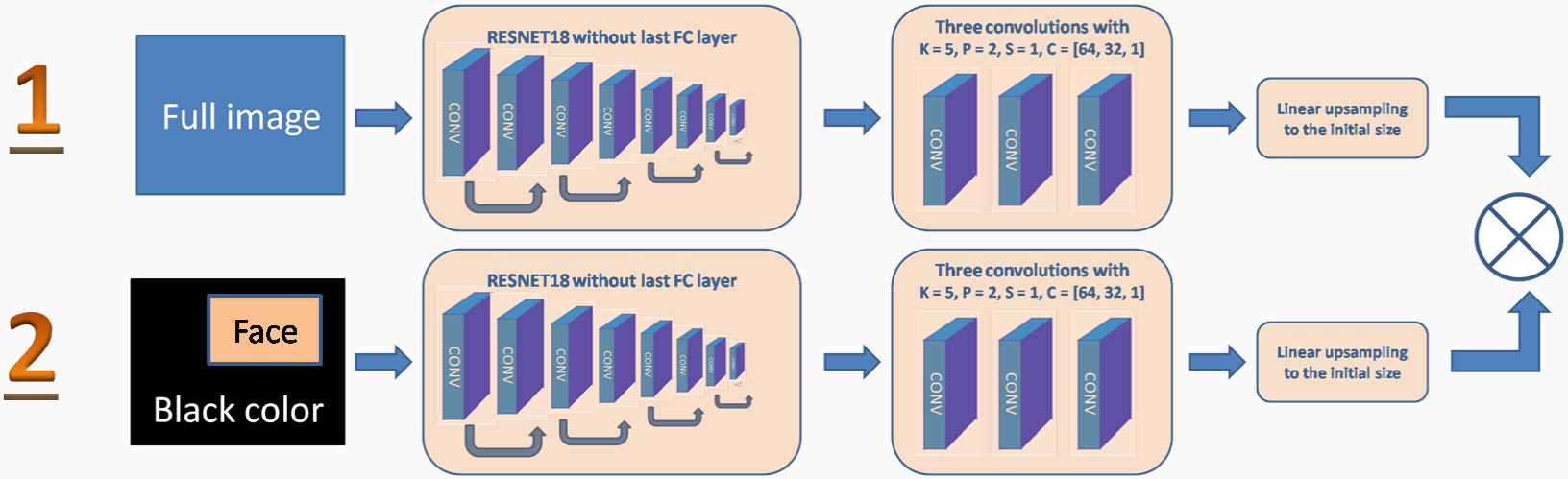
Gazeability of objects [4/6] – Results. Visual ablation analysis



2 networks were trained:

- **The Gaze network**: solving gaze estimation. Branches: *the Saliency-gaze* and the *Face* branch.
- **The Eyes network**: solving observer localisation. Branches: *the Saliency-eyes* and the *Object* branch.

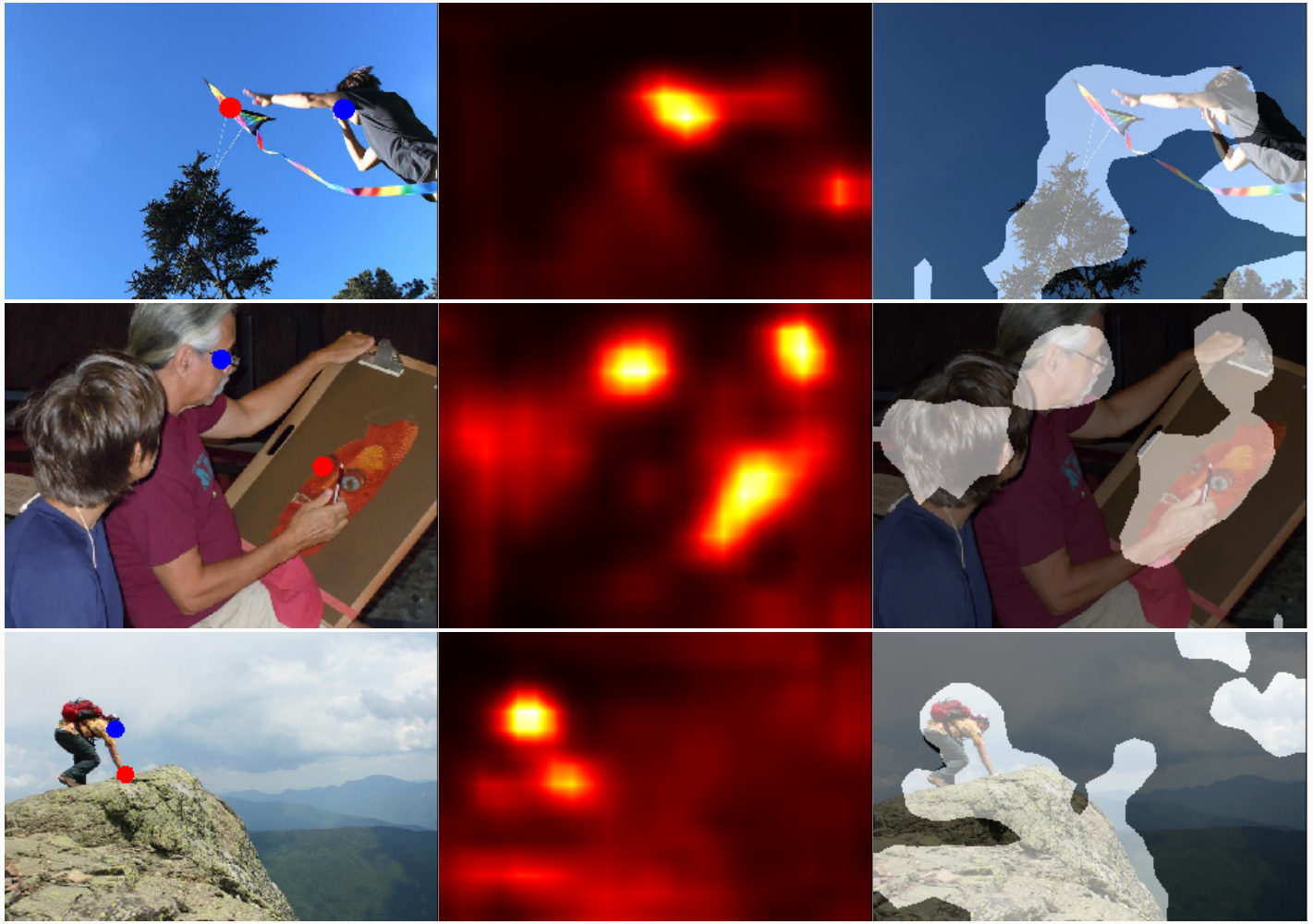
Gazeability of objects [4/6] – Results. Visual ablation analysis



The Gaze network:

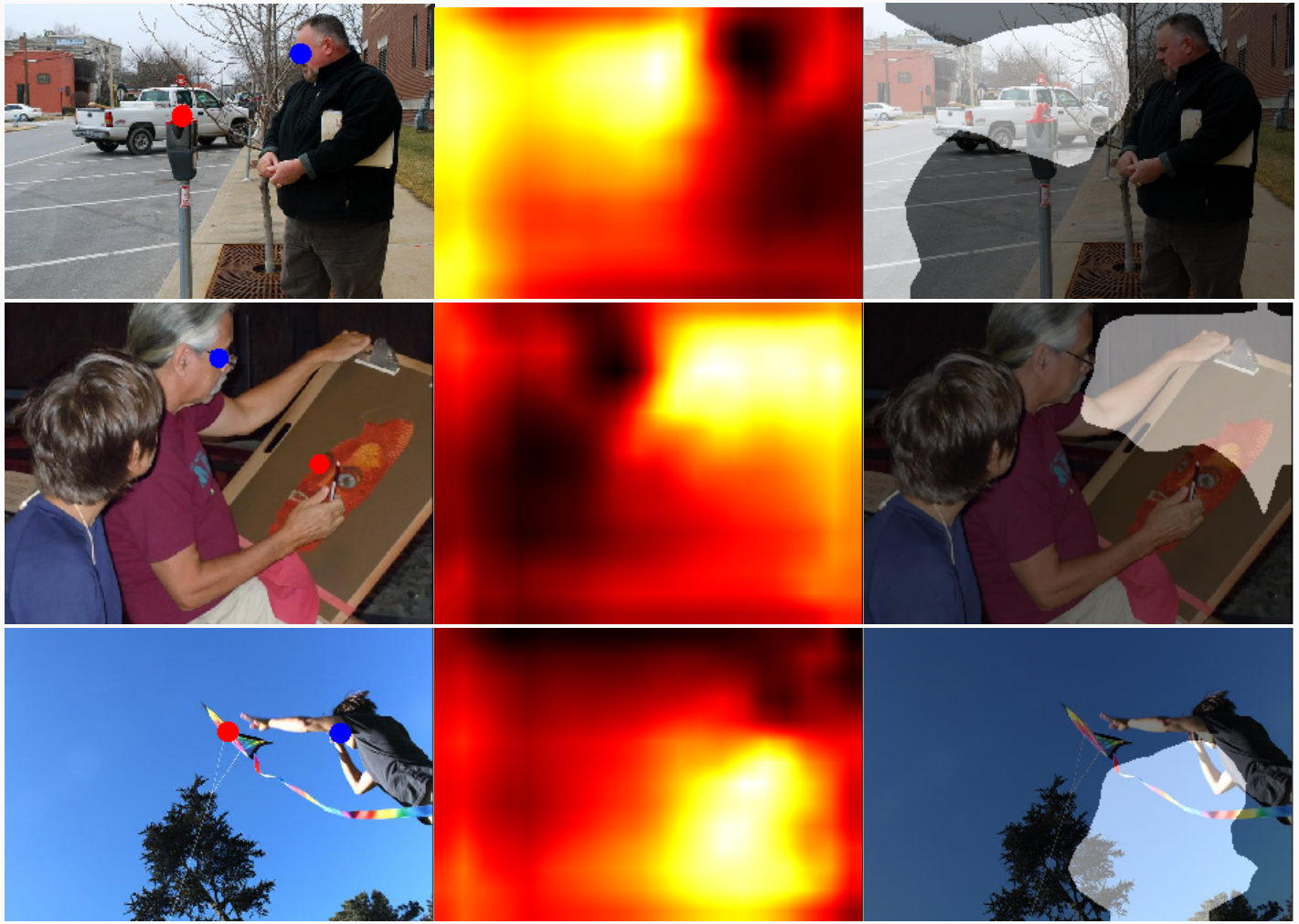
1. The Saliency-gaze branch
2. The Face branch

Gazeability of objects [4/6] – Visual results. Saliency-gaze branch



The network does saliency detection.

Gazeability of objects [4/6] – Visual results. Face branch



The network estimates gaze cone of the observer in 2D.

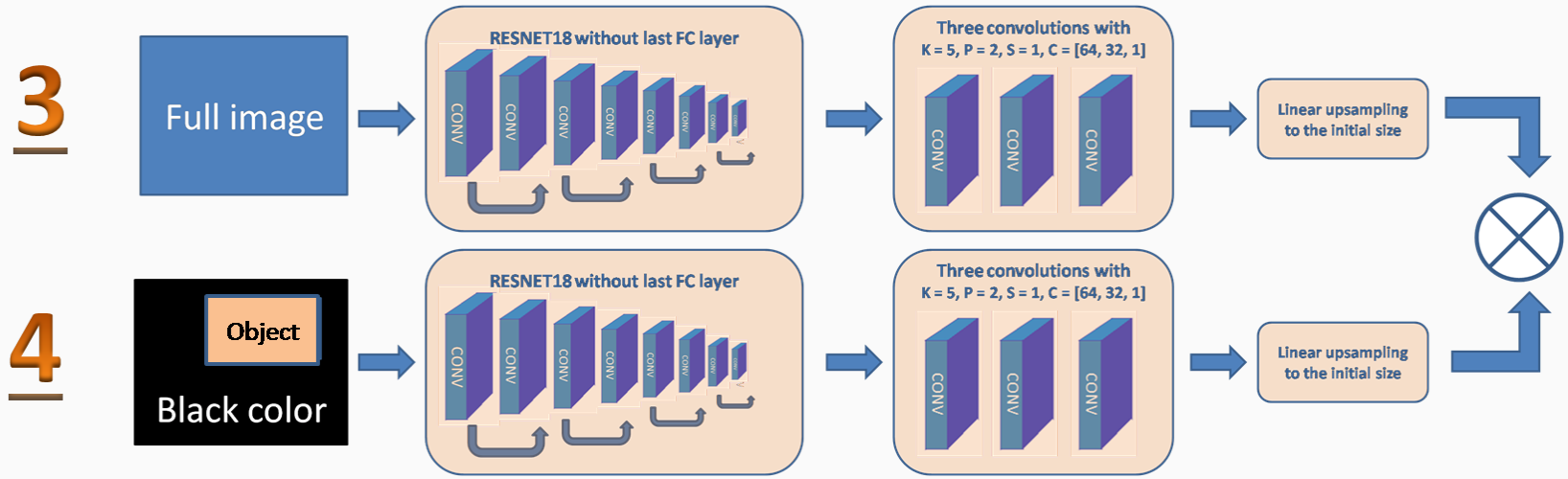
Gazeability of objects [4/6] – Quantitative results.

Separate assessment of the branches: 1 – Saliency-gaze, 2 – Face

Paths used				AUC (eyes)	AUC (gaze)	Dist (eyes)	Dist (gaze)	Acc (eyes)	Acc (gaze)
1				0.886	0.839	0.339	0.306	0.25	0.34
	2			0.478	0.792	0.210	0.299	0.30	0.25
1	2			0.839	**0.893**	0.332	**0.233**	0.18	**0.46**

- The Saliency-gaze and the Face branches contribute equally
- Heads of people also salient areas

Gazeability of objects [5/6] – Results. Visual ablation analysis



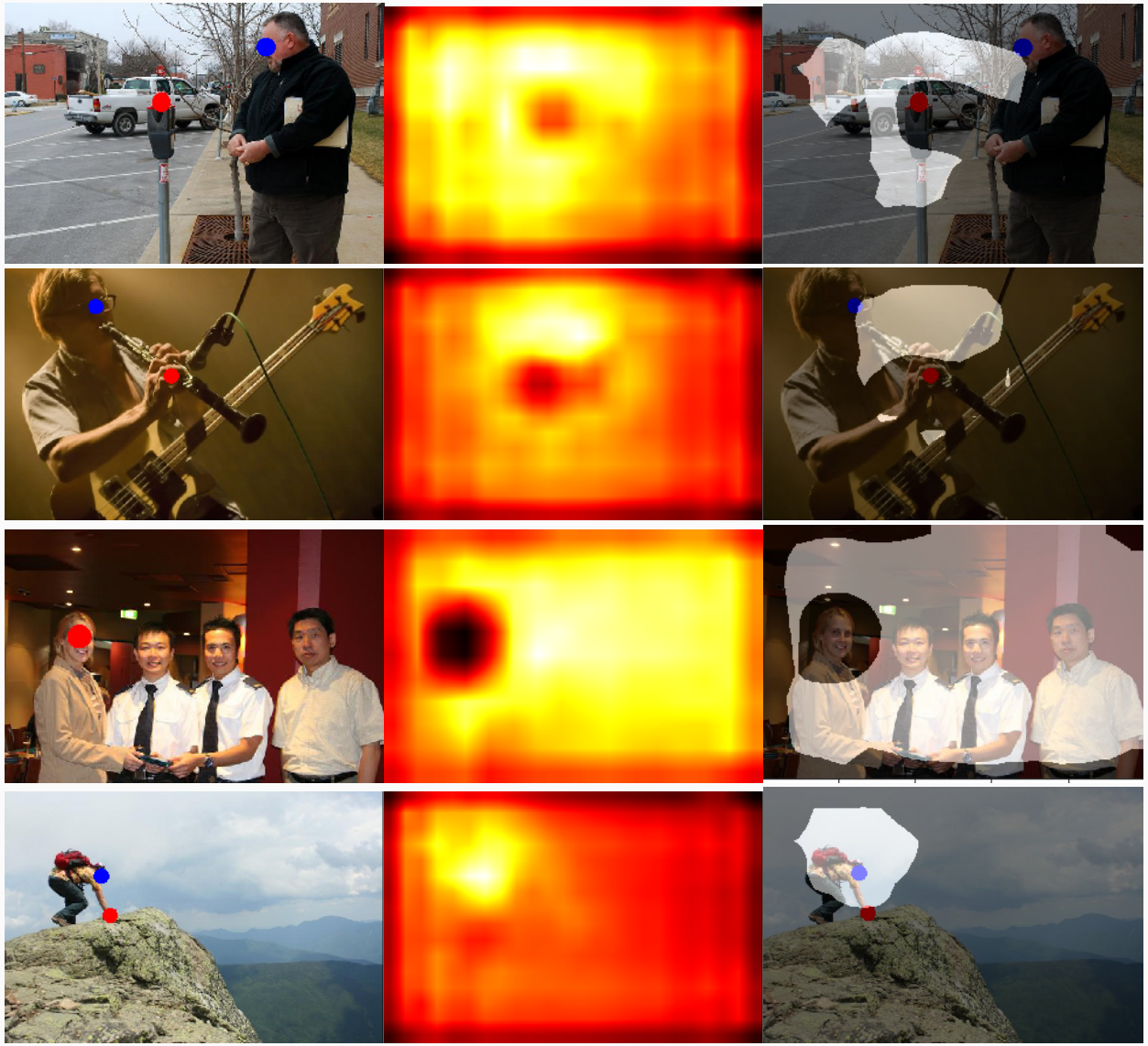
The Eyes network:

3. The Saliency-eyes branch
4. The Object branch

Gazeability of objects [5/6] – Visual results. Saliency-eyes branch

The branch searches for the heads of the observers.

Gazeability of objects [5/6] – Visual results. Object branch



The branch perfroms object gazeability estimation.

Gazeability of objects [5/6] – Quantitative results.

Separate assessment of the branches: 3 – Saliency-eyes, 4 – Object

Paths used				AUC (eyes)	AUC (gaze)	Dist (eyes)	Dist (gaze)	Acc (eyes)	Acc (gaze)
		3		0.964	0.580	0.223	0.404	0.52	0.15
			4	0.747	0.520	0.313	0.260	0.22	0.17
		3	4	**0.969**	0.606	**0.212**	0.409	**0.53**	0.13

- The saliency-eyes branch is dominant in the Eyes network

Gazeability of objects [5/6] – Quantitative results.

To check the impact of object gazeability, we use the DNN face detector from OpenCV.

Number of faces detected	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**more**
Proportion of images	**0.51**	**0.30**	**0.13**	**0.04**	**0.01**	**0.005**	**0.001**	**0.004**

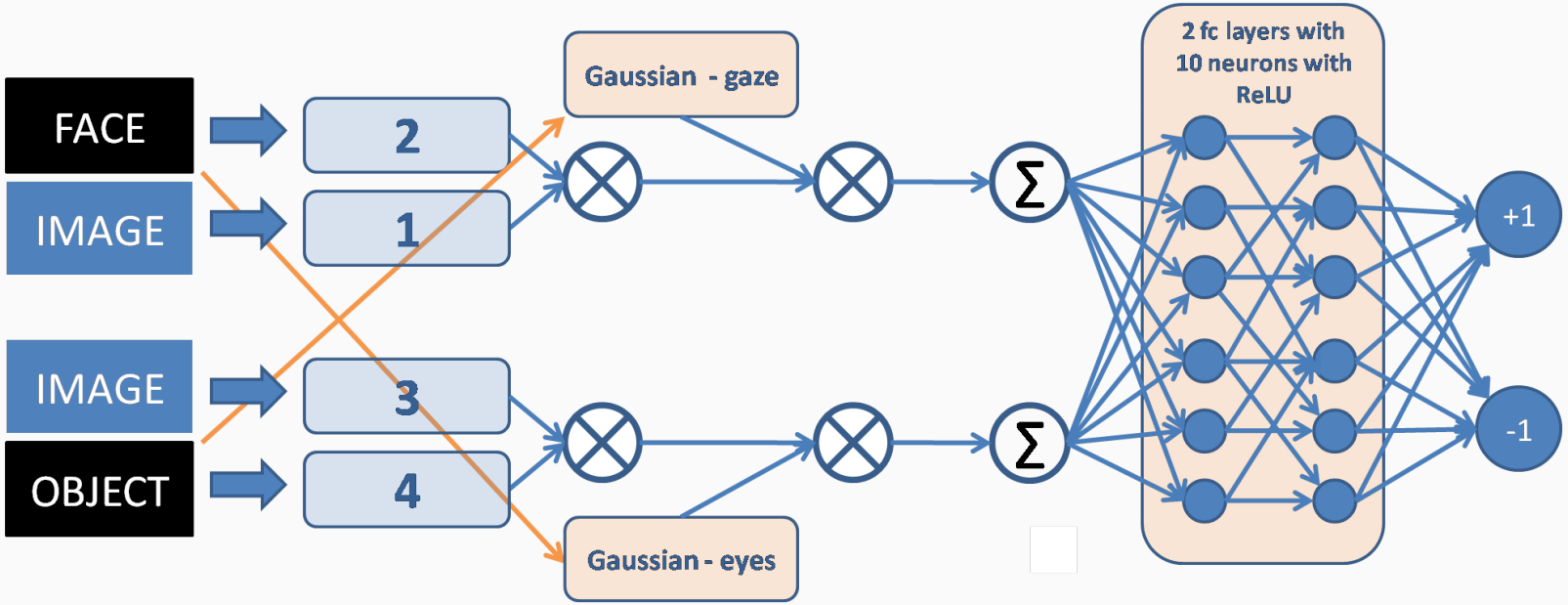
		Number of faces detected						
	Path used	0	1	2	3	4	5	6
Face Accuracy	only 3	0.69	0.86	0.59	0.39	0.32	0.25	0.26
	3 and 4	0.68	0.83	0.65	0.45	0.41	0.38	0.34

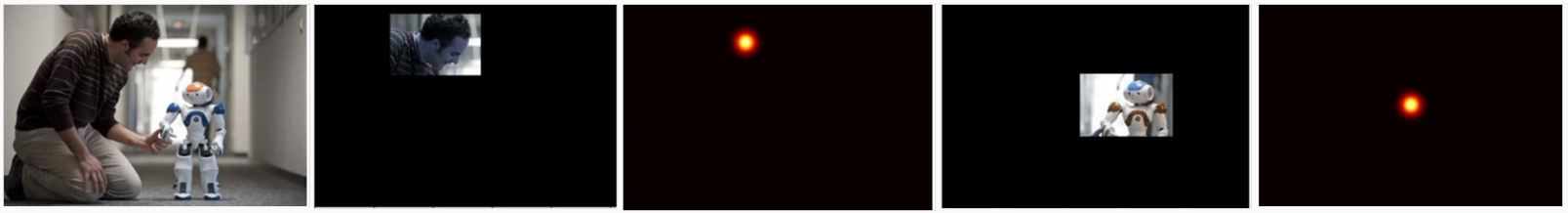
- When few people are present, it is enough to detect them
- When people are many, it is important to select the ones at a good position

Gazeability of objects [6/6] – Gazeability network.

How can we improve the quality of the gaze-following system?

- **New setting**: estimate the probability of the given person to look at the given object.





Gazeability of objects [6/6] – Gazeability network. Visual results

Binary classification problem:

- Positives: annotations from GazeFollow (blue and green)
- Negatives: head annotations from GazeFollow (blue), randomly generated gazed positions (red)

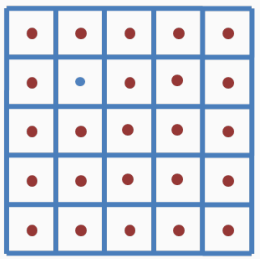
Gazeability of objects [6/6] – Gazeability network. Results

- The results of the network on the binary classification task:

	ROC AUC	Accuracy
Gazeability network	0.903	0.89
Simplified version (only Gaze network)	0.845	0.87

- The network can be used for gaze estimation. Predictions are the locations with the largest gazeability score for the given head.
- The results on gaze estimation:

	ROC AUC	Distance
Gazeability Network	**0.901**	**0.228**
The Gaze Network	0.893	0.233
The GazeFollow model	0.878	0.240



Gazeability of objects – Concluding remarks

- We managed to train a neural network to estimate the gazeability of a given object
- We designed a method to estimate the probability that a given observer looks at a given object
- Based on that solution, we improve the quality of the baseline on gaze-following

Outline of the presentation

- Introduction
- Study of Object Gazeability

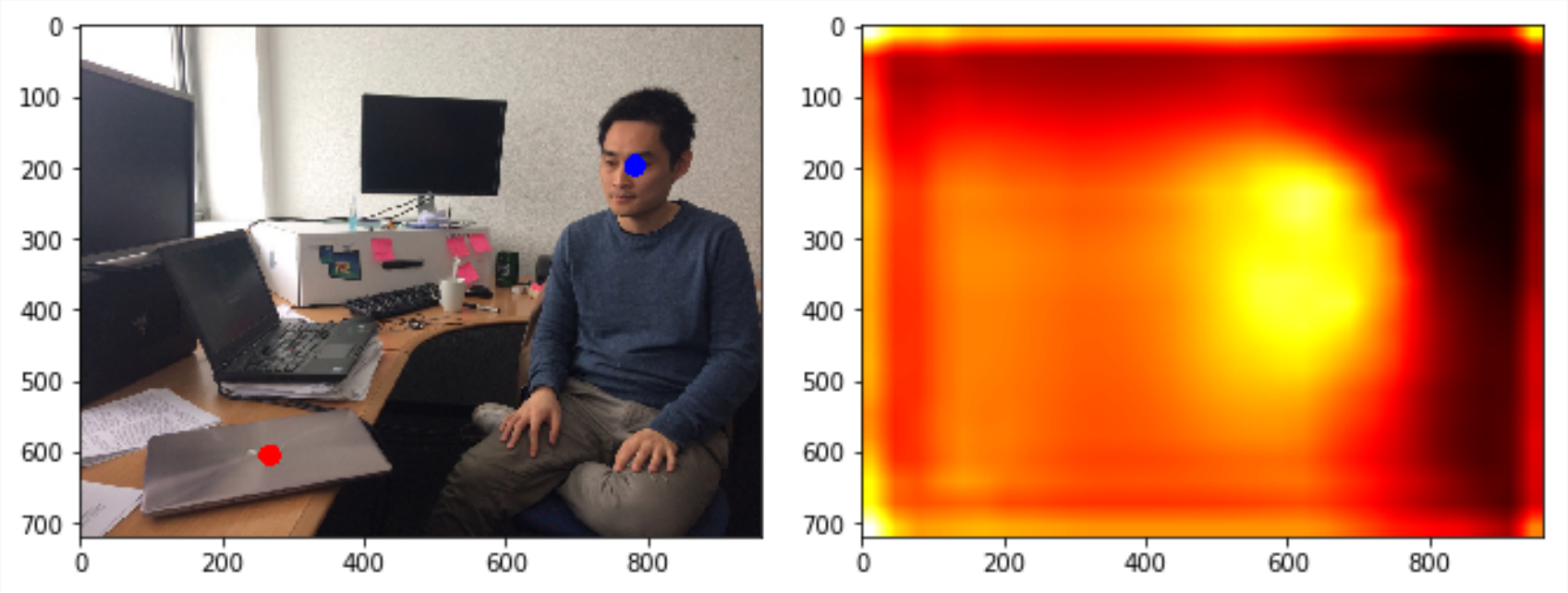
- You are here -

- 3D analysis of scenes for gaze prediction
- Conclusion

3D analysis of scenes for gaze prediction

3D analysis – Motivation

- 2D approaches may generate wrong predictions lying in the wrong plane of depth:



- We address this problem by introducing the 3D pipeline accounting on *depth extraction* and *3D head pose estimation*

3D analysis – Step 1. Depth extraction

- To build the 3D point cloud, we use the depth extractor proposed by [Lainaet al. 2017].





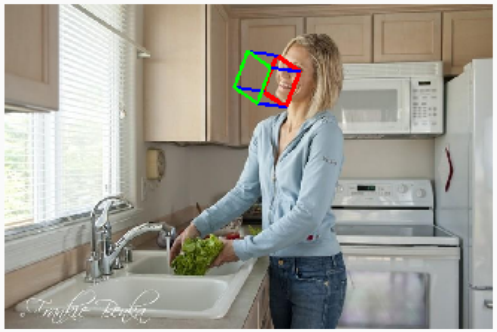
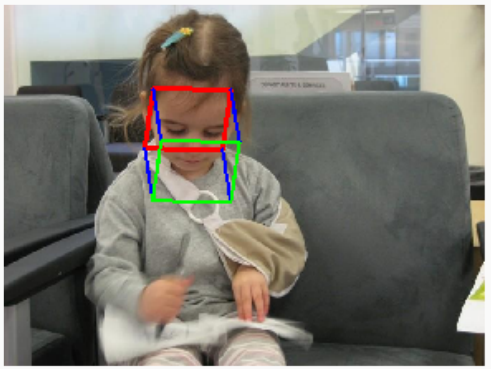



- Point coordinates corresponding to pixel X_i, Y_i:

$$x_i = \frac{X_i}{X_{\max}} \tan(\frac{65}{2}) D_i, \qquad y_i = \frac{Y_i}{X_{\max}} \tan(\frac{65}{2}) D_i, \qquad z_i = \sqrt{D_i^2 - x_i^2 - y_i^2}.$$

3D analysis – Step 2. 3D head pose

- To quantify the gaze cone of the observer in 3D, we use the 3D head pose estimator from [Ruizet al. 2018]
- The system returns the three Euler angles: yaw (ξ), pitch (ψ), roll (ρ).

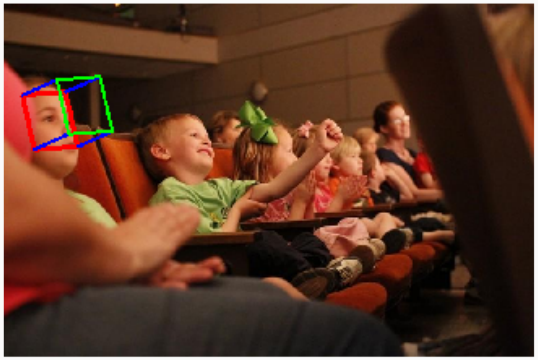
- The principal gaze direction is described by its directing vector:

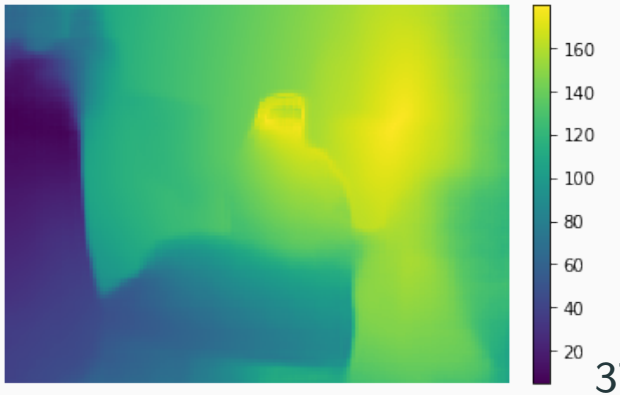
$$a_x = -\cos(\xi)\cos(\psi), \quad a_y = -\sin(\xi), \quad a_z = -\cos(\xi)\sin(\psi).$$

3D analysis – Step 3. Eyesight angles

- The deviation of each point of the point cloud from the center of the eyesight:

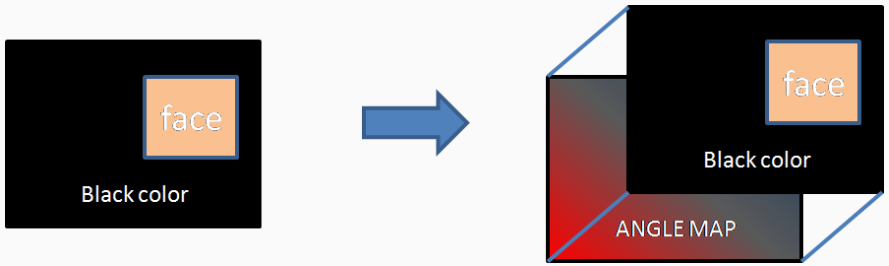
$$\phi_i = \arccos \frac{[a_x, a_y, a_z]^T \cdot [x_i - x_e, y_i - y_e, z_i - z_e]^T}{||[a_x, a_y, a_z]^T|| \cdot ||[x_i - x_e, y_i - y_e, z_i - z_e]^T||}.$$





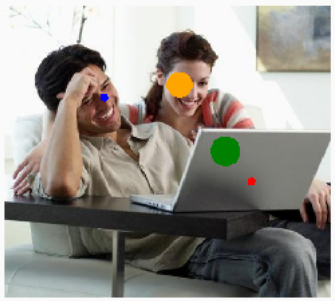
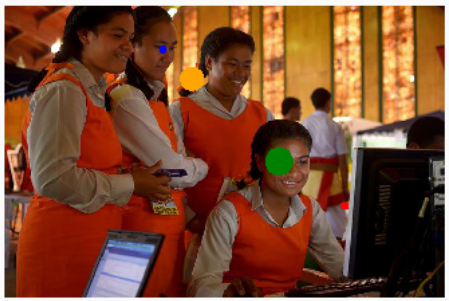
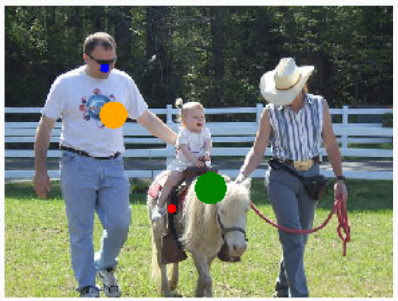
3D analysis – Results

- The computed eyesight angles can now be used as an additional feature:



- The network trained with depth information makes less mistakes:

Model	AUC (all images)	Distance (all images)	AUC (with detected faces)	Distance (with detected faces)
Only face image	0.891	0.233	0.908	0.217
Face image + depth	0.894	0.227	0.909	0.209
Face image + angles	-	-	**0.918**	**0.205**


Outline of the presentation

- Introduction
- Study of Object Gazeability
- 3D analysis of scenes for gaze prediction

- You are here -

- Conclusion

Conclusion

In this master thesis, we have studied gaze-following from the perspective of the scene understanding.

- A model accounting on gazeability of objects was designed,
- A 3D geometrical pipeline was proposed.

Moreover, the review of the literature and analysis of the existing approaches for gaze-estimation in videos was conducted, but was not included in the presentation.

Thank you for your attention!