

# Deep Head Pose: Gaze-Direction Estimation in Multimodal Video

Sankha S. Mukherjee and Neil Martin Robertson, *Senior Member, IEEE*

**Abstract**—In this paper we present a convolutional neural network (CNN)-based model for human head pose estimation in low-resolution multi-modal RGB-D data. We pose the problem as one of classification of human gazing direction. We further fine-tune a regressor based on the learned deep classifier. Next we combine the two models (classification and regression) to estimate approximate regression confidence. We present state-of-the-art results in datasets that span the range of high-resolution human robot interaction (close up faces plus depth information) data to challenging low resolution outdoor surveillance data. We build upon our robust head-pose estimation and further introduce a new visual attention model to recover interaction with the environment. Using this probabilistic model, we show that many higher level scene understanding like human-human/scene interaction detection can be achieved. Our solution runs in real-time on commercial hardware.

**Index Terms**—Convolutional neural networks (CNNs), deep learning, gaze direction, head-pose, RGB-D.

## I. INTRODUCTION

MODELING human head pose is a challenging problem in computer vision and signal processing. It is desirable because this headpose signal gives us important meta-information about communicative gestures [1], salient regions in a scene based on focus of attention [2], group detection, crowd behavioral dynamics and tracking [3], and anomaly detection. The grand aim of our work is to exploit the advanced signal acquired from head pose to achieve, what is called, “Social Signal Processing”. In domains where close level iris/eye tracking is not possible, human head pose is the most important feature in estimating human focus-of-attention. Head pose estimation has been studied in two separate and distinct domains, visual surveillance [4]–[7] and Human Computer Interactions (HCI) [8]–[10] with different methodologies required due to the difference in the quality of the input. In this work we develop a

Manuscript received April 21, 2015; revised September 08, 2015; accepted September 17, 2015. Date of publication September 28, 2015; date of current version October 20, 2015. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/K014277/1 and by the MOD University Defence Research Collaboration in Signal Processing. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Benoit Huet.

The authors are with the Visionlab at the Edinburgh Research Partnership in Engineering and Mathematics, Heriot-Watt University, Edinburgh EH14 4AS, U.K., and the University of Edinburgh, Edinburgh EH8 9YL, U.K. (e-mail: sm794@hw.ac.uk; n.m.robertson@hw.ac.uk).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2015.2482819

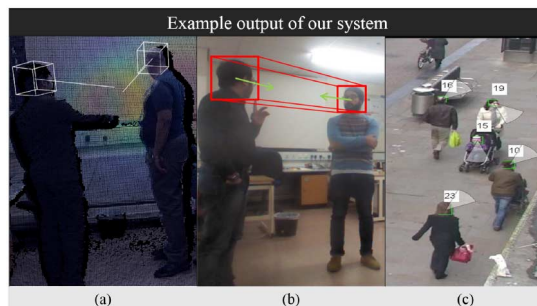


Fig. 1. Illustrative outputs of our system showing gazing direction estimation across a range of imaging modalities, resolutions, and applications: (a) visual attention modeling in the 3-D environment; (b) human-human and human-machine interaction recognition; and (c) gazing direction in low-resolution surveillance video, which may ultimately be used for tracking and anomaly detection.

new technique which unifies these research areas and exploits the multiple modalities of range images and colour images when it is beneficial so to do. The method is also highly robust and fast to compute as we demonstrate on data from both domains.

In the surveillance domain the nature of the problem requires the exploitation of priors such as walking direction [6] to augment the low resolution visual features. In the close range (i.e. higher resolution) domain of HCI, facial landmark detection approaches are employed for better accuracy [8]. However in HCI, the problem has been formulated with natural user interaction in mind, i.e., the user is always facing (near frontal) the sensor and is fairly close by (not more than 1-2 meters). Facial landmark based techniques typical of HCI cannot perform unconstrained head pose estimation at a distance. Furthermore, in most indoor interaction scenarios, the subjects are static and can be frequently occluded.

Hence the priors such as motion direction, body direction that are easily exploitable in a surveillance scenario may not be useful. Our focus in this paper is to introduce a system that addresses these issues by estimating the unconstrained head-pose by using a unified approach. Fig. 1 shows some illustrative outputs of our method.

### A. Motivation

Although the problem of head pose in two domains of HCI and surveillance have been solved with very different techniques, the underlying data is the same. As shown in Blanz *et al.* [11] human heads lie in a high dimensional manifold. See Fig. 2.

Any image of the head can be used to estimate parameters in this manifold. Occlusions such as hair, accessories like glasses,

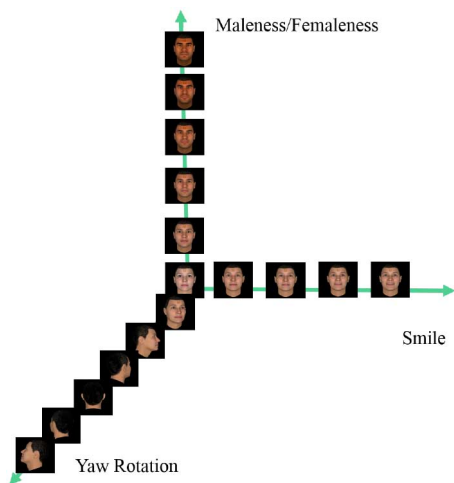


Fig. 2. Conceptual diagram showing different parameters controlling the appearance of the human head. The headpose can be considered as dimensions in this parametric space.

and/or low resolution make accurate estimation difficult. However since pose has a very high variance in the feature space, and thus a large eigenvalue principal component, this should allow us to recover the head pose in a holistic manner that spans the range of HCI to surveillance. Furthermore, in the surveillance domain the techniques rely on motion priors like walking direction to smooth the head pose estimation [5], [6]. This is valid only because most people *tend to look where they are going*. The drawback of such smoothing is that the information of the head pose signal itself is attenuated. As shown by Baxter *et al.* in [3] the cases of actual interest are when people deviate from this behaviour (i.e. look somewhere else). This information could be useful for anomaly detection or improving tracking and should not be smoothed out by a prior simply in order to achieve a more accurate tracking metric. Similarly, in the HCI domain, most techniques rely on the detection of facial landmarks. This is a valid assumption given the use-case scenarios. However this leaves a large gap in the applicability of such methods when it comes to achieving a reliable head-pose estimation in close range for non-frontal head pose. In summary, we address these problems by presenting the theory and implementation of a new technique that:

- 1) exploits high-resolution to low-resolution imageries and exploits multiple imaging modalities, i.e. RGB and depth where possible;
- 2) is independent of explicit facial landmark detection; and
- 3) does not require motion priors (“instantaneous”, i.e., only requiring single frames).

### B. Contributions of This Work

We have identified that there is a gap in the landmark free head-pose estimation research when it comes to unconstrained head-pose estimation in mid-low resolution which we address in this paper. Briefly, the scientific contributions of this paper are as follows: (a) Defining a machine learning framework for unified head pose estimation in RGB and/or RGB-D data that spans from high to very low resolutions; (b) Introduction of a regression loss that lets us pose the cyclic function (it is wrapped

in a sphere) in the Euclidean space by using vector decomposition of the unit directional vector. (c) Creation of a headpose dataset which is publicly available (contact lead author for access) that fills a major void in head pose research datasets by offering in one unified set, desirable properties in terms of modalities (RGB and Depth), constraints (all poses, not only frontal), quality (accurately labelled for regression) and at the same time one that spans from close to long range resulting in high to low quality data respectively. (d) Modeling human gaze and its spatial uncertainty from head-pose as a spherical Von-Mises Fisher distribution on a spherical manifold in  $\mathbb{R}^3$ ; (e) Defining person-person and person-scene interaction metrics and evaluating them on comprehensive open datasets.

## II. RELATED WORK

We now discuss the related work in two main threads: prior works in both head-pose estimation and deep learning.

### A. Head Pose Estimation

The pioneering work on low resolution head pose estimation was proposed by Robertson and Reid [4] which used a detector based on template training to classify head poses in eight directional bins. This approach is heavily reliant on skin colour detection. Subsequently this template-based technique was extended to a color invariant technique by Benfold *et al.* [5]. They proposed a randomized fern classifier for hair face segmentation for the template matching. This work was later improved upon by Siriteerakul *et al.* [12] using pair-wise local intensity and colour differences. However, in keeping with all template based techniques in head-pose estimation, these suffer from two major problems: first, it is non-trivial to localize the head in low resolution images; second, different poses of the same person may appear more similar compared to the same head-pose of different persons.

This led some researchers to propose representing head images in a different feature space that has more discriminatory property for head pose independent of persons. Non-linear regression approaches like Artificial Neural Networks [13], [14] and high-dimensional manifold based approaches [15], [16] try to estimate the head poses in a continuous range. Chen and Odobez [6] proposed the state-of-the-art method for unconstrained coupled head-pose and body-pose estimation in surveillance videos. They used multi-level Histogram of Oriented Gradients (HOG) [17] for the head and body pose features and extracted a feature vector for an adaptive classification using high dimensional kernel space methods. These techniques are quite general and do not depend on the heads being in near frontal poses unlike the HCI techniques. Nevertheless the high degree of error or uncertainties that arise from these methods, render them unsuitable for the tasks like fine grained human interaction or attention modeling.

On the other hand, on the HCI side of the problem the formulation is limited to 2 meter distance from the sensor along with near-frontal head-poses. An iterative closest point (ICP) based mesh fitting approach has been employed for head pose detection [9], [18]. In [10] the candidate head poses are rendered and matched to the input depth image and the 6 degree of freedom pose is solved by optimizing via particle swarm optimisation.

Fanelli *et al.* [8] used a randomized patch based decision forest regression for head pose regression. Work on head pose regression for scene and human interaction understanding has been presented [19]. This work focuses on head-pose regression and interaction detection in 2D movie/ tv-series scenes. While it is quite robust, this approach is limited in that it only works with yaw angles of  $\pm 90^\circ$ . However it does not depend on motion priors or specific facial landmark detections. Recently, manifold based metric learning methods have been applied to head pose estimation [20]. In another approach to manifold learning the spherical nature of the view manifold of objects is used as a strong prior [21]. Another approach reported in [22] uses reflection symmetry information in covariant features extracted from Gabor features. Features derived from local directional quaternary patterns (LDQP) have been used in conjunction with linear SVM successfully in high resolution RGB data [23].

### B. Deep Learning and Convolutional Neural Networks (CNN)

Recently, deep learning, especially CNNs have been shown to learn robust non-linear representations from input data and have been especially successful on images [24], [25] and audio [26]. This is in contrast to traditional computer vision pipelines where problem specific ad-hoc features like HOG [17] are extracted. These features would typically be used as input to machine learning framework such as support vector machines (SVM) to achieve classification or regression. The power of deep models lie in their ability to learn layers of non-linear transformations on the data [25]. The resurgence of these methods started with the successful introduction of a class of deep generative models called Deep Belief Networks (DBN) and their unsupervised training using Contrastive Divergence (CD) [27]. The power of a generative model, as shown by Tang *et al.* [28], lies in being able to reconstruct original images under noise or heavy occlusions [27]. CNNs [24] on the other hand are supervised, discriminative and have mostly surpassed the DBNs in terms of accuracy on large labelled datasets like the Imagenet [29].

CNNs have also been applied in the multimodal RGB-D domain. Lu [30] demonstrated early fusion of RGB-D channels and used transfer learning to initialise the weights of the green, blue and depth channels with filters learned from the depth channel. More recently it has been shown that this form of early fusion is not very helpful because the network can not propagate meaningful gradients across channels [31], [32]. Hence RGB-D networks are generally trained with late fusion where the modalities are learned separately and combined in the classifier phase [31], [32].

## III. THEORY AND METHODOLOGY

In this paper we do not concern ourselves with the problem of detecting heads. Instead we can adapt the output of any head detector and normalize the heads to  $256 \times 256$  as input to our algorithm. Once we have the normalized RGB-D heads as input the rest of the process can be briefly summarized as follows. First, if available, we encode the depth image using a scheme that we name DAE encoding which encodes the depth modality with three channels of depth, surface normal azimuthal and surface normal elevation angle as shown in Fig. 3 and is similar to

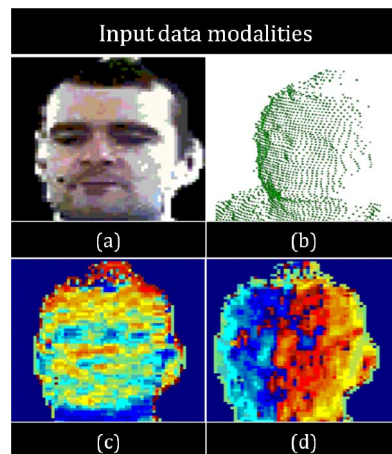


Fig. 3. Input modalities. (a) shows the RGB input. (b), (c), and (d) show the depth, surface normal elevation angle, and surface normal azimuthal angle, respectively, that form the three channels of the DAE encoding.

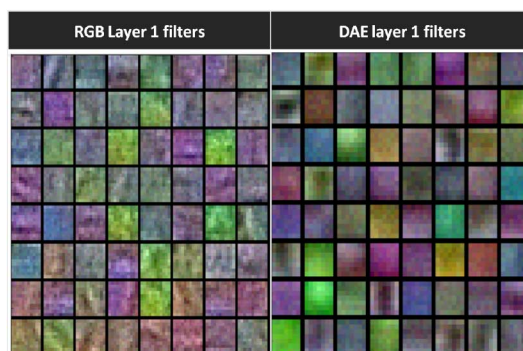


Fig. 4. Visualization of the first level of learned filters on both the RGB network and the depth network.

HHA of [31]. We do not encode the preferred gravity direction in DAE like HHA as all our heads are upright. These inputs are then used to train a CNN each for RGB and DAE and we call them RGB CNN and Depth CNN respectively. The combination of the posteriors of the two CNNs are called the RGB-D CNN.

### A. Convolutional Neural Networks

Convolutional neural networks belong to a class of fully supervised deep models that have proven to be very successful in a wide variety of tasks. The power of CNNs lie in the ability to learn multiple levels of non linear transforms on the input data using labeled examples through gradient descent based optimizations. The basic building blocks of CNNs are fully parametrized (trainable) convolution filter banks (as shown in Fig. 4) that convolve the input to produce feature maps (as can be seen from Fig. 5), non-linearities (like sigmoid or Rectified Linear Units/ReLU), pooling layers/downsampling layers (e.g. max pooling, mean pooling etc.) that downsample the feature maps, and fully connected layers. CNNs in particular through their multiple levels of convolution and pooling achieve a high degree of translation invariance in their features. Recent studies from Simonyan and Zisserman [33] have shown that deeper models with smaller filters achieve great expressive power in terms of learning powerful features from data in

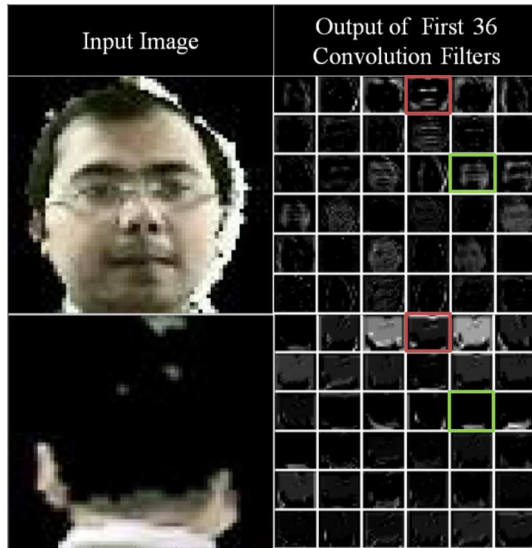


Fig. 5. Visualization of the features extracted after the first level of convolution. We show only the first 36 channels out of the 64 channels. It is easy to see that some filters are bringing out the facial landmarks (top red, where they are detected, and bottom red, where they are not), whereas others have learned skin maps (indicated in green), among other real facial features.

tasks like object recognition on large scale datasets like the Imagenet [34]. As the model go deeper the number of weights/parameters or the networks grow significantly. It then becomes imperative to use large scale labelled training data to train these networks. However one should note that the number of parameters in the convolution layers are orders of magnitude lower than the fully connected layer [35]. Hence by having more convolution layers helps alleviate the problem of this parameter explosion while retaining the expressive properties on the deep models. One such model is the recently introduced Googlenet model [24]. We train two CNNs on the RGB and depth modalities based on this architecture [24]. This architecture has the state-of-the-art results on the Imagenet dataset [34]. In our experiment the same network also gave the best results on our task. The advantage of this network lies in that it is very deep but has a lot less parameters (around 5 million) compared to other contemporary networks like the VGG-16 [33] which has more than 130 million parameters. This lets us train the networks using considerably less training data. We modified the network by changing the Rectified Linear Unit non-linearities (RELU) with Parametric Rectified Linear Unit (PRELU) and their corresponding weight initialization introduced in [36]. The non-linearities are defined as follows:

$$RELU(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases} \quad (1)$$

$$PRELU(x) = \begin{cases} x, & \text{if } x > 0 \\ mx, & \text{if } x \leq 0 \end{cases} \quad (2)$$

where  $m$ , the slope in the negative  $x$  is a learned free parameter.

The reason the PRELU activations are better than their RELU counterpart lies in the fact that PRELU activations have non zero outputs and non zero gradients in the negative values. This makes them easier to propagate gradients from. Whereas in case

of RELU, if some neuron's output becomes less than equal to zero, its gradients also vanish and it hampers learning through gradient descent. The motivation for doing it is that this small change, without increasing the number of parameters of the network significantly improves the accuracy (see [36]).

We also exploit the ability of CNNs to learn from multiple types of labels for the same kind of underlying data to achieve a valid representation learnt on the data. Since there are few explicit head-pose regression datasets, we initialize the training of models with classification into 8 head pose classes spanning 360 degrees. The representative head-pose classes are shown in Fig. 6. We learn an initial representation that is then transferred to the regression network and fine tuned for regression. Fig. 6 also shows how the CNN features separate easily in only two dimensions whereas the HOG feature that is used in other techniques including [37], [6], [19] is nowhere near as effective. It can also be seen from Fig. 4 and 5 that the network learns filters, some which could have been developed by intuition, where as other features are not as intuitive but effective nonetheless. It is interesting to note that the feature space embedding presented in Fig. 6 shows that the CNN learns the implicit circular geometry of the view manifold from the data itself. This is in contrast to [21] where this shape is imposed as a prior assumption. However due to imaging noises and low resolution they might not lie in an ideal circle. Besides, ideal circular distribution may or may not be ideal for a classifier as can be seen from Figs. 7 and 8. Hence, it is our belief that an end to end approach without prior assumption leads to better results for classification.

For regression we expect to see a similar distribution that is more evenly spread out on the manifold instead of forming clusters. Fig. 7 shows the output scatter plot of the first two LDA components of our fine-tuned features on regression on our dataset.

The regression output is then combined heuristically to obtain a probabilistic attention distribution which we parameterise as a Von-Misses Fisher distribution. This distribution captures two important properties of the head pose regressor output. First, it inherently models the regressor output confidence directly into the distribution concentration parameter  $\eta$ ; second, it also models the inherent irreducible uncertainty in every gaze tracking technique where eye balls are not tracked. We have performed experiments to determine the mean discrepancy between eye and head-pose to model this phenomenon.

We now discuss each of these steps in detail in the subsections that follow.

## B. DAE Depth Encoding

For depth data it is important to encode some spatial and surface information into the data itself, as shown by Gupta *et al.* [31]. We follow a similar approach however we do not encode the inferred gravity (vertical) direction, because in our case the heads are always upright and this parameter would yield no more information. We do however encode the surface normal azimuthal angle and the surface normal elevation angle along with the depth data to form three channels as can be seen in Fig. 3.

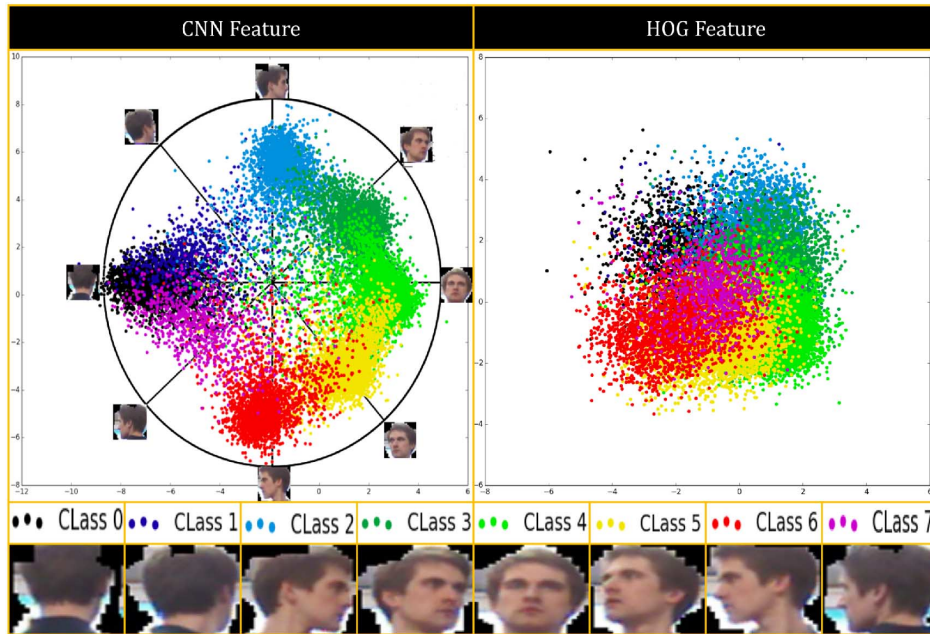


Fig. 6. LDA projected scatter plot of the clusters of the head pose classes after initial training for classification. We compare it with the HOG feature [17] which is most common in the comparable literature. Not only are our clusters well separated, they maintain the approximate closed topology of the circular head-pose manifold. The clusters have their mean near the pose class angles and spread around the circumference of the manifold. This validates our choice of transferring this network to the regression task.

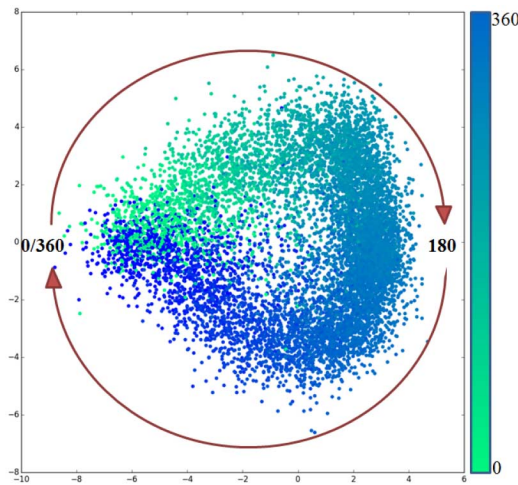


Fig. 7. LDA projected scatter plot of regression features on our dataset with a color map that spans the range of 0–360 degrees. The features maintain the manifold.

Surface normals have proved to be a very useful feature for object recognition [38]. We compute the surface normals via

$$\overrightarrow{N_{X_i, Y_j}} = \frac{\overrightarrow{\partial_X Z_{ij}} \times \overrightarrow{\partial_Y Z_{ij}}}{\left\| \overrightarrow{\partial_X Z_{ij}} \times \overrightarrow{\partial_Y Z_{ij}} \right\|} \quad (3)$$

where  $\overrightarrow{N_{X_i, Y_j}}$  is the normalized normal vector at  $X_i, Y_j, Z_{ij}$  which in turn are the real world coordinate at depth image point

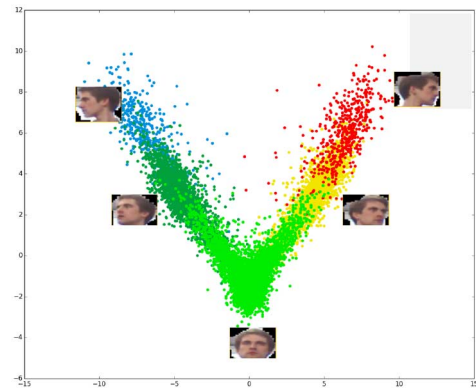


Fig. 8. LDA projected scatter plot of the regression features on the BIWI dataset [8]. This shows the headpose regression features thinly form around the frontal pose manifold. This dataset is relatively easy as it is very high resolution and contains only frontal/near-frontal head poses.

$U_i, V_j$  and  $\overrightarrow{\partial_X Z_{ij}}$  is the X derivative and  $\overrightarrow{\partial_Y Z_{ij}}$  the Y derivative at point  $X_i, Y_j$ . To compute the derivatives we use implicit filtering techniques as described in [39]. Implicit filtering techniques are much more accurate than the standard morphological derivative as can be seen in Fig. 9. Implicit filtering also involves larger neighbourhoods for computing more accurate gradients. Considering all these benefits we chose to use the one parameter family of implicit differentiation with the frequency domain transfer function defined as the following spatial domain equation:

$$\begin{aligned} & \beta f'_{i-2} + \alpha f'_{i-1} + f'_i + \alpha f'_{i+1} + \beta f'_{i+2} \\ &= c \frac{f_{i+3} - f_{i-3}}{6h} + b \frac{f_{i+2} - f_{i-2}}{4h} + a \frac{f_{i+1} - f_{i-1}}{2h} \quad (4) \end{aligned}$$

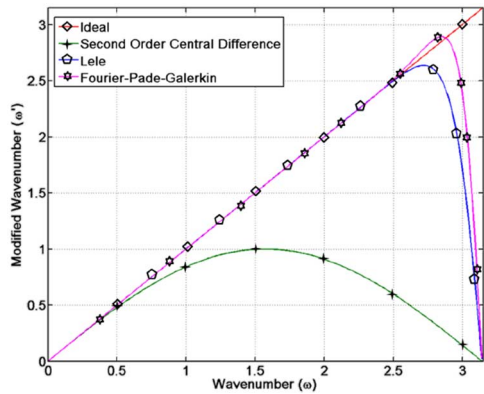


Fig. 9. Benefit of implicit differentiation is shown in this graph. It can be seen that the frequency response of the implicit Lele’s and Fourier-Pade-Galerkin schemes [39] better approximate the original derivative compared to the explicit second order central difference scheme.

where  $f'_i$  and  $f_i$  are the  $x$  derivative and the function value respectively at  $x = i$ . and its corresponding frequency domain counterpart is defined as follows:

$$H(\omega) = j \frac{a \times \sin \omega + (b/2) \times \sin 2\omega + (c/3) \times \sin 3\omega}{1 + 2\alpha \cos \omega + 2\beta \cos 2\omega} \quad (5)$$

where  $\alpha, \beta, a, b, c$  are user chosen parameters.

The  $y$  derivative can be computed similarly. There are many standard parameter choices for  $a, b, c, \alpha$ , and  $\beta$ . Here we use the Lele coefficient values [39].

### C. Fine-Tuning for Regression

The classification network is turned into a regression network by replacing the last Softmax layer with an Euclidian loss layer that measures the L2 distance of the prediction from the target. To activate the fine-tuning on regression on the head-pose data the following must be considered. The regression problem is ill-posed for the linear Euclidean manifold where we compute the regression L2 loss. This is because the normalized regression label goes from 0 to 1 where 0 is the back of the head to 0.5 that is for front facing to 1 (360 degree) that is again back of the head. Now the distance between the angle 0.1 and 0.9 should be 0.2 on the circular manifold. In the stated example the heads look very similar, however the loss function penalizes the network by having an error of 0.98, hence the gradients for weight update are large and these force large changes. Ideally the loss function would be defined as:

$$L = \begin{cases} \frac{1}{2}(t - o)^2, & \text{if } t - o < 0.5 \\ \frac{1}{2}((1 - (t - o))^2), & \text{if } t - o > 0.5 \end{cases} \quad (6)$$

where  $t$  is the target angle and  $o$  is the output of the network. However this function is not everywhere differentiable (with a discontinuity at  $t - o = 0.5$ ). In order to perform gradient descent the loss must be differentiable w.r.t the weights. To overcome this issue, instead of using the angles for regression, we use the X,Y coordinate of the unit vector pointing in that angle, the problem can be posed on the linear Euclidean manifold again. So instead of a single number we have a pair. For both Yaw and Pitch this same technique can be easily extended

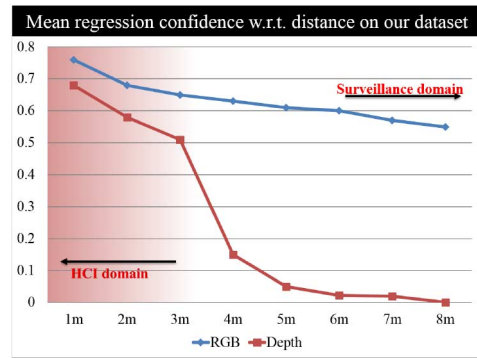


Fig. 10. Quality of depth data degrades rapidly while RGB stays much more reliable with distance.

to use the X,Y and Z coordinates of the head pose vector in 3-D. The network fine-tuned for regression should have features that are thinly spread along the manifold. We see this expected result in Fig. 8 where we plot the features projected to two dimensions using Linear Discriminant Analysis (LDA) on the Biwi dataset [8]. We also plot the same using our dataset in Fig. 7.

### D. Fusion of RGB and Depth Modalities

Whenever available, both RGB and depth give complementary information that can be combined to achieve an overall information gain. Apart from that depth information can further be exploited to compute the scene interaction/ attention metric in 3D that maps the head pose based attention to the 3D environment as can be seen from Fig. 1(a). Hence whenever possible, we average the class posterior scores from both the RGB and depth classifiers. However we note that depth information quality is highly dependent on distance from the sensor. Also from our experiments we have found out that the back of the head depth images are extremely noisy.

Fig. 10 shows the reliability of the RGB vs Depth information as a function of distance from the sensor (in this work we used both the Kinect and Kinect v2 sensors). We compute the confidence of the RGB and Depth information from the relative error with respect to ground truth. From our experiments we have seen that unless the distance of the detected head is taken into account, depth information is not very reliable after 3.5 meters as far as headpose is concerned. Hence if depth data is available and the detected head is less than 3.5 m distant, we average the output of the RGB and Depth models. Otherwise we only use the RGB information (e.g. in the surveillance domain). As can be seen from Fig. 11, depth information also degrades rapidly for non frontal poses but is very useful close HCI domain data.

### E. Regression Confidence Estimate

We determine the regression confidence by combining the regression angle output on the yaw angle with the classifier posterior on the angles. For this we train a Softmax classifier with a granularity of 1 degree (360 classes) on top of the final regression network while keeping the rest of the network weights constant. This enables the computation of the variance of the posterior to estimate the confidence of the regression. Fig. 12 shows the output of the classifier posterior along with regression.

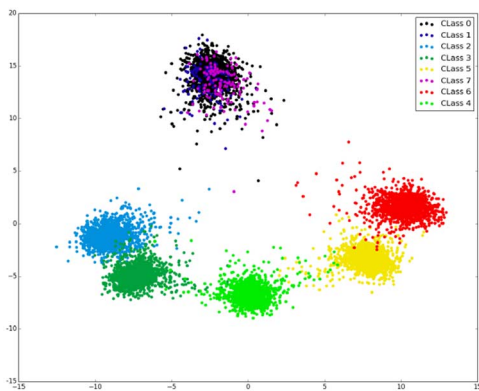


Fig. 11. LDA scatter of depth information shows that it is not very reliable for back of the head poses. This is in line with our expectation as hair does not reflect the depth sensor infrared illuminant very well and this often results in very noisy and sparse data. However, the depth information is quite good for the frontal poses.

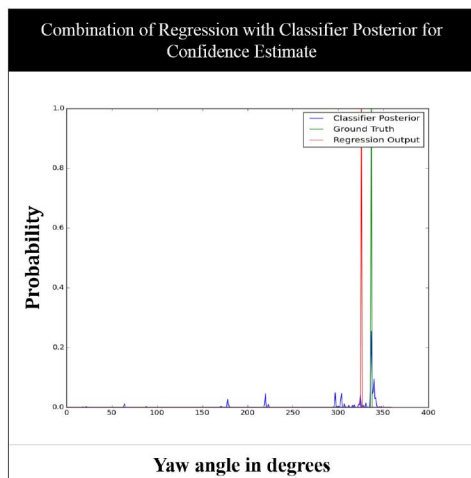


Fig. 12. To estimate the confidence of the regression model into the attention metric, we use the same network to get a posterior distribution on the 360 degrees. We show the regression output, the ground truth, and the probability distributions.

#### IV. VALIDATION OF THE TECHNIQUE

We have generated a dataset using the Kinect and Kinect 2 sensors where we recorded 46 people (32 males, 14 females) freely moving around with various head-poses in front of the sensor. To get accurate head pose ground truth data we used a discreet (actually hidden) wearable miniature X-BIMU IMU sensor which provides the head orientation as a quaternion. We then recorded each individual for one minute moving in the field of view with varying distance (2 m - 8 m). We annotated the head in each frame and associated the IMU data with it in each frame. We acquired around 1500 frames for each person giving a dataset of the order of 68000 training examples. This dataset, which we will release publicly for all academic researchers addresses a vital gap in head pose research, i.e. it is the first of its kind head pose estimation dataset which is multimodal (RGB-D), is accurately labeled for regression, covers all poses (not only constrained to frontal poses), and also spans a wide range of depths leading to both high quality and low quality/noisy data.

To maximise the training corpus, we gathered data from multiple sources that had similar underlying distributions. Datasets annotated for unconstrained face recognition, facial landmark detection, expression detection all have facial data under various poses. The different head pose datasets that we used are the Oxford town centre dataset [5], the BIWI Kinect head-pose dataset [40], the Caviar shopping centre dataset [41], the HIIT Head Orientation dataset along with the IDIAP head-pose dataset [42]. It should be highlighted that the different datasets have different annotations; some of them have real-valued ground truth, others have 6-8 classes spanning the  $360^\circ$ . The datasets also vary in resolution from very high (BIWI) to very low (Caviar). To compare with the other high resolution RGB methods we also report our result on the CMU Multi-PIE dataset [43].

#### A. Experimental Setup

We train one network for RGB and Depth each. This is done to unify the problem of both HCI and surveillance domains. Typically, one might adapt the networks for each domain, however from our initial experiments we found that including both high and low resolution imagery in the training set improved classification performance on the low resolution inference while the high resolution inference results were more or less the same. The convergence rate during training was faster as well. We think this is because the high resolution images help the network estimate the underlying model better and that translates into better parameter estimation for low resolution and/or noisy images. For training and validation we split the combined dataset in a ratio of 70:30 randomly across several trial runs and averaged the mean squared error. For training we used a dropout rate of 20% on before every fully connected layer. We jittered the input images by mirroring them (with corresponding change in groundtruth) scaling the bounding box and cropping them with scales 0.75, 0.9, 1.5, 1.8, 2.0, and 2.5. For all scales greater than 1, we also translated the images randomly by 20% in both directions. This was done to improve scale invariance along with mitigating the effects of poorly-aligned or partially-occluded head detections. We used a modified version of the deep learning framework Caffe [44] to train our network. We translated the centroid of each head to (0, 0, 0) in 3D Euclidean space and uniformly re-sampled the point cloud to an organized  $256 \times 256$  set. For re-sampling we used bi-cubic interpolation for the RGB values and nearest neighbour interpolation for the XYZ values. To obtain the mean inherent variance due to eye balls (the true focus of attention is somewhat independent of head pose), we set up an experiment with where we tracked the difference between the absolute head-pose (using the IMU) and the focus of attention of the eye using the Gazeport eye tracker, which has a resolution of  $0.5^\circ$  degrees at upto 30 cm distance. We computed the mean variance for 11 people. This provides us with an interesting insight to the problem. In conclusion, head-pose error less the mean error of  $12.35^\circ$  does not make any sense for the application of *true visual attention* estimation without tracking the additional dimensional freedom provided by the eyeballs. In order to gain a good understanding of human attention model

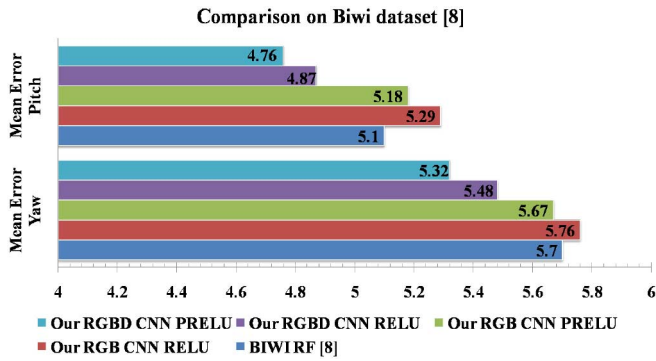


Fig. 13. Comparison of our method on the BIWI dataset with respect to the random forest (RF) algorithm [8]. Our RGB+Depth CNNs (both RELU and PRELU) outperform the HCI technique without explicitly tracking facial landmarks. Here in this range we see the tangible benefit of having depth information along with RGB data.

without eyeball tracking, further studies into human gaze pattern with respect to scene saliency and semantic contextual information would be needed.

We selectively fused the RGB and Depth modalities based on availability and quality of the depth data as shown in Fig. 10. We only fuse the depth classification if the detected head is less than 3.5 meters in distance, otherwise the reliability of the depth data falls off rapidly as can be seen directly from the lower curve in Fig. 10.

## B. Results

Here we present the comprehensive validation of our technique on both HCI and surveillance domains.

1) *Validation on BIWI Kinect Headpose Dataset [8]*: The data in this dataset has been captured very close to the sensor and does not contain non-frontal poses. Here the output of our RGB and Depth models are averaged to get the result. This data lets us compare our general technique to a finer grained HCI technique as presented in [8]. The comparative results are shown in Fig. 13. We use mean angular error as the metric which is the same used in comparable literature [8]. In both pitch and yaw we outperform the best method [8], which has the advantage of explicit landmark detection, by 7%. It should be noted that while we do not detect landmarks explicitly, from Fig. 5, it is clear that the CNN has now learned landmark detection automatically. However as can be seen from Fig. 5 the network detects landmarks whenever necessary implicitly along with other non obvious features. We also see that depth information actually improves the results in this range when combined with RGB.

2) *Validation on Our Dataset*: One weakness of the BIWI dataset is that it does not contain non frontal or distant head-pose data. To overcome this, and to show the power of our technique we report the results obtained on our dataset which is far more challenging. Our two baseline methods are the ‘‘Here’s Looking at You Kid’’ (HLYK) [19] which uses only the RGB data and the Random forest (RF henceforth) based approach [8] which uses only the depth and normal data. We outperform both the techniques by a significant margin as shown in Fig. 14. We reduce the relative error by 40% to that of our closest competing technique [19].

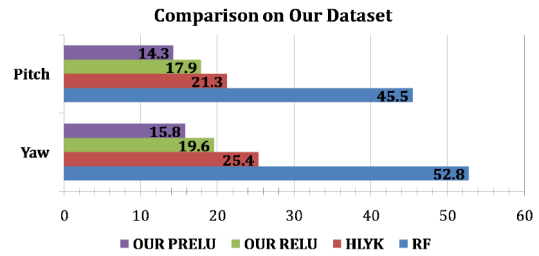


Fig. 14. Mean squared errors (MSE) of the RF [8] and HLYK [19] techniques, compared to ours on our dataset.

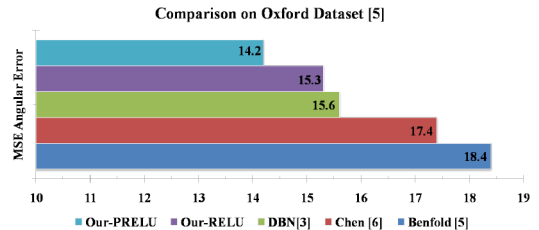


Fig. 15. Mean squared error on the Oxford dataset. Here we compare our regression output with the Benfold [36] and the Chen [6] techniques.

3) *Validation on Low-Resolution Surveillance Dataset*: For the low resolution surveillance domain dataset, we report our results on the Oxford and the Caviar datasets. In these datasets we classify the head pose into 8 equally spaced ( $45^\circ$ ) angular bins as shown in Fig. 6. For comparison with [6] and Benfold [37] we use the Oxford dataset in which both have reported results. One consideration has to be made while comparing because [6] reported the mean square error (MSE) which they derived from a weighted combination of their 8 class classifier output multiplied with the bin angles as  $\sum_{i=1}^8 p_i \vec{\eta}_{\theta_i}$  where  $p_i$  is the classifier output value for the class  $i$  and  $\vec{\eta}_{\theta_i}$  is the unit vector in that angular direction. Fig. 15 shows the comparison between our method with the previous state-of-the-art results. In terms of MSE we have achieved the best published results. The margin alone does not give the true picture of performance. We therefore present the confusion matrices on the Oxford and Caviar datasets, as shown in Fig. 16.

On the Oxford dataset, for comparison, we also show the output confusion matrix of the Benfold algorithm [5] along with our confusion matrix as shown in Fig. 16.

4) *Validation on Multi-PIE Dataset [43]*: The Multi-PIE dataset consists of 337 subjects, under 15 view angles and 19 illumination conditions. This is a close range high resolution RGB dataset. We compare our method against two state-of-the-art techniques on this dataset 1. LDQP [23] and 2. circle23Sphere [21]. As shown in Table I we outperform both competing techniques [23] and [21] in terms of Mean Angular Error(MAE) by a significant margin without any training on this dataset.

5) *Comparison Between PRELU and RELU*: In all our experiments PRELU activation outperformed RELU consistently. As can be seen from Fig. 15, in case of low resolution surveillance domain dataset [5], we gain  $> 1^\circ$  improvement in angular error by using PRELU. On our challenging RGB-D dataset the effect is even more pronounced with an improvement of around



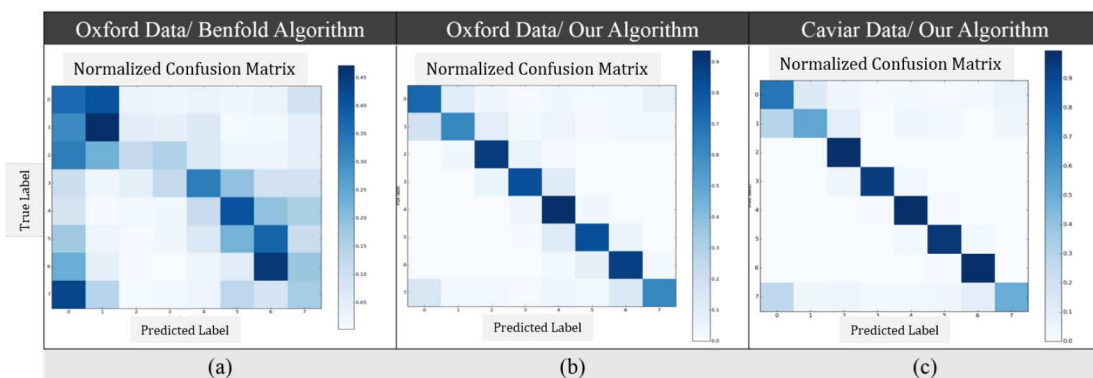


Fig. 16. Confusion matrix comparing the methods the results. (a) Oxford data, benfold algorithm [37]. (b) Oxford data, our RGB CNN. (c) Caviar data, our RGB CNN. On both datasets we have the state-of-the-art results by far .

TABLE I  
RESULTS ON THE MULTI-PIE DATASET

Method	Mean Angular Error in $^{\circ}$
LDQP[23]	7.3
circ23D[21]	5.8
Our RELU	4.56
Our PRELU	4.2

$2.9^{\circ}$  as seen in Fig. 14. Results on the CMU Multi-PIE dataset as reported in Table I suggests that PRELU provides an additional reduction in MAE of  $0.36^{\circ}$ . Finally in Fig. 13, on the Biwi dataset where error rates are already pretty low, we get less improvement ( $0.2^{\circ}$ ), but consistent improvement nonetheless by using PRELU over RELU.

6) *Discussion*: We validated our approach on two datasets. In the first HCI domain BIWI dataset [8] our more general technique came very close in terms of accuracy compared to the Random Forest method employed there. While this was expected, we want to argue that without specifically tracking the eye balls, an error less than  $\pm 12.35^{\circ}$  carries no meaning by itself as has been established by our experiment on eye gaze and head-pose variance. As long as the error is less than that, as is the case for both the techniques on this dataset, any two techniques are equivalent for true focus of attention (as in HCI).

On our, more challenging dataset, there are a few observations worth noting. As the distance increases and the quality of depth data decreases, we see that the error in the depth feature based techniques have a larger gradient than the HOG based technique. This is because the loss in colour resolution is not as high compared to the loss in quality of the depth data. This suggests that better depth reconstruction techniques should improve results further. However interpolation schemes like bi-cubic interpolation produced significant artefacts around the edges and further degraded the results. So we used nearest neighbour interpolation for resampling the head point cloud. We now believe that a “data driven” head depth reconstruction will be the way forward and warrants further investigation in future work. We show the qualitative results on different datasets in Fig. 17.

## V. EXPLOITING HEAD-POSE AS A SOCIAL SIGNAL

We use our robust head-pose estimation technique to further infer meta information regarding human centric scene understanding i.e. we wish to know what people are looking at in the

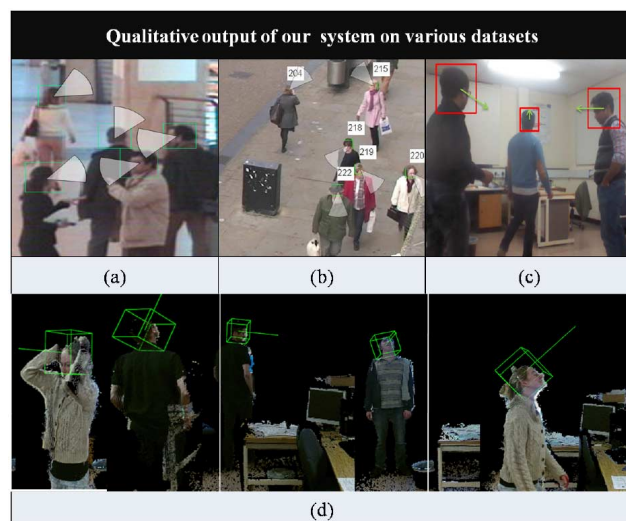


Fig. 17. Qualitative output of our headpose estimation system on various datasets. (a) Caviar dataset. (b) Oxford Dataset. (c) Our RGB dataset. (d) Our low-resolution RGB-D dataset.

real world, not merely the image plane. To this end we first define a “human attention metric” based on the regression output.

### A. Probabilistic Attention Metric

While pure head pose angle is important, we note that it carries little meaning by itself if there is no object at which the person is gazing. If we model the head-pose as a spread of attention with a mean direction and a uncertainty spread that depends upon regressor confidence that is computed by taking the variance of the classifier output (we also compute the  $360^{\circ}$  classification result along with the regression output), along with the inherent uncertainty due to not tracking the eye, we can gain a lot more useful information. Our aim is to achieve gaze estimation as a spatial probability distribution in an unified framework that can be used for both gaze estimation and interaction detection. This is distinct from approaches defined in literature. In [18] head pose is used for estimating gaze through a fixed sized disc surrounding the intersection point of the head pose ray and the object/camera plane. This approach does not incorporate the confidence of the head pose estimate to peak or diffuse the gaze estimate that our technique proposes. Whereas the LAEO system [19], while useful for interaction detection, lacks

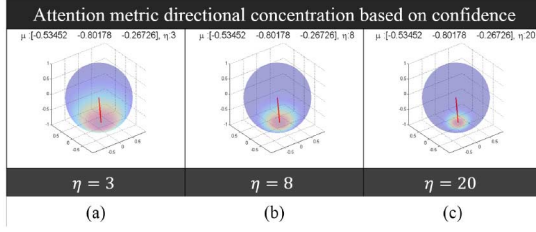


Fig. 18. Von-Mises Fisher distribution visualized on a unit sphere. The mean direction  $\mu$  is represented by the red line and the factor  $\eta$  represents the concentration of the distribution.

the ability to project the headpose estimate into gaze estimate. Our proposed approach, the Attention Metric (AM) solves both these problems in a unified fashion.

To define the field of attention given the Regressor output of the yaw ( $\theta(t)$ ) and pitch ( $\phi(t)$ ) head angles and their corresponding variances ( $\sigma_1(t)$ ,  $\sigma_2(t)$ ) for each frame, we turn to the field of directional statistics. We define a unit 2-D spherical probability distribution manifold in the 3-D space using the Von Mises–Fisher distribution [45]. This distribution is analogous to a 2-D normal distribution but wrapped around a 2 dimensional unit sphere in  $\mathbb{R}^3$ . In general for a  $(p - 1)$  dimensional sphere in  $\mathbb{R}^p$  the von Mises-Fisher distribution for the  $p$ -dimensional unit vector  $\mathbf{x}$  is defined as

$$f_p(\mathbf{x}; \mu, \eta) = C_p(\eta) \exp(\eta \mu^T \mathbf{x}) \quad (7)$$

where  $\eta \geq 0$  is the concentration factor (inversely proportional to the variance  $\sigma$ ) and  $\|\mu\| = 1$  is the unit vector in the direction of the mean and  $C_p(\eta)$  is the normalization factor defined as

$$C_p(\eta) = \frac{\eta^{p/2-1}}{(2\pi)^{p/2} \times J_{p/2-1}(\eta)} \quad (8)$$

where  $J_v$  denotes the modified Bessel function of the first kind and order  $v$ . In our case of  $\mathbb{R}^3$  or  $p = 3$  It reduces to

$$C_3(\eta) = \frac{\eta}{4\pi \sinh(\eta)}. \quad (9)$$

Fig. 18 shows the Von Mises-Fisher distribution for various  $\eta$ .

In our particular case we compute the mean direction unit vector  $\mu$  from the yaw and pitch angles in spherical coordinates, and also the concentration factor  $\eta$  assuming isotropic variance in both yaw and pitch angles as

$$\mu = \frac{1}{\sqrt{1 + \theta^2 + \phi^2}} \begin{bmatrix} 1 \\ \theta \\ \phi \end{bmatrix} \quad \eta = \frac{1}{\sqrt{\sigma_1'^2 + \sigma_2'^2}}. \quad (10)$$

The  $\sigma_1'$  and  $\sigma_2'$  are the sum of the regressor variance ( $\sigma$ ) and the inherent mean uncertainty (E- constant due to no eye tracking) in standard deviation units.

In case one needs to preserve anisotropic variances in both directions one can use the Kent distribution [45] which preserves those properties. However from our experience, we decided not to use it (keeping in mind its higher computational complexity). So our final attention metric for person  $i$  is defined as

$$AM_i(\mathbf{x}, t; \mu_i, \eta_i) = \frac{\eta_i}{4\pi \sinh(\eta_i)} \exp(\eta_i \mu_i^T \mathbf{x}). \quad (11)$$

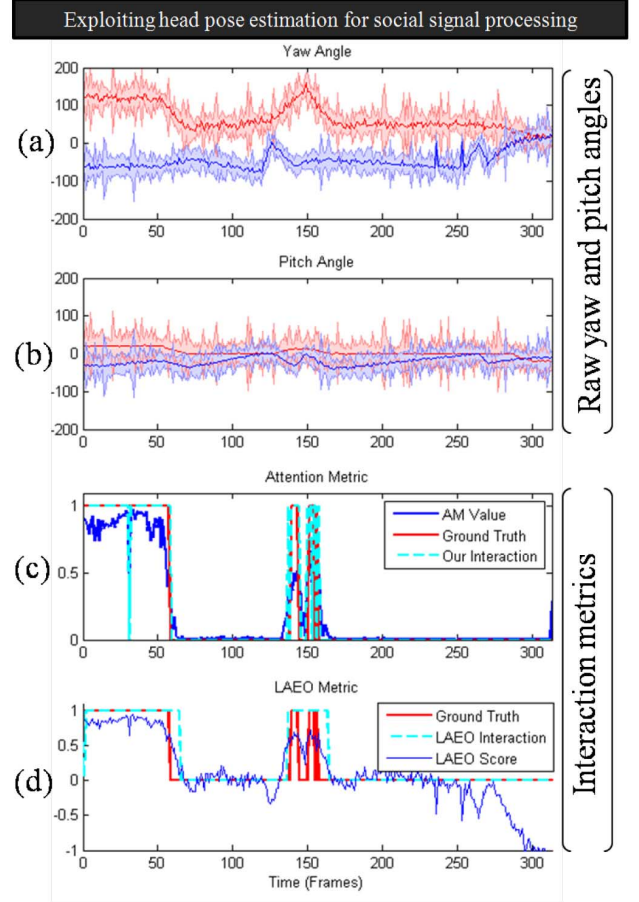


Fig. 19. In this figure, we show the use of the head pose angles with the interaction metrics LAEO [19] and AM to do interaction detection. (a) and (b) show the yaw and the pitch angles along with their 95% confidence intervals, of two heads in a sequence of two people interacting. (c) shows the output of our interaction metric (IM) (in blue) and interaction detection (dotted cyan), and (d) compares the LAEO [19] metric (blue) and its corresponding interaction detection (dotted cyan). The ground truth for interaction is shown for reference in red in both (c) and (d) and the signal is binary (the interaction is either happening or not). IM clearly outperforms LAEO.

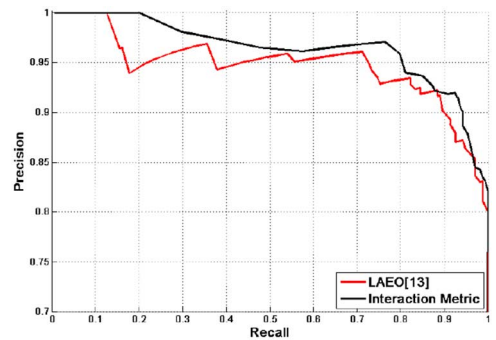


Fig. 20. Precision recall curve comparing our attention metric to the LAEO metric on our dataset.

To detect interaction between any two people ( $i, j$ ) we multiply two attention matrices ( $AM_i$  and  $AM_j$ ) together computed at the location of the other person's head. Hence the interaction metric (IM) for a pair of people ( $i, j$ ) with their corresponding head positions  $\mathbf{x}_i$  and  $\mathbf{x}_j$  is defined as

$$IM_{ij} = \frac{AM_i(\mathbf{x}_j) \times AM_j(\mathbf{x}_i)}{r_{ij}} \quad (12)$$

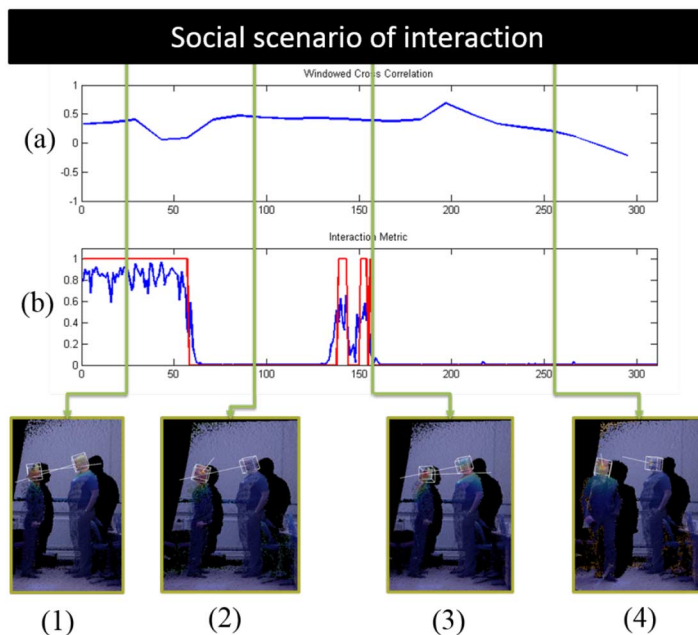


Fig. 21. We show the interaction and WCC head pose signals. The binary ground truth for interaction is shown in red. The raw head pose signals are same as 17 (a) and (b). The scenario can be described with the four snapshots as follows. (1) Two people are talking facing each other, and from (b) the IM can be observed to be high while from (a) the WCC is not observed high. (2) One person looks away towards the direction of the camera which is followed by a drastic fall in the IM in (b), while the WCC in (a) falls while the two heads behave differently and stabilizes. (3) The person looks back intermittently and we see the corresponding change in IM. (4) Finally, the person walks away with the other person looking at the same place. This makes the WCC fall drastically. The peak WCC is achieved around frame 200 when both of them look at the general direction of the camera.

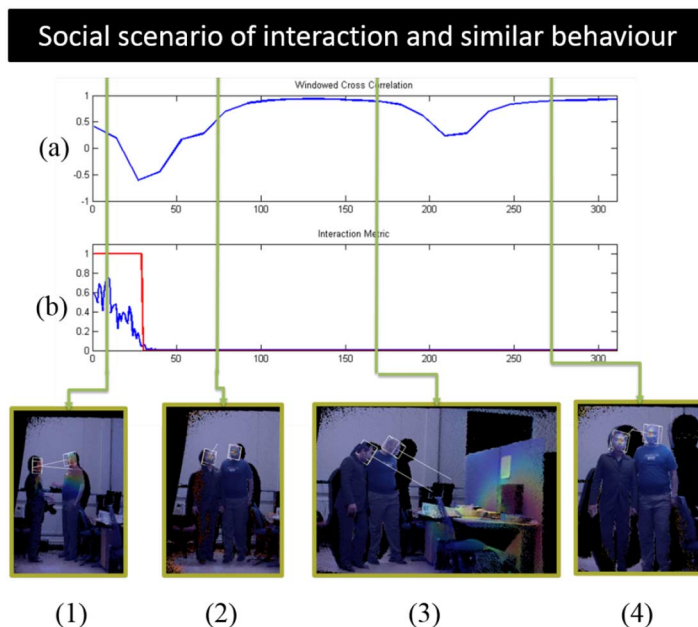


Fig. 22. In this scenario, two people are interacting as can be seen from (1). This results in the corresponding IM and WCC signals in (b) and (a), respectively. Then in (2), they start walking together in the same direction facing the camera. This makes the WCC signal go up. The WCC signal stays high when they look at the same object of interest together in (3). Finally, in (4) they walk away together looking towards the camera. The dip in the WCC signal near frame 220 is caused when one person walks away before the other.

where  $r_{ij}$  is the euclidean distance between the pair of heads.

Fig. 19 shows the output of our Interaction metric along with interaction detection on our dataset. For comparison we also show the HLYK interaction detection scheme, i.e. looking at each other (LAEO) as reported in [19]. We also show the raw yaw and pitch angles for both persons. In both the interaction detection signals, namely IM and LAEO the interaction ground

truth is plotted in red, and the IM and LAEO signals are plotted in blue. To detect interactions from IM we can simply specify a threshold above which interaction is detected. This is the only free parameter in the IM scheme. We cross validated the parameter for various values and found that setting this IM threshold to 0.32 results in highest accuracy. In contrast, LAEO requires three free parameters, the aperture of the viewing cone  $\phi$ , the

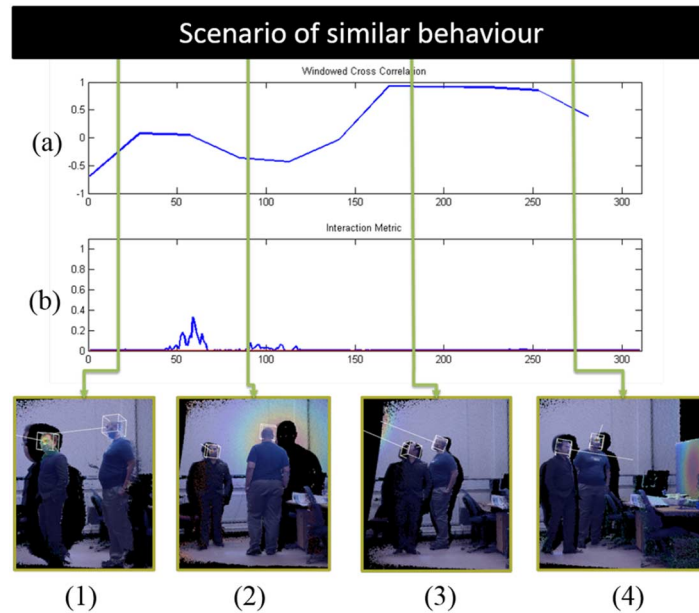


Fig. 23. In this scenario, the two people are not behaving similarly at the beginning and are looking at different things at different times. In (3) they are attracted by the same thing on the wall on the left and look at it together. This makes their head pose signal to become highly correlated as can be seen from (a). Finally, they walk away their separate ways and we see a drop in the WCC signal.

temporal window for smoothing  $T$  and the interaction threshold  $\tau$ . We computed LAEO using the best reported values for these parameters from [19]. It is note worthy that IM is bound between  $[0,1]$  allowing a probabilistic interpretation of the same, whereas the LAEO signal is not bounded. From Fig. 20, where we show the precision-recall curves comparing both IM and LAEO, it is evident that IM outperforms LAEO consistently across all parameter choices.

We show another instance of our interaction metric in Fig. 21. In this instance there are two people who are interacting in the beginning (high IM signal), then one person looks away towards the camera while the other person keeps looking at the said person (low IM signal), near the middle of the sequence they interact intermittently, and finally one person walks away. Both the binary ground truth for interaction (red) and the IM signal (blue) are shown.

Apart from showing interaction metric we also show another social signal metric called windowed cross correlation (WCC henceforth) [46]. This signal measures the similarity between any pair of time series head pose signals (within some time window; leading or lagging) and can be used to detect group behavior.

To further show our system we consider the scenarios shown in Figs. 22 and 23 by using the both the interaction metric (IM) and WCC signals. In the case of Fig. 22, the scenario starts with two people looking with each other. During this period we see that the IM signal is indicative of the scenario. Then both the persons start walking together in the same direction. This leads to a high valued WCC signal and zero IM signal. This state of the signals persists as both of them look together into the same object of interest. Finally, one person walks away before the other causing a drop in the WCC signal, which again goes up as the other person joins moments later. All this while the IM signal is zero or near-zero as no interaction is taking place.

In the case of Fig. 23, the scenario represents two people loitering in a common area until suddenly something attracts their attention and both look towards the same thing. This leads to a low WCC signal at the beginning which is followed by a high WCC signal when both of them look towards the same thing. Finally when they move apart from the scene we see a corresponding drop in the WCC signal.

*Discussion:* The IM signal was quantitatively evaluated in Fig. 20 and compared against the state-of-the-art metric LAEO as described in [19]. The IM signal with significantly less number of free parameters to tune (one, namely the interaction threshold) outperformed LAEO in all scenarios. Subsequently from all the qualitative scenarios described in Figs. 19, 21, 22, and 23, we see that both the IM and WCC signals are intuitive, and both provide key evidences that can contribute towards higher level behaviour inference in social signal processing.

## VI. FUTURE WORK AND CONCLUSION

In this paper we presented a novel approach towards human attention modeling via head-pose estimation. We unified the low-resolution and high-resolution application domains and outperformed the best reported methods in each. We showed that our approach achieves state-of-the-art results on unconstrained head pose estimation on RGB-D point clouds. In future we plan to do a detailed expansion of the different components of the algorithm. It is interesting to note that our eyes have an additional degree of freedom, so it may not make sense to improve accuracy to less than  $\pm 12.35^\circ$  without eyeball tracking. To achieve better performance the authors believe a hybrid approach where semantic information about salient regions in a scene have to be probabilistically fused with the regressor.

## ACKNOWLEDGMENT

The authors would like to thank Nvidia for donating the Tesla K40 GPU that helped them with this research. They would also like to thank Microsoft for providing access to the Kinect developer program.

## REFERENCES

- [1] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [2] S. R. Langton, H. Honeyman, and E. Tessler, "The influence of head contour and nose angle on the perception of eye-gaze direction," *Perception Psychophys.*, vol. 66, no. 5, pp. 752–771, 2004.
- [3] R. H. Baxter, M. J. Leach, S. S. Mukherjee, and N. M. Robertson, "An adaptive motion model for person tracking with instantaneous head-pose features," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 578–582, May 2015.
- [4] N. Robertson and I. Reid, "Estimating gaze direction from low-resolution faces in video," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 402–415.
- [5] B. Benfold and I. Reid, "Colour invariant head pose classification in low resolution video," in *Proc. BMVC*, 2008, pp. 1–10.
- [6] C. Chen and J.-M. Odobez, "We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 1544–1551.
- [7] T. Siriteerakul, Y. Sato, and V. Boonjing, "Estimating change in head pose from low resolution video using LBP-based tracking," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst.*, Dec. 2011, pp. 1–6.
- [8] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 617–624.
- [9] K. A. F. Mora and J.-M. Odobez, "Gaze estimation from multimodal Kinect data," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2012, pp. 25–30.
- [10] P. Paderleris, X. Zabulis, and A. Argyros, "Head pose estimation on depth data based on particle swarm optimization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2012, pp. 42–49.
- [11] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interactive Techniques*, 1999, pp. 187–194.
- [12] T. Siriteerakul, D. Sugimura, and Y. Sato, "Head pose classification from low resolution images using pairwise non-local intensity and color differences," in *Proc. 4th Pacific-Rim Symp. Image Video Technol.*, 2010, pp. 362–369.
- [13] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley, "Head pose estimation on low resolution images," in *Multimodal Technologies for Perception of Humans*. New York, NY, USA: Springer, 2007, pp. 270–280.
- [14] R. Stiefelhagen, "Estimating head pose with neural networks—Results on the Pointing04 ICPR Workshop evaluation data," in *Proc. Pointing04 ICPR Workshop Int. Conf. Pattern Recog.*, 2004.
- [15] V. N. Balasubramanian, J. Ye, and S. Panchanathan, "Biased manifold embedding: A framework for person-independent head pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–7.
- [16] C. BenAbdelkader, "Robust head pose estimation using supervised manifold learning," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 518–531.
- [17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.
- [18] D. Cazzato, M. Leo, and C. Distanto, "An investigation on the feasibility of uncalibrated and unconstrained gaze tracking for human assistive applications by using head pose estimation," *Sensors*, vol. 14, no. 5, pp. 8363–8379, 2014.
- [19] M. J. Marín-Jiménez, A. Zisserman, and V. Ferrari, "Here's looking at you, kid. detecting people looking at each other in videos," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 1–12.
- [20] Q. Wang, Y. Wu, Y. Shen, Y. Liu, and Y. Lei, "Supervised sparse manifold regression for head pose estimation in 3D space," *Signal Process.*, vol. 112, pp. 34–42, 2015.
- [21] X. Peng, J. Huang, Q. Hu, S. Zhang, A. Elgammal, and D. Metaxas, "From circle to 3-sphere: Head pose estimation by instance parameterization," *Comput. Vis. Image Understand.*, vol. 136, pp. 92–102, 2015.
- [22] B. Ma, A. Li, X. Chai, and S. Shan, "Covga: A novel descriptor based on symmetry of regions for head pose estimation," *Neurocomput.*, vol. 143, pp. 97–108, 2014.
- [23] B. Han, S. Lee, and H. S. Yang, "Head pose estimation using image abstraction and local directional quaternary patterns for multiclass classification," *Pattern Recog. Lett.*, vol. 45, pp. 145–153, 2014.
- [24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, 2014 [Online]. Available: <http://arxiv.org/abs/1409.4842>
- [25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105, 2012.
- [26] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
- [27] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [28] Y. Tang, R. Salakhutdinov, and G. Hinton, "Robust Boltzmann Machines for recognition and denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2264–2271.
- [29] J. Deng, K. Li, M. Do, H. Su, and L. Fei-Fei, "Construction and analysis of a large scale image ontology," *Vis. Sci. Soc.*, vol. 186, 2009.
- [30] L. A. Alexandre, "3D object recognition using convolutional neural networks with transfer learning between input channels," in *Proc. 13th Int. Conf. Intell. Autonomous Syst.*, 2014, pp. 889–898.
- [31] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 345–360.
- [32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, 2014 [Online]. Available: <http://arxiv.org/abs/1411.4038>
- [33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014 [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, 2015 [Online]. Available: <http://arxiv.org/abs/1502.03167>
- [35] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *CoRR*, 2014 [Online]. Available: <http://arxiv.org/abs/1404.5997>
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," *CoRR*, 2015 [Online]. Available: <http://arxiv.org/abs/1502.01852>
- [37] B. Benfold and I. Reid, "Unsupervised learning of a scene-specific coarse gaze estimator," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2344–2351.
- [38] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao, "Histogram of oriented normal vectors for object recognition with a depth sensor," in *Proc. Asian Conf. Comput. Vis.*, 2013, pp. 525–538.
- [39] A. Belyaev, "Implicit image differentiation and filtering with applications to image sharpening," *SIAM J. Imaging Sci.*, vol. 6, no. 1, pp. 660–679, 2013.
- [40] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, 2013.
- [41] B. Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 1918–1925.
- [42] D. Tosato, M. Spera, M. Cristani, and V. Murino, "Characterizing humans on Riemannian manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1972–1984, Aug. 2013.
- [43] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.
- [44] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

- [45] K. V. Mardia and P. E. Jupp, *Directional Statistics*. Hoboken, NJ, USA: Wiley, 2009, vol. 494.
- [46] S. M. Boker, J. L. Rotondo, M. Xu, and K. King, "Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series," *Psychological Methods*, vol. 7, no. 3, p. 338, 2002.



**Sankha S. Mukherjee** received the B.E (Hons.) degree from the University of Burdwan, Burdwan, India, in 2011, and is currently working toward the Ph.D. degree at Heriot-Watt University, Edinburgh, U.K.

He was a Visiting Researcher at the Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, from 2012 to 2013. His main research interests include multimodal machine learning and deep learning.

Mr. Mukherjee was the recipient of numerous awards for research, including the University Post Graduate Research Prize from Heriot-Watt University and the Techtot 2009 award from Techtot, India.



**Neil Martin Robertson** (SM'10) received the M.Sci. degree from Glasgow University, Glasgow, U.K., in 2000, and the D.Phil. degree from Oxford University, Oxford, U.K., in 2006.

From 2000 to 2007, he worked with the U.K. Scientific Civil Service, DERA, Malvern, U.K., and then was with QinetiQ, Worcestershire, U.K. He held a 1851 Royal Commission Fellowship with the Robotics Research Group, Oxford University, from 2003 to 2006. He is currently the Principal Investigator with the Visionlab, Heriot-Watt University,

Edinburgh, U.K. He also co-leads the EPSRC Edinburgh Centre for Robotics, Heriot-Watt University, and the EPSRC/DSTL University Defence Research Centre, Heriot-Watt University. His research interests include human behavior recognition, computer vision, and multi-modal sensor fusion.

Dr. Robertson is an Honorary Fellow of the School of Engineering, University of Edinburgh.