



Deep Head Pose

Fast and robust gaze direction estimation in colour and depth images

Sankha Subhra Mukherjee, James Watt PhD scholar

Visionlab, Heriot-Watt University



Real problems in search of a solution

- ▶ Automatic sports video analysis
 - ▶ What did referee see?
 - ▶ Where is the player going to pass next?
- ▶ Gestures vs. actions
 - ▶ Gestures must have an intended recipient (Kendon et al.)
 - ▶ HRI: Is this person “commanding” the robot or simply waving his hand?
- ▶ Wide area tracking of people
 - ▶ People tend to go where they are looking (outdoors)
 - ▶ Deviations from this are interesting (anomalies)
- ▶ Intention of people mediated by gazing direction
 - ▶ Autonomous cars: has that pedestrian seen the oncoming vehicle?
 - ▶ Security: is this a threat or simply someone who is lost?

Prediction of intent - quickly

Machine may need to make “instantaneous, real-time” decisions

Head-pose is a stronger cue than trajectory in this case



Prediction of intent from a single image

Scenario: Vehicle approaching crossing with exit door opening immediately onto crossing. Footage gathered from low-res camera inside vehicle.



Intention not-to-cross mediated by the instantaneous gaze

Prediction of intent from a single image



Intention to cross could be inferred and signal used by vehicle (etc.)

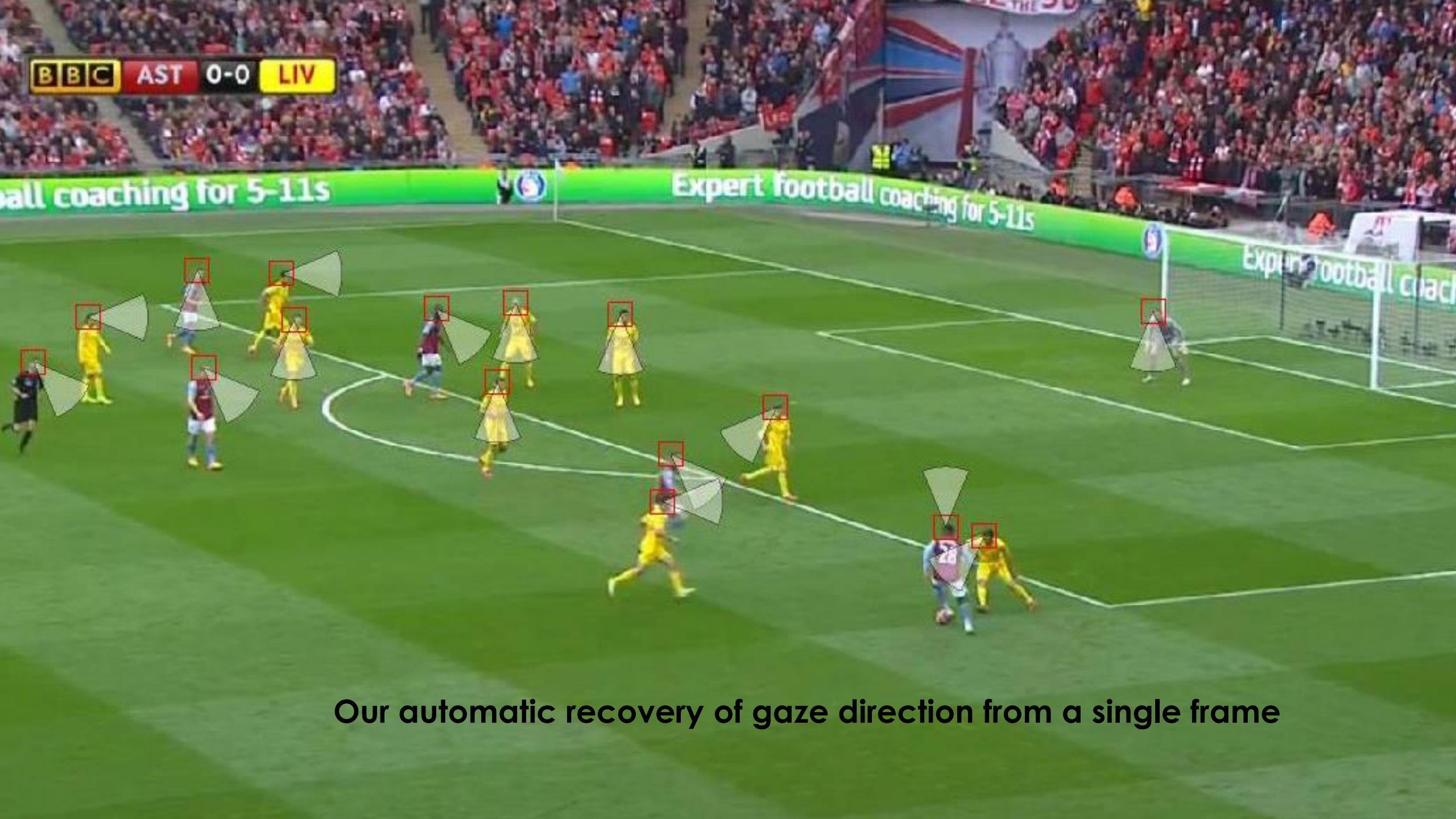
Penalty ... or not?

Has the referee seen this incident?





Actual manual annotation by BBC studio analyst as presented on MoTD



BBC AST 0-0 LIV

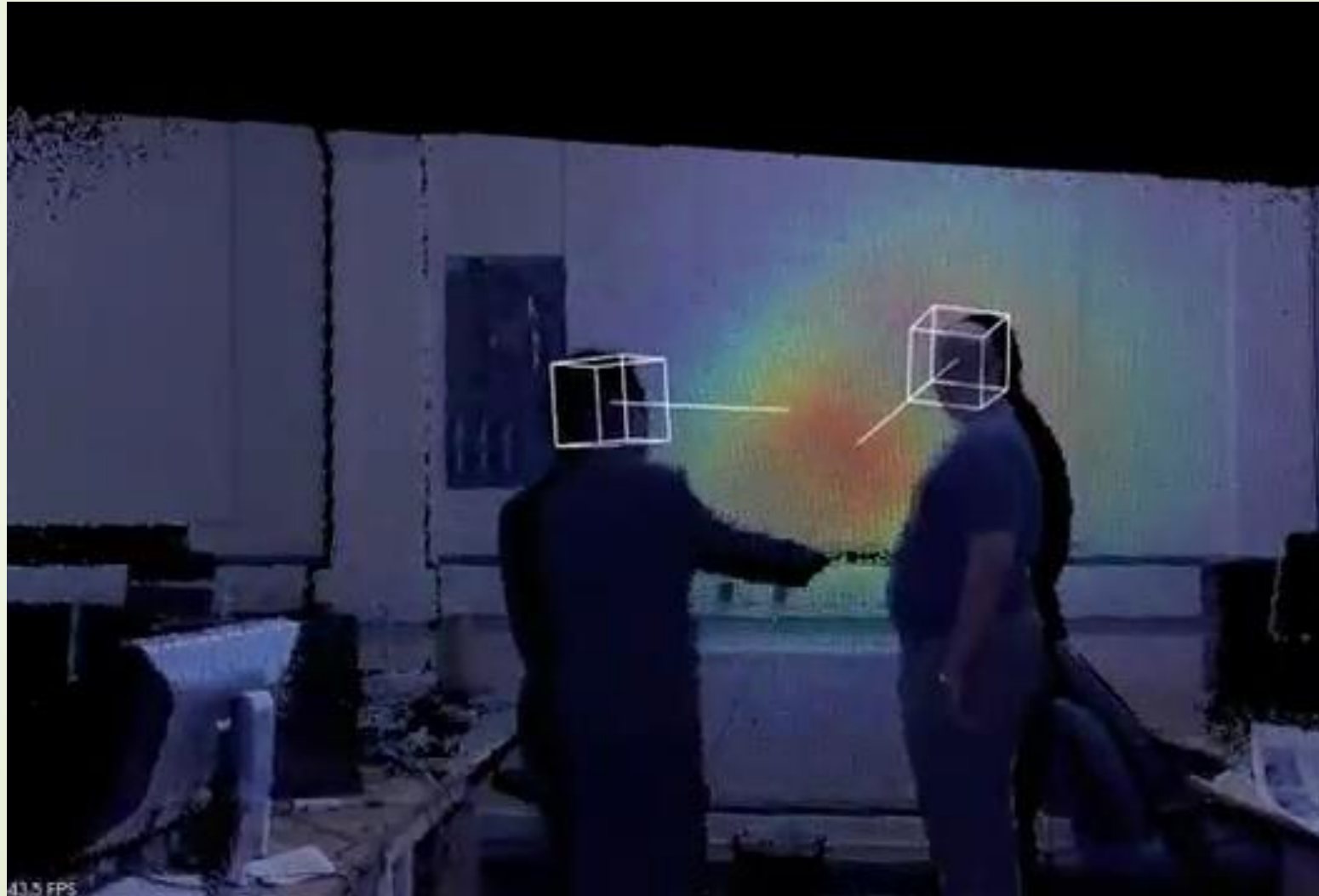
Expert football coaching for 5-11s

Expert football coaching for 5-11s

Expert football coaching for 5-11s

Our automatic recovery of gaze direction from a single frame

Human visual attention





Where are the gaps in the research?

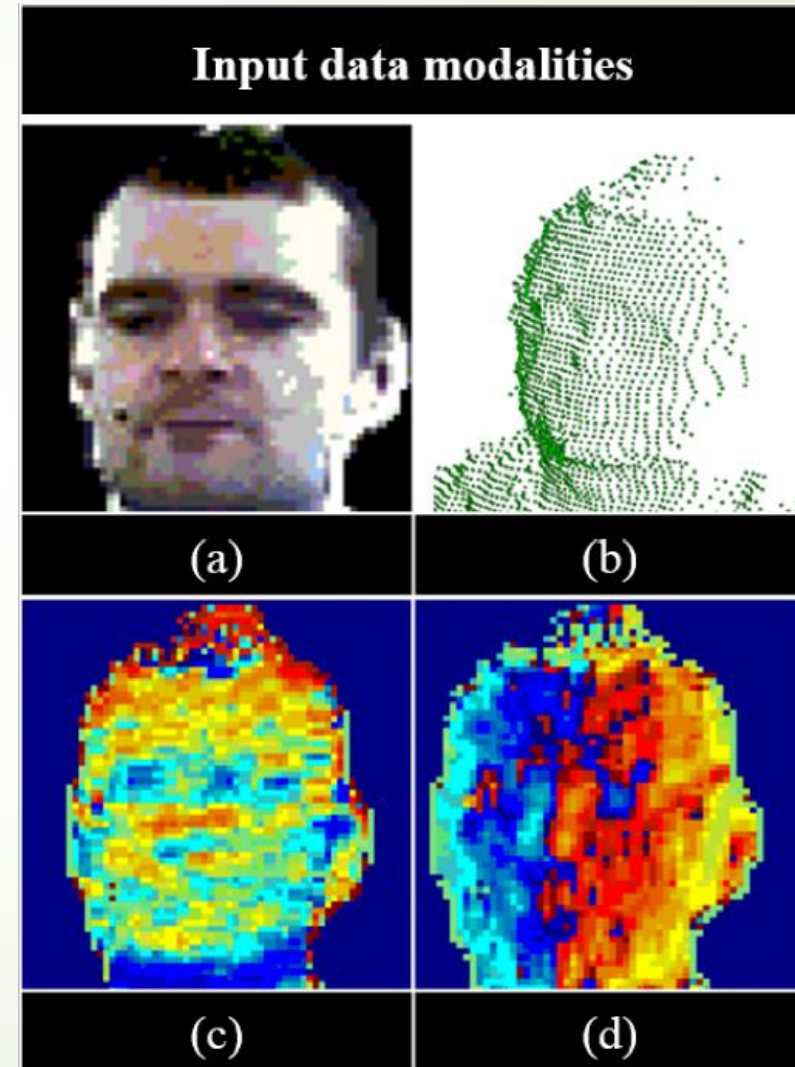
- ▶ Previous methods often fail if decision with single frame is required
 - ▶ Motion smoothing often not possible, or unreliable (Benfold & Reid)
 - ▶ Skin colour or local gradients are not highly separable into classes (Robertson et al.)
- ▶ HCI (high-res, depth) and surveillance (low-res, RGB) treated differently
- ▶ Computational complexity is a problem for leading methods
 - ▶ On-line adaptive Kernel methods (Chen et al.)
 - ▶ GPR (Marin-Jimenez et al.) ... are too slow
- ▶ **More robust and faster feature extraction needed**

Classification of gaze by ConvNets

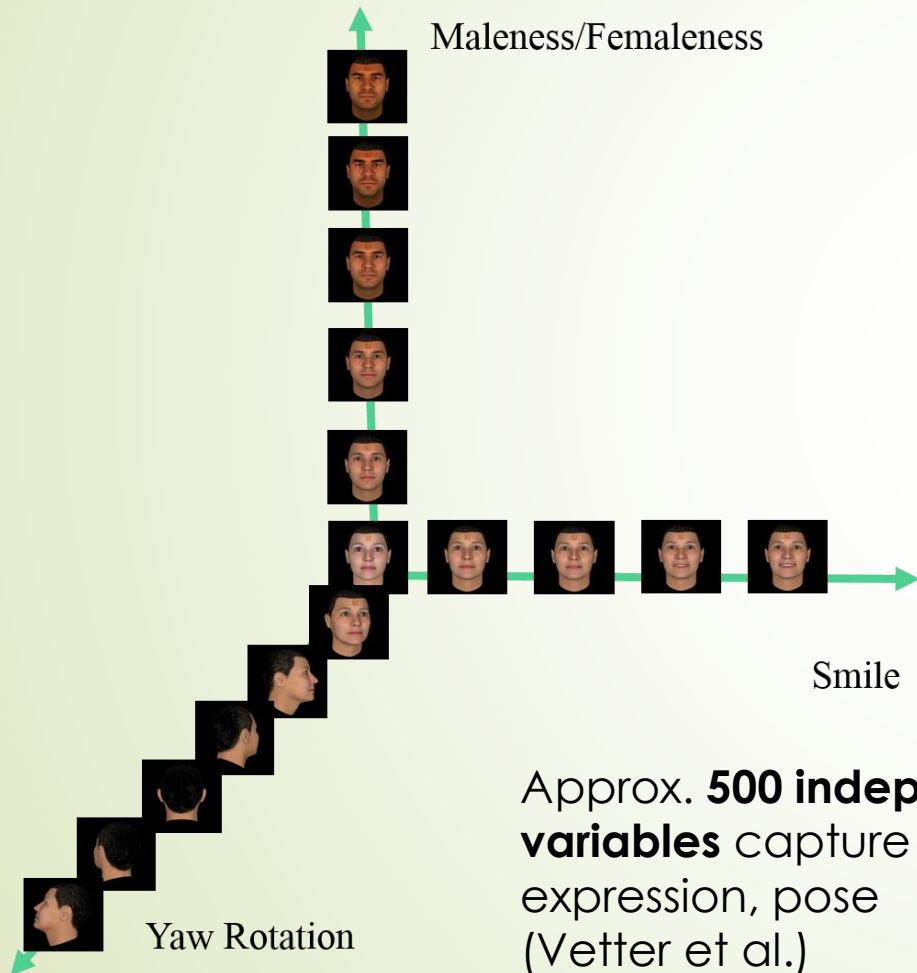


Data sources (e.g. Kinect 2)

- (a) RGB
- (b) Depth point cloud,
- (c) surface normal **E**levation angle
- (d) surface normal **A**zimuthal angle



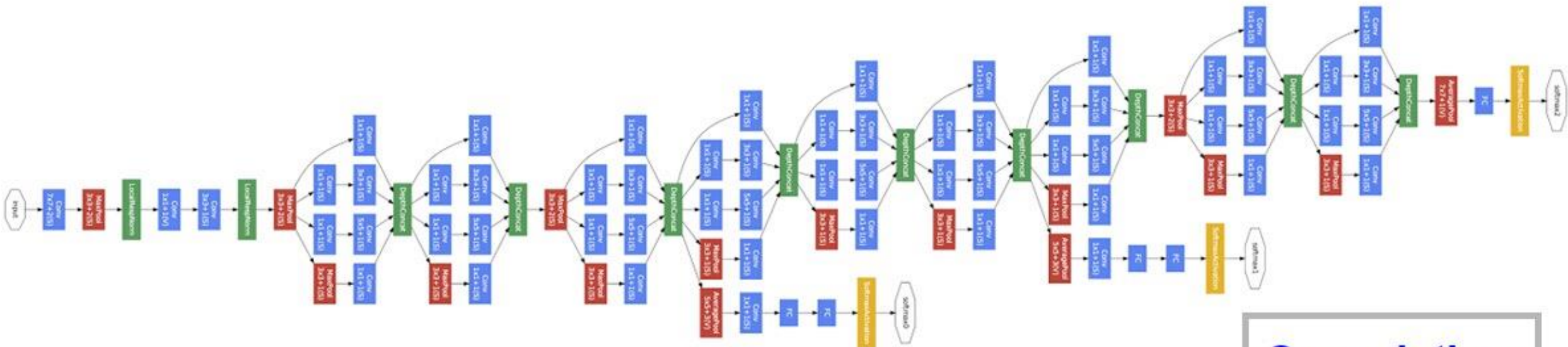
Parametric human head space



Approx. **500 independent variables** capture identity, expression, pose (Vetter et al.)

- Challenges involved in training the CNN
 - Define the architecture
 - Combine RGB & D
 - Classification to regression
 - Best models may overfit e.g. VGG16 (140m parameters)
 - Use a deep network with fewer params
 - GoogleNet (5m)

The GoogLeNet CNN architecture



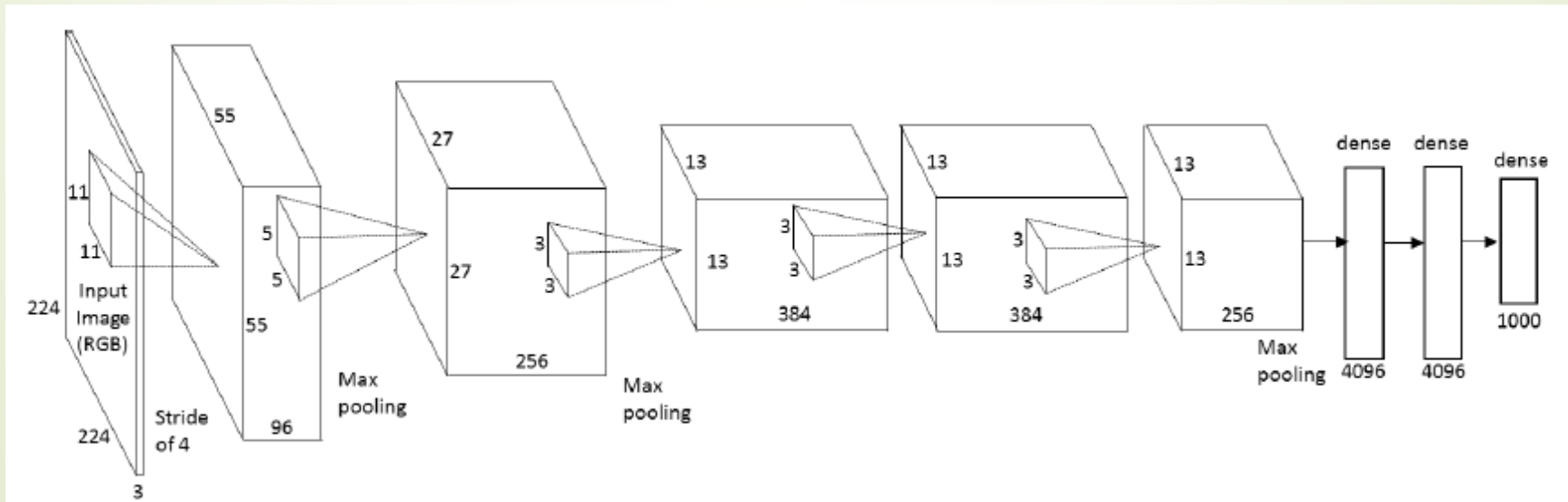
~5 million
parameters

<http://arxiv.org/abs/1409.4842>

Convolution
Pooling
Softmax
Other

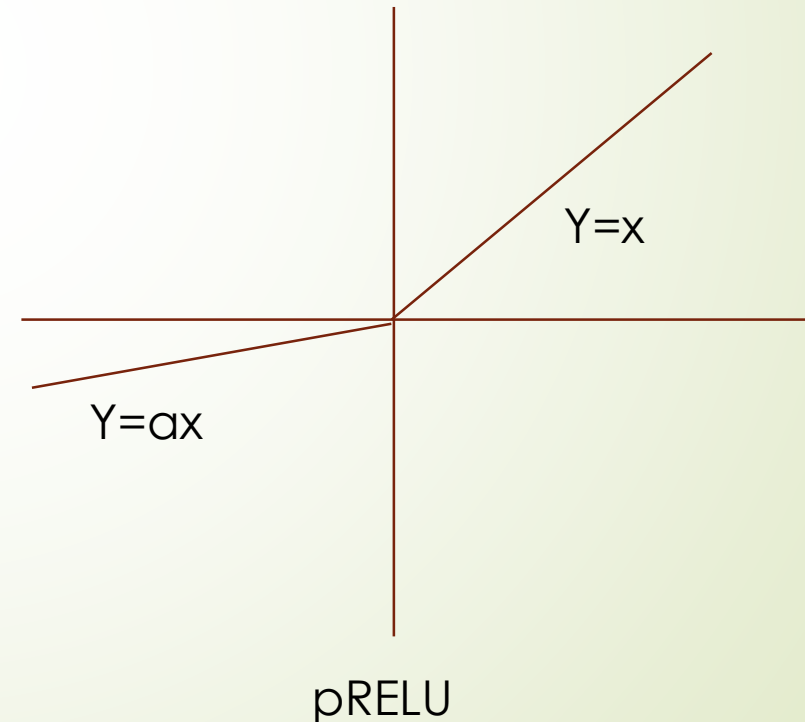
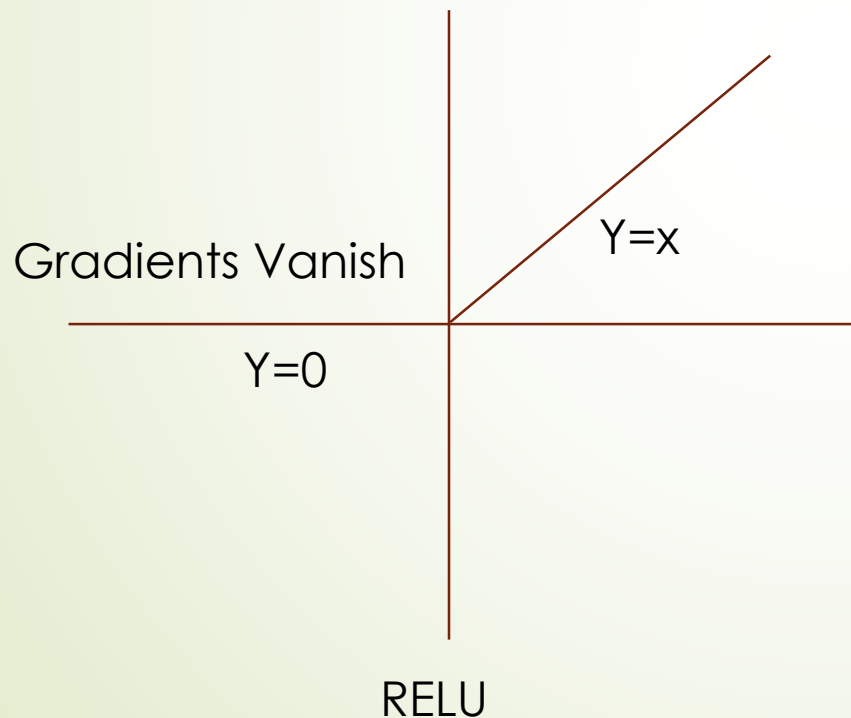
Building blocks

- Convolution layers - translational invariance
- Downsampling/Pooling (Average, Max, Stochastic etc.)
- Nonlinearities (via Sigmoid, Tanh, Rectified Linear etc.)
- Fully connected layer(s)
- Loss functions to train the network using back propagation



Modifications

- Recent results suggest RELU non-linearities may hamper learning if the activations of neurons go to zero.
 - prevent this problem with Parametric RELU (Microsoft research, 2014)
- Gives us ~2.5 % pt increase in performance**



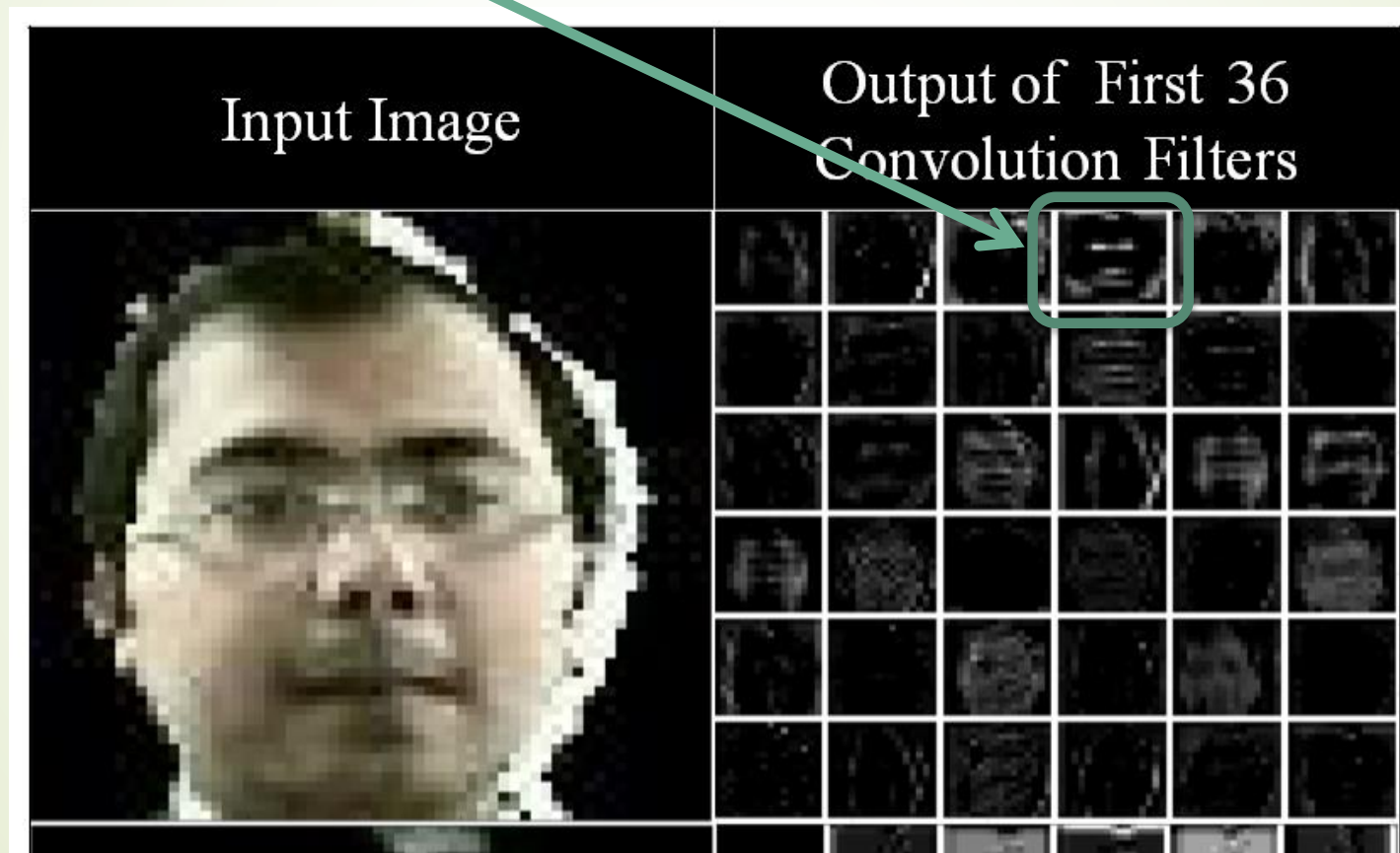


Multimodal combination

- ▶ **Early fusion of Depth and RGB does not work**
 - ▶ The network can not propagate meaningful gradients to different modalities at the same time (see also Gupta 2014)
- ▶ **Late fusion** i.e. compute a final model average from two independent CNNs
- ▶ Statistics:
 - ▶ Training: 3 days on commercial GPU (Tesla K40)
 - ▶ Training samples: ~125,340
 - ▶ Tuning: 2 days
 - ▶ Testing: 20 heads per frame at 25 fps

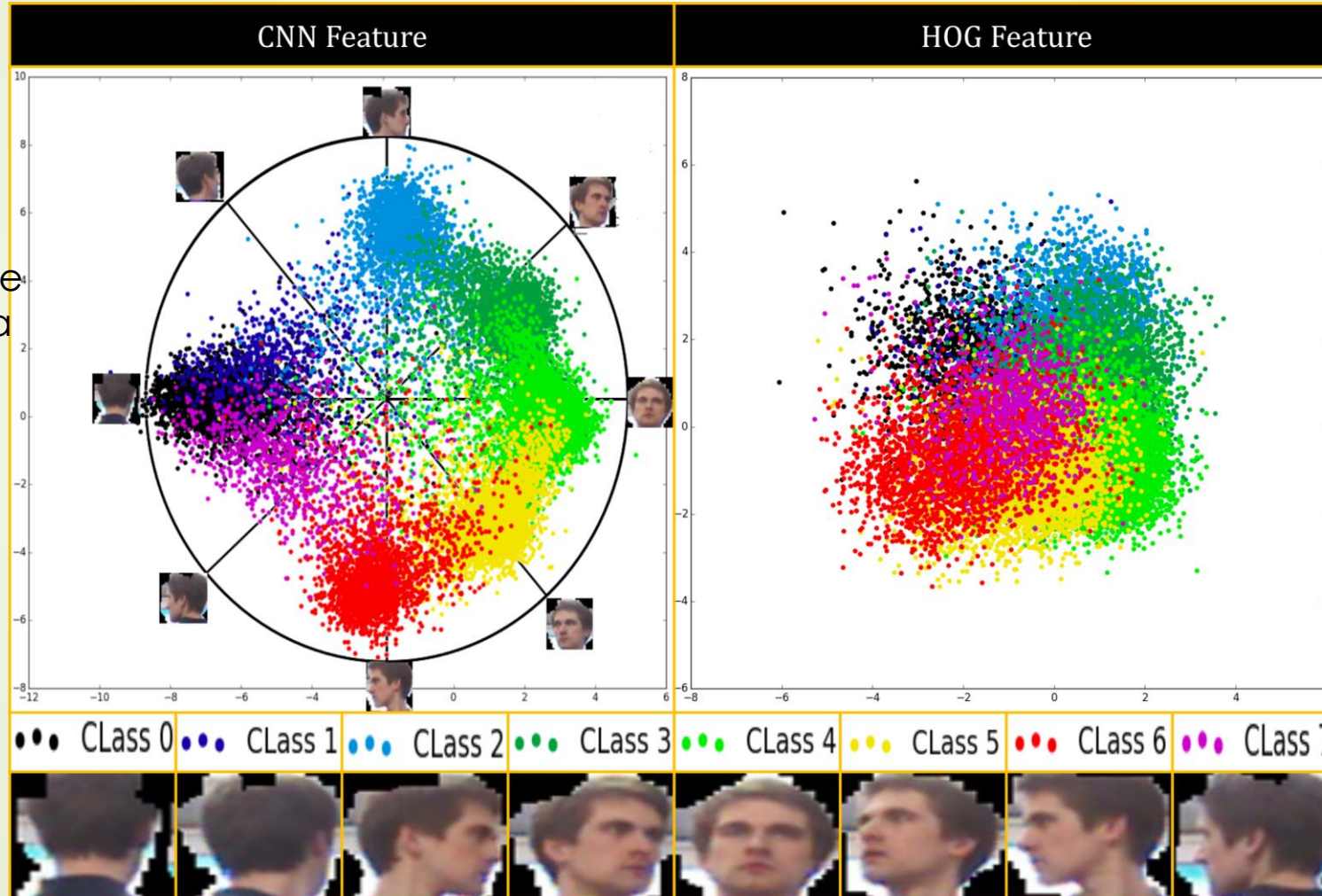
Learned network features (RGB)

- ▶ Skin map, landmarks (eyes nose) ... plus many more ...



How good are these features?

LDA scatter plots of the high-dimensional data



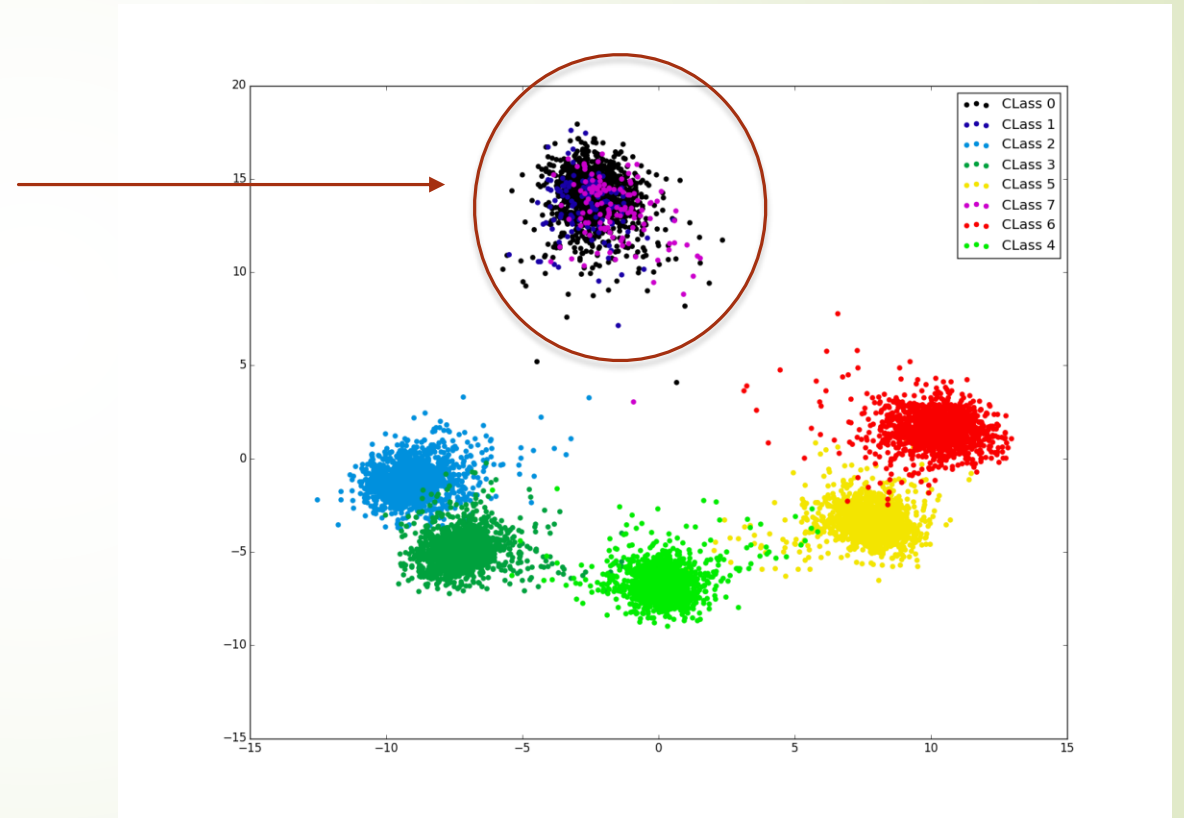
HOG features are clearly a poorer choice (as shown in the LDA projection of the high-dimensional data)

The learned CNN classes reveal the underlying circular manifold of the data

The depth CNN features

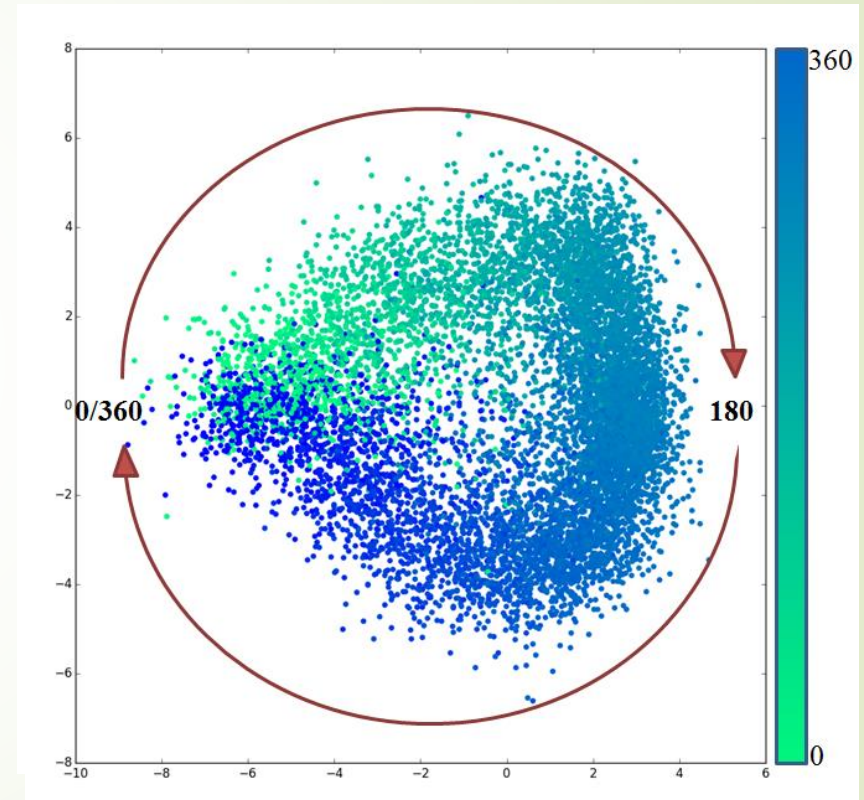
Classes towards the back of the head are much more uncertain in depth.

This exposes the limitations of the sensor



From Classification to Regression

- ▶ Linear manifold L2 loss regression destroys the topology of the features.
 - ▶ Pose lies on a circular manifold.
- ▶ Provide the Euclidean vector components of the head-pose vector instead of the raw angles.
- ▶ For recovering both yaw and pitch angles we pose the regression problem using (X, Y, Z) of the 3D unit pose vector.



circular manifold recovered

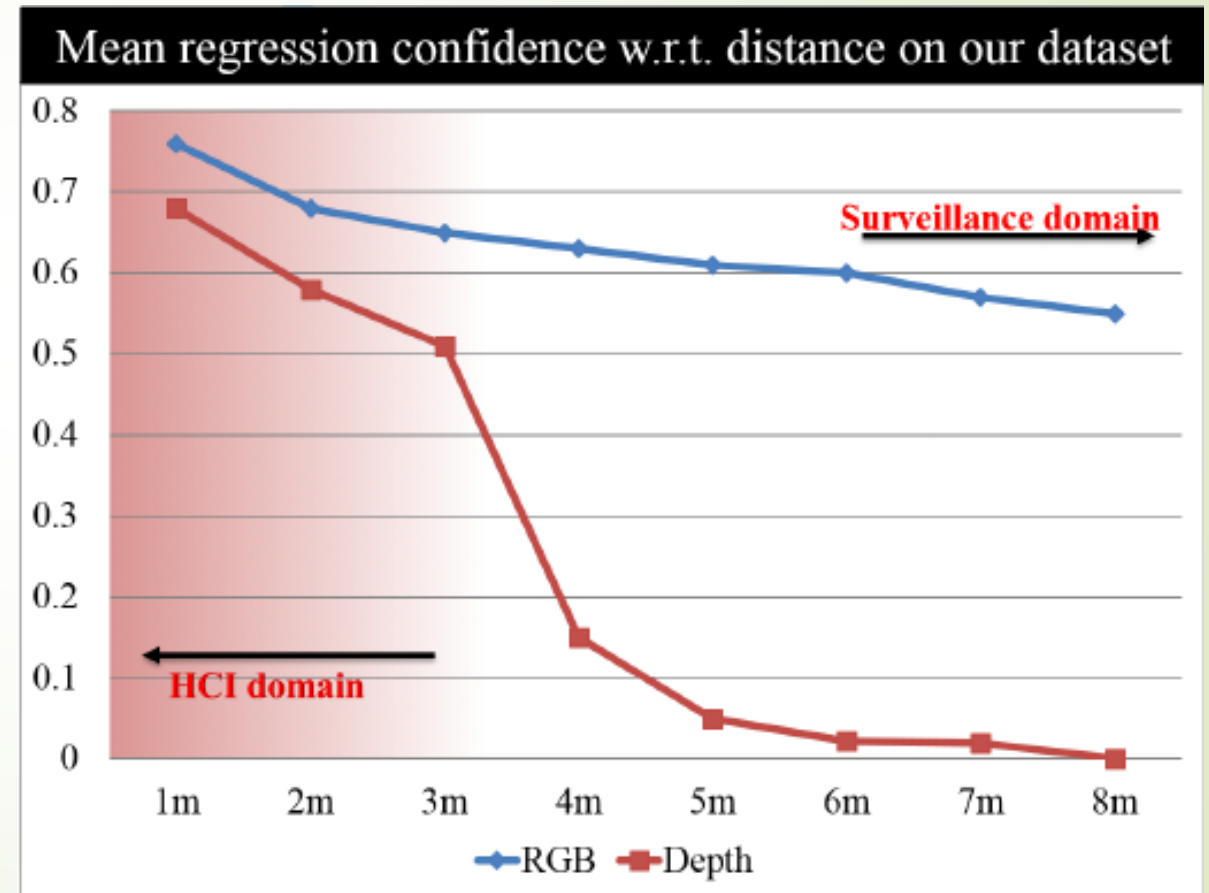



30% off

IC
W

Regression confidence estimate

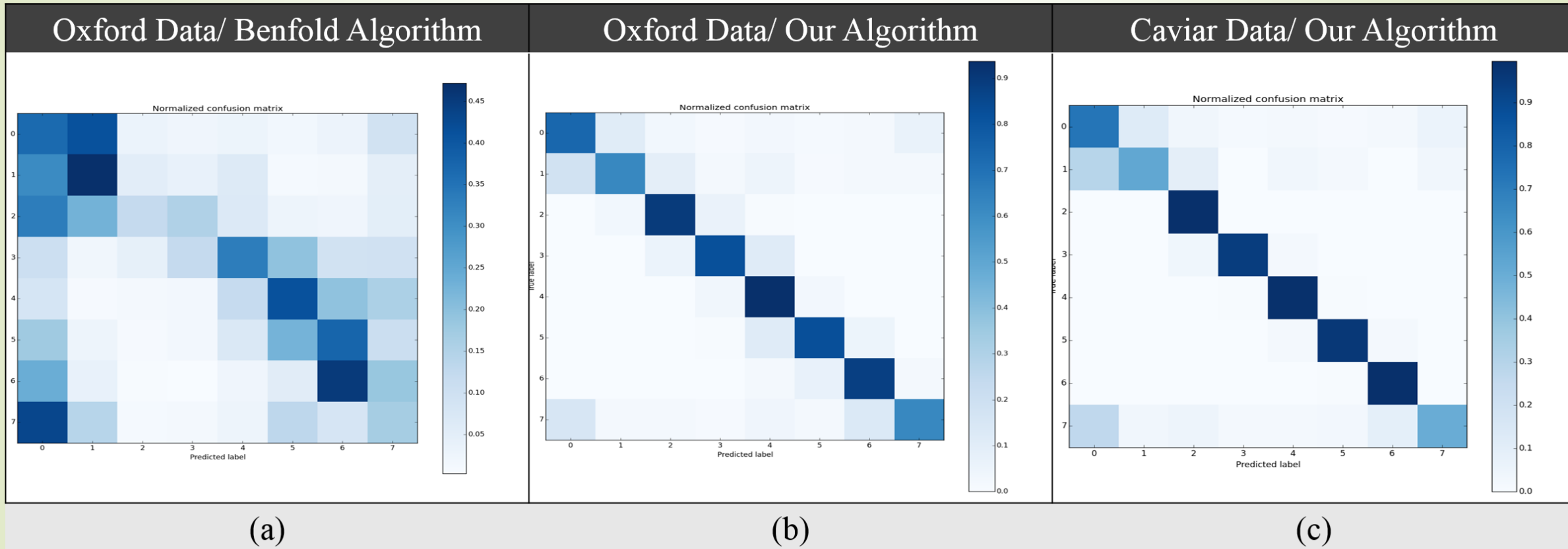
- ▶ We estimate regression confidence by computing the variance of the classifier outputs which gives a PDF over pose classes
- ▶ This classifier is a 360 class (one per degree) SOFTMAX classifier learnt on top of the regression features





Results

Results on RGB low-resolution datasets

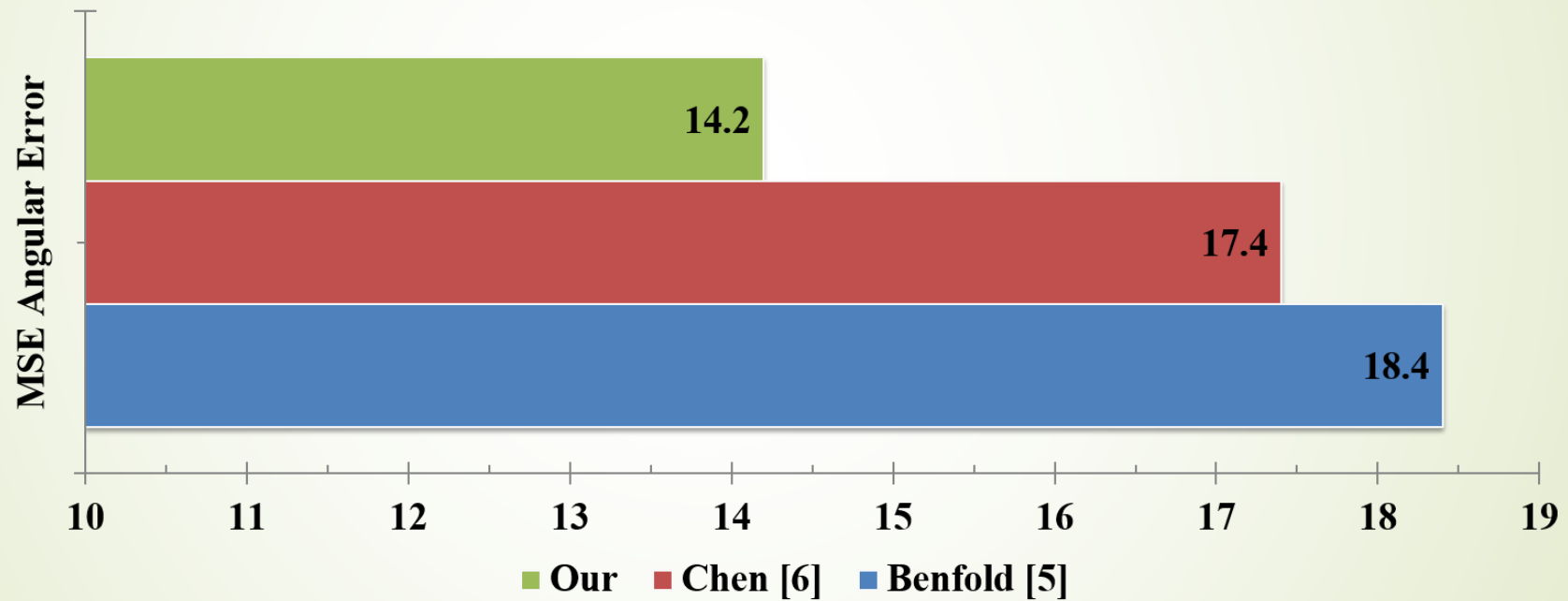


Sparser is better.

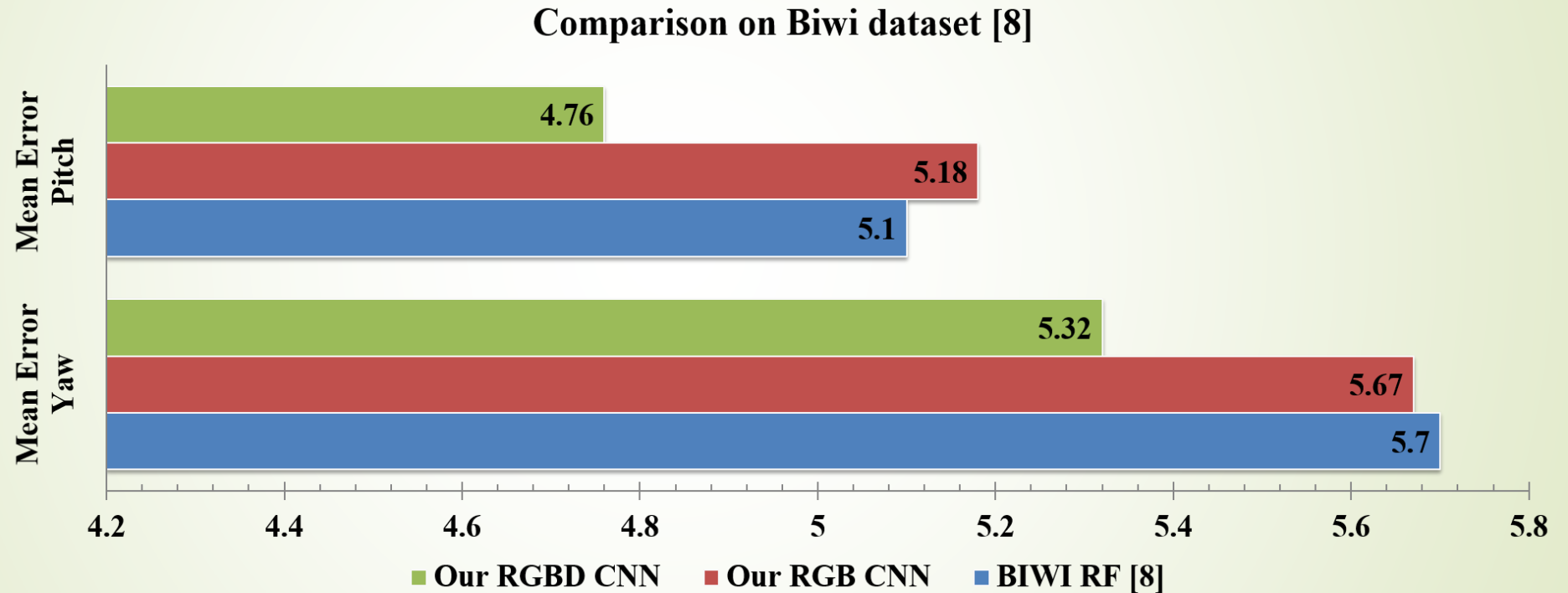
(a) previous (b) ours (c) ours

Our results (RGB) vs best published

Comparison on Oxford Dataset [5]



Our results on multi-modal data



In the HCI, high resolution domain, multiple modalities do improve accuracy



Our contributions

- ▶ New theory for feature computation
 - ▶ Unified the HCI and surveillance head-pose classification problem
 - ▶ Achieved regression on pose via deep learning (CNNs)
 - ▶ From 8 classes (of 45 degree bins) to 360-degree confidence
- ▶ Better and faster
 - ▶ “Instantaneous” classification e.g. from single frame
 - ▶ 20 heads classified at 25fps
 - ▶ We do not rely on body pose or motion as smoothing priors on the head pose
- ▶ New applications
 - ▶ Improved interaction metrics for human-human interaction in video
 - ▶ We show how to exploit the feature for better tracking in surveillance



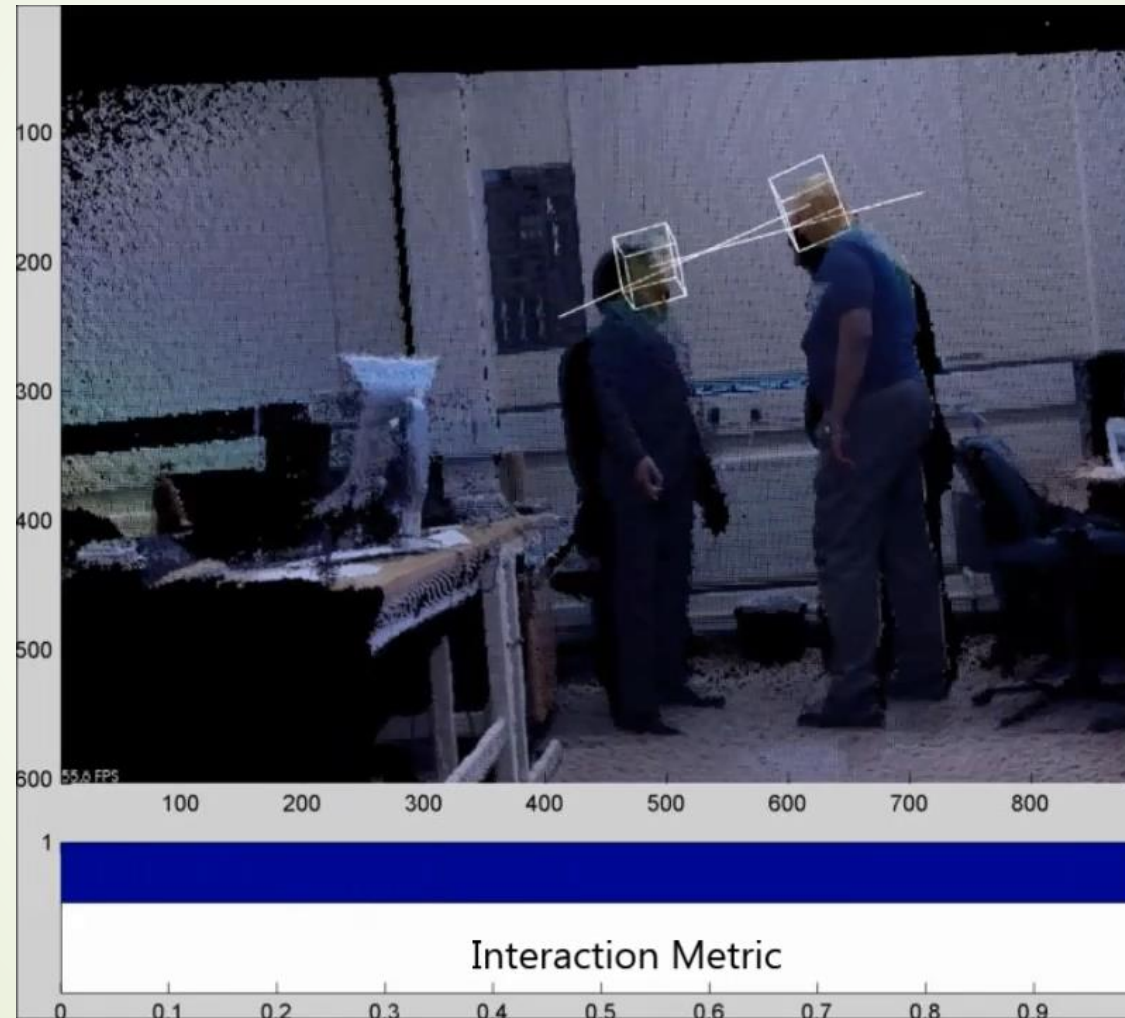
Applications



1. Human attention modelling

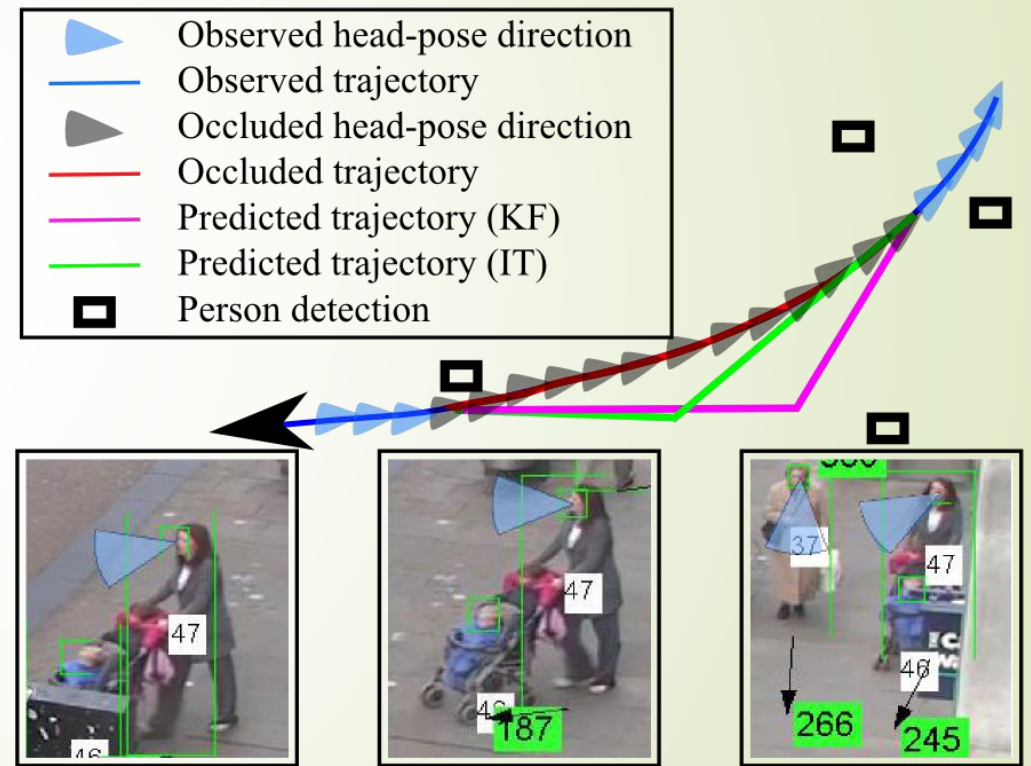
- ▶ Model the attention using a spherical normal distribution (Von Misses).
- ▶ Set the confidence estimate of the model to peak or diffuse the distribution accordingly.
- ▶ We may project the attention heat map to a 3D scene
 - ▶ In real-time
 - ▶ Aggregate for change detection
- ▶ Extracted Signals
 - ▶ Looking at each other (Pair wise Interaction probability)
 - ▶ Looking at the same thing (Windowed cross correlation)
 - ▶ Visual attention "heatmap" in 3D

Interaction metric computation



2. Intentional tracking

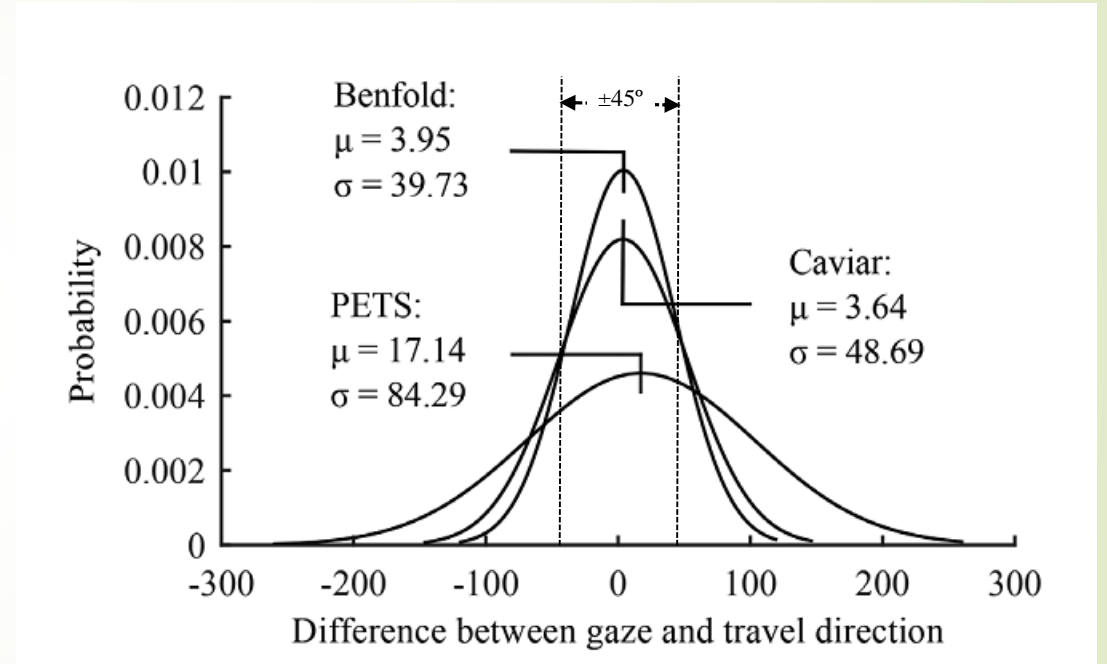
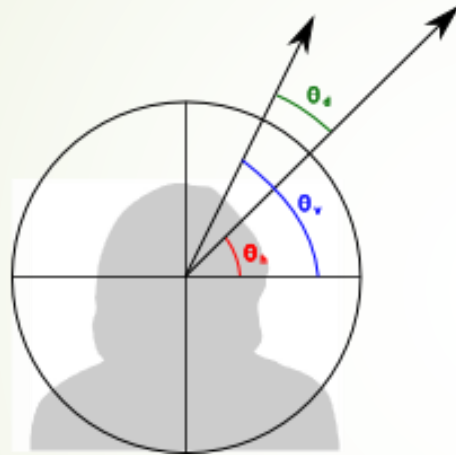
- ▶ Head pose can provide both spatial and social context
- ▶ Some gazing patterns are anomalous only in certain places
- ▶ Hypothesis:
 - ▶ We can use head-pose to build better person trackers
 - ▶ Focus initially on track discontinuities
 - ▶ Change in direction



Motivating example

The signal derived from gaze

- θ_h Head pose direction
- θ_v Body velocity direction
- θ_d Deviation

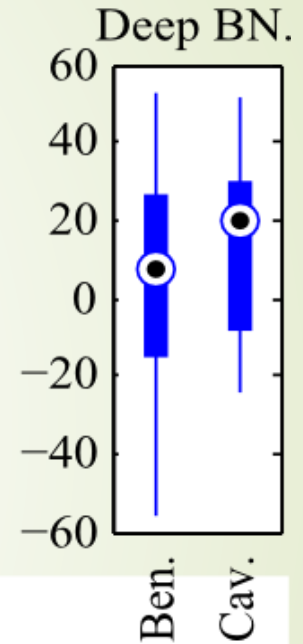


We extract the deviation from heading

Most people do look where they are going.
Deviations are interesting - they help us track better

With automatic head pose

- ▶ On real data:
 - ▶ Oxford town centre (Ben) + 7.2% in LL
 - ▶ CAVIAR (shopping mall) + 19.5% in LL
- ▶ Out performs standard KF under all conditions
 - ▶ Even vs. learned KF parameters via EM
- ▶ Tracking metrics are more informative
 - ▶ 63% reduction in MSE cf. ground truth





Where next?



Future work

- ▶ New theory: Multimodal signal processing
 - ▶ Multi-task Fully Convolutional Neural Networks (FCNN) for detection and classification map of arbitrary size in one pass through the network.
 - ▶ Distil the knowledge from the soft-targets of the already trained headpose networks into the FCNN during training. Gain similar performance at a fraction of the computational cost.
 - ▶ Joint audio-video classification of people and behaviours through CNN-LSTM
 - ▶ Exploit networks of sensors
- ▶ Novel hardware platforms: real-time adaptive deep learning
 - ▶ Embed CNN and Recurrent Neural Networks on FPGA for smart camera networks e.g. autonomous vehicles
 - ▶ Current EPSRC shortlisted proposal with Queen's Belfast, IBM, Xilinx
- ▶ Direct application:
 - ▶ Security, surveillance, human behaviour inference
 - ▶ Social scene understanding
 - ▶ Human robot interaction



Related recent publications

- ▶ Deep Head Pose: gaze-direction estimation in multimodal video, S.Mukherjee, N.M.Robertson, IEEE Trans. Multimedia (in press), 2015
- ▶ An adaptive motion model for person tracking with instantaneous head-pose features, R.Baxter, M.Leach, S.Mukherjee, N.M.Robertson, IEEE Signal Processing Letters, 2014
- ▶ Instantaneous real-time head-pose at a distance, S.Mukherjee, R.H.Baxter, and N.M.Robertson, IEEE Int. Conf. Image Processing, 2015
- ▶ Tracking with intent, R.Baxter, M.Leach, N.M.Robertson, Proc. IEEE Conf Sensor Signal Processing for Defence, 2014
- ▶ Detecting Social Groups in Crowded Surveillance Videos Using Visual Attention, M.V. Leach, R.Baxter, N.M.Robertson, E.Sparks, IEEE Int.Conf. Computer Vision and Pattern Recognition Workshops, Workshop on Computational Models of Social Interaction and Behaviour, 2014