Synthetic Thermal Image Generation for Human-Machine Interaction in Vehicles

Richard Blythman
Heliaus Research Group
DTS, Xperi
Galway, Ireland
richard.blythman@xperi.com

Eoin O'Connell
Heliaus Research Group
DTS, Xperi
Galway, Ireland
eoin.oconnell@xperi.com

Michael O'Byrne
Heliaus Research Group
DTS, Xperi
Galway, Ireland
michael.obyrne@xperi.com

Cian Ryan
Heliaus Research Group
DTS, Xperi
Galway, Ireland
cian.ryan@xperi.com

Amr Elrasad
Heliaus Research Group
DTS, Xperi
Galway, Ireland
amr.elrasad@xperi.com

Paul Kielty
School of Engineering
National University of Ireland
Galway, Ireland
p.kielty3@nuigalway.ie

Mohamed Moustafa
School of Computer Science
National University of Ireland
Galway, Ireland
m.moustafa1@nuigalway.ie

Joe Lemley
Heliaus Research Group
DTS, Xperi
Galway, Ireland
joe.lemley@xperi.com

***Abstract*—Thermal infrared imaging holds promise for human-machine interaction in vehicles owing to superior performance in low-light and low-visibility conditions, and the potential for monitoring human psycho-physiological state. However, the shortage of large-scale 2D thermal image datasets and public benchmarks has hindered progress of deep-learning-based solutions. To tackle this problem, we develop a pipeline for creating a synthetic thermal image dataset. Firstly, 3D models of human heads are generated from uncalibrated TIR images (without additional visible or depth images) using photogrammetry techniques. A synthetic dataset of 100k images of 640x480 resolution are then generated by rendering each of the five 3D models for a range of head poses, camera positions and backgrounds using commercial animation software. The effectiveness of the approach is evaluated using a number of deep learning algorithms that may enable human-machine interaction such as head pose estimation and face detection. The neural networks are trained on the new synthetic thermal dataset, before fine tuning on real world data where possible.**

***Index Terms*—deep learning, driver monitoring systems, face detection, head pose estimation, synthetic datasets, thermal infrared imaging**

I. Introduction

Quality of experience (QoE) measures the overall acceptability of an application or service, from the perspective of the end user [1]. The proliferation of cameras in the cabins of automobiles presents an opportunity to acquire feedback for enhancing the QoE for driver and passengers.

Face and head pose estimation are important elements of human behavioural analysis, that have attracted attention from the human-computer interaction, computer vision, and virtual reality communities. Head pose, for example, determines the human visual field and can act as a pre-processing step to aid in gaze estimation and attention modelling. Real-time estimates of head pose enable new ways to manipulate multimedia content and interact with users.

Thermal infrared (TIR) imaging has the potential to extract physiological parameters to be used in driver state estimation such as fatigue, drowsiness, aggressiveness, emotion, stress and cognitive load estimation [2]. It is useful in driver monitoring applications owing to the lack of illumination variance and the ability to penetrate smokes, aerosols, dust, and mists more effectively than visible radiation. However, there is still a general lack of publicly-available thermal image datasets. We take the first steps towards addressing this in the current paper by making the following contributions:

- We demonstrate that it is possible to use photogrammetric techniques - such as structure from motion (SfM) and multi-view stereo (MVS) - to generate 3D thermal models of humans from uncalibrated TIR images without additional visible or depth images
- We develop a scalable pipeline for generating synthetic datasets of thermal images of humans using rendering software
- We train and evaluate deep learning algorithms for face detection and head pose estimation on the new synthetic dataset for the application of human-machine interaction in vehicles

II. Related Work

A. Human-Machine Interaction in Vehicles

While some experts predict that fully autonomous vehicles (AVs) are still at least a decade away [3], an increasing amount of functions are being passed to the vehicle, as described by the classifications on autonomy [4]. Levels 0 to 2 cover no driving automation to partial automation that still requires the drivers full attention. AVs with Levels 3 to 4 automation will enable users to disengage from the driving task for extended periods and vehicles with Level 5 automation can perform self-driving under all conditions without any takeover from drivers during a journey.

Human-machine interaction in vehicles consists of operation of the primary vehicle controls by the driver and the interaction

and communication of the occupants with a variety of in-vehicle systems and applications. Currently, the design of in-vehicle systems is largely concerned with providing feedback, information and support to drivers while driving the car manually. In future, this interaction will cover takeover behaviour as well as performing other non-driving related tasks such as interacting with multimedia (perhaps even watching movies projected over the front windshield [5]).

For the foreseeable future, there will likely be a diverse set of non-driverless vehicles, offering different levels of automation functionality as well as services in entertainment and comfort, and the QoE of the occupants in using these will be affected by a number of influence factors [1]. With respect to human influence factors, studies have shown that older drivers display safer and more cautious behaviour in takeover situations [6]. During non-driving periods, younger drivers are likely to use electronic devices while older drivers become more heavily engaged in non-driving related tasks and prefer to talk to other people [7]. As well as age, factors such as gender, socio-cultural and economic background, mood and emotional state are likely to play a significant role. Furthermore, context influence factors such as an occupant's position in space and the conditions in the cabin (e.g. lighting) are also important for high-quality interactions. Vision-based techniques have the potential to provide useful feedback on some of these factors. For example, real-time estimates of head pose enable new ways to manipulate multimedia content and interact with users.

B. Face Detection and Head Pose Estimation

Deep learning has transformed the field of computer vision, achieving state-of-the-art performance on a number of tasks. While real data with ground-truth labels remains the gold standard for training, the maturity of 3D rendering engines means that generated datasets can improve the generalisation performance of deep learning algorithms. This is due to the ability to synthesise images over a high range of parameters, including orientation, background and illumination.

Data augmentation is distinguished from synthetic data generation by the intention to modify real data rather than create new synthetic data using computer graphics techniques. Since CNNs implicitly consider the surrounding context of the objects, the performance should be improved when the object placements are based on semantic and geometric context of the scene. Alhaija et al. [8] augmented real backgrounds with synthetic cars consistent with the scene. Georgakis et al. [9] found that augmenting some hand-labeled training data for object detection with synthetic object models blended into real indoor scene backgrounds achieved comparable performance to using much more hand-labeled data. Vazquez et al. [10] performed pedestrian detection using a hybrid dataset of real and synthetic images. Hattori et al. [11] superimposed pedestrians on real backgrounds taking into account the perspective geometry of the scene for the task of detection.

Manually annotating data with orientation angles is difficult and time consuming and synthetic data can be used to produce very accurate annotations [12]. An early paper by Su et al. [13] showed that object viewpoint estimation algorithms trained on purely synthetic data could outperform the state of the art. When tested on real data, Gupta et al. [14] found that their algorithm outperformed alternatives trained on real data. More recently, Tremblay et al. [15] achieved state-of-the-art results for object pose estimation by training a deep neural network on purely synthetic data. Movshovitz et al. [16] found that realism of the data was important.

Synthetic 3D models of people and faces are much harder to create than models of basic objects. Face models such as 3DMM (3D morphable model) [17] have been used to reduce the dimensionality of predictions, to help deal with issues such as occlusion and to perform data augmentation. Zhu et al. [18] use a profiling method based on this simplified model to predict depth information, before varying the yaw angle to synthetise training samples of faces across large poses. Ruiz et al. [19] learn to directly regress head pose from images by train on a large synthetic dataset. The generalization capacity of the network was illustrated by successfully testing on various real datasets without any fine tuning.

3D datasets for face and head pose use a combination of photogrammetric, depth and stereo imaging techniques [20]. The BIWI dataset [21] was collected by acquiring about 15,000 frames of RGB-D video of numerous subjects across different head poses using a Kinect v2 device. Pose annotations were produced by fitting a 3D model to the point cloud of each individual. Wang et al. [22] found that existing head pose datasets contained limited number of samples and variation, and partially incomplete annotation. They created a synthetic dataset of human head pose and trained a deep neural network that combines the strengths of classification and regression methods.

The ability to improve generalisation and address the issue of dataset bias (e.g. in terms of gender and race) is highly desirable. Gerig et al. [23] and Kortylewski et al. [24] pre-train neural networks for face detection, pose and landmark detection using large quantities of synthetic data before fine-tuning on real world data. It was found that generalisation performance was improved, and the damage of dataset bias was largely reduced.

III. Synthetic Dataset Generation

The impressive results of the previous section have been largely confined to visible images and there are no 2D datasets of real thermal images with head pose annotations, or 3D datasets of humans with thermal textures. The aim of this section is take the first steps in this direction and to provide a proof of concept for a pipeline for generating synthetic thermal image data.

A. 3D Modelling using Photogrammetry

Photogrammetry for thermal images is not well developed and commercially-available software tends to target visible images. Furthermore, the re-purposing of existing visible image processing infrastructure towards TIR images has not been fully explored. TIR images possess statistical regularities

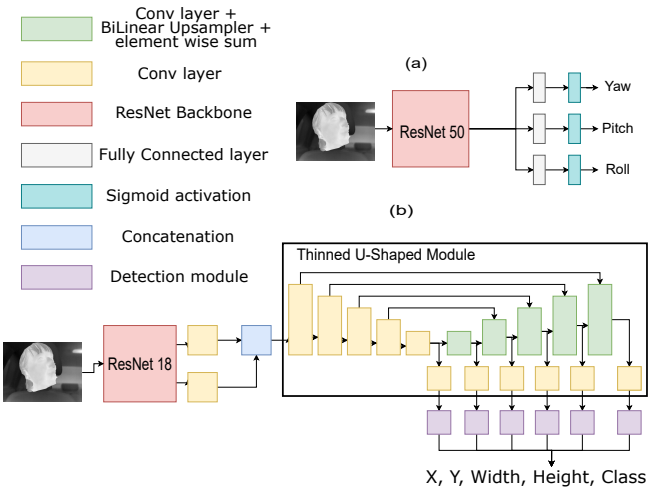

Fig. 1. (a) Head pose estimation network using ResNet-50 as backbone and 3 fully connected layers followed by sigmoid activations to regress yaw, pitch and roll. (b) M2Det architecture for face detection using ResNet-18 as the model backbone and include a single TUM to extract feature maps at varying scales. Each extracted feature map is passed to a detection module, each of which is responsible for detecting different bounding box sizes.

similar to those of visible light images [25] and have several advantages - such as illumination invariance - that should be beneficial for reconstruction. In fact, this characteristic of thermal information has been exploited to improve the robustness of camera pose estimation of other modalities [26].

Of course, some characteristics of thermal images present problems for image processing techniques For example, the lesser number of points in areas with little temperature variation can lead to higher reconstruction error (although this should be less problematic for human skin). Furthermore, edges are less visible in the TIR with respect to the visual spectrum. Perhaps most significantly, the lower resolution of commercial thermal cameras leads to lower 3D geometric resolution. As a result, a number of studies have opted to fuse images from depth sensors with thermal images for 3D modelling applications [27], [28]. We demonstrate that it is possible to use out-of-the-box photogrammetric techniques - such as SfM and MVS - to generate 3D thermal models of humans from uncalibrated TIR images without additional visible or depth images.

To start, 3D thermal head models were generated from a sequence of thermal images captured from various arbitrary viewpoints around five subjects' heads. A full rotation at a constant speed was captured and approximately 100 frames were extracted from the footage for each subject. It was found that 3D model reconstruction was improved by first masking the background using edge detection and thresholding. The masked images were loaded into the commercial software 3DF Zephyr (developed by 3D Flow) for the 3D thermal model construction. 3DF Masquerade was used to generate a mask for use in 3DF Zephyr by setting the green in the segmented images as the background. For the reconstruction, it was found that using between 60 and 120 images yielded the best results. Straying too high from that range would produce models with high noise, while going lower would lead to a higher image rejection rate by 3DF Zephyr. Additionally, including a lower number of images with subjects facing the camera was found to lead to better results in term of model cohesiveness. The 3D models were exported as .obj files with the thermal data overlaid as the material texture.

B. Image Generation

Commercial software for rendering thermal radiation is not yet publicly available. Thermal radiation shares some similarities with the visible and near-infrared spectrums in that thermal energy reflects off objects. This makes it possible to perform ray tracing simulations to track the reflections of thermal radiation where the emissivities of surfaces are known [29]. Unlike the other wavebands however, objects themselves are light sources in the infrared spectrum as a result of the thermal energy that they emit. To simulate thermal emission from a synthetic object requires more than just the shape of the surface and hence 3D thermographic imaging likely holds more promise than thermal modelling for generating synthetic datasets.

Blender, an open-source software, was used for the generating thermal images for each 3D model. The 3D models were imported and the head poses were parameterised by three values that approximate the range of motion of the head [30] $(\theta_p, \theta_y, \theta_r)$, $\theta_p \in [-60°, 60°]$, $\theta_y \in [-30°, 30°]$, $\theta_r \in [-20°, 20°]$, where θ_p, θ_y and θ_r represent the pitch, yaw, and roll, respectively. A Python layer was built on top to allow for generating large quantities of synthetic data automatically. Several augmentation operations were applied in order to generate a large and varied dataset of thermal head poses. The position of the heads in 3D space could vary by ± 10 cm in all directions with respect to the camera. Random non-linear scaling operations were applied to the five prototype head poses such that all axis could be scaled independently by a random scaling factor within a $\pm 20\%$ range of their original size. The camera focal length f was randomly chosen $f \in$ [15 mm, 35 mm]. Ambient lighting was used to control the average brightness of the thermal heads. Ambient lighting contributes equally to all parts of the thermal head, and therefore, it does not change the relative thermal information between various points on the thermal head. In contrast, directional light sources would have caused uneven illumination of the thermal heads. The heads were overlaid on a randomly chosen background image taken from a set of real-world thermal images acquired from inside a car cabin. Perspective view was used for realism. Gaussian noise was added to the background thermal images as well as the texture materials of the 3D models. In total, a dataset of 100,000 synthetic thermal images was created.

IV. Deep Learning Pipeline Human-Machine Interaction in Vehicles

This section presents the training and evaluation of deep convolutional neural networks on the new synthetic thermal

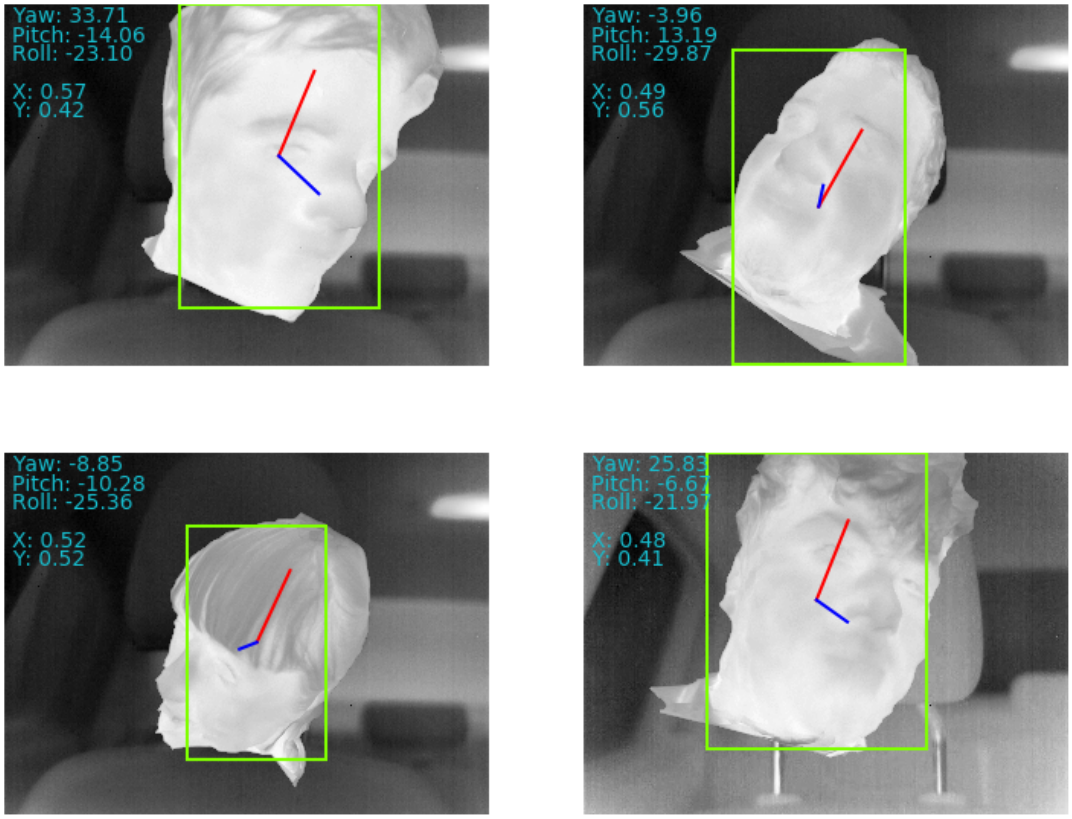

Fig. 2. Examples of synthetic thermal images from the dataset, with detected face and head pose estimates.

image dataset for tasks associated with human-machine interaction in vehicles.

A. Face Detection

The first step in the vision pipeline involves face detection with the purpose of providing the approximate 2D position of the head in the image. We develop our object detection algorithm based on the model outlined in [31] (known as M2Det). We use ResNet-18 to serve as the model backbone. Features are extracted from the backbone at two scales and fused before being input to a single Thinned U-shape Module (TUM). A TUM is an encoder-decoder pair used to extract features at varying scales. These extracted features each feed into a detection module which consists of two convolutions responsible for regression and classification. Each detection module is responsible for detecting different object scales. This makes it especially proficient at detecting faces. Figure 1(b) demonstrates the model architecture. Images are resized to 416 $\times$ 240 and standard image augmentations including flipping, cropping and distortions to hue, saturation and brightness, are applied. The model is pretrained on the COCO dataset for 160 epochs and trained a further 30 epochs on the synthetic thermal dataset [32]. Training is performed on two 12 GB NVIDIA 2080 GPUs with a batch size of 192 images (96 per GPU). The learning rate is set to 0.004 with a 5 epoch warm-up.

Evaluation is performed on a small manually annotated thermal dataset of real faces. The same form of masking used for the 3D modelling is first applied to each image. Figure 3 presents an example of the output from the face detection model. The performance of the face detection algorithm is shown in Figure 4 as a Receiver Operating Characteristic (ROC) curve. Though the evaluation dataset is small at 30 images the algorithm successfully detects faces with an intersection over union (IoU) greater than 0.6 for all examples. The bounding box predictions are passed on to the next stage of the pipeline.

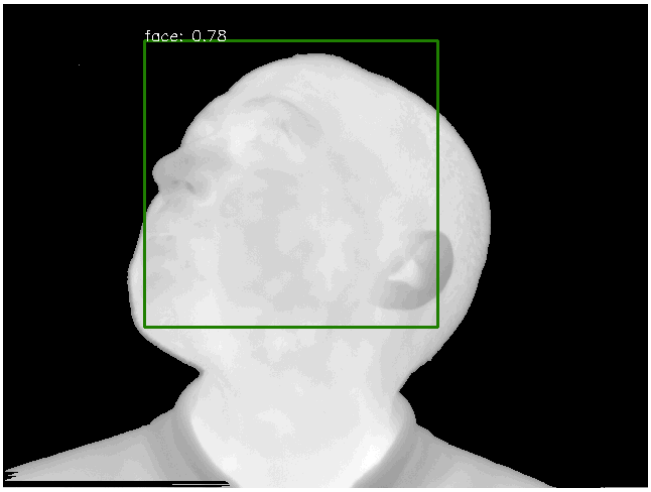

Fig. 3. Example output of object detection algorithm on real thermal data.

B. Head Pose Estimation

Head pose estimation is an important component of human-machine interaction systems that gives an indication of attention. The algorithm for head pose estimation is based on direct regression of head pose Euler angles from images, as opposed to multi-stage methods that make use of facial landmarks. We use the multi-loss approach of Hopenet [19], where the authors found that regressing head pose Euler angles directly from images did not perform well on synthetic training data. They proposed a separate loss for each angle, where the three signals are backpropagated into the network to improve learning. Each of these seperate loss functions uses a combination of classification and regression terms:

$$\mathcal{L} = H(y, \hat{y}) + \alpha \cdot MSE(y, \hat{y}) \tag{1}$$

where H and MSE are the cross-entropy and mean-squared error loss, respectively. Posing the problem as a classification task in discrete space exploits the stability of the softmax layer and cross-entropy, while the inclusion of a regression term improves the fine-scale accuracy.

The images are first resized to 320 $\times$ 240 pixels. The backbone of the architecture is a ResNet-50 [33] consisting of an input stem and four stages. The input stem is composed of a 7$\times$7 convolution with 64 filters and a stride of 2. This is followed by a 3$\times$3 max pooling layer with a stride of 2. The residual blocks in each of the four stages are bottleneck structures. The output layer consists of three branches of 2 fully-connected layers (one for each Euler angle), with the prediction discretised into one of 66 bins.

As previously mentioned, there are currently no datasets available online of real thermal images with head pose annotations. As a result, the network is trained and tested on synthetic data. The training set consists of 4 subjects while the final subject is withheld for testing. Training is performed using stochastic gradient descent for 150 epochs with a learning rate of 10^{-4} and $\beta_1 = 0.9$. We experiment with different orders of magnitude of α before finding that the small value of 0.001 produces optimal results.

TABLE I
MEAN-ABSOLUTE ERROR OF HEAD POSE EULER ANGLES IN TERMS OF YAW, PITCH AND ROLL. * EVALUATED ON REAL VISIBLE DATASET.

	Yaw	Pitch	Roll	MAE
Multi-Loss ResNet-50	4.9°	6.8°	5.2°	5.6°
Ruiz et al. [19]*	3.3°	3.4°	3.0°	3.2°
Wang et al. [22]*	4.76°	5.48°	4.29°	4.8°

Figure 2 shows outputs of the deep learning pipeline on the training set. Table I presents the mean-absolute error of head pose Euler angles in terms of yaw, pitch and roll, evaluated on the test set. The mean errors are similar in magnitude to existing studies (although these studies evaluate on larger real datasets). While tasks like detection and segmentation should be easier in the TIR waveband, it is possible that the lack of texture in thermal images makes head pose estimation more challenging. Nonetheless, the ability to analyse temperature variations over regions of the face and infer psycho-physiological state makes thermal imaging worthy of future research endeavours.

V. CONCLUSION

This work has investigated the use of synthetic thermal image data of humans for training face detection and head pose estimation algorithms for human-machine interaction in vehicles. We achieve a mean-absolute error of 5.6° over the three Euler angles, and a 98% percent F1 score with a

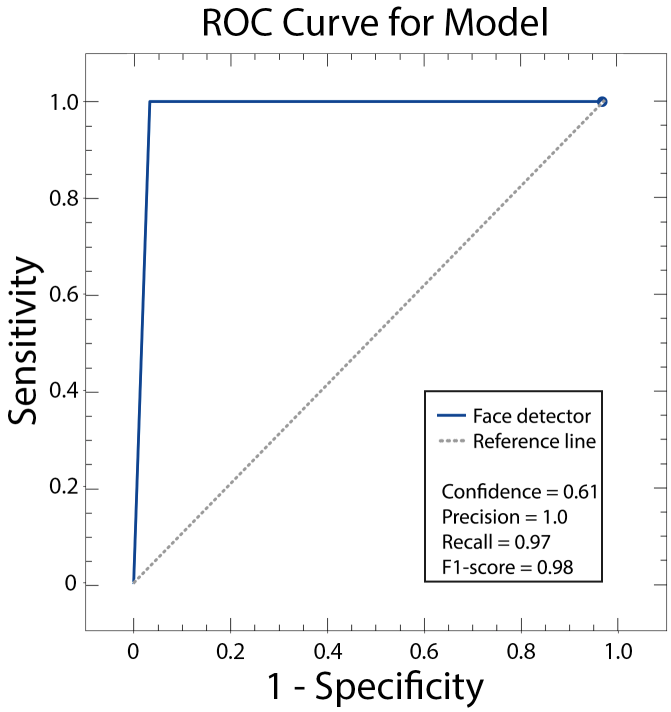

Fig. 4. Performance of face detection model applied to real-world thermal imagery. The optimal F1 score achieved at a confidence level of 0.61.

detection threshold of 60% IOU (albeit for small synthetic and real validation sets, respectively).

Since photogrammetry for thermal images is not well developed, a number of studies opt to fuse thermal images with depth for 3D modelling applications, although it is more cost effective to use thermal images alone. In the current work, we demonstrate that TIR images are similar enough that photogrammetry techniques alone are sufficient for reconstruction. To the best of our knowledge, scanned 3D objects with thermal textures have not been used for generating synthetic datasets of 2D thermal images. While our synthetic data currently lacks some realism, it is highly scalable with the potential to provide large quantities of training data to improve generalisation.

Head pose and face detection can act as pre-processing steps to aid in gaze estimation, attention modelling and psychophysiological state estimation to determine vehicle takeover and for interacting with controls or multimedia. This presents many opportunities for enhancing the QoE of the occupants as they engage with various primary and secondary services.

ACKNOWLEDGMENT

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826131. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Germany, Ireland, Italy.

REFERENCES

[1] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, "Factors influencing quality of experience," in *Quality of Experience*. Springer, 2014, pp. 55–72.

[2] S. Ioannou, V. Gallese, and A. Merla, "Thermal infrared imaging in psychophysiology: Potentialities and limits," *Psychophysiology*, vol. 51, no. 10, pp. 951–963, 2014.

[3] T. Litman, *Autonomous vehicle implementation predictions*.

[4] S. O.-R. A. V. S. Committee *et al.*, "Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems," *SAE Standard J*, vol. 3016, pp. 1–16, 2014.

[5] M. A. Cuddihy and M. K. Rao, "Autonomous vehicle entertainment system," Mar. 1 2016, US Patent 9,272,708.

[6] M. Körber, C. Gold, D. Lechner, and K. Bengler, "The influence of age on the take-over of vehicle control in highly automated driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 39, pp. 19–32, 2016.

[7] H. Clark and J. Feng, "Age differences in the takeover of vehicle control and engagement in non-driving-related activities in simulated driving with conditional automation," *Accident Analysis & Prevention*, vol. 106, pp. 468–479, 2017.

[8] H. A. Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 961–972, 2018.

[9] G. Georgakis, A. Mousavian, A. C. Berg, and J. Kosecka, "Synthesizing training data for object detection in indoor scenes," *arXiv preprint arXiv:1702.07836*, 2017.

[10] D. Vazquez, A. M. Lopez, J. Marin, D. Ponsa, and D. Geronimo, "Virtual and real world adaptation for pedestrian detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 797–809, 2013.

[11] H. Hattori, V. Naresh Boddeti, K. M. Kitani, and T. Kanade, "Learning scene-specific pedestrian detectors without real data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3819–3827.

[12] S. I. Nikolenko, "Synthetic data for deep learning," *arXiv preprint arXiv:1909.11512*, 2019.

[13] H. Su, C. R. Qi, Y. Li, and L. J. Guibas, "Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2686–2694.

[14] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, "Inferring 3D object pose in RGB-D images," *arXiv preprint arXiv:1502.04652*, 2015.

[15] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," *arXiv preprint arXiv:1809.10790*, 2018.

[16] Y. Movshovitz-Attias, T. Kanade, and Y. Sheikh, "How useful is photo-realistic rendering for visual learning?" in *European Conference on Computer Vision*. Springer, 2016, pp. 202–217.

[17] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[18] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 146–155.

[19] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2074–2083.

[20] M. D. Breitenstein, D. Kuettel, T. Weise, L. Van Gool, and H. Pfister, "Real-time face pose estimation from single range images," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[21] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437–458, 2013.

[22] Y. Wang, W. Liang, J. Shen, Y. Jia, and L.-F. Yu, "A deep coarse-to-fine network for head pose estimation from synthetic data," *Pattern Recognition*, vol. 94, pp. 196–206, 2019.

[23] A. Kortylewski, A. Schneider, T. Gerig, C. Blumer, B. Egger, C. Reyneke, A. Morel-Forster, and T. Vetter, "Priming deep neural networks with synthetic faces for enhanced performance," 2018.

[24] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[25] T. R. Goodall, A. C. Bovik, and N. G. Paulter, "Tasking on natural statistics of infrared images," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 65–79, 2015.

[26] A. O. Müller and A. Kroll, "On the temperature assignment problem and the use of confidence textures in the creation of 3D thermograms," in *2015 9th International Conference on Sensing Technology (ICST)*. IEEE, 2015, pp. 223–228.

[27] A. O. Muller and A. Kroll, "Generating high fidelity 3-D thermograms with a handheld real-time thermal imaging system," *IEEE Sensors Journal*, vol. 17, no. 3, pp. 774–783, 2016.

[28] Y. Cao, B. Xu, Z. Ye, J. Yang, Y. Cao, C.-L. Tisse, and X. Li, "Depth and thermal sensor fusion to enhance 3D thermographic reconstruction," *Optics Express*, vol. 26, no. 7, pp. 8179–8193, 2018.

[29] D. Robinson, R. Simpson, J. Parian, A. Cozzani, G. Casarosa, S. Sablerolle, and H. Ertel, "3D thermography for improving temperature measurements in thermal vacuum testing," *CEAS Space Journal*, vol. 9, no. 3, pp. 333–350, 2017.

[30] V. F. Ferrario, C. Sforza, G. Serrao, G. Grassi, and E. Mossi, "Active range of motion of the head and cervical spine: a three-dimensional investigation in healthy young adults," *Journal of Orthopaedic Research*, vol. 20, no. 1, pp. 122–129, 2002.

[31] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2Det: A single-shot object detector based on multi-level feature pyramid network," 2018.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014, pp. 740–755.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.