Real time single image dehazing & soil removal for advanced driving assistance systems using convolutional neural networks

Wajahat Akhtar*, Sebastian Carreno and Serio Roa-Ovalle

***Abstract*— Removal of atmospheric haze and soil from single images captured via monocular cameras is very challenging and computational ill pose phenomena in digital and Advanced Driving Assistance Imaging Systems. Such models are not applicable to run on a reasonable automotive embedded platform due to the deepness in the network. The challenge is to re- duce the number of layers while achieving the same performance in order to make it embedded friendly. In this paper, we propose a new Convolutional Neural Network (CNN) based model, which inspires EVD-Net j-level fusion and AOD-Net for real-time single image dehazing and soil removal for Nvidia Jetson embedded platforms. The CNN model is designed based on a reformulated atmospheric scattering model for haze by Koschmieder [1] and dirt removal by Eigen, Krishnan and Fergus [2], called Haze and Soil Removal Convolutional Neural Networks (HSRCNNs-Net). There exist different haze removal techniques in the literature among them AOD-Net, CANDY, DehazeNet, and MSCNN. This is a the first fully end-to-end model for real-time single image *haze* and *soil removal* for image enhancement for ADAS. The model is trained and tested using our own dataset created during the research for both haze and soil removal. Furthermore, a pre-trained Faster R-CNN model was used to verify the performance difference between hazy and soil images as compared to clean images. Finally, we witnessed a great improvement especially in object detection and image quality. Its design and performance makes it applicable to different scenario including ADAS, medical imaging, night imaging and underwater imaging. The lightweight mobile CNN allows easy cascading with other neural networks. The model was tested and evaluated using different public datasets such as RESIDE.**

I. Introduction

Advanced safety automobiles (ASAs) geared up with vision-based advanced driver assistance systems (ADAS) cameras have become rapidly greater prevalent in the automobile industry [1]. The development of such sophisticated driving force has helped the driving experience to transform altogether. In the last few decades, we have witnessed a huge transition in ADAS sector from automobile mobility to e-mobility.

Advanced Driver Assistance Systems (ADAS) is a rising upcoming technology to improve road safety, autonomous driving, driver comfort and to reduce energy consumptions.[2]

This work was supported Adasens Automotive GmbH, Lindau , Germany.

W. Akhtar is a Master student in the field of Computer Vision and Robotics. wajahat.akhtar@ficosa.com

Sebastian Carreno and Sergio Roa-Ovalle are senior software developer at Adasens Automotive GmbH, Lindau, Germany sebastian.carreno@ficosa.com

[1]Future of autonomous cars is explained here https://toshiba.semicon-storage.com/ap-en/application/automotive/safety-assist/image-recognition.html

[2]The article explains about ADAS http://telematicswire.net/bmw-driving-into-the-automotive-future-with-adas/

Fig. 1: Proposed model for haze (Top left) : Input real hazy image, (Top right) : generated dehazed image.
Proposed model for soil (bottom left) : Input real image captured with soil camera and (bottom right) : generated clean image.

The ADAS uses single and multiple monocular cameras for autonomous parking, surround view system, edge-based lane and pedestrians detection. Object such as automobiles tracking, traffic sign detection and recognition are among the image processing applications of ADAS to ensure reliability. However, the detection and recognition qualities are strongly affected by frequent haze such as aerosols in the atmosphere [3] and Soil on camera lens e.g dust, sand, silt and clay [4]. Therefore, haze and soil removal have become a notable problem in ADAS for Autonomous cars. The presence of haze in the atmosphere due to the poor weather conditions have allowed the images acquired by cameras to suffer from poor quality and scene visibility. The light scattered by haze and soil can deteriorate not only the aesthetic beauty of the scene but also occludes important salient features in the images which significantly reduces the performance of the algorithm used in ADAS which need to ensure the reliability in different object detection algorithms. People in past have proposed different methods for haze removal, most recently with the advancement in deep learning computer vision has become an attractive field of research in ADAS, achieving the best state-of-the-art results especially in object classification and recognition. A common problem that exist between all the previous methods was the computational cost which made them unsuitable for ADAS. In this paper, we proposed a method to perform not only real time single image haze removal using CNN but also soil removal for embedded platforms. We designed a CNN based model for

both haze and soil removal using two different mathematical formulations.

II. Research context

Haze is traditionally an atmospheric phenomenon in which images captured under bad conditions dust, smoke, and other dry particulates obscure the clarity of the sky [3] whereas soil is a black or dark brown material typically consisting of a mixture of organic remains, clay, and rock particles [4] that normally occur on the cameras installed in ADAS and surveillance outdoor vision system(SOVs).

The first mathematical model for the formation of haze was formulated by [1] and later reformulated [5] 2, which is widely used by almost all the method proposed in the literature, shown as follows:

$$I(x) = J(x)t(x) + \alpha(1 - t(x)) \tag{1}$$

clean image generation module.

$$J(x) = K(x)I(x) - K(x) + b \tag{2}$$

The model incorporates two parts 1 and 3: the attenuation of transmitted light $t(x)$ in 1, which is the scene transmission map, and the haze absorption β 2, which is the scattering coefficient of the atmosphere which represents the ability of a unit volume of atmosphere to scatter light in all directions [6].$I(x)$ is the observed hazy image, $J(x)$ is the actual scene irradiance, α is the ambient or atmospheric light formed by the scattering of the environmental illumination and linked to the quantity of light illuminating the scene. x denotes an individual pixel location in the image. Whereas the reformulated model $K(x)$ is the integration of both α and $t(x)$ with variable b used as constant bias. As stated earlier, the scene transmission, is a function of depth and is given by:

$$t(x) = e^{-\beta d(x)} \tag{3}$$

Here, $d(x)$ is the depth of the scene point corresponding to the pixel location x

Whereas, for soil removal we follow the mathematical model initially presented by [4] as shown in 4 for soil lens artifact and later reformulated by [2] as shown in Equation 5 :

$$I(x) = I_0(x).a(x) + c.b(x) \tag{4}$$

Above here, I_0 is the clean image, $alpha(x)$ as the attenuation map(camera dependent). $c(x)$ represents aggregate of Outside Illumination and is scene dependent. Whereas $b(x)$ is the scattering Map and is also camera dependent.

$$I' = p\alpha D + I(1 - \alpha) \tag{5}$$

[3]The definition for haze is explained here https://en.wikipedia.org/wiki/Haze

[4]The definition for soil is explained here https://en.wikipedia.org/wiki/Soil

I represents the original clean image, I' as generated a noisy image. α is a transparency mask the same size as the image, and D is the additive component of the soil, also the same size as the image. p is a random perturbation vector in RGB space, and the factors $p\alpha D$ are multiplied together element-wise as discussed in [2].

Deep Convolutional Neural Networks (DCNN) have shown record-shattering performances in a variety of computer vision problems. Recently CNNs have been used for image dehazing and soil removal to produce better quality and clean images. However, there were many major issue and problems. When considering supervised methods there was a lack of sufficiently and correctly labeled data. Once the modeled are trained they were not portable to embedded platform. Whereas in this paper we try to overcome most of the aforementioned drawbacks by designing a method to generate clean images, with better quality and real time implementation for embedded platforms. We also have developed a technique to create a real dataset for both haze and soil.

A. Traditional Methodologies

In general, there exist three kinds of methodologies in literature for haze removal : **Multiple Images** [6]–[9] **Single Image** [10]–[17] and using **Deep learning** [5], [18]–[21]. Deep Learning for solving ill posed image dehazing is quite recent(2016) whereas for soil removal there first work was done back in 2013 by [2].

Earlier methods such [6], [8] used multiple images under different weather conditions and degree of polarization to perform haze removal. While other [22] approaches resorted to estimate atmospheric scattering model parameters with the empirical Dark Channel Prior. [17] provided a method to enhance the local contrast of the images based on the study that haze free images have higher contrast to non-hazy images. However [10], [23] presented a method to remove haze from images captured from moving vehicle camera. Recently this problem was addressed by [5], [18]–[21] using deep learning.

There exist a common problem among these methods, firstly the all are computational expensive except(AOD-Net). Secondly among all the methods in literature very few could design models to be used for dynamic scenarios specially in ADAS. According to the study by [24] and [25] none of these methods could produce high quality images except [18]. Due to their limitations and limited practical applicability these methods are not been used in ADAS. We try to solve most of the above problems by presenting a novel end-to-end deep learning model to generate haze and soil free images.

1) Contribution: The main contribution in this paper are summarized as follows:

1. HSRCNN-Net is a first real time single image haze and soil removal CNN architecture, which directly generates clean haze and soil free image with better quality, estimating attenuation and scattering parameters jointly. Whereas most of the method use multiple images with significant large computational cost.

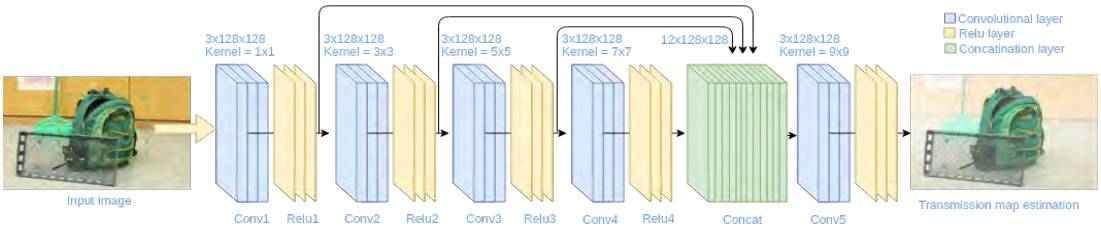
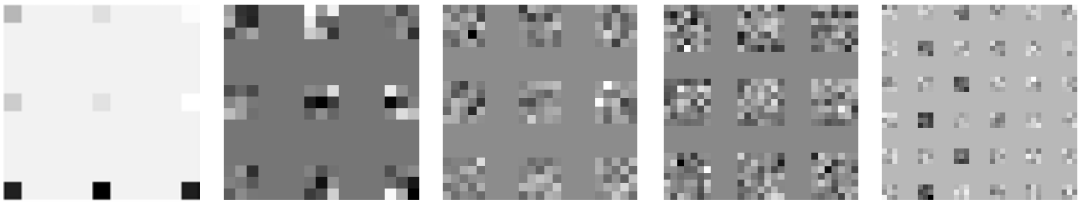


Fig. 2: The proposed architecture of HSRCNN-Net. HSRCNN-Net is constructed by 5 convolutional layers, 1 concatenation layer and Relu activation functions to estimate the transmission maps.



(a) Conv1 (b) Conv2 (c) Conv3 (d) Conv4 (e) Conv5

Fig. 3: Layer visualization of Proposed HSRCNN (left-right) : "conv1"-"conv5" layers and their kernels.

2. A unique setup with two monocular cameras is designed to record real time soil and non-soil images. The setup acquires real images with label dataset for training our soil model. It solves the problem of labeling the data captured from soil lens. As upto date there exist no single labeled public dataset for both synthetic and real images.

3. A novel technique was also established to generate synthetic dataset using real soil on cameras lens. Different soil samples were created and images were acquired using the samples. The soil was extracted from the images and then used as a mask for creating synthetic datasets.

4. All the available dataset in literature only use homogeneous haze to generate hazy images which is less realistic. Whereas, we created a method to generate synthetic dataset which contain homogeneous and non-homogeneous haze for training HSRCNN. The images are divided into patches and haze is generated using different hyper parameter for each patch as compared to state-of-the-art methods, where synthetic dataset were created using homogeneous haze only.

III. Model Architecture

In this work, we formulate the constrained problem of Real time single image dehazing and soil removal for ADAS for generating a high quality haze and soil free image from a degraded hazy and soil carrying input image. We propose a novel end-to-end convolutional neural network (CNN) called HSRCNN-Net. We use two different mathematical equations each for haze and soil with architecture comprises of same deep convolutional neural network (CNN) module, by estimating transmission map and global atmospheric light. The CNN architecture is designed based on the inspiration from j-level fusion of EVD-Net [26] and AOD-Net [5]. To generate clean image $J(x)$ it estimates $K(x)$ from input image $I(x)$, followed by a clean image generation module that utilizes $K(x)$ as its input-adaptive parameters to estimate clean image $J(x)$ [5] as shown in 2. Whereas for soil we use mathematical expression as given in 5. The estimation of K is significant for our model HSRCNN-Net as it estimates both haze and depth levels as shown in figure 2. Our model follows solely a standard CNN model as in [19]. Each convolutional layer applies a kernel composed of $w*h*d$ coefficients, with w defining the width, h as the height and d as the depth of the hidden convolutional layers. The depth of the layers depend on the number of activation maps in the layers. Each layer is followed by an activation function to introduce nonlinearity as discussed in [27] [19]. The first layer called $conv1$ takes an input single RGB image of $size = h*w$ and d, stride of $size = 1$ with kernel $size = 1$ results in 3 different activation maps, followed by layer "$relu1$" to introduce nonlinearity. The second layer takes $conv1$ as its input, with stride and pad $size = 1$ and kernel $size = 3$ generating 3 activation maps, followed by a "$relu2$" layer. Similarly "$conv3$" and "conv4" layers were created with kernel $size = 5$, and 7. Inspired from [21], which concatenates the coarse-scale network features we create layer 5 of of HSRCNN-Net, which concatenates "conv1","conv2","conv3" and "conv4" called "$concat1$" generating three channel R-G-B output, followed by a last convolutional layer "$conv5$" of kernel $size = 9$. The output of the layer "$conv5$" i.e $'K(x)'$ is the estimated transmission map with global atmospheric light. Which is the then used as a prior to generate a clean image in both and soil removal cases using 2 and 5 as could be seen in figure 2 with layer visualization of each kernel in each layer as shown in Figure 3.

A. Dataset creation and Training

Training for both haze and soil model was performed using Deep learning framework caffe [28].

Haze model : As there exist no benchmark datasets for haze and its corresponding non-hazy images except [29]

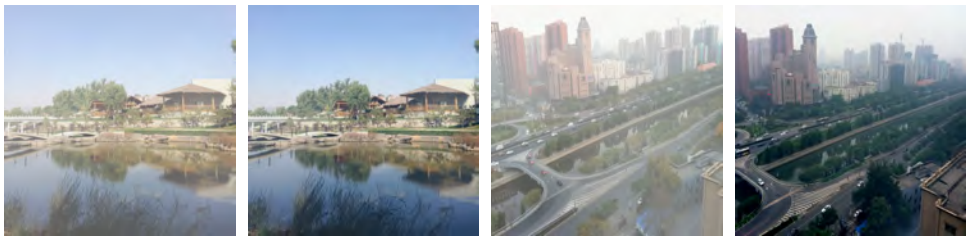
TABLE I: Average full and no reference evaluation results of dehazing on Synthetic Objective Testing Set (SOTS).

Metrics	DCP	GRM	CAP	NLD	DehazeNet	MSCNN	AOD	HSRCNN-Net
PSNR	16.62	18.86	19.05	17.29	21.14	17.57	19.06	22.24
SSIM	0.8179	0.8553	0.8364	0.7489	0.8472	0.8102	0.8504	0.862
Time	1.62	83.6	0.95	9.89	2.51	2.60	0.65	0.28

TABLE II: Average subjective score, with full and no reference evaluation results of dehazing on Hybrid Subjective Testing Set (HSTS).

Metrics	DCP	GRM	CAP	NLD	DehazeNet	MSCNN	AOD	HSRCNN-Net
PSNR	14.84	18.54	21.53	18.92	24.48	18.64	20.55	23.24
SSIM	0.7609	0.8184	0.8726	0.7411	0.9153	0.8168	0.8973	0.892
Time	1.62	83.6	0.95	9.89	2.51	2.60	0.65	0.28

(a) Foggy image (b) Clean image (c) Foggy image (d) Clean image



(e) Foggy image (f) Clean image (g) Foggy image (h) Clean image

Fig. 4: Experimental results of HSRCNN haze removal on public datasets.

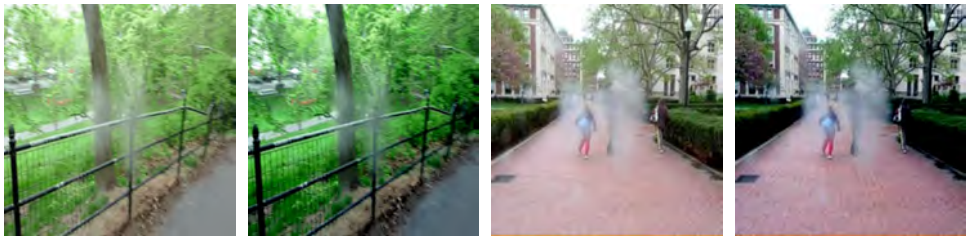
which only uses homogeneous haze. Therefore we decided to create a dataset which contains both homogeneous and non homogeneous haze with different levels. The nature of haze was inspected after studying natural images, as haze was non-homogeneous in nature and its concentration is not constant over the image space(the fog might be denser over a body of water due to its vaporisation). A Synthetic dataset of fifty thousand training and twenty thousand validation non overlapping hazy images was generated using our Automotive and region segmented SUN2012 dataset of cleaned images. Synthetic haze was added to each segmented region as been explained in [5]. The training data was converted into hdf 5 format as explained in [5]. Weights are initialized using Gaussian random variables with Relu neuron as stated in [27] and [19] it performed effective then BRelu. The base learning rate was set to $base_lr : 0.000001$ with $lr_policy : "step"$. The model is trained with a batch size of hundred taking five hundred iteration to complete one epoch, In total the model converged in less then ten epoch (i.e five hundred total iterations) using Stochastic Gradient Descent "SGD".

Soil model : Similar to haze, there exist no benchmark or public datasets for images with and without soil on camera lens. To create a dataset with ground truth a novel simple technique is designed to extract soil from images taken from monocular cameras. Different soil samples are created and the soil is extracted, to create a labeled dataset with and without soil from real images. A dataset of thirty thousand labeled images for training and ten thousand non overlapping soil images for testing are generated, as explained in Section 3.1 of [2]. The training data is first converted into $hdf5$ format as presented in [5]. Weights are initialized using Gaussian random variables with Relu neuron as for haze. The base learning rate, learning policy, step size and batch size is set accordingly. The model is trained using Tesla P100-PCIE and tested real time on Nvidia Jetson TK1 pro.

[5]http://machinelearninguru.com/deep_learning/data_preparation/hdf5/hdf5.html

[5]http://machinelearninguru.com/deep_learning/data_preparation/hdf5/hdf5.html

(a) Soil image (b) Clean image (c) Soil image (d) Clean image



(e) Soil image (f) Clean image (g) Soil image (h) Clean image

Fig. 5: Experimental results of HSRCNN soil removal on public datasets.

Fig. 6: HSRCNN haze removal performance evaluation using FastRCNN, (left-right) : Input hazy image, generated clean image by HSRCNN, FastRCNN applied on hazy input with recognition rate of 0.96, 0.34 and 0.43, FastRCNN applied to generated clean image with better recognition rate such as 0.96, 0.51, 0.45, 0.33, and 0.63.

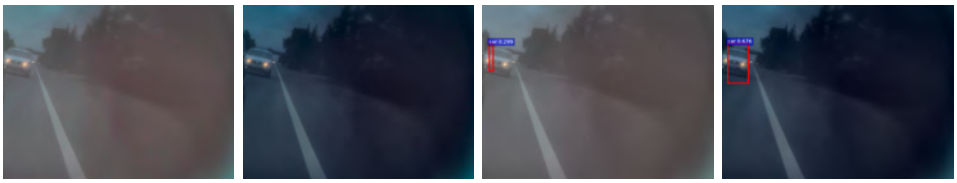


Fig. 7: HSRCNN soil removal performance evaluation using FastRCNN, (left-right) : Input real image captured with soil lens, generated clean image by HSRCNN, FastRCNN applied on soil input results in a recognition rate of 0.299, FastRCNN applied to clean image results with a recognition rate of 0.670.

IV. Experimental Results

In this section we compared our proposed model with several state-of-the-art methods using CNNs. As discussed, two different datasets are generated one for soil and one for haze. To evaluate our algorithm for haze removal we use synthesized benchmark dataset caleed RESIDE [29]. To conduct a fair test we compute PSNR and SSIM [29]. SSIM computes errors beyond pixel level and reflects human perception. The qualitative results achieved for haze removal are shown in 4 whereas, for soil removal in 4 . Table I and II depicts that our model produces promising results both in terms of peak signal to noise ratio (PSNR) and structural similarity index(SSIM).. To performed some further experiments we used public single images for soil and haze as shown in figure 1 and Figure 5. Lastly, Fast-RCNN is applied to further compare and verify the performance of HSRCNN on haze images from RESIDE. It is found that, the performance of the image tested significantly improved its quality for both haze and soil as shown in Figure in 6. To check the robustness of the model, different test images are taken at varying lighting scenarios. To compare HSRCNN-net the model was also tested with various methods, such as Fast Visibility Restoration (FVR) [30], Dark Channel Prior (DCP) [31], Boundary Constraint and Contextual Regularization (BCCR) [32], Color Attenuation Prior (CAP) [33], Non-local Image Dehazing (NLD) [34], Dehaze-Net [19], Multi-Scale Convolutional Neural Networks (MSCNN) [21] and All in One Dehazing network (AOD) [5] .

V. Conclusions and Future Work

The paper proposes two main components: an optimized network design to estimate the transmission map, and a mathematical model each for haze and soil to generate single clean image. Initially, we aimed to explore atmospheric haze removal techniques under different weather conditions using CNNs. Nevertheless, we decided to further pursue this promising approach for real-time single image soil removal, by building our own proof of concept.

1) HSRCNN-Net is the first real-time single image haze and soil removal CNN architecture, Whereas most of the method use multiple images with significant large computational cost.
2) The speed achieved during testing of our proposed model is marked as 0.28 sec, which outperforms the best score [5] in the literature.
3) A unique setup with two monocular cameras was designed to record real time soil and non-soil images.
4) A new technique for soil extraction is established to generate synthetic dataset using real soil on cameras lens.
5) A realistic was applied to generate homogeneous and non-homogeneous haze.
6) The model is evaluated and compared with state-of-the-art methods using public and our own automotive dataset, as shown in Figure 7. Moreover our network is tested on different real and synthetic datasets with different lighting conditions to prove the robustness. Lastly the model is tested with FastRCNN to check the recognition performance of the network.
7) A few sample input and output images are presented in Figure 5, 6 and 4 to show the impact of our network layers, HSRCNN. As seen, it is clearly visible by human eye as well as the quality of the images are enhanced with better SSIM and peak signal to noise ratio. Our quantitative results are shown in Table I, II.

HSRCNN contains optimized model architecture, fast speed and reduce number of layers, a unique model suitable for embedded platforms. The light weight structure allows it to be used in different fields e.g medical imaging and robotics. In future we aim to design a new model called joint HSRCNN to jointly remove haze and soil from single image with one mathematical formation.

References

[1] Koschmieder, *H. Theorie der horizontalen sichtweite.* Beitr. Zur Phys. d. freien Atm 171-181, (1924).

[2] D. Eigen, D. Krishnan, and R. Fergus, "Restoring an image taken through a window covered with dirt or rain," pp. 633–640, Dec 2013.

[3] S. Hwang and Y. Lee*, "Sharpness-aware evaluation methodology for haze-removal processing in automotive systems," *IEIE Transactions on Smart Processing and Computing*, vol. 5, pp. pp. 390–394, Dec. 2016.

[4] J. Gu, R. Ramamoorthi, P. Belhumeur, and S. K. Nayar, "Removing image artifacts due to dirty camera lenses and thin occluders," vol. 28, 12 2009.

[5] B. Li, X. Peng, Z. Wang, J.-Z. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[6] S. G. . N. S. K. Narasimhan, "Contrast restoration of weather degraded images," *IEEE Trans. Pattern Anal. Mach. Intell.(TPAMI)*, vol. 25, pp. 824–840, 2003.

[7] S. K. Nayar and S. G. Narasimhan, "Vision in bad weather," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 820–827 vol.2, 1999.

[8] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar, "Instant dehazing of images using polarization," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, pp. I–325–I–332 vol.1, 2001.

[9] S. Shwartz, E. Namer, and Y. Y. Schechner, "Blind haze separation," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1984–1991, 2006.

[10] N. Hautiere, J. P. Tarel, and D. Aubert, "Towards fog-free in-vehicle vision systems through contrast restoration," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2007.

[11] S. G. Narasimhan and S. Nayar, "Interactive deweathering of an image using physical models," in *IEEE IEEE Workshop on Color and Photometric Methods in Computer Vision, In Conjunction with ICCV*, October 2003.

[12] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, vol. 27, pp. 72:1–72:9, Aug. 2008.

[13] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2341–2353, Dec 2011.

[14] J. P. Tarel and N. Hautire, "Fast visibility restoration from a single color or gray level image," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 2201–2208, Sept 2009.

[15] K. Nishino, L. Kratz, and S. Lombardi, "Bayesian defogging," *Int. J. Comput. Vision*, vol. 98, pp. 263–278, July 2012.

[16] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *2013 IEEE International Conference on Computer Vision*, pp. 617–624, Dec 2013.

[17] R. T. Tan, "Visibility in bad weather from a single image," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.

[18] K. Swami and S. K. Das, "CANDY: Conditional Adversarial Networks based Fully End-to-End System for Single Image Haze Removal," *ArXiv e-prints*, Jan. 2018.

[19] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, pp. 5187–5198, Nov 2016.

[20] Z. Ling, G. Fan, Y. Wang, and X. Lu, "Learning deep transmission network for single image dehazing," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2296–2300, Sept 2016.

[21] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *European Conference on Computer Vision*, 2016.

[22] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, pp. 2341–2353, Dec. 2011.

[23] K. Ma, W. Liu, and Z. Wang, "Perceptual evaluation of single image dehazing algorithms," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 3600–3604, Sept 2015.

[24] C. Ancuti, C. O. Ancuti, and C. D. Vleeschouwer, "D-hazy: A dataset to evaluate quantitatively dehazing algorithms," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2226–2230, Sept 2016.

[25] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, ECCV'12, (Berlin, Heidelberg), pp. 746–760, Springer-Verlag, 2012.

[26] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "End-to-end united video dehazing and detection," *CoRR*, vol. abs/1709.03919, 2017.

[27] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, (USA), pp. 807–814, Omnipress, 2010.

[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *CoRR*, vol. abs/1408.5093, 2014.

[29] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Reside: A benchmark for single image dehazing," *arXiv preprint arXiv:1712.04143*, 2017.

[30] J. P. Tarel and N. Hautire, "Fast visibility restoration from a single color or gray level image," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 2201–2208, Sept 2009.

[31] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2341–2353, Dec 2011.

[32] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *2013 IEEE International Conference on Computer Vision*, pp. 617–624, Dec 2013.

[33] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Transactions on Image Processing*, vol. 24, pp. 3522–3533, Nov 2015.

[34] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1674–1682, June 2016.

[35] R. Girshick, "Fast r-cnn," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, (Washington, DC, USA), pp. 1440–1448, IEEE Computer Society, 2015.

[36] M. Sulami, I. Glatzer, R. Fattal, and M. Werman, "Automatic recovery of the atmospheric light in hazy images," in *2014 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–11, May 2014.

[37] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Transactions on Image Processing*, vol. 24, pp. 3522–3533, Nov 2015.

[38] Z. Li and J. Zheng, "Edge-preserving decomposition-based single image haze removal," *IEEE Transactions on Image Processing*, vol. 24, pp. 5432–5441, Dec 2015.

[39] G. Bi, J. Ren, T. Fu, T. Nie, C. Chen, and N. Zhang, "Image dehazing based on accurate estimation of transmission in the atmospheric scattering model," *IEEE Photonics Journal*, vol. 9, pp. 1–18, Aug 2017.

[40] W. Wang, X. Yuan, X. Wu, Y. Liu, and S. Ghanbarzadeh, "An efficient method for image dehazing," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2241–2245, Sept 2016.

[41] A. Galdran, J. Vazquez-Corral, D. Pardo, and M. Bertalmo, "Fusion-based variational image dehazing," *IEEE Signal Processing Letters*, vol. 24, pp. 151–155, Feb 2017.

[42] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, ECCV'12, (Berlin, Heidelberg), pp. 746–760, Springer-Verlag, 2012.

[43] C. D. V. Cosmin Ancuti, Codruta O. Ancuti, "D-hazy: A dataset to evaluate quantitatively dehazing algorithms," in *IEEE International Conference on Image Processing (ICIP)*, ICIP'16, 2016.

[44] P. Westling and H. Hirschmller, "High-resolution stereo datasets with subpixel-accurate ground truth," 09 2014.

[45] J.-P. Tarel, N. Hautière, A. Cord, D. Gruyer, and H. Halmaoui, "Improved visibility of road scene images under heterogeneous fog," in *Proceedings of IEEE Intelligent Vehicle Symposium (IV'2010)*, (San Diego, California, USA), pp. 478–485, 2010. http://perso.lcpc.fr/tarel.jean-philippe/publis/iv10.html.

[46] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[47] H. Xu, J. Guo, Q. Liu, and L. Ye, "Fast image dehazing using improved dark channel prior," in *2012 IEEE International Conference on Information Science and Technology*, pp. 663–667, March 2012.

[48] Y. Shuai, R. Liu, and W. He, "Image haze removal of wiener filtering based on dark channel prior," in *2012 Eighth International Conference on Computational Intelligence and Security*, pp. 318–322, Nov 2012.

[49] C. Ancuti, C. O. Ancuti, C. D. Vleeschouwer, and A. C. Bovik, "Night-time dehazing by fusion," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2256–2260, Sept 2016.

[50] P. Agrawal, R. Girshick, and J. Malik, "Analyzing the performance of multilayer neural networks for object recognition," in *Computer Vision – ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), (Cham), pp. 329–344, Springer International Publishing, 2014.

[51] J. DENG, "A large-scale hierarchical image database," *Proc. of IEEE Computer Vision and Pattern Recognition, 2009*, 2009.

[52] N. Joshi and M. F. Cohen, "Seeing mt. rainier: Lucky imaging for multi-image denoising, sharpening, and haze removal," in *2010 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–8, March 2010.

[53] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, Nov 1998.

[54] D. Differta and R. Möllera, "Insect models of illumination-invariant skyline extraction from uv and green channels," 2015.

[55] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 567–576, June 2015.

[56] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, June 2010.

[57] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, June 2017.

[58] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. S. Huang, "Unitbox: An advanced object detection network," *CoRR*, vol. abs/1608.01471, 2016.

[59] S. McCloskey, "Masking light fields to remove partial occlusion," in *2014 22nd International Conference on Pattern Recognition*, pp. 2053–2058, Aug 2014.

[60] P. Svoboda, M. Hradis, D. Barina, and P. Zemcík, "Compression artifacts removal using convolutional neural networks," *CoRR*, vol. abs/1605.00366, 2016.

MSc. Thesis VIBOT: Convolutional Neural Networks (CNN) for segmenting and quantifying endomicroscopic images of the lung

With the supervision of Dr. Antonios Perperidis

1st Savinien Bonheur
Université de Bourgogne & Universitat de Girona & Heriot-Watt University
Edinburgh, Scotland
savinien.bonheur@gmail.com

***Abstract*—In the following master thesis will be studied the feasibility of using deep learning approaches to evaluates cellular load within lungs endomicroscopics images through semantic segmentation. Indeed, while Endomicroscopic (OEM) imaging allows real-time imaging of the lung at its distal end, it generates large and hard to interpret dataset leading to laborious and subjective interpretations. To assess the OEM analysis automation feasibility, a labelled dataset is constructed and two state-of-the-art Convolutional Neural Network (CNN) architectures are trained (relying on data augmentation) several times using different training parameters, data pre-processing and architecture modifications. Therefore in this thesis we will firstly, study, using the U-Net architecture, the influence of input downscaling on the network performances. Secondly, investigate the loss function importance by comparing the well known weighed softmax loss with the generalized Dice loss (to our knowledge, a first in multi-class segmentation). Thirdly, explore our second network (ENet) architecture and design choices. Fourthly, evaluate the repeatability of our dataset annotations and use it to show the close-to-human performances of the different algorithms developed as well as the feasibility and success of our objective.**

I. Introduction

Lungs diseases have been a growing concern in the medical field. Indeed, in direct contact with exterior pathogens, lungs present an acute sensibility to bacterial infection. To characterise such infection, the current methods rely on the succession of slow detection procedures, biopsies and lab culture growth. Presenting inconfort and risk for the patient (hence limiting measurement repetition and therefore a simple monitoring of the condition evolution), these methods must be improved to allow, not only a more systematics uses of lungs analysis, but also the repeatability of lungs diagnostic in time to permit the following of lungs pathologies evolution. In this optics, Proteus aim to develop a quick *in vivo*, *in situ* approach to lung analysis through the uses of optical Endomicroscopics imaging, a novel technique allowing to image the lungs down to the alveolar ducts and sacs (see Figure 1). Endomicroscopy is a recent method combining the advantages of microscopy and endoscopy. In fact, this method allow the acquisition of histopathologic like (i.e high resolution) images through a real time, *in vivo*,*in situ* and minimally invasive method. To achieve this feast, a fibre bundle endomicroscope is inserted, through the endoscope working channel, within the patient lungs. Compounded of tens of thousands fibres jointed with a proximal illumination units (Laser or LED), endomicroscope produces an honeycomb like images (due to the fibres bundle) of the naturally fluorescent (collagen, elastin and cells are fluorescent at some wavelength such as 500nm) lungs structures as well as theirs surrounding. By its thinness, the endomicroscope allows the exploration of the lungs alveolar ducts and sacs, in real time and in high resolution. Due to those advantages, endomicroscopy should see itself being largely adopted and developed (such as in [16], [3]) in the upcoming years to create new ways of diagnosing.

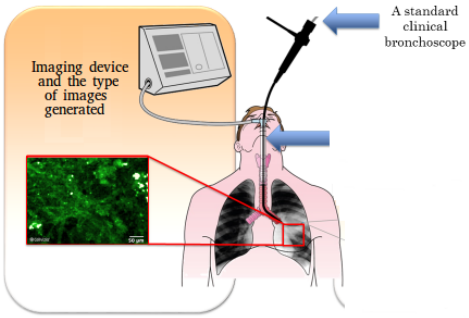
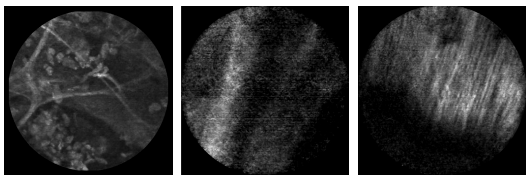


Fig. 1: Lungs endomicroscopics system. Adapted from [1].

For example, the aim of this paper, the automatic evaluation of lung cellularity within endomicroscopics images, could be an useful tool to assist the clinician diagnostic. In fact, in [31], lungs of smoker and non smokers are compared through endomicrocopics images and show a significant correlation (> 0.7) between the smoking frequency and the number of macrophages (a type of white blood cell trapping microbes and foreign particles) and, in [24], link between cellular load and pathologies have been noticeably studied and proven for acute cellular rejection.
Few studies were carried on the problems of cellular load

estimation. In [23] the automatic assessment of pulmonary nodule malignancy was investigated through local binary pattern and sift descriptor. Meanwhile in methods such as [22], [21] pattern matching have been explored to asses the lungs bacterial load (and cellular load in [21]). In this study, we decided to approach the evaluation of cellularity as a segmentation problem. Indeed, the segmentation of cells would allow clinicians to access patients lungs cellular load through the percentages of cell pixels in each frame. Moreover, the segmentation of cells could allow the development of further approaches by being embedded as the first step of a more complex tasks (i.e cell numbering, cell size evaluation, etc..). If no semantic segmentation algorithm have been applied to OEM images (to our knowledge), several publications tackled this task in other medical modalities. In [20], an heavily symmetric architecture, using skip connections and trained with a weighed loss function (to compensate class imbalance), segment images of neuronal structures in electron microscopic stacks. Inspired by this work, [14] developed a symmetric architecture, and, training it with a Dice loss function (for a Dice score improvement of 0.13 over the same architecture trained with a weighed softmax loss), to segment 3D MRI images. Meanwhile, ensemble of networks [10], [30] are employed to join networks predictions in intraoperative CLE (Confocal laser endomicroscopy) images and coloscopie videos respectively. Due to the novelty of endomicroscopic imaging and the originality of our objective, no dataset of cells segmented endomicroscopics images are available. A manually annotated (by a clinical investigator with substantial prior experience in pulmonary OEM images) dataset of segmented endomicroscopic images is thus gathered before seeing its annotation repeatability assessed for future reference.

II. Methodologies



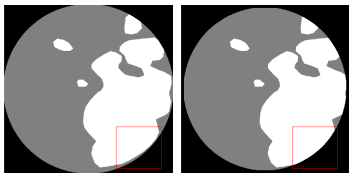
(a) Frame suitable for analysis (b) Out of focus frame (c) Motion blurred frame

Fig. 2: Different kind of recording noise.

Caused by the limited depth of field of endomicroscopic imaging (creating blurry to noise only images), and the motion artifacts introduced by the moving imaged tissues and the probe, a significant part (up to 25%) of the captured images are not suitable for analysis (see Figure 2). The problem was successfully investigated and addressed (with an overall sensitivity up to 93% and an overall specificity up to 98.6%) in several publications [17], [15], [2]. Considering the uninformative frames problem solved, we built a dataset of 378 informative (i.e without artifacts, blurriness, etc..) OEM frames, captured on 8 different patients, and splitted such as the validation test and the testing test, while being representative of the lungs variety, do not share any patients with the training set (see Table I, II, III).

A. Data labelling

Normalized by frame to compensate irregular dynamic ranges (consequence of the images being reconstructed by joining several optic fibres with different transmission coefficient), each frame is labelled manually (the cell pixels are selected),using Matlab 2017b, by an annotator with substantial prior experience in pulmonary OEM images . To mitigate annotation errors, the measured images border are eroded with a disk shaped element of size 10 (hence compensating for the difficulties of following borders), and the cell designated pixel are morphologically opened with a disk element of size 5 (removing small mislabelled cells pixels). For our purpose (i.e quantifying cells in endomicroscopic images), the subsequent binary annotated frames are transformed into three distinct semantic regions (see Figure 3):
-Cells
-Measured background (elastin and air)
-Padding



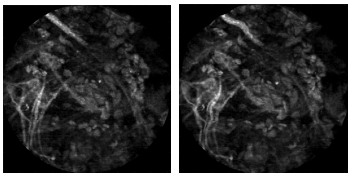
(a) Before erosion (b) After erosion

Fig. 3: An endomicroscopic images and its eroded counterpart. Cells pixel: white , Measured background: grey, Padding: black.

While considering the padding as a distinct region doesn't have any physical meaning, this trick allow us to easily use the endomicroscopic round shaped images with any popular CNN architecture and without the loss of information that image cropping would have induced. Moreover, the padding constance (all the value are zeros) as well as the experiment described in Appendix A lead us to believe in the innocuity of this practice.

B. Data augmentation

It is usual to artificially enlarge a dataset by applying label-preserving transformations. While it cannot replace a wide dataset, data augmentation reduce over-fitting by promoting networks generalisation capabilities. Based on the assumption that the measured endomicroscopic images will always be centered in the network input frame, we applied the following data augmentation techniques.



(a) Original frame (b) Elasticly deformed frame

Fig. 4: Example of elastic deformation.

1) Elastic deformations: Operating, the lungs contract and dilate elasticly. Those elastic deformations stretch the observed image in several directions simultaneously. To simulate this effect and augment realisticly our dataset, we applied elastic deformation to our dataset in a similar fashion as in [26], [12], [20].

Training: Each patient dataset are expanded to reach 100 frames by randomly selecting images and applying elastic deformation to them.

Testing and Validating: Each frames of each patient undergo elastic deformation (thus doubling the number of frames per patient).

2) Rotations and mirroring: As Neural Networks aren't intrinsically rotation invariant [7], [6] we applied the following transformations on our previously elasticly augmented dataset:

Training and Testing: Every frames is mirrored along the x axis and added to the augmented dataset (doubling the size of the augmented dataset). Then the augmented dataset is rotated by increment of 45 degrees (multiplying the size of the dataset by a factor of 8)

Validation: Due to time restriction, the validation set was solely augmented by a mirroring along the x axis followed by a single rotation of 90 degrees of the mirrored frames. Those two consecutive transformations, applied at once, are meant to create a single output image as different as possible from the input image.

The following Tables (I, II, III) describe the numbers of frames used from each patient, for each dataset, before and after data augmentation.

Training dataset			
Patient ID	Number of frames	Number of frames after elastic deformation	Number of frames after rotation and mirroring
90	60	100	1600
106	100	100	1600
172	80	100	1600
191	50	100	1600
Total	290	400	64000

TABLE I: Training dataset images origins and numbers before and after data augmentation.

Validation dataset			
Patient ID	Number of frames	Number of frames after elastic deformation	Number of frames after rotation and mirroring
63	11	22	44
96	10	20	40
118	10	20	40
124	10	20	40
Total	41	82	164

TABLE II: Validation dataset images origins and numbers before and after data augmentation.

Testing dataset			
Patient ID	Number of frames	Number of frames after elastic deformation	Number of frames after rotation and mirroring
63	20	40	640
96	5	10	160
118	10	20	320
124	6	12	192
Total	41	82	1312

TABLE III: Testing dataset images origins and numbers before and after data augmentation.

C. Training

We decided to use the U-Net network [20], for its performance and its design adapted to small training dataset, and the ENet network [18] (applied on medical images in [32]) for its reasonable dataset size exigency (in the 10^3 order), its speed [34], and its efficiency [5].

Although many ways of experimenting with CNN exist (i.e changing the number of layers, the number of kernels per layer, the kernels size, the loss functions, etc ...). We, as one of the first papers applying deep learning on endomicroscopics images, decided to explore less networks architecture modification (i.e changing the number of layers, the kernels size, etc ..) and more the influence of what we believe to be more transferable and application driven knowledges. Indeed, while we do experiment with transfer learning (where the result depends greatly on the dataset on which the network was firstly trained on), we mostly investigated the trade off between speed (ENet) and accuracy (U-Net), the importance of the loss function (and therefore the task the network actually solve), the influence of different feature sizes, and the networks behaviours. Nevertheless, we experimented with simple modification of ENet in order to simplify its training so that later research might employ it more easily.

All the experiments are carried using the Caffe framework [11] and the adadelta [33] (for the ease of use brought by its single parameter) update rules. As it is too time-consuming to evaluate the loss function over the full dataset, two commons subsampling solutions were considered, grouping images in batches and increasing the network output size (as the error is evaluated at each pixel we can consider each pixel as a sample). To maximize the use of our GPU (a GEFORCE 740M) we, as in [20], privileged the use of wider outputs rather than batches. We settled, in all our experiments, on an U-Net output size of 100x100 pixels (our GPU biggest supported output) and an 96x96 pixels output for ENet (the closest to 100x100 we could considering the ENet architecture). Each image is splitted in non overlapping output titles before being used in training.

1) Training parameters experiments:

a) Input downsampling: Neighbourhood pixels are highly correlated within themselves. This semantic redundancy leads us to believe than downscalling (within boundaries) the network inputs will have an effect close to increasing the output size during training (hence leading to a better dataset approximating). Moreover, due to its pixels correlation, the downscalled images will present a better signal to noise ratio.

We will test the influence of downscalling by comparing the U-Net architecture trained with no downscalling, and downscalling factors of 2, 4 and 6. The images and theirs corresponding groundtruths are downscalled and concatenated (to avoid patches consisting mostly of padding) before being fed into the network in 286x286 pixels patches (with U-Net, an input of 286x286 pixels leads to an output of 100x100 pixels).

b) Loss function: In a neural network the loss function, or cost function, defines the metrics which will be improved upon (by adjusting the network weight). As our dataset classes are imbalanced (see Figure 5), we chose loss functions for which the class imbalance can be incorporated as a parameter. Indeed, in non-class weighted metrics, class imbalance usually detriments the less represented classes as they contribute very little to the overall metrics. The weighted cross entropy loss and the general Dice loss are therefore investigated as they both propose solutions to the well known class imbalance problem and are the extensions of well known formula.

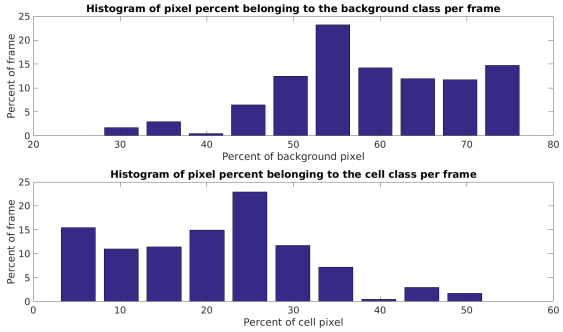


Fig. 5: Quantity of pixels belonging to the background class and the cell class.

c) Weighted Cross entropy loss: Noticeably used in [20], the weighted cross entropy loss is an extension of the widely used cross entropy loss [13], [4], [29] and can be expressed as :

$$E = -\sum_{label,l} w_l \sum_{voxels,i} g_{li} log(\frac{exp(x_{li})}{\sum_{voxels,j} exp(x_{lj})}) \quad (1)$$

where :
w_l is the class l weight such as $\sum_l w_l = 1$
x_{li} is the network score at pixel i for the label l
and g_{li} is the groundtruth at pixel i for the label l

This formula evaluate the weighted (to counteract class imbalance bias) difference between the true probability distribution (the ground-truth distribution) and the network's inferred probability distribution.

d) Generalized Dice loss: Firstly introduced in [14], the Dice loss function is shown leading to better results than the weighed softmax loss. However, the introduced Dice loss is only suitable for binary classification. It is defined as :

$$DL = 1 - 2\frac{\sum_{voxels,i} p_i g_i}{\sum_{voxels,i} p_i^2 + \sum_{voxels,i} g_i^2} \quad (2)$$

Where :
-p_i is the network inferred foreground probability at pixel i
-g_i is the groundtruth foreground probability at pixel i

Approaches has been made to extend this metrics to multi-class segmentation. For example, in [25] they multiply the Dice score of each class by a weight (several weight definitions are investigated) before averaging the weighted Dice to compute an overall score. Nevertheless the lack of mathematical justification in the weighting scheme incorporation make us believe that a more robust solution exists. To address this issue we, in a similar fashion as [28], adapted the formulation developed in [8] to allow the use of the Dice loss in multi-class segmentation problems. We define the Generalized Dice loss as :

$$GDL_v = 1 - 2\frac{\sum_{labels,l} \alpha_l \sum_{voxels,i} p_{li} g_{li}}{\sum_{labels,l} \alpha_l (\sum_{voxels,i} p_{li}^2 + \sum_{voxels,i} g_{li}^2)} \quad (3)$$

From the different proposition presented in [8], we chose the class weighting factor α_{lg} as the inverse of each class surface so that each class impact equally the overall metric. Therefore : $\alpha_l = \frac{1}{\sum_{voxels,i} g_{li}}$
.

As demonstrated in Appendix B, the derivative can be defined as :

$$\frac{\partial GDL_v}{\partial p_{lk}} = 2\alpha_l \frac{2p_{lk}\sum_l \alpha_l \sum_i p_{li} g_{li} - g_{lk} \sum_l \alpha_l (\sum_i p_{li}^2 + \sum_i g_{li}^2)}{(\sum_l \alpha_l (\sum_i p_{li}^2 + \sum_i g_{li}^2))^2} \quad (4)$$

When implemented, we pass the network output through a softmax layer to ensure segmentation scores belonging to the interval [0,1]. Moreover, we normalize α_{lg} such as $\sum_l \alpha_{lg} = 1$. We also added a small constant ϵ to the GDL_v and its derivative to solve the division by zero issue.

2) Architecture experiments:

a) ENet decoder: ENet and its decoder relatives shortness (compared to its encoder), as well as the network two steps training [18] (first the encoder is trained, then the whole network), made us wonder about the decoder actual role. To explore its behaviour we will, in one experiment (as in [4] [19]), replace its decoder by a billinear upsampling (see Figure 6) and, in another experience, train ENet at once rather than in two steps.

b) Networks comparison: To compare the two network architectures, we will, for both of them, with nearly equal output size (100x100 against 96x96), no input downscalling, and a weighted softmax loss layer, apply:

Name	Type	Output size
initial		$16 \times 256 \times 256$
bottleneck1.0	downsampling	$64 \times 128 \times 128$
$4\times$ bottleneck1.x		$64 \times 128 \times 128$
bottleneck2.0	downsampling	$128 \times 64 \times 64$
bottleneck2.1		$128 \times 64 \times 64$
bottleneck2.2	dilated 2	$128 \times 64 \times 64$
bottleneck2.3	asymmetric 5	$128 \times 64 \times 64$
bottleneck2.4	dilated 4	$128 \times 64 \times 64$
bottleneck2.5		$128 \times 64 \times 64$
bottleneck2.6	dilated 8	$128 \times 64 \times 64$
bottleneck2.7	asymmetric 5	$128 \times 64 \times 64$
bottleneck2.8	dilated 16	$128 \times 64 \times 64$
Repeat section 2, without bottleneck2.0		
Bilinear_up	upsampling	$16 \times 256 \times 256$
fullconv		$C \times 512 \times 512$

Fig. 6: The ENet architecture with a bilinear upsampling replacing its decoder (for an 512x512 input). Adapted from [18]

-a training with randomly initialized weights
-a shallow fine-tuning (where the weight are frozen as shown in Figure 7, 8)
-a deep fine-tuning (initialized with trained shallow fine-tuned networks)

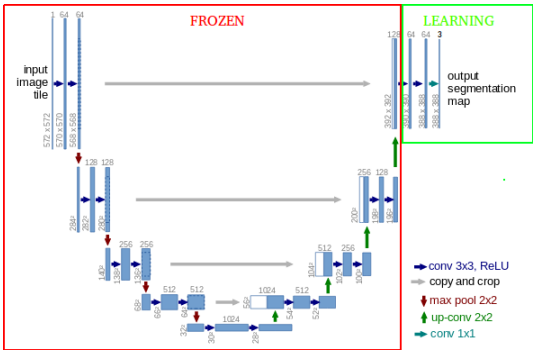


Fig. 7: U-Net shallow-Fine tuning architecture

III. Results and Discussion

Along theirs training, neural networks achieve different results as their weights evolve. To select the best version of each training, we test the networks at different number of iterations on our validation test. Indeed, due to possible over-fitting, training a network for too long can reduce its performances. To evaluate the different training states, we measured several metrics (see below) against the number of training iterations before selecting each network best training iteration.

The following metrics will be used :
-pixel accuracy= $\frac{\sum_i n_{ii}}{\sum_i t_i}$
-mean accuracy= $\frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i}$
-mean IoU (intersection over union) = $\frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$

Name	Type	Output size
initial	FROZEN	$16 \times 256 \times 256$
bottleneck1.0	downsampling	$64 \times 128 \times 128$
$4\times$ bottleneck1.x		$64 \times 128 \times 128$
bottleneck2.0	downsampling	$128 \times 64 \times 64$
bottleneck2.1		$128 \times 64 \times 64$
bottleneck2.2	dilated 2	$128 \times 64 \times 64$
bottleneck2.3	asymmetric 5	$128 \times 64 \times 64$
bottleneck2.4	dilated 4	$128 \times 64 \times 64$
bottleneck2.5		$128 \times 64 \times 64$
bottleneck2.6	dilated 8	$128 \times 64 \times 64$
bottleneck2.7	asymmetric 5	$128 \times 64 \times 64$
bottleneck2.8	dilated 16	$128 \times 64 \times 64$
Repeat section 2, without bottleneck2.0		
bottleneck4.0	upsampling	$64 \times 128 \times 128$
bottleneck4.1		$64 \times 128 \times 128$
bottleneck4.2		$64 \times 128 \times 128$
bottleneck5.0	upsampling	$16 \times 256 \times 256$
bottleneck5.1		$16 \times 256 \times 256$
fullconv	LEARNING	$C \times 512 \times 512$

Fig. 8: ENet shallow fine-tuning architecture

-mean Dice coefficient= $\frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji}}$
-frequency weighted IoU = $\left(\sum_k t_k\right)^{-1} \sum_i \frac{n_{i_i}}{t_i + \sum_j n_{ji} - n_{ii}}$
-Pearson's correlation coefficient for the percentage of cell pixel per frame: $\text{Corr}_c = \frac{n_{cl} \sum xy - \sum x \sum y}{\sqrt{(n_{cl} \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)}}$
where :
-n_{ij} is the number of pixels of the class i predicted to belong to the class j.
-n_{cl} the total number of classes
-$t_i = \sum_j n_{ij}$ the total number of pixels belonging to the class i
-$y = \frac{t_i}{\sum_i t_i}$ the percentage of cell pixels per frame
-$x = \frac{n_{ii}}{\sum_i t_i}$ the percentage of cell pixels inferred per frame

Although quite similar, (Dice coefficient= $\frac{2IoU}{(1+IoU)}$), we decided to measure both the intersection over union metric (for its intuitiveness) and the Dice coefficient (for its commonness in medical imaging). We also decided to use the Pearson's correlation coefficient to evaluate the quality of our networks cell percent evaluation (as this will be the information used by the clinician).

A. Repeatability

Annotator	Pixel accuracy	Mean accuracy	Mean IoU	Frequency weighted IoU	Mean dice	Corr_c
Without preprocessing	0.9177	0.9179	0.8271	0.8617	0.8958	0.7824
With preprocessing	**0.9189**	**0.9180**	**0.8274**	**0.8638**	**0.8960**	**0.7826**

TABLE IV: Analysis of the annotation repeatability.

Table IV shows, through a slight improvement of the annotation repeatability , the validity of the pre-processing applied to the dataset (modifying the frames cell percent by a mean value of, out of 48 images, 0.01 percent).

IV. UNET

A. Parameters experiments

U-Net input downscaling factor	Pixel accuracy	Mean accuracy	Mean IoU	Frequency weighted IoU	Mean dice	$Corr_c$
x1	0.8739	0.8676	0.7778	0.7873	0.8655	0.6843
x2	**0.8924**	**0.8827**	**0.8126**	**0.8089**	**0.8915**	0.8524
x4	0.8899	0.8812	0.8020	**0.8089**	0.8840	**0.8648**
x6	0.8600	0.8498	0.7515	0.7681	0.8481	0.8257

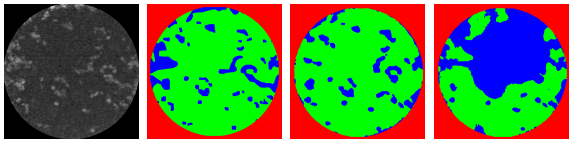
TABLE V: Comparison of different input downscaling factors on U-Net performances.

1) Input downscaling influence: The metrics in Table V shows better result for a input downscaling of factor 2 and factor 4. We believe those improvements to be the consequences of both the noise reduction and the access to a broader fields of view while training. In fact, with downscaled images, the network trains itself on less semanticly redundant images (roughly equivalent to an increase in patch of batch size and thus in performance). Due to a lack of time, no experiment were carried to compensate this increase of semantic information and conclude about the best trade off between noise and features size. Despite those considerations, a downscaling of factor 4 seems to be the most adapted to our limited computation power (the network converge ≈ 2.67x faster) and to future clinical application (the heavier the downscaling, the faster can the offline inference be).

U-Net loss function	Pixel accuracy	Mean accuracy	Mean IoU	Frequency weighted IoU	Mean dice	$Corr_c$
Weighted Softmax	**0.8899**	**0.8812**	**0.8020**	**0.8089**	**0.8840**	**0.8648**
Generalized Dice	0.8777	0.8674	0.7864	0.7893	0.8731	0.6499

TABLE VI: Comparison between the weighed softmax loss and the generalized Dice loss to train an U-Net architecture with a downscaled by 4 input.

2) Loss function importance: The above result (Table VI), obtained with a downscaling factor of 4 on the U-Net architecture, shows the superiority of the weighed softmax loss over the generalized Dice loss. This difference of metrics is due to the generalized Dice loss trained U-Net performing poorly on low cell percents images (see Figure 9 and 10). While this behaviour reasons are unknown and should be investigated (through retraining and the use of different weighting schemes), it leads us to consider the use of the weighed softmax loss as more appropriate to cells segmentation.



(a) Input Image (b) Groundtruth (c) Weighed softmax loss (d) Generalized Dice loss

Fig. 9: Network segmentation comparison.

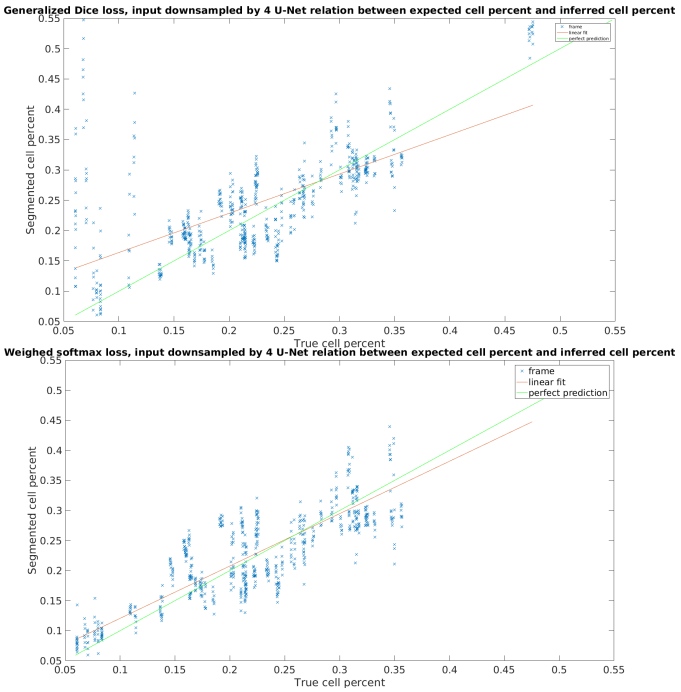


Fig. 10: Loss function comparison through the prediction of cell percent.

B. Architecture experiments

ENet	Pixel accuracy	Mean accuracy	Mean IoU	Frequency weighted IoU	Mean dice	$Corr_c$
Bilinear decoder	0.8102	0.8191	0.7081	0.6850	0.8175	0.6580
Trained at once	0.8617	0.8534	0.7679	0.7626	0.8608	0.7780
Regular	**0.8631**	**0.8548**	**0.7697**	**0.7638**	**0.8626**	**0.8004**

TABLE VII: ENet decoder influence.

1) Decoder influence: In all metrics, the two steps trained original ENet architecture, by outperforming its variants, shows the non negligible contribution of the decoder. Nevertheless, the experiment showed than an ENet trained at once demonstrates near equivalence to its two steps trained counterpart while converging significantly faster (200 000 iterations against 350 000 iterations). This gain in time and practicality (only one network to train instead of two) as well as its performances make us believe that prototyping would be worth pursuing using a trained at once ENet.

Network	Pixel accuracy	Mean accuracy	Mean IoU	Frequency weighted IoU	Mean dice	$Corr_c$
U-Net from scratch	**0.8739**	**0.8676**	0.7778	**0.7873**	0.8655	0.6829
U-Net shallow fine-tuned	0.8353	0.8326	0.7292	0.7183	0.8377	0.6683
U-Net deep fine-tuned	0.8729	0.8648	**0.7848**	0.7802	**0.8716**	0.7479
ENet from scratch	0.8631	0.8548	0.7697	0.7638	0.8626	**0.8004**
ENet shallow fine-tuned	0.8353	0.8326	0.7292	0.7183	0.8377	0.6683
ENet deep fine-tuned	0.8639	0.8535	0.7628	0.7694	0.8573	0.7991

TABLE VIII: Comparison between U-Net and ENet architectures trained with a weighed softmax loss and no downscaling.

2) Networks comparison: Surprisingly, U-Net while being slower (on 50 images of output size 516x516 pixels, U-Net mean inference time is $\sim$ 11.19s and ENet mean inference time is $\sim$ 2.85s using an Intel i7-3630QM and multi-threading) lead to better overall metrics and less noisy segmentation (see

Figure 11, 12 and Tables VIII, X, IX). However, we can remark ENet clear superiority at evaluating the cellular load through more constant estimations (see Figure 11). Globally, we can observe the clear cell correlation improvement of deep fine-tuning (we believe this improvement to be the cause of more general filters being developed in transfer learning) over the others form of training (deep fine-tuned ENet must be compare to its trained at once counterpart). The great amelioration of U-Net correlation score when deep fine-tuned leads us to believe in U-Net overfitting its training dataset. This observation is supported by the superiority of ENet, a network using a dropout strategy to reduce overfitting [27], [9], correlation score.

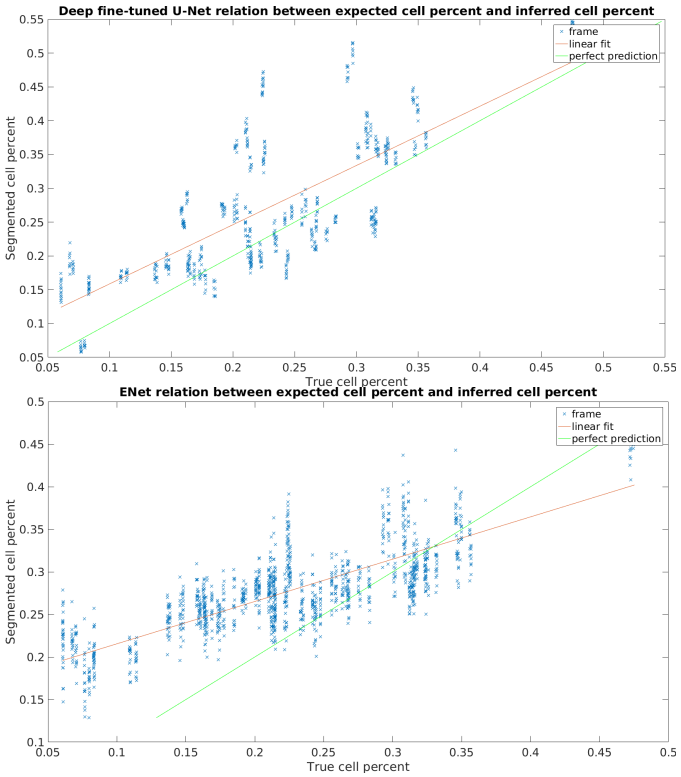


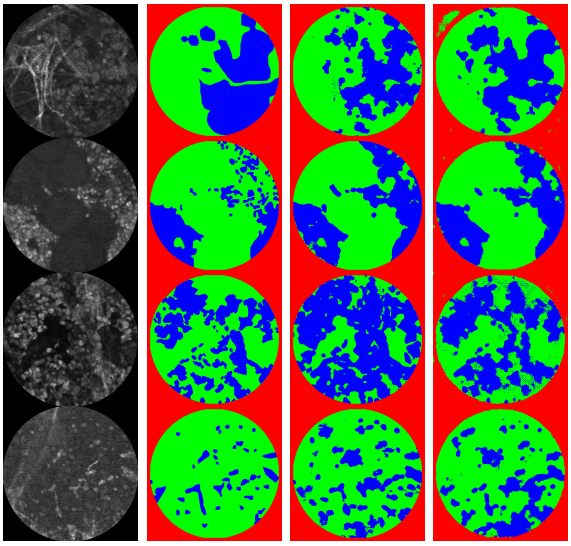
Fig. 11: Network function comparison through the prediction of cell percent.

	Padding	Background	Cells
Padding	18.2572	13.9070	12.2881
Background	13.7371	18.8026	17.1282
Cells	12.4048	16.3769	17.9623

TABLE IX: Deep fine-tuned U-Net testing dataset log confusion matrix.

	Padding	Background	Cells
Padding	18.2488	14.4400	11.9503
Background	14.5034	18.7791	17.2097
Cells	13.1917	16.2906	17.9745

TABLE X: ENet testing dataset log confusion matrix.



(a) Measured frame (b) Groundtruth (c) Deep fine-tuned U-Net (d) ENet

Fig. 12: Network segmentation comparison. Cells pixel: blue , Measured background: green, Padding: red

V. Conclusion

On one hand, based on a recent imaging technique and reaching for new goals, our dataset, annotated by only one clinician, is small and rely heavily on data augmentation. On the other hand, CNN and others machine learnings approaches are highly dependant of the quality and quantity of data available to train them. This discrepancy between best practice and actual practice is detrimental to our results and should therefore be addressed in future works. Nevertheless, the close-to-human cell correlation scores (better to worse depending on the network and its training parameter) for most of the developed architectures make us believe in the application of semantic segmentation to evaluate cellular loads within endomicrocopics images through deep-learning. Indeed, if improvement could be made (the human annotator still marginally outperform our best developed algorithm worst metric by 6.79 relative percent) and experiments carried , we have demonstrated the feasibility of such approach. Moreover, the two applied networks, far from opposing themselves, could be applied on different objectives. In fact, ENet seems well adapted to be used in real-time segmentation and U-Net, trained to reduce overfitting, could be the bases of more complex offlines application.

Altogether, we consider our studies on the uses of CNN for cellular loads evaluation within endomicroscopic images of the lungs a success. Indeed, a strong linear correlation between the annotator cell percent evaluation and the best developed algorithm has been demonstrated (outperforming the annotator repeatability by 10.5 relative percent thanks to a score of 0.8648) despite the annotation subjectivity arising from a limited image depth field, noisy images and tissues and cells superposition.

References

[1] Molecular imaging (https://proteus.ac.uk/clinical/molecular-imaging/).

[2] Y. Altmann P. McCool J. Westerfeld et al. A. Perperidis, A. Akram. Automated detection of uninformative frames in pulmonary optical endomicroscopy (oem).

[3] E. Scholefield M. Bradley K. Dhaliwal B. Mills, A. R. Akram. Optical screening of novel bacteria-specific probes on ex vivo human lung tissue by confocal laser endomicroscopy.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561, 2015.

[5] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *CoRR*, abs/1605.07678, 2016.

[6] G. Cheng, P. Zhou, and J. Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, Dec 2016.

[7] Taco S. Cohen and Max Welling. Group equivariant convolutional networks. *CoRR*, abs/1602.07576, 2016.

[8] W. R. Crum, O. Camara, and D. L. G. Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, Nov 2006.

[9] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[10] Mohammadhassan Izadyyazdanabadi, Evgenii Belykh, Michael Mooney, Nikolay Martirosyan, Jennifer Eschbacher, Peter Nakaji, Mark C. Preul, and Yezhou Yang. Convolutional neural networks: Ensemble modeling, fine-tuning and unsupervised semantic localization. *CoRR*, abs/1709.03028, 2017.

[11] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, page 2012.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[14] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *CoRR*, abs/1606.04797, 2016.

[15] Paul McCool Jody Westerfeld David Wilson Kevin Dhaliwal Stephen McLaughlin Antonios Perperidis Mohammad Rami Koujan, Ahsan Akram. Multi-class classification of pulmonary endomicroscopic images.

[16] T. R. Choudhary N. McDonald M. G. Tanner et al. N. Krstaji, A. R. Akram. Two-colour widefield fluorescence microendoscopy enables multiplexed molecular imaging in the alveolar space of human lung tissue.

[17] Ahsan Akram Jody Westerfeld David Wilson Kevin Dhaliwal Stephen McLaughlin Antonios Perperidis Oleksii Leonovych, Mohammad Rami Koujan. Texture descriptors for classifying sparse, irregularly sampled optical endomicroscopy images.

[18] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *CoRR*, abs/1606.02147, 2016.

[19] Atif Riaz, Muhammad Asad, S. M. Masudur Rahman Al-Arif, Eduardo Alonso, Danai Dima, Philip Corr, and Greg Slabaugh. Fcnet: A convolutional neural network for calculating functional connectivity from functional mri. In Guorong Wu, Paul Laurienti, Leonardo Bonilha, and Brent C. Munsell, editors, *Connectomics in NeuroImaging*, pages 70–78, Cham, 2017. Springer International Publishing.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[21] K. Dhaliwal S. Seth, A. R. Akram and C. K. I. Williams. Estimating bacterial and cellular load in fcfm imaging.

[22] K. Dhaliwal S. Seth, A. R. Akram and C.K.I. Williams. Estimating bacterial load in fcfm imaging.

[23] P. McCool J. Westerfeld D. Wilson et al. S. Seth, A. Akram. Assessing the utility of autofluorescence-based pulmonary optical endomicroscopy to predict the malignant potential of solitary pulmonary nodules in humans.

[24] Sohan Seth, Ahsan R. Akram, Kevin Dhaliwal, and Christopher K. I. Williams. Estimating bacterial and cellular load in fcfm imaging. *Journal of Imaging*, 4(1), 2018.

[25] Chen Shen, Holger R. Roth, Hirohisa Oda, Masahiro Oda, Yuichiro Hayashi, Kazunari Misawa, and Kensaku Mori. On the influence of dice loss function in multi-class organ segmentation of abdominal CT using 3d fully convolutional networks. *CoRR*, abs/1801.05912, 2018.

[26] P. Y. Simard, D. Steinkraus, and J. C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 958–963, Aug 2003.

[27] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. 15:1929–1958, 06 2014.

[28] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *CoRR*, abs/1707.03237, 2017.

[29] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[30] Nima Tajbakhsh, Suryakanth Gurudu, and Jianming Liang. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. 2015:79–83, 07 2015.

[31] Luc Thiberville, Mathieu Salan, Samy Lachkar, Stphane Dominique, Sophie Moreno-Swirc, Christine vever bizet, and Genevive Bourg-Heckly. Human in vivo fluorescence microimaging of the alveolar ducts and sacs during bronchoscopy. 33:974–85, 02 2009.

[32] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. *CoRR*, abs/1709.00382, 2017.

[33] M. D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. *ArXiv e-prints*, December 2012.

[34] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. *CoRR*, abs/1704.08545, 2017.

Appendix A
Experiment: Influence of segmenting two classes against three classes

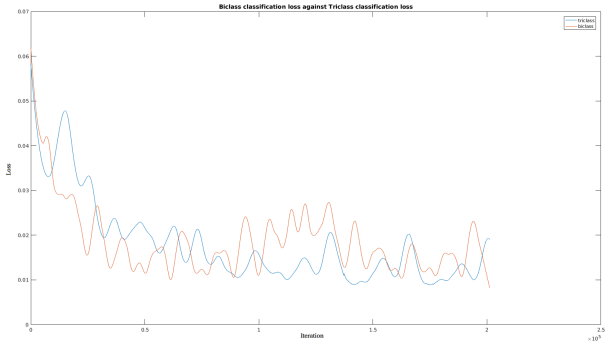


Fig. 13: Comparison between U-Net two classes training and U-Net three classes training

Our choice of adding a third classes to the segmentation task, while convenient, raises the question of its on the

overall segmentation performances. To investigate this new class impact on performances, we specially trained the U-Net architectures by automatically setting the padding pixels (the pixels corresponding to the class we added for convenience) to theirs correct labels. By being perfectly labelled, these classes does not influence the weights patterns to which the network converge to segment the two others class. As our special trained U-Net and the normally trained U-Net both are trained with the same initialization and the same hyperparameters, we can directly compare both loss function with theirs training loss. As we can see in Figure 13, the loss of the network against the number of iteration shows the innocuousness of the third added class to the overall classification.

APPENDIX B
GENERALIZED DICE LOSS DERIVATION

Let's recall the Generalized Dice Loss :

$$GDL_v = 1 - 2\frac{\sum_{labels,l} \alpha_l \sum_{voxels,i} p_{li} g_{li}}{\sum_{labels,l} \alpha_l (\sum_{voxels,i} p_{li}^2 + \sum_{voxels,i} g_{li}^2)}$$

Where is the inverse of each class volume such as $\alpha_l = \frac{1}{\sum_{voxels,i} g_{li}}$
g_{li} is the groundtruth for the label l at the voxel i
p_{li} is the network output for the label l at the voxel i

The partial derivative is then :

$$\frac{\partial GDL_v}{\partial p_{lk}} = \frac{\partial 1}{\partial p_{lk}} - \frac{\partial 2\frac{\sum_l \alpha_l \sum_i p_{li} g_{li}}{\sum_l \alpha_l (\sum_i p_{li}^2 + \sum_i g_{li}^2)}}{\partial p_{lk}}$$
$$= -\frac{\partial 2\frac{\sum_l \alpha_l \sum_i p_{li} g_{li}}{\sum_l \alpha_l (\sum_i p_{li}^2 + \sum_i g_{li}^2)}}{\partial p_{lk}}$$

Let's define f and g such as :

$$\frac{\partial GDL_v}{\partial p_{lk}} = -2\frac{\partial \frac{f}{g}}{\partial p_{lk}} = 2\frac{f\frac{\partial g}{\partial p_{lk}} - g\frac{\partial f}{\partial p_{lk}}}{g^2}$$

Therefore :

$$f = \sum_l \alpha_l \sum_i p_{li} g_{li}$$

$$g = \sum_l \alpha_l (\sum_i p_{li}^2 + \sum_i g_{li}^2)$$

and :

$$\frac{\partial f}{\partial p_{lk}} = \frac{\partial \sum_l \alpha_l \sum_i p_{li} g_{li}}{\partial p_{lk}} = \alpha_l g_{lk}$$

$$\frac{\partial g}{\partial p_{lk}} = \frac{\sum_l \alpha_l (\sum_i p_{li}^2 + \sum_i g_{li}^2)}{\partial p_{lk}} = 2\alpha_l p_{lk}$$

Finally the partial derivative is:

$$\frac{\partial GDL_v}{\partial p_{lk}} = \frac{f\frac{\partial g}{\partial p_{lk}} - g\frac{\partial f}{\partial p_{lk}}}{g^2}$$

$$= 2\alpha_l \frac{2p_{lk}\sum_l \alpha_l \sum_i p_{li} g_{li} - g_{lk}\sum_l \alpha_l (\sum_i p_{li}^2 + \sum_i g_{li}^2)}{(\sum_l \alpha_l (\sum_i p_{li}^2 + \sum_i g_{li}^2))^2}$$

We can remark than, if $p_{li} = g_{li}$:

$$\sum_l \alpha_l \sum_i p_{li} g_{li} = \sum_l \alpha_l \sum_i g_{li}^2$$

$$\sum_l \alpha_l (\sum_i p_{li}^2 + \sum_i g_{li}^2) = 2\sum_l \alpha_l \sum_i g_{li}^2$$

such as :

$$GDL_v = 1 - 2\frac{\sum_l \alpha_l \sum_i p_{li} g_{li}}{\sum_l \alpha_l (\sum_i p_{li}^2 + \sum_i g_{li}^2)}$$
$$= 1 - \frac{2\sum_l \alpha_l \sum_i g_{li}^2}{2\sum_l \alpha_l \sum_i g_{li}^2} = 1 - 1 = 0$$

and :

$$\frac{\partial GDL_v}{\partial p_{lk}} = 2\alpha_l \frac{2p_{lk}\sum_l \alpha_l \sum_i p_{li} g_{li} - g_{lk}\sum_l \alpha_l (\sum_i p_{li}^2 + \sum_i g_{li}^2)}{(\sum_l \alpha_l (\sum_i p_{li}^2 + \sum_i g_{li}^2))^2}$$

$$= 2\alpha_l \frac{2g_{lk}\sum_l \alpha_l \sum_i g_{li}^2 - 2g_{lk}\sum_l \alpha_l \sum_i g_{li}^2}{(2\sum_l \alpha_l \sum_i g_{li}^2)^2}$$

$$= 0$$

Deep learning architectures for stroke lesion segmentation and outcome prediction

Albert Clèrigues Garcia, Sergi Valverde, Arnau Oliver and Xavier Lladó

***Abstract*— Stroke lesions have two differentiated areas: the core, formed by irreversibly damaged tissue, and the penumbra, damaged tissue at risk that could be eventually healed and salvaged. Segmentation and differentiation of core and penumbra can help doctors to asses if the amount of potentially salvageable tissue outweighs the risks of recanalization surgery.**

In this master thesis, five different state of the art deep learning architectures from related biomedical tasks are used as baseline models to evaluate their application in stroke lesion segmentation and outcome prediction. Ensembles combining the outputs of several independently trained models are shown to minimise the effect of suboptimal training hyper-parameters. We show that ensemble models, built under specific conditions, do improve segmentation quality to a degree. A recent stroke lesion outcome prediction challenge shows automatic methods are still far from human level performance. The addition of registered atlases as additional modalities is proposed to compensate the lack of data with related knowledge. The use of this strategy improves the qualitative and quantitative results only on some models, which suggests some architectures can make better use of the additional information. Four different datasets, including three from already held challenges, involving the tasks of stroke lesion segmentation and lesion outcome prediction are used to evaluate the implemented models. Finally, the best methods on challenge datasets are submitted to their online platform for evaluation. We achieve state of the art results which place us among the three best methods in the three challenges.

I. INTRODUCTION

Stroke is a medical condition by which an abnormal blood flow in the brain causes the death of cerebral tissue. Stroke is the third most common cause of morbidity worldwide, after myocardial infarction and cancer, and is the leading cause of acquired disability. Ischaemic strokes happen due to due to insufficient blood supply and comprise 80% of stroke episodes. Fig. 1 shows the appearance of stroke lesions in several Magnetic Resonance Imaging (MRI) modalities. Once the symptoms of stroke have been identified, a shorter time to treatment is strongly correlated with a positive outcome. The philosophical premise underlying the importance of rapid stroke intervention was summed up as *Time is Brain!* in the early 1990s [1].

The infarcted tissue after an episode is divided into three regions depending on the potential for recovery, also referred as salvageablity, of the tissue involved: core, penumbra and benign oligemia (see Fig. 2). The core is formed by irreversibly damaged tissue, characterised by a fatally low vascularisation. The penumbra represents tissue with enough blood supply that can be eventually salvaged depending on factors such as revascularization, collateral blood supply, tissue resistance, etc. The benign oligemia is the area whose vascularity has been altered by the stroke but is not at risk of permanent damage.

In the affected area of the brain, the stroke lesion undergoes a number of disease stages that can be subdivided according to the time passed since stroke onset. These are divided into *acute* in the first 24 hours, *sub-acute* from day one up to the second week, and *chronic* from the second week onward.

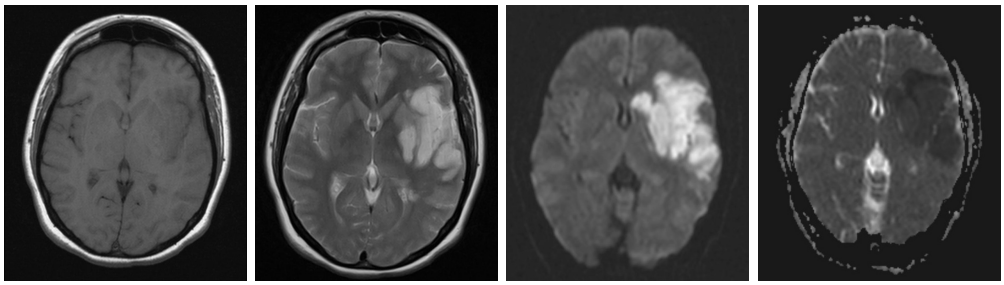
1) Swelling and shrinking: Lesion swelling is commonly observed soon after ischemic stroke, peaking at 3–5 days. Over time, the stroke lesion shrinks as the swelling reduces and tissue damaged by the injury is lost and replaced by cerebrospinal fluid (CSF) leaving an area of cerebromalacea with ex vacuo effect on adjacent structures. This effect refers to the deformation of surrounding tissue as it starts filling the area previously occupied by the infarcted area.

2) Spontaneous reperfusion: Spontaneous reperfusion is a physiological response attempting to restore blood supply in underperfused areas, occurring in about 20% of patients by 24 hours and 80% by 5 days. This response is mainly achieved by protraction of neighbouring vessels, if done promptly it can greatly alter tissue outcome during the upcoming 3 weeks.

3) No-reflow phenomenon: No re-flow phenomenon occurs when recanalization of the blocked artery fails to reperfuse the tissue capillaries. The absence of reflow is related to a bad outcome.

II. STATE OF THE ART

Methods for stroke lesion segmentation are often collateral, being originally designed for general brain lesion segmentation such as tumours, multiple sclerosis lesions, white matter hyperintensities... The complexity of the task, the clinical nuances and the lack of incentives are all factors that contribute to a sparse and diffuse state of the art for stroke lesion segmentation. However, the situation has evolved favourably for stroke imaging in the last few years with the proliferation of high quality labelled public datasets. The Ischemic Stroke Lesion Segmentation (ISLES) challenge [3] in 2015 included the sub-acute ischemic stroke lesion segmentation (SISS) and the acute stroke outcome/penumbra estimation (SPES) subtasks. The following two editions of the ISLES challenge in 2016 and 2017 focused on prediction of chronic lesion outcome from sub-acute images. The Anatomical Tracings of Lesions After Stroke (ATLAS R1.1) is another dataset [4] released in late 2017 includes a large number of samples of chronic stroke lesions in T1 images.



(a) T1 (b) T2 (c) DWI (d) ADC

Fig. 1: Example of stroke lesion, two days after the episode, caused by a Middle Cerebral Artery (MCA) occlusion in different MRI modalities. Case courtesy of Dr Sandeep Bhuta, Radiopaedia.org, rID: 6840. The lesion is seen as hyperintense on T2 and DWI images, where the visible extent of the lesion differs significantly, and hypointense in T1 and ADC modalities.

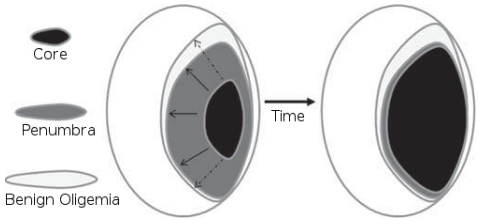


Fig. 2: Temporal evolution of an ischaemic stroke lesion [2].

In the ISLES 2015 challenge SISS sub-task, Kamnitsas et al. [5] used their fully convolutional architecture DeepMedic achieving the first position. The third place of SISS was awarded to Halme et al. [6] with an approach based on Random Decision Forests (RDFs). Maier et al. [7], the best method on the SPES sub-task, and McKinley et al. [8], the second best method, also made use of RDFs for penumbra segmentation. This kind of classifiers have excellent generalization properties, which has made them popular for difficult tasks with few training samples such as stroke lesion segmentation. However, random decision forests are essentially a cascade of simple classifiers acting on hand crafted features. Consequently, RDFs see their potential severely limited by the quality of the given features, which may vary for different tasks.

The ISLES 2016 and 2017 challenges changed their task with respect to the 2015 edition, from acute lesion segmentation to lesion outcome prediction. In the 2016 edition, among the top three methods one was based on RDFs and two on CNNs. In the upcoming 2017 edition, only CNNs were present among the top three methods. Recently, big advances have been made in techniques that minimise the downsides of deep learning methods for brain lesion segmentation.

III. METHODOLOGY

A. Implemented architectures

The deep learning architectures selected for evaluation are a varied representation of the state of the art in brain imaging. A combination of networks designed for tissue segmentation and lesion segmentation as well as other well-known networks will be evaluated.

- **U-Net** [9] is a U shaped network with 5 resolution steps that won the ISBI 2012 cell tracking challenge.
- **3D U-Net** [10] is a network based on the original U-net architecture extended for 3D with some other improvements.
- **uResNet** [11] is a U-shaped network with four resolution steps and residuals. Architectures using residuals are easier to optimize and can gain accuracy from considerably increased depth [12]. In this work, the original 2D architecture has been extended for 3D to better suit the problem of stroke lesion segmentation.
- **DMRes** [13] is an extension of the DeepMedic architecture, an 11-layer deep, multi-scale 3D FCNN, with residual connections was the overall winner of the BRATS 2016 and the ISLES 2015 SISS challenge.
- **CNN$_{\mathbf{MULTI}}$** [14] is a nine layers deep multiscale FCNN originally proposed for brain tissue segmentation.

B. Implemented pipeline

1) Training patch sampling: The implemented patch extraction pipeline is a mixture of the ones from uResNet [11] and DeepMedic [15] aimed not only to address class imbalance but lesion type imbalance. Class imbalance is also a concern for automatic lesion segmentation methods since it can lead to poor performance and unexpected mistakes. Stroke lesions in both acute and chronic stages are very heterogeneous. This pipeline aims to have a balanced patch representation of the lesion from each patient by ensuring the same number of patches is extracted from each. To ensure the desired number of positive patches is reached for patients with smaller lesions a combination of several patch extractions from the same lesion voxel and data augmentation is done. Data augmentation is performed with six anatomically feasible operations including horizontal and vertical axial mirroring and 90^{o}, 180^{o} and 270^{o} axial rotations. Hence, using these functions the number of patches can be augmented up to a factor of six. In practice, a goal number of patches to extract per patient is specified. Then, for each patient, 50% of the training patches are extracted with uniform sampling from the whole volume and the other 50% with positive sampling. A diagram summarising the whole process can be found in Fig. 3

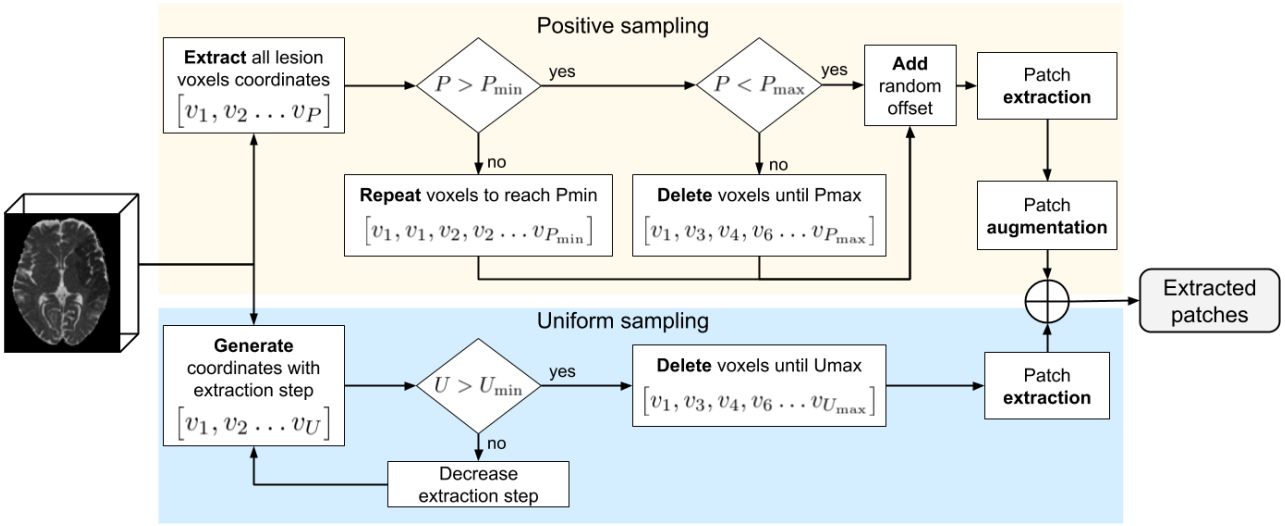


Fig. 3: Diagram of implemented patch sampling strategy.

2) Network training: Once a balanced training patch set is extracted, the weights of the networks will be trained with the same hyper-parameters to offer a fair comparison. To avoid costly grid search of training hyper-parameters, the Adadelta optimizer is used to train all networks which requires no manual tuning of a learning rate. Crossentropy is employed as the loss function given its gradient properties that ease convergence of the backpropagation algorithm. A *soft* Dice loss based on the Dice similarity coefficient (DSC) is used as the monitored metric for the Early stopping technique. The networks were trained with a batch size of 32. A global goal of 250 000 patches used for training each dataset is set that makes a compromise on the amount of samples and the training time. Finally, the crossvalidations will be done in 4 folds on all datasets, adjusting the amount of cases per fold accordingly.

3) Segmentation and post-processing: Once the network has been trained with a balanced training set we evaluate it with samples never used for training. To segment a volume with patch-based methods first the patches are extracted and forward passed through the net individually. They are sampled uniformly with a regular extraction step to make sure all the parts of the volume are forward passed through the net. Then, the segmented patches are combined in a common space with the same dimensions as the original volume, preserving their original spatial location, to produce the final segmentation. In our case, the combination is performed per voxel by averaging the class probabilities of the various segmentations. Finally, the post processing step involves a thresholding of the class probabilities followed by a connected component filtering by lesion volume. The variable threshold can compensate over/under confident networks while the minimum lesion size takes advantage of lesion priors to minimise false positives. An optimal combination of threshold and minimum lesion size maximising the average overlap is obtained through grid search for each task and architecture.

IV. Proposals

A. Ensembles

Recently, Kamnitsas et al. [16] won the BRATS 2017 brain tumour segmentation challenge with its method Ensemble of Multiple Models and Architectures (EMMA). The key idea being that averaging the outputs of individual models marginalizes the influence of the meta-parameters used for training. This effectively averages away the variance of any individual model, resulting in more robust and consistent performance. Given the good results achieved by EMMA, an ensemble will be built in a similar fashion with a selection the implemented methods with the aim of improving results. In practice, the performance of the trained models will be evaluated and an ensemble will be built with the best performing ones. In the end, the built *Ensemble UNET*, makes use of the 3D U-Net and 3D uResNet. The performance of both architectures is overall consistent across datasets and show similar confidence levels on their class probabilities. Given that both U-nets have compatible patch sizes, the input and output patch shape of the ensemble will be $24 \times 24 \times 8$ as seen in Fig. 4.

B. Atlas assisted lesion outcome prediction

Most of the presented methods to the ISLES 2017 challenge are adapted to stroke lesion outcome prediction from other brain imaging tasks. Methods performing this task need to capture the full correlation between acute images and final lesion extent, really weakened both spatially and temporally. For this, a network could be trained with a big enough representation of all the possible cases and its evolution. However, the number of required images for this approach grows exponentially with the number of factors influencing the evolution. To compensate the lack of data, a model using additional knowledge, not contained in the training images, could be used. In other words, either you have enough data representing the full set of correlations or you build a model that uses additional knowledge to *fill the gap*. In our case, *a*

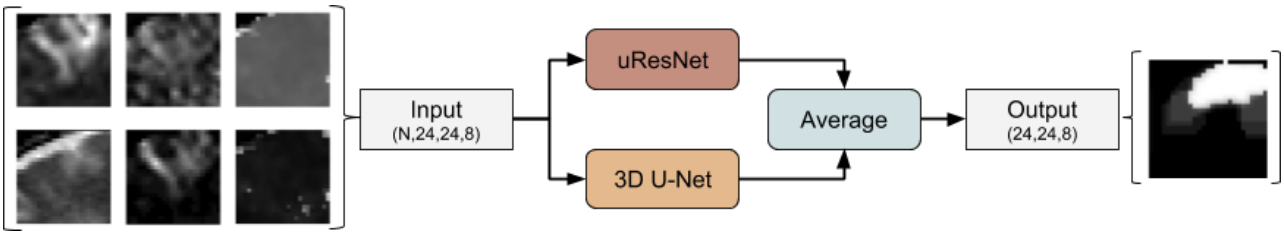


Fig. 4: Diagram of implemented Ensemble UNET including the 3D extension of uResNet [11] and the 3D U-Net [10].

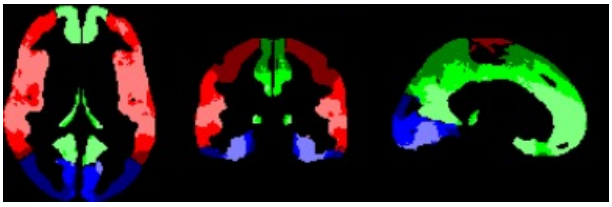


Fig. 5: Atlas of blood supply territories [17]. ACA (green), MCA (red) and PCA (blue) refer to the standard flow territories perfused by the bilateral anterior, middle and posterior cerebral arteries respectively.

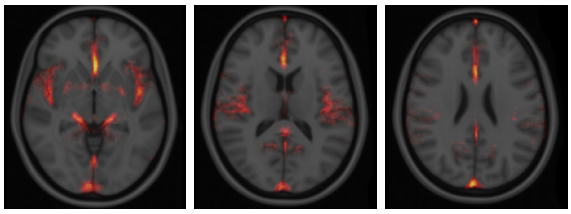


Fig. 6: Vessel density atlas [18] axial slice series in a red to yellow colormap, overlayed on top of the structural MNI 152 template for reference.

priori information about vascular and other processes can be fed as additional knowledge to CNNs using *atlases*. For our purpose, an atlas can be described as an image representing an average model of a structure. Furthermore, an atlas can include richer information, such as local image statistics and the probability that a particular spatial location has a certain label. In our case, atlases with information about brain vascularity are of especial interest:

- **Blood supply territories atlas** depicted in Fig. 5 was created based on the prints of vascular territories by Tatu et al. in 1998, using arterial transit time (ATT) data from a population of elderly with hypertension [17]. The labels from the atlas correspond with the ramifications of the Anterior, Middle and Posterior cerebral arteries which are the three main arteries that irrigate the brain. Establishing the arterial territories for each part of the brain is a non-trivial task which is open to interpretation.
- **Vessel density atlas** [18] depicted in Fig. 6 is a freely available digital atlas of frequency of vascularisation. It was obtained from N = 38 healthy participants scanned with the time-of-flight magnetic resonance technique. It essentially encodes the probability of finding a vessel in a particular spatial location.

Atlases can be integrated in a CNN framework as an additional input modality. In this way, the network can consider the whole neighbourhood and extract local and global statistics as features. In practice, the atlases are already registered to the MNI 152 T1 structural template. We perform a linear registration of the template to the patient ADC image and obtain the transform, which will be used to register the atlases to the patient's space. Finally, patches are extracted for training and testing containing all co-registered modalities across the channel dimension including the atlases.

V. EVALUATION

A. ISLES 2015 SISS results

The evaluation metrics for each of the architectures can be found in Table I. The best overlap is achieved by the Ensemble UNET model with 63.7 $\pm$ 25.3 %, not far from the inter-rater overlap for the SISS dataset of 70.0 $\pm$ 20.0 % [3]. The ensemble of 3D u-nets has a comparable performance to uResNet and is significantly better than the 3D U-Net ($p < 0.05$). The richer correlations found in the bigger 3D neighbourhood probably make the network more sensitive to probable lesions. However, overall the average overlaps of 62.1% of 3D U-Net and 61.5% of uResNet are marginally higher than the 60.1% achieved by the U-Net.

Challenge evaluation results: The best approach in the SISS dataset was the 3D uResNet with an average crossvalidation overlap of 61.5%. The results situate our implementation on the **third place** out of 34 entries of the ongoing leaderboard as seen in Table II.

B. ATLAS R1.1 results

The evaluation metrics for each of the architectures can be found in Table I. The 3D U-Net achieves the best result with an overlap of 72.4 $\pm$ 21.0 %, which is not far from the inter-rater overlap of 76.0 $\pm$ 14.0 % for this dataset [4]. The 3D U-net is significantly better than all other tested networks ($p < 0.05$). The bigger size of this dataset means a more complete representation of the correlations between input and output is present. Our hypothesis is that the doubling of channels on the 3D U-Net increases the capability to capture the bigger number of correlations present in this dataset. This capability is not apparent with smaller datasets as there are is a limited amount of information to capture. The two FCNN architectures perform significantly worse than the u-nets, probably due to the training procedure less suited for this

TABLE I: Results of 4 folds crossvalidation on ATLAS R1.1, ISLES 2017 and ISLES 2015 SISS and SPES.

	SISS		**ATLAS R1.1**		**SPES**		**ISLES 2017**	
Architecture	DSC (%)	HD	DSC (%)	HD	DSC (%)	HD	DSC(%)	HD
uResNet	61.5 ± 25.3	36.2 ± 27.6	62.5 ± 24.0	56.2 ± 25.0	77.6 ± 17.6	14.1 ± 10.3	38.7 ± 22.2	26.2 ± 20.9
3D U-Net	62.1 ± 25.5	38.1 ± 26.8	**72.4 ± 21.0**	48.3 ± 26.6	77.8 ± 17.8	12.2 ± 5.6	38.8 ± 23.0	26.0 ± 20.0
U-Net	60.1 ± 27.4	32.2 ± 28.0	63.0 ± 22.4	78.1 ± 20.8	**78.2 ± 15.3**	**11.7 ± 6.1**	33.6 ± 22.4	34.8 ± 24.2
DMRes	54.6 ± 28.6	47.1 ± 28.9	33.6 ± 26.2	89.8 ± 15.4	71.6 ± 18.8	14.8 ± 9.6	34.9 ± 20.6	32.2 ± 21.9
CNNmulti	51.0 ± 30.2	56.7 ± 35.6	38.4 ± 26.6	95.0 ± 14.3	75.8 ± 16.4	13.5 ± 8.6	30.7 ± 22.2	25.9 ± 17.3
Ensemble UNET	**63.7 ± 25.2**	**31.7 ± 26.1**	68.8 ± 23.4	**32.1 ± 26.2**	78.1 ± 18.5	11.7 ± 6.1	**40.0 ± 22.8**	**24.8 ± 19.9**

TABLE II: Ongoing benchmark leaderboard of the ISLES 2015 SISS challenge testing set evaluation.

Ranking	**Username**	**DSC**	**HD**
1	kamnk1	0.59 ± 0.31	39.6 ± 30.7
N/A	zhanr6	0.58 ± 0.31	38.9 ± 35.3
N/A	**clera1 (ours)**	0.56 ± 0.29	34.4 ± 27.1
2	fengc1	0.55 ± 0.30	25.0 ± 22.0
3	halmh1	0.47 ± 0.32	46.3 ± 34.8

TABLE III: Ongoing benchmark leaderboard of the ISLES 2015 SPES challenge testing set evaluation.

Ranking	**Username**	**Approach**	**DSC**
1	mckir1	RDF [8]	0.82 ± 0.08
2	maieo1	RDF [7]	0.81 ± 0.09
N/A	**clera1 (ours)**	U-Net	0.80 ± 0.10
3	robbd1	RDF [20]	0.78 ± 0.09

kind of architectures. Overall, CNN_{MULTI} performs significantly better than DMRes ($p<0.05$) with 5% more average overlap. The Ensemble UNET made by averaging the results of 3D U-Net and uResNet is significantly worse than 3D U-Net ($p<0.05$). This is the only task where the Ensemble UNET model does not improve results as compared with the individual networks.

C. ISLES 2015 SPES results

The evaluation metrics for each of the architectures can be found in Table I. All the networks obtained a high DSC as compared with the other considered tasks in this document. This is due to the systematic way in which the gold standard was generated, which establishes a simple numerical correlation between the intensities and output label. The original U-Net achieves the maximum overlap at an average of 78.2%. The 3D U-Net and uResNet achieve 77.8% and 77.6% of average overlap respectively, marginally lower and with around 2.5% more variability. The more consistent results achieved by the U-Net are due to the smaller, less confounding 2D neighbourhood that allows for more generalizable and robust features. The FCNNs achieve the lower overlap with 71.6% for the DMRes and 75.8% for CNN_{MULTI}, which is the closest to the u-nets. The Ensemble UNET achieves a higher average overlap, minimising the segmentation errors of individual networks and improving the overall segmentation consistency.

Challenge evaluation results: The best approach was the 2D U-net architecture, which was submitted for evaluation. The results situate our implementation on the **third place** out of 12 entries of the ongoing leaderboard with an average overlap of 80% as seen in Table III. In the SPES challenge, just one out of the seven presented methods was CNN based and it was the worst performer, with mainly RDFs in the top positions. Three years later, the advances in deep learning mean CNNs can offer similar generalizability properties as RDFs, achieving comparable results.

TABLE IV: Average metrics of 9-fold crossvalidation on ISLES 2017 dataset with and without the BST and VD atlases registered as an additional modality.

Architecture	**DSC (%)**	**HD**
uResNet	38.7 ± 22.2	26.2 ± 20.9
uResNet + atlas	40.5 ± 23.6	26.5 ± 21.7
3D U-Net	38.8 ± 23.0	26.0 ± 20.0
3D U-Net + atlas	38.1 ± 23.8	30.3 ± 24.4
U-Net	33.6 ± 22.4	34.8 ± 24.2
U-Net + atlas	35.1 ± 22.8	33.9 ± 25.9
DMRes	34.9 ± 20.6	32.2 ± 21.9
DMRes + atlas	34.3 ± 21.7	25.4 ± 18.1
CNN_{MULTI}	30.7 ± 22.2	25.9 ± 17.3
CNN_{MULTI} + atlas	32.3 ± 21.0	28.5 ± 19.5
Ensemble UNET	40.0 ± 22.8	**24.8 ± 19.9**
Ensemble UNET + atlas	**41.1 ± 23.5**	27.7 ± 24.0

D. ISLES 2017 results

The evaluation results of each model in the lesion outcome prediction task can be seen on Table I. The Ensemble UNET achieves the best results with a 40% average overlap, around 1% higher than the individual methods. Again, the presence of several failed cases widens the metrics variance difficulting a finer analysis of the performance differences between the models. In this case, the 3D u-nets have significantly better results ($p<0.05$) as compared with the 2D u-net, which had comparable performance in the ISLES 2015 SISS and SPES tasks. The networks greatly benefit from the richer spatial correlations of 3D neighbourhoods for the task of stroke lesion outcome prediction. This suggests that the features needed for accurate prediction have a wide spatial influence. The FCNN architectures show significantly lower results as compared with the 3D u-nets, with a performance comparable to the 2D U-Net.

E. ISLES 2017 with atlases results

The addition of vascular related atlases marginally improved the quantitative and qualitative results of some networks. From the evaluated networks uResNet, CNN_{MULTI} and U-Net all show an increase in average overlap of around

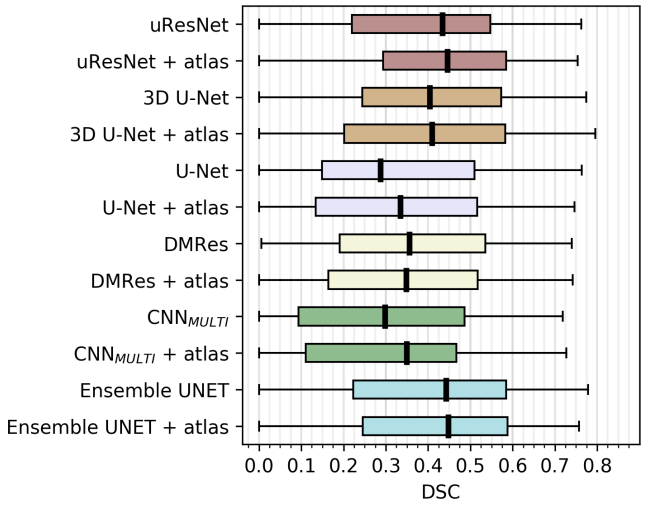


Fig. 7: Detailed overlap results of the ISLES 2017 dataset evaluation with and without the BST and VD atlases registered as an additional modality.

TABLE V: Ongoing benchmark leaderboard of the ISLES 2017 challenge testing set evaluation.

Ranking	User	DSC	HD
N/A	pinta1	0.36 ± 0.22	30.6 ± 14.0
N/A	**clera2 (ours)**	0.34 ± 0.21	32.5 ± 13.8
1	kwony1	0.31 ± 0.23	45.3 ± 21.0
2	lucac1	0.29 ± 0.21	33.9 ± 16.8
3	mokmc2	0.32 ± 0.23	40.7 ± 27.2

1.6%, which also translates in higher median values. On the other hand, the 3D U-Net and DMRes models have marginally less overlap as compared with the baseline. This suggests something in the architectural design of these networks can better handle with the increasing number of modalities. In the case of the 3D u-nets, uResNet improved its performance while 3D U-Net marginally decreases. The main architectural difference of the 3D U-net with respect to uResNet is the doubling of convolutional layers per step and the absence of residuals. One of these factors or a combination of both might be preventing the 3D U-Net from effectively utilizing the additional information. The uResNet model sees an increase in its lower quartile with respect to the baseline, as seen in Fig. 7, indicating a better and more consistent segmentation of some harder cases. The Ensemble UNET model achieves the highest average overlap with 41.1%. Despite the worse performance of 3D U-Net in this case, the ensemble still improves results as compared with the individual results of each model.

Challenge evaluation results: The best approach was the 3D uResNet in combination with the two registered atlas , which was submitted for evaluation. The results situate our implementation on the **second place** out 27 entries of the ongoing leaderboard with an average overlap of 34% as seen in Table V.

VI. CONCLUSIONS AND FUTURE WORKS

In this master thesis, we made an extensive quantitative and qualitative evaluation of state of the art deep learning architectures applied to stroke imaging tasks. More specifically, the developed work can be summarised as follows.

1) Reviewed the state of the art on methods for stroke image processing and related biomedical imaging tasks.
2) Implemented five state of the art deep learning architectures for stroke lesion segmentation and outcome prediction.
3) Implemented a training patch extraction pipeline suited for lesion segmentation tasks that addresses class and lesion type imbalance and makes use of anatomically feasible data augmentation functions.
4) Proposed of the addition of atlas as registered modalities to improve stroke lesion segmentation, which showed marginally better results only on some networks. We also explored the use of ensemble models that improved results by reducing false positives and refining the segmentation borders.
5) Evaluated the implemented deep learning architectures, trained with the same hyperparameters and pipeline, and the proposed improvements.
6) Compared our implementation against state of the art methods. Our implementation achieved state of the art performance, ranking *third* in the ongoing leaderboard of both ISLES 2015 SISS and SPES and *second* on the ISLES 2017.

We hope the experience and insights gained during the development of this work will serve for the participation in the upcoming ISLES 2018 challenge. The insights obtained from the response of different architectures to the same dataset allows us to determine what strategies are effective and which not. The addition of atlases as additional modalities as well as what architectural features can effectively utilize the given information will be further studied. Despite the change of focus of ISLES 2018, the strategy of atlas assisted deep learning may also be of help for lesion segmentation. The creation of more stroke lesion related atlases from images of available datasets will be explored too. Finally, a more deliberate ensemble strategy can be implemented once a number of suited architectures for the task have been found.

REFERENCES

[1] Camilo R. Gomez. Editorial: Time is brain! *Journal of Stroke and Cerebrovascular Diseases*, 3(1):1–2, jan 1993.

[2] Bernd F. Tomandl, Ernst Klotz, Rene Handschu, Brigitte Stemper, Frank Reinhardt, Walter J. Huk, K.E. Eberhardt, and Suzanne Fateh-Moghadam. Comprehensive Imaging of Ischemic Stroke with Multisection CT. *RadioGraphics*, 23(3):565–592, may 2003.

[3] Oskar Maier et al. ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Medical Image Analysis*, 35:250–269, jan 2017.

[4] Sook-Lei Liew et al. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. *Scientific Data*, 5:180011, feb 2018.

[5] Konstantinos Kamnitsas, Liang Chen, Christian Ledig, Daniel Rueckert, and Ben Glocker. Multi-Scale 3D Convolutional Neural Networks for Lesion Segmentation in Brain MRI. In *Proceedings of ISLES (SISS) challenge*, 2015.

[6] Halla-Leena Halme, Antti Korvenoja, and Eero Salli. ISLES (SISS) challenge 2015: Segmentation of stroke lesions using spatial normalization, Random Forest classification and contextual clustering. *Proceedings of ISLES (SISS) challenge*, pages 31–34, 2015.

[7] Oskar Maier, Matthias Wilms, and Heinz Handels. Random forests for acute stroke penumbra estimation. In *Proceedings of ISLES (SPES) challenge*, 2015.

[8] Richard Mckinley, Levin Häni, Roland Wiest, and Mauricio Reyes. Segmenting the ischemic penumbra: a spatial Random Forest approach with automatic threshold finding. In *Proceedings of ISLES (SPES) challenge*, 2015.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234—-241, may 2015.

[10] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. pages 424–432. Springer, Cham, oct 2016.

[11] R. Guerrero, C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M.C. Valdés-Hernández, D.A. Dickie, J. Wardlaw, and D. Rueckert. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage: Clinical*, 17:918–934, jan 2018.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, jun 2016.

[13] Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V. Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. DeepMedic for Brain Tumor Segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2016.*, pages 1–12. Springer, Cham, oct 2017.

[14] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, apr 2017.

[15] Konstantinos Kamnitsas, Christian Ledig, Virginia F.J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, feb 2017.

[16] Konstantinos Kamnitsas, Wenjia Bai, Enzo Ferrante, Steven McDonagh, Matthew Sinclair, Nick Pawlowski, Martin Rajchl, Matthew Lee, Bernhard Kainz, Daniel Rueckert, and Ben Glocker. Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2017.*, pages 450–462. Springer, Cham, 2017.

[17] H. J. M. M. Mutsaerts, J. W. van Dalen, D. F. R. Heijtel, P. F. C. Groot, C. B. L. M. Majoie, E. T. Petersen, E. Richard, and A. J. Nederveen. Cerebral Perfusion Measurements in Elderly with Hypertension Using Arterial Spin Labeling. *PLOS ONE*, 10(8):e0133717, aug 2015.

[18] Roberto; Viviani, Julia; C. Stingl, and Universitäts Klinikum Ulm. Vessel Density Atlas. https://www.uniklinik-ulm.de/psychiatrie-und-psychotherapie-iii/forschung-studien/clinical-neuroimaging.html.

[19] Chaolu Feng, Dazhe Zhao, and Min Huang. Segmentation of Stroke Lesions in Multi-spectral MR Images Using Bias Correction Embedded FCM and Three Phase Level Set. In *Proceedings of ISLES (SISS and SPES) challenge*, 2015.

[20] David Robben, Daan Christiaens, Janaki Raman Rangarajan, Jaap Gelderblom, Philip Joris, Frederik Maes, Paul Suetens, Oskar Maier, Mauricio Reyes, David Robben, Daan Christiaens, Janaki Raman Rangarajan, Jaap Gelderblom, Philip Joris, Frederik Maes, and Paul Suetens. ISLES Challenge 2015: A voxel-wise, cascaded classification approach to stroke lesion segmentation. In *Proceedings of ISLES (SISS) challenge*, pages 254–265. Springer, Cham, oct 2015.

[21] Youngwon Choi, Yongchan Kwon, Myunghee Cho Paik, Beom Joon Kim, and Joong-Ho Won. Ischemic Stroke Lesion Segmentation with Convolutional Neural Networks for Small Data.

[22] Christian Lucas and Mattias P Heinrich. 2D Multi-Scale Res-Net for Stroke Segmentation.

[23] Tony C W Mok, Albert C S Chung, I Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, A Courville, and Y Bengio. Deep Adversarial Networks for Stroke Lesion Segmentation.

Particle Detection with Acoustic Waves

Thomas DREVET

***Abstract*— Particles detection is a process that is getting more and more interest for applications mainly related to disease detection (Cancer, AIDS). Several methods were developed, but there is a lack of works and studies on some aspects of the particles detection techniques: for instance, the existing methods are all based on very complex setups interacting with Bulk Acoustic Waves.**
My work is oriented about investigations on those aspects: try to build a technique based on Surface Acoustic Waves, perform particles in suspension detection, use a less complex setup. It implies that I should answer different interrogations: what should be the properties of my setup, which particle model, which detection technique? During the time that lasts my thesis, I investigate on those problematics using a simulation software, COMSOL, during two studies: one based on the frequency domain to know more about what is happening inside a microchannel in presence of a particle, and one based on time domain to simulate what a particles detection should look like with my setup and my parameters.

I. INTRODUCTION

Particles detection techniques are widely used in different fields such as medical or industries [1], in order to perform different tests (detecting flaws, medical diagnostics). Those techniques can be under different format using either chemical reactions for destructive tests, either mechanical properties (acoustic waves) for non-destructive tests. All of those techniques can be improved (higher sensitivity) in order to have more accurate results. The work presented here is oriented around one precise technique, based on Surface Acoustic Waves (SAWs). My work focused on the detection of particle inside a 'Lab on Chip' device.
My thesis was designed around two problematics:

- What is the detection capability of SAWs for particle in suspension? Can we differentiate particle mechanical properties ?
- What are the parameters requirements (frequency, channel size, particle size, particle material) to obtain exploitable results ?

In order to answer these two problematics, I used COMSOL Multiphysics, a simulation software, designed to solve finite physics analysis. This software is widely used in academia and industries and can be applied for several applications such as electrical, mechanical or chemical. The choice of using this tool was due to its capability to perform the different types of analysis required for my investigations: possibility to work in frequency and time domains, easy to play on parameters/resolution.

My thesis is oriented around two investigations, based on:

- frequency domain: study of the changes in the acoustic pressure field with different type of particles
- time domain: study of acoustic response in time with different type of particles

II. LITERATURE

A. Detection: Back-scattering and Transmission

The design of the 'Lab on Chip' device (two InterDigital Transducers (IDTs) set parallel to each other and around the microchannel) is illustrated on Figure 1. It permits to record the data thanks to two methods: Back-scattering or Transmission. Those two techniques are commonly used in acoustic detection, especially for flaw detection in materials (non destructive tests) [2]. One of my objective is to access which method is the most appropriate for the characterization of particles in suspension. For this, I will model both methods in COMSOL and compare the signal received in both configuration.

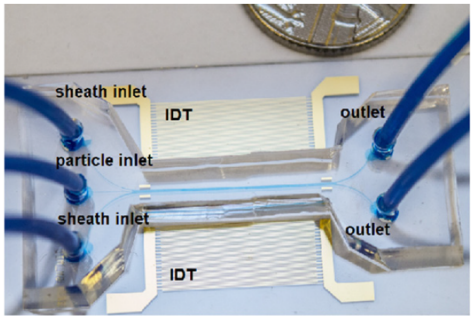


Fig. 1. The 'Lab on Chip' device, *image taken from [3]*

1) Back-scattering Technique: The back-scattering is an approach of the pulse-echo technique. First introduced by Sergei Sokolov in 1937, the acoustic pulse-echo technique is directly inspired from the military RADAR system [4]. The main principle is to send an acoustic pulse that will reflect on everything it encounters. The pulse will meet and then go back to the emitter which will also serve as a detector. The obtained signal will contain data from what it crossed (nature and location). The figure 2 illustrates the way the back-scattering technique can be applied in the case of the 'Lab on Chip' device, and what I can extract from it. In order to extract the content of the signal, the objective is to compute the back-scatter function BSTF(f), following this formula:

$$BSTF(f) = \frac{R_{signal}(f)}{R_{coef} * R_{ref}(f)}$$

with $R_{signal}(f)$ the Fourier Transform of the back-scatter signal, R_{signal} the reflection coefficient and $R_{ref}(f)$ the Fourier Transform of the input signal. The objective with

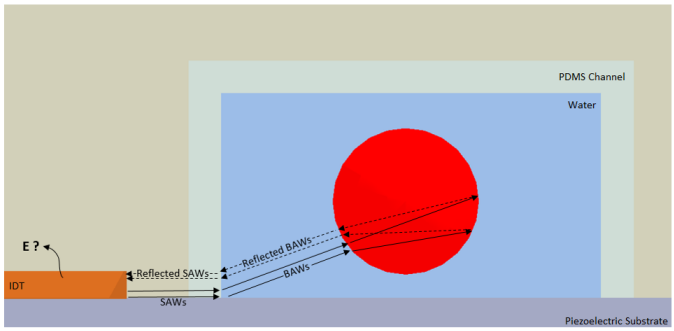


Fig. 2. The back-scattering technique applied to the 'Lab on Chip' device

the obtained curve is to see if we can use it as 'fingerprint' tool, to detect the type of particle we are facing.

2) Transmission Technique: In this second technique instead of emitting and receiving the signal with the same IDT, we emit the pulse with one and we receive the output signal with the opposite one, as shown on figure 3. As for the Back-scattering technique, the goal here is to catch the signal modified by the medium it goes through (material, location). For this method, we are computing the Measured Scattering-

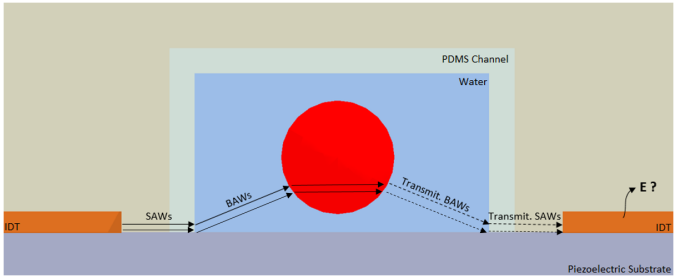


Fig. 3. The transmission technique applied to the 'Lab on Chip' device

Cross Section σ(f), following this formula:

$$\sigma(f) = R^2 \frac{|S(f) - S_{np}(f)|^2}{|S_{ref}(f)|^2}(2d)^2$$

with S(f) the signal obtained with the particle, S_{np} the signal obtained in the situation where there is no particle in the channel, S_{ref} the signal emitted by the IDT, *R* the radius of the particle and *d* the distance IDT/particle. The obtained curve should be close to the curve inferred by the particle model, as for the other techniques.

B. The Faran Particle Model

The concept of particle model will be important to simulate, in true condition, what is happening (physical parameters, positioning of elements) and to compare my results with the theory. Those models are used to compute the emitted frequency and afterwards the comparison tool for my results. There are several particle models which are existing and that are currently used by different detection methods (Single Cell, Anderson). Among them, I decided to study more deeply one of them that is the Faran model[5]. The Faran model designed by James Faran, in 1951, to model the sound scattering of solid cylinders and spheres. This mathematical model is still commonly used to estimated and solve the solution of the way an acoustic wave is interacting with a cylinder or a sphere (particle in 2 or 3D). Despite the model oldness, it is still very used to estimate different elements in acoustic scattering cases as explained in [6], for ultrasonic scattering of bones.

C. The 'Lab on Chip' device with Surface Acoustic Waves

1) The 'Lab on Chip' device: The 'Lab on Chip' device is a small instrument composed by a Lithium Niobate ($LiNbO_3$) substrate, a piezoelectric material that permits the generation and the travelling waves propagation. On this substrate, two InterDigital Transducers (IDTs) are set parallel to each other, between them, a microchannel in Polydimethylsiloxane (PDMS) is bounding in. PDMS is commonly used in microfluidic applications for its low reflection [7]. This type of device can be used in applications such as particles separation (Nam *et al.* [8]), particles trapping (Chen *et al.* [9]) or particles sensing (Go *et al.* [10]).

2) SAWs Generalities and their Solid Propagation: The study from Lord Rayleigh in 1885, about waves propagation [11], put in light the existence of a new type a waves that propagate in elastic solids with interesting properties, called the Surface Acoustic Waves (SAWs). Composed by a longitudinal and a transverse (shear) component, they present, whatever the material (isotropic or not), different velocities [12]: the first propagating parallelly to the SAW propagation axis and the second normal to the surface. The figure 4 shows this propagation: the longitudinal component along the x-axis (direction of the wave) and the shear one along the y-axis.

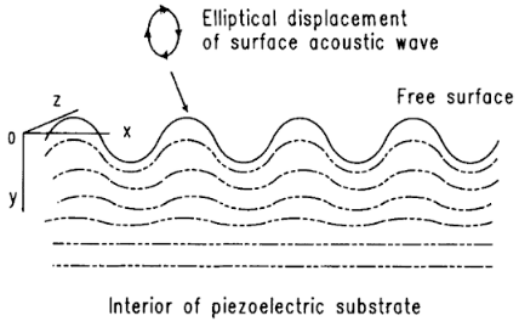


Fig. 4. SAWs propagation, *image taken from [12]*

The Surface Acoustic Waves are generated by applying a voltage to two Interdigital Transducers (IDTs), placed on a piezoelectric material [12]: one of them is used as converter between the voltage signal and the mechanical SAWs and the second one do the contrary to work as receiver, as shown on figure 5.

3) The SAWs/Liquid Interation: The SAWs propagation is a key element in my thesis: you will see the importance of the propagation of the SAWs facing a liquid. As explained in [13], the waves are strongly absorbed (where the amplitude is decaying) depending on the 'quantity' of liquid crossed,

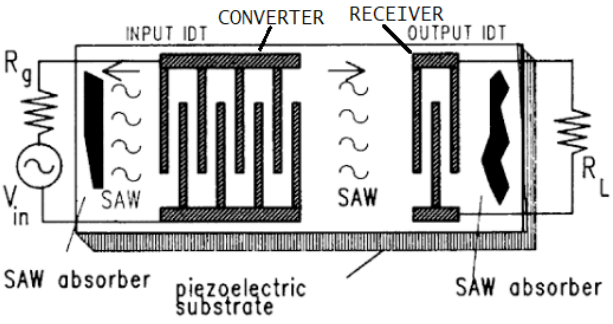


Fig. 5. SAWs generation by IDTs, *image taken from [12]*

becoming Leaky SAWs (LSAWs). The figure 6 shows the wave propagation inside the 'Lab-on-Chip' device, when the two IDTs are emitting SAWs.

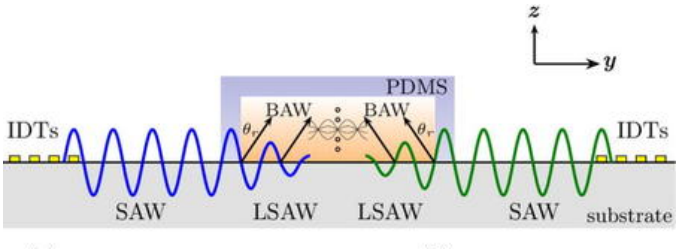


Fig. 6. SAWs in 'Lab on Chip' device, *image taken from [3]*

This phenomenon also creates a finite pressure difference, that leads to the generation, directly above the substrate, of longitudinal sound waves (a Bulk Acoustic Waves) into the liquid. The created waves will have a diffraction angle Θ_R, called the Rayleigh angle:

$$\Theta_R = arcsin(\frac{v_l}{v_S})$$

with, v_S the sound velocity of the substrate and v_l the sound velocity of the liquid. Another phenomenon is appearing thanks to the interaction SAWs/liquid: an acoustic radiation pressure appears in the direction in which the sound propagate himself within the liquid, creating some droplets. At high SAW amplitudes (high generation frequency or setup not well designed), those droplets become very deformed, preventing to see well what is happening inside the channel. An example of droplets is shown on figure 7.

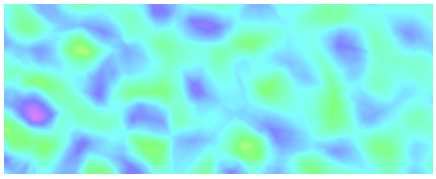


Fig. 7. Example of Droplets created by high SAW amplitude, *image done using COMSOL*

D. Existing Particles Detection Techniques

Particles detection has been existing for years and several techniques have been developed. Kishor *et al.* [14], from 2016, is a technique that is very close to my research question. As my work, the technique is based on 'the lab on chip' device: The main difference comes from the acoustic waves generation: as explained in my work, I generate SAWs with one of my IDT and I obtain my data either through a transmission with my second IDT or via a back-scattering. Here, they are generating bulk acoustic waves using a pulsed laser pointed on the micro channel that will become SAWs in order to be measured by the IDTs. The different tests they have done using a blue dye, are able to detect them very successfully.

There are several techniques that are not using a single device as the 'Lab on Chip' one. One of them is the method presented by Baddour *et al.* [15]. They are using a pulser, that through a transducer, will generated high frequency sound waves. Those waves, as in a SAW device, will be recorded after passing to a channel, by a transducer. The processing of their data is very interesting because it is near to the one I am using: they are also using a scattering signal, but it is another model, the 'Single Cell' one, mainly used for cell analysis. Their work gives a results close to their theory but it is not adapted to their application (do not fit well with all the type of cells they are working on). They also have issues with resonances and interferences that are highly visible in their output work. There is another interesting method, the one from Falou *et al.* [16]. This technique is close to the previous one in terms of methodology (scattering) even if they are using the Anderson model instead of the 'Single Cell' one. The main difference comes from their different setup usage: In it, they are using a custom made structure in plexiglass containing an ultrasound transducer, pointing a window served to let pass the optical informations, formed at this location. Those data will then be catch up by a microscope, situated under the window. This setup, quiet complex to build, is giving them very good results on the particles they are testing (polystyrene micro-particles, cells).

E. Particles Detection Applications

Those particle detection techniques can be used for different applications. For instance, they can be used as tracker as shown in different research papers from Jochen Guck, as Urbanska *et al.* [17]. In this publication, they are using cytometry to track the lineage of cells during days (cytometry is a measurement technique for cells that allows to get their characteristics based on , it is often use for different tests as Cancer or AIDS diagnostics [1]).

III. Model Presentation

A. The Virtual Model from the 'Lab on Chip' device

As explained previously, there are several methods to perform particles detection (Falou *et al.*, Kishor *et al.*). Their setups are very complex/expensive, and my work is using a simpler setup and surface acoustic waves what it hasn't been done yet. I decided to work with the basic model from the 'Lab on Chip' device, shown on figure 8.

My investigations are realized using COMSOL, with a 2.5 dimension model from the device realized by Gergely Simon,

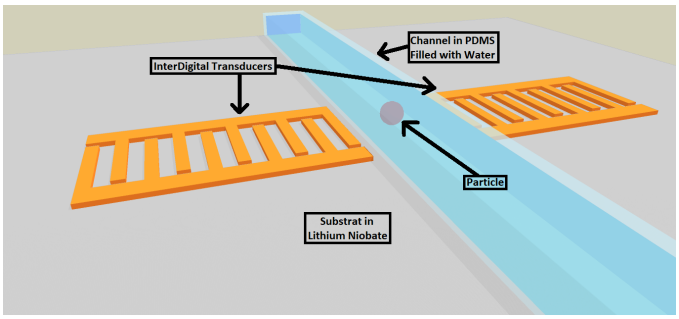


Fig. 8. 3D Diagram from the 'Lab on Chip'

shown in the figure 9. A 2.5 dimension model is an linear extension from a 2D one, transforming it into a king of three dimension model. This type of model permits to speed up the simulation in many cases and it allows, for instance, to include a particle.

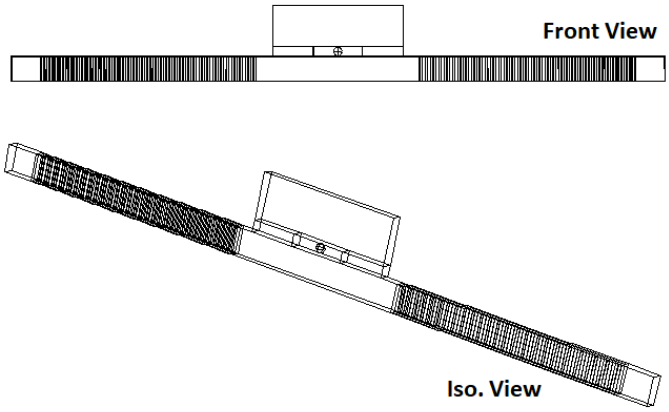


Fig. 9. Front and Isometric views from the COMSOL Model

B. The Parameters Choice

1) Size of Particles and Particles Materials: For the tests with a particle, I included a single particle centered in the channel. The choice of using a single particle has been taken because it permits to use easily the Faran Model that works in my case. The different tests I realized during my studies are mainly oriented around the way particles interact with the acoustic field in the frequency domain study and influence the data recorded at the IDTs in the time domain study. In order to check the size influence, I decided to do tests with two sizes: 50μm and 25 μm for the sphere radius, allowing to have an overview of the output modifications that can differ with sizes evolution. The second particle parameter that I played with, during my tests, is its material. It is important to test several materials in order to see if we are able to detect correctly the nature of the particle. I summarized all the materials and their properties of the particles in the table I.

2) IDTs Frequency: For my tests, I decided to fix the frequency following the Faran model [5]. To test the possible frequencies, I used a Matlab code, that allows to draw

2*Name	Density (g/cm^3)	Cp (m/s)	Cs (m/s)	Acoustic Impedance ($g/cm^2 - sec \times 10^5$)
Glass	2.32	5640	3280	14.5
Liver Cell	1.07	1549	1.03	1.65
Polystyrene	1.06	2325	1120	2.94
Steel	7.85	5960	3235	45.5

TABLE I
PARTICLE MATERIAL PROPERTIES

attenuation for all the frequencies, depending on the situation (particle size/material, environment properties). The figure 10 shows the curve from the situations with a polystyrene particle one of 50 μm radius. The objective is to match the spikes from the studies output and the local minimum from those models. I decided to work with two different frequencies for my studies: 30 MHz for the frequency domain study and 50 MHz for the time domain one.

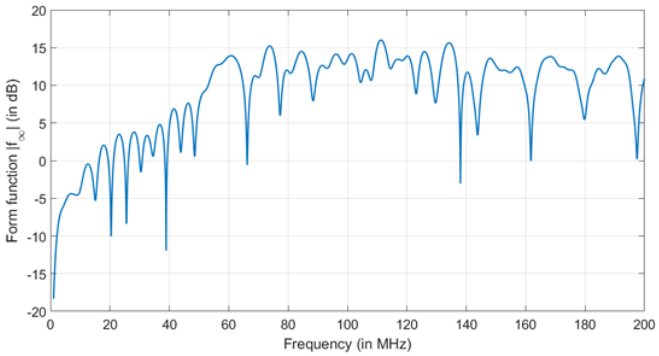


Fig. 10. Faran Model simulation for a polystyrene particle (50 μm radius), *done using Matlab*

IV. FREQUENCY DOMAIN STUDY

A. Presentation of the Study

The frequency domain study is the first one I realized: as said in introduction, its goal is to evaluate the way particles impact the detections within microchannel from the device. To do so, we check the acoustic pressure field from the channel obtained using the COMSOL simulations. I ran one simulation without any particle and four with a 50 μm particle with a different material (from the table I) each time to test the properties influence of them (Density, Acoustic Impedance, Wave Speeds) and a last test with a smaller particle 25 μm to test the influence of the particle size.

B. Presentation of the Test Results

A simulation realized in the frequency domain takes around 6 hours to be performed an Intel Core i5-6500 64 bits with 16 Go RAM. COMSOL can give different informations depending on what we want to study or verify. For this frequency domain study, I mainly work on the acoustic pressure field (in Pa) happening in the microchannel. For instance, the figure 11 shows how the full result are displayed in COMSOL, where we have a view of the microchannel

containing, in this case a particle, at the center of it. I can work with the full 2D data, but in order to speed up the analysis, I decided to only work in 1 dimension by taking only the middle line of the channel. I exported those reduced data to Matlab in order to have more tools for the data processing and for the plotting.

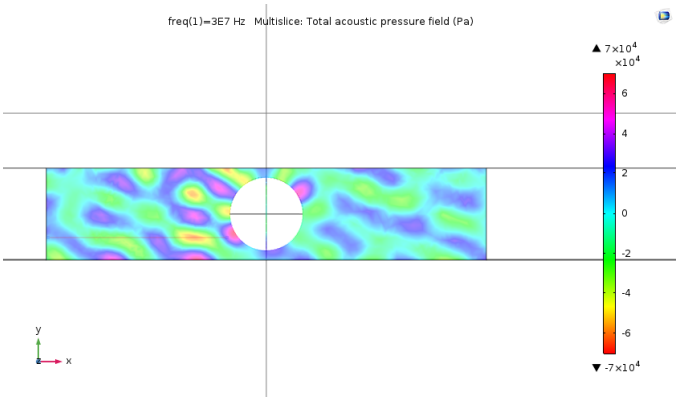


Fig. 11. Acoustic Pressure Field - Frequency Domain test - 50 μm particle of Steel (30MHz), *done with COMSOL Multiphysics*

The figure 12 shows the central line acoustic pressure for different tests regarding the particle material. The figure 13 shows the results from the tests I did about the particle size.

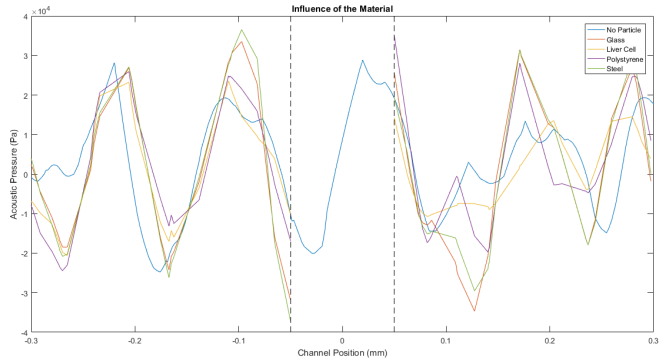


Fig. 12. Frequency Domain study, Particle Material Influence (30MHz), *plot done with Matlab*

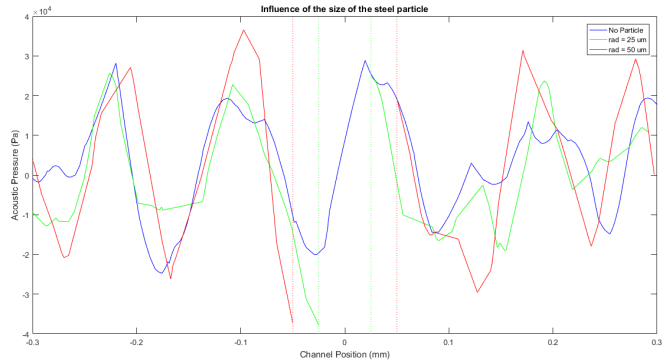


Fig. 13. Frequency Domain study, Particle Size Influence (30MHz), *plot done with Matlab*

From what we see from the two graphs, we can't see any clear link between parameters as the wave-speeds or the density, and curve shape as the amplitude or a time-delay. The only link that might be done is between the Pressure-Wave speed and the amplitude of the font before facing the particle. To be sure about this relation, I would need to do more tests (more particles with different properties).
For the particle size influence, it is also the same conclusion that we can do: we can observe that the particle size has an influence on the curve amplitude, but it is not clear enough (not enough data to confirm the hypothesis).

V. TIME DOMAIN STUDY

A. Presentation of the Study

This second study was the longest. As previously stated, the aim of it is to check if we are able to detect the particle inside the microchannel. To do that, I will use the 'Lab-on-Chip' model and work with a IDT frequency of 50 MHz. I will follow the same process as for the frequency domain study: I realized different simulations, for testing how the particle properties and the particle size are influencing the voltage received by the IDTs, thanks to the acoustic waves displacements. I will realize the tests with the same particles (size and material), as the frequency domain one. To perform such work, I will emit a signal with 17 of 18 electrodes from the IDT (the 17th farthest electrodes from the channel). This signal follows the mathematical formula $s(x) = sin(freq * 2*\pi*x)$, with freq being the IDT frequency (50 MHz), from 0 to 33*10^{-9} seconds. The shape of the signal is shown on figure 14.

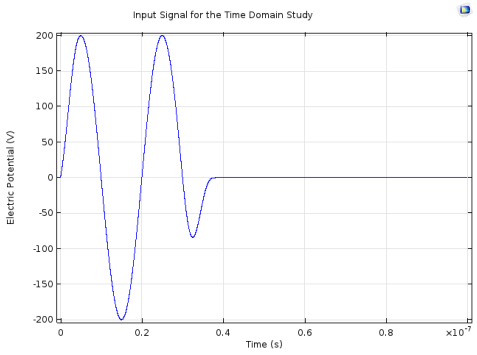


Fig. 14. Input signal for the time domain study, $s(x) = sin(freq*2*\pi*x)$ with freq = 50MHz

The last electrode from the emitting IDT (the closest to channel) is used as a receptor, as long as the closest electrode to the channel from the opposite IDT, permitting to test the two types of detection techniques

B. Presentation of the Test Results

At first, those tests have an important issue: they are very time consuming. One simulation in the time domain, with an Intel Core i5-6500 64 bits and 16 Go RAM, takes around 2 days (50 hours), and the time required to export the data (few hours), limiting the number of simulations I could run during

the time that lasted my thesis. A time domain simulation gives different information on what is happening: as for the frequency domain study, we can have access to the pressure acoustic field (in Pa) that takes place inside the microchannel. In this study, we can observe the displacement of the waves in the water, for each iteration of the simulation, as shown on figure 15.

C. Applying the Detection Techniques

1) The Back-scattering technique: We should use the data in order to compute the back-scattering technique, by computing the background difference. The background difference allows to quantify the impact of the particle presence on the signal. To obtain it, I do the difference between the signal containing the particle and the signal with no particle to have the background for the studied particle. Then, I compute the Fast Fourier Transform (FFT) to pass it into the frequency domain. Finally, I use the formula from the back-scatter function BSFT(f). To assert my results, I should compare the obtained curve with the one obtained by the Faran model simulation. The obtained curve for the 50 μm radius Polystyrene particle is shown on figure 16.

We can observe that the curve I am getting is very far from what I should obtain (curve realized by the Faran Model 10): I should see a link between the spike locations from the BSTF and the local minimum from the Faran curve (at 30 MHz, 40 MHz, 50 MHz). But in this situation, there is absolutely no relationship. It means that I either did a mistake during my work (simulations or computations), either the strategy I decided to follow is not working.

2) The Transmission technique: From those data, we can apply the FFT on the signal without the particle, and on all the signal obtained in the tests where there are particles. When we have them, we can use the formula from the Scattering-Cross section σ(f).
Then, I should follow the same logic as for the previous technique, by analyzing the results I am getting with the Faran model. The figure 17 shows the results I am getting for the transmission technique (50 μm radius Polystyrene particle).

The transmission technique gives a one spike curve at 50 MHz. When I analyze this curve with the Faran model one, we have the matching spike but there are two that are missing (30 and 40 MHz). It means that our results are incomplete. I can deduce that we face the same issue as for the back-scattering technique: I have either a modeling or processing error.

We can compare the results got from those two techniques, to see which one is working the best in our setup. From the results I obtained, we can say that the transmission technique seems to give a better matching (we have one spike on the three, compared to the zero from the back-scattering). But both methods are giving results that are too far from the Faran model. I don't know from the results I have right now, if the strategy is good or not: I have to check if I did no major mistake in my work, or the technique I built is just not doable.

D. Testing the Noise limit for the detection

We can perform another test with the electrical potential recorded by the IDTs, with the time domain data: it is possible to determine the level of noise for which the detection starts to be impossible. This study will have an important role, when the process will be tried in real condition, to determine the quality of the obtained results.
Those tests are realized using Matlab: I add white Gaussian noise on the recorded data, using the *awgn* function, with a different Signal to Noise Ratio (SNR). The SNR corresponds to the level of noise on the signal, computed by dividing the power of a signal (meaningful information) by the power of background noise (unwanted signal). Most of the time, we define it in dB by applying a $\log_{10}$ on the ratio and then multiplying the obtained value by 10. Once I add noise on my tested signal, I try to compare it to all the original signal (signals obtained for every particles I tested in the time domain study) by computing the errors, leading to a match. My full study is to run for different SNR (-30 dB, -22 dB, -16 dB, -9 dB, -3 dB, 0 dB) for all the particles I tested, an important number of times (1 million). I present in the table II, the results I obtain for the 50 μm radius Polystyrene particle.

SNR (dB)	-30	-22	-16	-9	-3	0
No Particle	12.9	3.0	0	0	0	0
50 μm Polystyrene	**45.4**	**76.3**	**98.1**	**100**	**100**	**100**
25 μm Polystyrene	16.1	8.9	1.0	0	0	0
50 μm Glass	12.2	5.3	0.3	0	0	0
50 μm Liver Cell	13.3	6.5	0.5	0	0	0

TABLE II
NOISE LIMIT STUDY FOR THE 50 μM RADIUS POLYSTYRENE PARTICLE (RESULTS IN %)

The results obtained here, are showing that after an SNR of -18 dB, the correct matching drops under 95 %, showing that the noise will start to be an issue with a noise that corresponds to a SNR around -16 dB, when we will realize our test in the 'real' world. Those results are around same value for all the type of particles I worked on (more or less 2-3 dB).

VI. CONCLUSION

A. Conclusion

My work, on the detection of particles in suspension, had to check the capability of SAWs to realize this job. To do that, I used the 'Lab-on-Chip' device (a COMSOL version of it), to see how the particle properties (size, material) are influencing some physical aspects (acoustic pressure field, electrical potential). Those aspects were tested inside two studies: one in the frequency domain that focus on the way the particle is modifying the acoustic pressure field, and one in the time domain focused more on the detection using the electric potential. This first study permitted to put in light that the pressure-wave speed of the material and the particle size seems to have the biggest influence on what is happening. The second study dealt with detection issues,

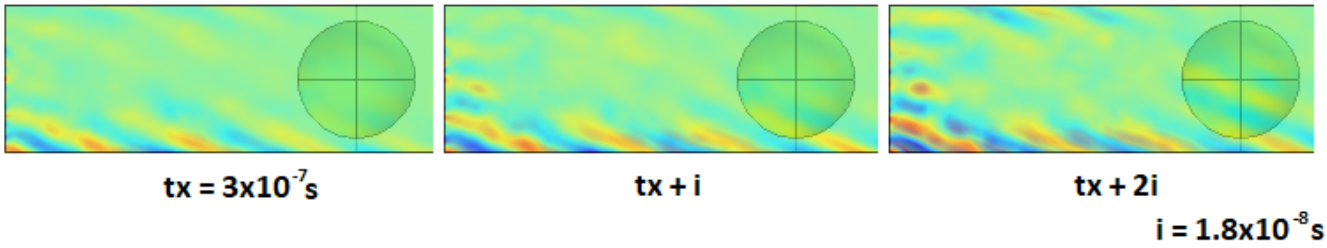


Fig. 15. The pressure acoustic field inside the microchannel for some iterations (50 μm Polystyrene, 50 MHz), *done with COMSOL Multiphysics*

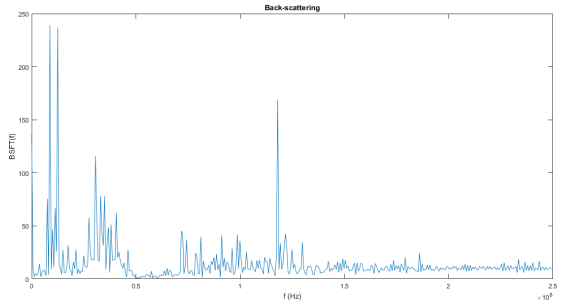


Fig. 16. The back-scattering technique applied to a 50 μm radius Polystyrene particle, *done with Matlab*

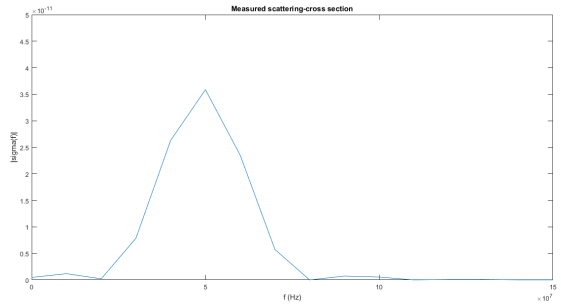


Fig. 17. The transmission technique applied to a 50 μm radius Polystyrene particle, *done with Matlab*

where I compared the back-scattering and the transmission technique, to quantify them. At this point, I am not sure about my results that have to be confirm but the transmission technique looks to be better. I also work on the noise limit that is acceptable, showing that until a noise of -18 dB, we can still use our data. My overall work allowed to study different parameters and see what suit the most to have exploitable results: The physical model I choose (the Faran model) helped me to find the optimal frequency (30-50 MHz); some tests I realized permits me to determine also the channel size (in order to avoid the appearance of droplets), the best detection technique, the noise limit (SNR = -16dB).

B. Future Work and Improvements

There are some points that can be improve: First, we should confirm the results from the studies by realizing new tests, with new parameters: try new size for the particles, test new materials. If the results are confirmed, we can compute the influence of the particle on what the sensor is recording from the IDT: the sensor should be able to detect the particle, with its precision (the sensor that equip the 'Lab-on-Chip' device from the lab has a 12 bits resolution). It should be computed from the results for the detection techniques and the sensor properties. When the simulations are giving results that are good enough, we can start to do tests in 'real condition' with the 'Lab-on-Chip' device, to confirm the results obtained using the COMSOL simulations, and the process we decide to follow. There are some choices that I did during my work that can be modify to enhance my results; the main one is the 'physical' model choice. The Faran model is the existing one that fit the most my situation but it presents a major default: in this model, we consider the particle as being surrounded by nothing (air), whereas in my setup, the particle is inside a PDMS channel, containing water. Optimally, we would need to design our own model to fit 100 % our situation, but it requires to spend a lot of time to develop it (get the Mathematical knowledge, understand fully the existing models, design and test the new model). The second point that can improve my result quality is to work with a true 3D model of the 'Lab-on-Chip' device (instead of the 2.5D model). It would allow to ensure independently the particle way of influencing what it is happening in the microchannel, without all the limitations of the 2.5D simulation model, and to perform new type of studies (shape of the IDTs, channel orientation)

REFERENCES

[1] S. H. M., "The evolution of cytometers," *Cytometry Part A*, vol. 58A, no. 1, pp. 13–20. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.10111

[2] G. Wróbel and S. Pawlak, "A comparison study of the pulse-echo and through-transmission ultrasonics in glass/epoxy composites," vol. 22, 06 2007.

[3] G. Simon, M. A. B. Andrade, J. Reboud, J. Marques-Hueso, M. P. Y. Desmulliez, J. M. Cooper, M. O. Riehle, and A. L. Bernassau, "Particle separation by phase modulated surface acoustic waves," *Biomicrofluidics*, vol. 11, no. 5, p. 054115, 2017. [Online]. Available: https://doi.org/10.1063/1.5001998

[4] S. Singh and A. Goyal, "The origin of echocardiography," *Tex Heart Inst J.*, vol. 34, no. 4, pp. 431–438, 2007.

[5] J. J. F. Jr., "Sound scattering by solid cylinders and spheres," *The Journal of the Acoustical Society of America*, vol. 23, no. 4, pp. 405–418, 1951. [Online]. Available: https://doi.org/10.1121/1.1906780

[6] K. A. Wear, "Ultrasonic scattering from cancellous bone: A review," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 55, no. 7, pp. 1432–1441, 07 2008.

[7] A. Mata, A. Fleischman, and S. Roy, "Characterization of polydimethylsiloxane (pdms) properties for biomedical micro/nanosystems." *Biomedical Microdevices*, vol. 7, no. 4, pp. 281–293, 12 2005.

[8] J. Nam, H. Lim, D. Kim, and S. Shin, "Separation of platelets from whole blood using standing surface acoustic waves in a microchannel," *Lab Chip*, vol. 11, pp. 3361–3364, 2011. [Online]. Available: http://dx.doi.org/10.1039/C1LC20346K

[9] Y. Chen, S. Li, Y. Gu, P. Li, X. Ding, L. Wang, J. P. McCoy, S. J. Levine, and T. J. Huang, "Continuous enrichment of low-abundance cell samples using standing surface acoustic waves (ssaw)," *Lab Chip*, vol. 14, pp. 924–930, 2014. [Online]. Available: http://dx.doi.org/10.1039/C3LC51001H

[10] D. B. Go, M. Z. Atashbar, Z. Ramshani, and H.-C. Chang, "Surface acoustic wave devices for chemical sensing and microfluidics: a review and perspective," *Anal. Methods*, vol. 9, pp. 4112–4134, 2017. [Online]. Available: http://dx.doi.org/10.1039/C7AY00690J

[11] L. Rayleigh, "On waves propagated along the plane surface of an elastic solid," *Proceedings of the London Mathematical Society*, vol. s1-17, no. 1, pp. 4–11, 1885. [Online]. Available: + http://dx.doi.org/10.1112/plms/s1-17.1.4

[12] C. Campbell, *Surface Acoustic Wave Devices and Their Signal Processing Applications*. Elsevier Science, 2012. [Online]. Available: https://books.google.co.uk/books?id=nBJex-_wQdoC

[13] A. Wixforth, C. Strobl, C. Gauer, A. Toegl, J. Scriba, and Z. Guttenberg, "Acoustic manipulation of small droplets," vol. 379, pp. 982–91, 09 2004.

[14] R. Kishor, F. Gao, S. Sreejith, X. Feng, Y. P. Seah, Z. Wang, M. C. Stuparu, T.-T. Lim, X. Chen, and Y. Zheng, "Photoacoustic induced surface acoustic wave sensor for concurrent opto-mechanical microfluidic sensing of dyes and plasmonic nanoparticles," *RSC Adv.*, vol. 6, pp. 50 238–50 244, 2016. [Online]. Available: http://dx.doi.org/10.1039/C6RA06849A

[15] R. E. Baddour, M. D. Sherar, J. W. Hunt, G. J. Czarnota, and M. C. Kolios, "High-frequency ultrasound scattering from microspheres and single cells," *The Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 934–943, 2005. [Online]. Available: https://doi.org/10.1121/1.1830668

[16] O. Falou, M. Rui, A. E. Kaffas, J. C. Kumaradas, and M. C. Kolios, "The measurement of ultrasound scattering from individual micron-sized objects and its application in single cell scattering," *The Journal of the Acoustical Society of America*, vol. 128, no. 2, pp. 894–902, 2010. [Online]. Available: https://doi.org/10.1121/1.3455795

[17] M. Urbanska, M. Winzi, K. Neumann, S. Abu Hattum, P. Rosendahl, P. Mller, A. Taubenberger, K. Anastassiadis, and J. Guck, "Single-cell mechanical phenotype is an intrinsic marker of reprogramming and differentiation along the mouse neural lineage," vol. 144, pp. 4313–4321, 12 2017.

OPHA*: planning feasible 3D paths for AUVs

Daniel González-Adell, Pedro Patrón, Juan David Hernández, Yvan Petillot

***Abstract*— This paper presents Orthogonal Planes Hybrid A* (OPHA*), a novel technique that expands the Hybrid A* algorithm to plan 3D paths in an orthogonal planes fashion. Such approach is targeted to autonomous underwater vehicles constrained by minimum turning radii (vertical and horizontal). Moreover, it seeks to reduce the steering differences between control architectures when deployed into different torpedo-shaped AUVs.**

The approach has been validated in two simulated underwater environments: one cluttered and known and a second one unknown by performing real-time re-planning. The method has been tested with three different control architectures from real vehicles inside SeeByte's Neptune © Autonomy Framework. Furthermore, OPHA* has been proven to perform in significantly less time when compared to a RRT-based approach operating with the same limitations.

Whilst developed for non-holonomic underwater vehicles, the approach could be extended or used in other domains.

I. INTRODUCTION

Three-dimensional (3D) path planning is not a recent problem as in multiple robotics applications (in platforms like autonomous underwater vehicles (AUVs), unmanned aerial vehicles (UAVs) or industrial manipulators) robots are forced to perform tasks in a three-dimensional space. Additionally, if the robot has motion constraints in one or more of its axis, such path needs to be drivable/feasible to be followed with accuracy. Additionally, the waypoint following technique can vary among different vehicles, and a new path planner should provide an easy way to deal with such constraint.

Not much research has addressed the problem of path planning for non-holonomic AUVs. Even less investigations have been carried out to take into account motion constraints of such vehicles in an unknown 3D environment.

Hernandez *et al.* combined two underwater vehicles to perform close-proximity surveys [4]. The same author presented an asymptotic optimal RRT (RRT*)-based motion planner for a torpedo-shaped AUV moving in a 2D workspace that used concepts of lazy-collision checking and anytime algorithms [5], and validated it in an online fashion in an unknown environment.

Fig. 1. Nessie VII, a torpedo-shaped AUV simulated to validate OPHA* in an unknown environment. Source: OSL, Heriot-Watt University.

Petres et al. presented the FM* method, which takes into account vehicle kinematics and ocean currents to plan a continuous path from a sampled environment. Moreover, Petillot *et al.* used sequential quadratic programming to plan a path from sonar images taken by forward-looking multibeam sonar mounted on an AUV, and proved the validity of such approach in simulation [9]. Nevertheless, the two last approaches did not validate their capability of performing real-time re-planning.

In the ground domain, Dolgov *et al.* used the Hybrid A* algorithm [12] in a semi-structured environment to obtain a drivable path over a lattice which was later improved via non-linear optimization. They verified their approach in a real vehicle in the Darpa Urban Challenge [1], [2].

In order to deal with underwater scenarios like the ones mentioned, this paper contributes to the state-of-the-art of motion planning by presenting a path planner for AUVs operating under motion constraints in 3D workspaces. The technique, named Orthogonal Planes Hybrid A* (OPHA*), aims as well to reduce tracking variations among different vehicles thanks to its orthogonal planes planning fashion. This approach has been integrated in SeeByte's Neptune Autonomy Framework [7].

II. ORTHOGONAL PLANES HYBRID A*

This section exposes Hybrid A* and the proposed approach OPHA* as an evolution of the first. Its node generation equations are detailed.

A. 2D planning with Hybrid A*

The Hybrid A* algorithm is a variant of the A* algorithm that takes into account continuous coordinates inside a discrete lattice over the workspace. Each

This work was supported by SeeByte Ltd.

Daniel González-Adell is with SeeByte Ltd., Edinburgh, Scotland. daniel.gonzalez@seebyte.com

Pedro Patrón is with SeeByte Ltd., Edinburgh, Scotland. pedro.patron@seebyte.com

Juan David Hernández is with the Department of Computer Science, Rice University, Houston, TX, USA. juandhv@gmail.com

Yvan Petillot is with the Institute of Sensors, Signals and Systems, Heriot-Watt University of Edinburgh, Edinburgh, Scotland. y.r.petillot@hw.ac.uk

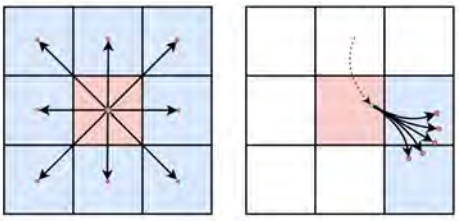


Fig. 2. Left: discrete search in A*, which allows to travel between cells' centers. Right: continuous search in Hybrid A*, which considers minimum turning radius and orientation of each node while expanding it.

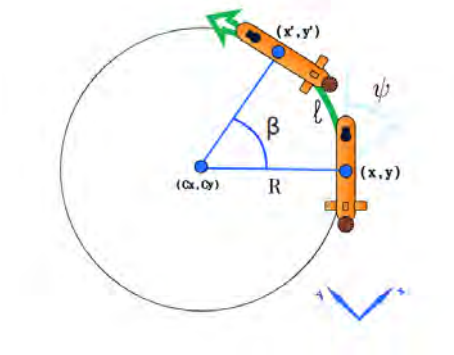


Fig. 3. Hybrid A* child node generation for a turning radius R on a left turn maneuver with a travelled distance l. The green arrow under the vehicle indicates the curve of expansion. It is assumed that vehicle steers around its gravity center.

node configuration is composed of continuous position (x, y) in the 2D plane and orientation ψ with respect to the world frame. The goal search relies on expanding at every iteration the node with minimum combined cost (that is, its path cost plus the heuristic distance to the goal node). In this expansion, child nodes are generated at the end of a circular motion with a predefined set of turning radius and at a given travelled distance. These are multiples of the minimum turning radius of the vehicle. This is graphically shown in figure 2.

During that exploration stage, for each of the nodes to be expanded and given a minimum turning radius R_{min} to be satisfied, the computation of each of the child states of a parent state (i.e. its expansion, shown in figure 3) follows:

1) The location of the turning center is computed as:
$$C_x = x + R * S * sin(\psi) \quad (1)$$
$$C_y = y - R * S * cos(\psi) \quad (2)$$
where $[C_x, C_y]$ is the location in the plane of movement of the rotation center for a particular turning radius R = α*R_{min} (multiple of R_{min}) and $S \in \{1, -1\}$ denotes the side towards which the vehicle turns in such plane. It is assumed that the vehicle steers from its gravity center.
2) The computation of the corresponding steering angle β at the expanded child node with respect to the parent node position is computed as:
$$\beta = l/R \quad (3)$$
where l is the travelled distance from parent to child node defined.
3) The state of the child node for the previous turning radius is computed as:
$$x' = C_x - R * S * sin(\psi - (S * \beta)) \quad (4)$$
$$y' = C_y + R * S * cos(\psi - (S * \beta)) \quad (5)$$
$$\psi' = (\psi - (S * \beta)) mod(2\pi) \quad (6)$$

This procedure is used for the rest of turning radii, where each of the values is a multiple α*R_{min} of the minimum turning radius R_{min} given by the AUV specifications.

In the particular case that the heading angle β is lower than a predefined threshold th, the trajectory from the current node to the child is assumed to be a straight line with the same heading. Then, the child state is computed as:
$$x' = x + l * cos(\psi) \quad (7)$$
$$y' = y + l * sin(\psi) \quad (8)$$
$$\psi' = \psi \ mod(2\pi) \quad (9)$$

Every child node generated in a collision-free location of the map is added to an *Open List* O. Once all the child nodes have been generated from a parent node, the node with minimum combined cost in the list O becomes the one to be expanded in the next iteration.

This pipeline is repeated until the goal node is expanded. At such moment the path is computed back by a reverse search in the list of all nodes previously expanded. Such a list, also known as *Closed List*, containes each node expanded and its parent node.

B. 3D planning with OPHA*

OPHA* (whose algorithm is shown in algorithm 1) is an extension of the Hybrid A* algorithm to plan three-dimensional (3D) paths by taking into account minimum turning radii R_{Hmin} (horizontal) and R_{Vmin} (vertical) as well as the maximum operational pitch of the vehicle θ_{max}. This approach uses two orthogonal planes: the first one is parallel to the sea surface (if AUV horizontal) and is centered at the robot's position; the second plane is orthogonal to the first one at that position. Both planes are tilted by the node's orientation. New states are generated analogously to the Hybrid A* method detailed, but in both planes. The equations that describe the constraints associated to the motion in Y axis is now used for the vertical motion (Z axis). Figure 4 shows the generation of states in both planes.

When calculating a path for a vehicle that moves in a 3D workspace (e.g. an underwater environment), the

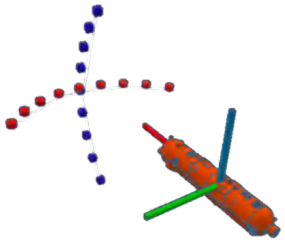


Fig. 4. OPHA* expansions in XY and XZ orthogonal planes (red and blue nodes respectively) at a configuration node tilted by its orientation. Each node corresponds to a multiple of the minimum turning radius for each plane. AUV in orange.

way a dynamic controller has to achieve a waypoint can slightly vary among vehicles. In other words, assuming that the circular trajectory that the AUV will follow will be the desired one can lead to collisions. Moreover, these motion controllers typically reside in the manufacturer's main vehicle computer (MVC) [3] (i.e. the *frontseat*) and can rarely be accessed. For that reason, the dimensionality of the workspace over which the vehicle moves is reduced to the two dimensions of one of the orthogonal planes. This is expected to lead to a more accurate tracking of the path defined. Also, such assumption allows to use the planner in more than one vehicle with different *frontseat* interfaces.

Additionally, an extra constraint is set: in order to prevent implicit roll ϕ motion, performing a yaw motion at a given state is only allowed when the pitch on the AUV at that particular state is lower than a threshold θ_{th_ψ}.

Then, the added planning constraints on the yaw and pitch motions are of the form:

$$\theta_{q_i} \leq \theta_{max} \parallel \theta_{max} \geq 0 \quad \forall q_i \in Q \tag{10}$$

$$\psi_{q_i} \neq 0 \iff \theta_{q_i} \leq \theta_{th\_\psi} \parallel \theta_{th\_\psi} \geq 0 \quad \forall q_i \in Q \tag{11}$$

where q_i is any state generated during the planning phase and Q is the set of accepted states.

III. FRAMEWORK FOR ONLINE RE-PLANNING

A. General architecture

The behavior of the joint architecture (visually shown in figure 5) between the path planner proposed and SeeByte's Neptune Autonomy framework is described as the following key points:

- A Neptune behavior sets an initial configuration q_{start} and a goal location (without orientation) q_{goal}.
- The motion planner module receives both configurations and plans an OPHA* path taking into account R_{Hmin}, R_{Vmin}, θ_{max}, θ_{th_ψ} and the updated octomap of the environment. The goal area is defined as a sphere of radius th_{goal}.

Algorithm 1 OPHA* algorithm

$n_s \leftarrow (q_{s,c}, \theta_{s,c}, \psi_{s,c}, q_s, 0, h(q_s, G), -)$ ▷ Starting node
$UNVISITED \leftarrow PriorityQueue()$
$UNVISITED.insert(n_s)$
$VISITED \leftarrow 0$
$nodeToExpand \leftarrow n_s$
while $nodeToExpand \neq G$ **do**
  $nodeToExpand \leftarrow UNVISITED.pop()$
  $VISITED.append(nodeToExpand)$
  $isYawAllowed \leftarrow (nodeToExpand.\theta_{s,c}) < \theta_{th\_\psi}$
  $D \leftarrow computeHeadings(isYawAllowed)$
  for all $\delta \in$ D **do**
    n$\leftarrow nodeToExpand.neighbor(\mu(\delta))$ ▷ $\mu(\delta)$: motion primitive for angle δ
    if $|$ n.$\theta_{s,c}| > \theta_{max}$ **then**
      $break$
    if n$\in C_{free}$ *and* n$\notin VISITED$ **then**
      $compared = False$
      for all $node \in UNVISITED$ **do**
        if n== $UNVISITED.node$ **then**
          $compared = True$
          if $UNVISITED.node.g >$ n.g **then**
            $UNVISITED.node \leftarrow$ n
          $break$
      if !$compared$ **then**
        $UNVISITED.insert($ n$)$
$path \leftarrow ComputePath(VISITED)$

- The mission handler receives the full waypoints list, establishes a connection to the vehicle interface (operated by Neptune) and iteratively sends the next waypoint to it as soon as the waypoint achievement signal is received from the same module.
- Inside the vehicle interface, the dynamic controller of the vehicle (interfaced as a waypoint follower controller) steers the AUV towards the waypoint and notifies the mission handler whenever the waypoint has been achieved in order to get the next one.
- The vehicle maps the environment with the equipped rotating multibeam sonar as it advances through the environment.
- In case the path followed is discovered to be occluded at any location, the re-planner module (which performs a continuous path collision-checking) immediately sets a new planning query from the next waypoint from the old path (which the AUV still has not reached) to the goal.

B. Mapper and simulated environment

A UWSim [10] simulated device (a vertically mounted multi-beam sonar) produces the sonar data

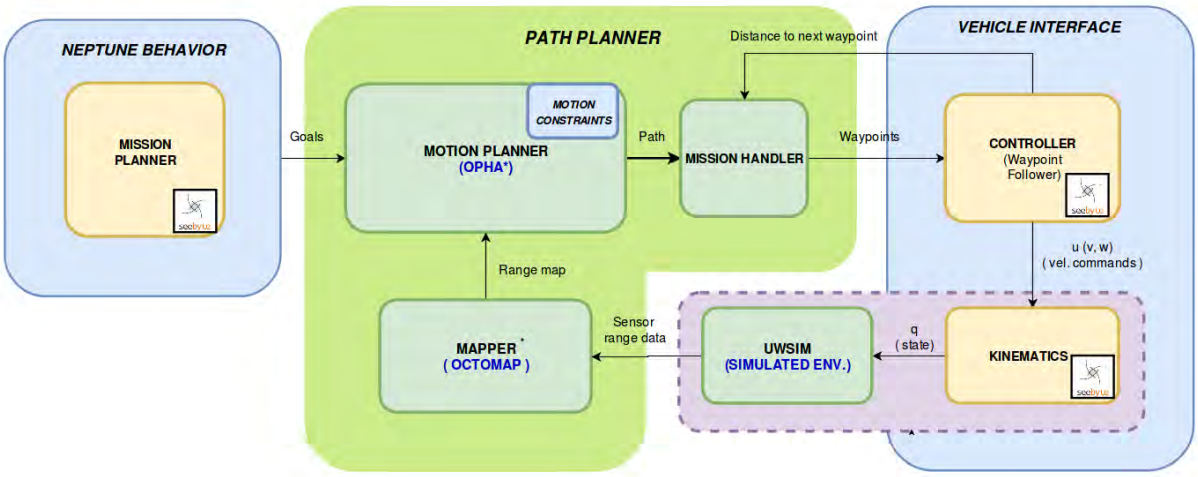


Fig. 5. Architecture designed for the online (re-planning) implementation integrated in SeeByte's Neptune Autonomy Framework.

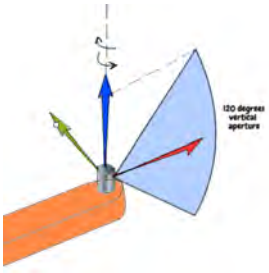


Fig. 6. Sonar beam simulated configuration with aperture of 120 degrees (no longer horizontal but vertical aperture as the sensor is mounted on the vehicle rotated 90 degrees). Sonar joint allows rotation around the sonar's new vertical axis (blue arrow).

over a *.3ds mesh* simulated environment. This sensor has, in the position placed, a vertical aperture of 120 degrees. Additionally, the joint that holds the sonar allows a rotation of 30 degrees to each side, so the sensor continuously sweeps the environment with acoustic waves inside this rotation range. This configuration is shown in figure 6.

C. Collision-checking

A new configuration is accepted as valid if and only if all the intermediate points from the parent node to itself (included) are collision-free. Also, given the vehicle is not understood as a point in the 3D space but as an object of concrete size, the collision-checking must be performed for the whole volume of the AUV and for its orientation at each of the intermediate points. Figure 7 illustrates this.

This collision checking has been achieved by using the flexible collision library (FCL) [8] and *Octomap* [6] library altogether inside the robot operating system (ROS) [11] framework. The vehicle has been modeled as a square prism in all tests.

D. Online re-planning

The re-planning sub-module performs a collision-checking for all the main and intermediate waypoints

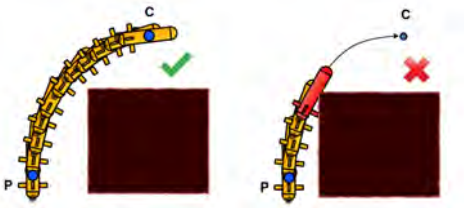


Fig. 7. Examples of accepted (left, collision-free) and discarded (right, non-collision-free) expansions from parent (P) node to child (C) node for a given turning radius. Each position of the AUV corresponds to the intermediate point that is checked for collision.

Fig. 8. Left: path before discovering occlusion (path as black curve). Right: re-planning a new path after discovering the occlusion (new path as blue curve and starting node for re-planning as red circle). Borders of object mapped as purple lines.

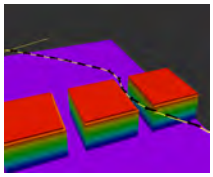
of the path with the same collision-checking approach explained. In the case of a newly discovered path occlusion at any of the checked points, the module triggers a *pause* signal that stops the waypoint-feeding from the mission handler to the vehicle interface. The re-planner then sends a new planning query from the next waypoint to achieve of the old path to the defined goal. This prevents the AUV from passing over the new query's start node while the re-planning is still being computed.

IV. RESULTS

A. Comparison with RRT-based approach

The approached designed has been compared with Orthogonal Planes RRT (OPRRT), a variant of

Fig. 9. LAUV and REMUS 600 AUVs. Sources: OceanScan and Hydroid.



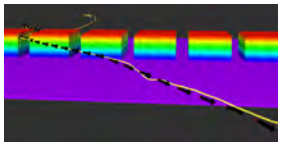


Fig. 10. An example of an offline computed OPHA* path successfully followed by the Remus 600 dynamic controller (trajectory followed as yellow arrows).

the rapidly-exploring random tree (RRT) with non-holonomic constraints method that performs in the orthogonal planes fashion introduced in this work. Such a comparison is done on the base of computational time required and states expanded for each of the planners and for the following set of parameters:

In the case of OPRRT being a stochastic method, 10 goal-biased path computations sampling with a probability of 5% the goal node have been performed and averaged into a final result.

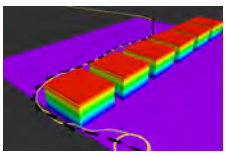
From the results shown it can be extracted that the RRT methodology sees its capacity of rapidly finding a path by sampling the environment severely damaged when restricted to expanding its nodes in two orthogonal planes and with the additional constraints exposed in section II-B, instead of in the whole 3D space. Moreover, it is clear that the OPHA* planner finds a 3D path in such conditions expanding many less nodes than the sampling-based technique. The computational time required for the grid-based method is significantly lower as well.

B. *Offline planning and trajectory tracking*

The control architectures from three different AUVs (figures 1 and 9) inside SeeByte's Neptune Autonomy Framework have been used to successfully follow offline-planned paths in non-cluttered (figures 10 and 11) and cluttered environments (figure 12).

C. *Online planning*

The re-planning procedure has performed as expected and all obstacles encountered during the survey shown in figure 13 have been avoided. An example of such an avoidance is shown in figure 14, where a re-planning sequence is performed to avoid the new obstacle encountered. In that situation, false sonar measurements have caused unnecessary re-plannings. Further sonar filtering techniques could be used to reduce their frequency of occurrence. The AUV Nessie VII has been used in the online tracking of the path.



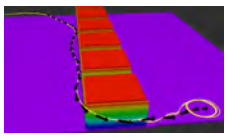
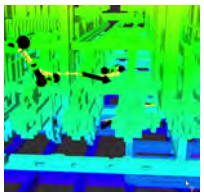


Fig. 11. An example of an offline computed OPHA* path successfully followed by the LAUV dynamic controller (trajectory followed as yellow arrows). AUV loitering around final waypoint as the path gets close to its end.



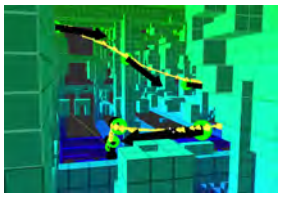


Fig. 12. An example of an OPHA* path and trajectory successfully followed (as yellow arrows) in the cluttered oil and gas structure by Nessie VII dynamic controller. Left: entering the structure. Right: spiraling down to get under the structure.

V. CONCLUSIONS AND FUTURE WORK

A. *Conclusions*

The purpose of the approach presented is twofold: first to allow to perform path planning in a 3D workspace explicitly taking into account non-holonomic constraints; second to limit the motion controller to perform vertical or horizontal curves instead of curves in three dimensions when steering towards a waypoint, allowing the planner to be deployable in different vehicles.

Real dynamic controllers from three different torpedo-shaped vehicles have been effectively interfaced to follow offline OPHA* paths computed thanks to the combination with SeeByte's Neptune Autonomy Framework. In particular, the controller from Nessie VII AUV has been used to explore an unknown environment by performing a survey in an online fashion, i.e. autonomously performing 3D obstacle avoidance while moving towards the goals set.

Fig. 13. The result of a survey close to the simulated seabed with the AUV Nessie VII. Online re-planning performed in order to avoid the structures encountered. Sonar false measurements appear above of the mapped structures (on the right side). Survey goals as blue circles and trajectory followed in red (on the left side).

OPRRT Iteration	Path Length (m)	Time (s)	States expanded
1	140	9.8487	1340
2	228	9.5439	1226
3	117	1.5912	285
4	123	2.3125	370
5	198	10.5902	1207
6	117	1.1213	246
7	120	0.8053	198
8	117	1.3072	292
9	117	1.1447	297
10	201	4.8929	717
OPRRT Average	147.8	4.3158	618

	Path Length (m)	Time (s)	States Expanded
OPHA*	105	0.3061	53

TABLE I

OPHA* AND OPRRT PERFORMANCE FOR THE SAME QUERY IN AN UNCLUTTERED ENVIRONMENT. MEASURED ON AN ACER ASPIRE 5755G I7-2630QM 8GB. USING WEIGHTED EUCLIDEAN AS HEURISTIC IN OPHA*. OPHA* OUTPERFORMS NOT ONLY THE AVERAGE, BUT ALL COMPUTATIONS OF THE RRT-BASED METHOD CONSTRAINED TO THE LIMITATIONS EXPLAINED IN SECTION II-B.

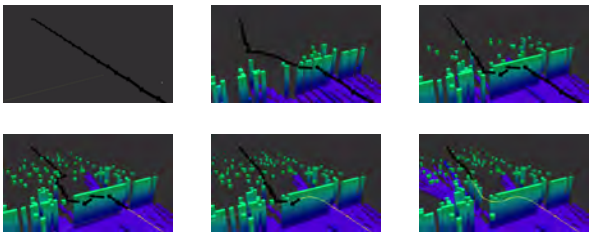


Fig. 14. From left to right and from top to bottom: five online re-planning procedures with OPHA*. Bottom-left re-planning is triggered by the sonar false measurements (as sparser voxels). Re-planned paths as black arrows and trajectory followed in yellow. The obstacle is successfully avoided.

A comparison between the OPHA* algorithm and OPRRT, a 3D RRT algorithm under the same constraints, i.e. the absolute pitch limitation and the allowance of performing a horizontal turn based on the current state's pitch, has been carried out. This has proved the advantage in speed that the first method has over the second in such restricted situations by achieving planning rates that typically oscillate between 1 and 20 Hz even in the simulated cluttered environment of an *oil and gas* sub-sea structure.

B. Future work

Testing the approach in a real-world scenario would be desirable, as ocean currents have not been taken into account in the planning procedure.

REFERENCES

[1] Dmitri Dolgov, Sebastian Thrun, Michael Montemerlo, and James Diebel. Practical search techniques in path planning for autonomous driving. *Ann Arbor*, 1001(48105):18–80, 2008.

[2] Dmitri Dolgov, Sebastian Thrun, Michael Montemerlo, and James Diebel. Path planning for autonomous vehicles in unknown semi-structured environments. *The International Journal of Robotics Research*, 29(5):485–501, 2010.

[3] Donald P Eickstedt and Scott R Sideleau. The backseat control architecture for autonomous robotic vehicles: A case study with the iver2 auv. *Marine technology society journal*, 44(4):42–54, 2010.

[4] Juan D Hernández, Guillem Vallicrosa, Eduard Vidal, Èric Pairet, Marc Carreras, and Pere Ridao. On-line 3d path planning for close-proximity surveying with auvs. *IFAC-PapersOnLine*, 48(2):50–55, 2015.

[5] Juan David Hernández, Eduard Vidal, Guillem Vallicrosa, Enric Galceran, and Marc Carreras. Online path planning for autonomous underwater vehicles in unknown environments. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 1152–1157. IEEE, 2015.

[6] Armin Hornung, Kai M. Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots*, 2013. Software available at http://octomap.github.com.

[7] SeeByte Ltd. SeeByte Neptune: Payload control architecture goal based mission planning and real-time autonomy engine for unmanned maritime systems to plan and execute well known patterns of behavior. URL: http://www.seebyte.com/oceanography/neptune.

[8] Jia Pan, Sachin Chitta, and Dinesh Manocha. Fcl: A general purpose library for collision and proximity queries. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3859–3866. IEEE, 2012.

[9] Yvan Petillot, I Tena Ruiz, and David M Lane. Underwater vehicle obstacle avoidance and path planning using a multi-beam forward looking sonar. *IEEE journal of oceanic engineering*, 26(2):240–251, 2001.

[10] Mario Prats, Javier Pérez, J Javier Fernández, and Pedro J Sanz. An open source tool for simulation and supervision of underwater intervention missions. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 2577–2582. IEEE, 2012.

[11] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009.

[12] Nathan Richards, Manu Sharma, and David Ward. A hybrid a*/automaton approach to on-line path planning with obstacle avoidance. In *AIAA 1st Intelligent Systems Technical Conference*, page 6229, 2004.

Visual Place Recognition through Reading Scene Texts

Ziyang Hong and Sen Wang

***Abstract*— Visual place recognition is a fundamental problem in mobile robot navigation application. Sparse feature based visual place recognition has been dominant in the research community for long time, however it becomes fragile when it try to deal with extreme perceptual changes. In contrast to traditional feature based methods, we tackle the visual place recognition from a novel perspective. When describing a place or a scene, we use scene text as our primitive, scene text is invariant to illumination changes and discrimnative. Our proposed system takes the spatial coherence between scene text landmarks into account. Our experiment result shows that the proposed system outperform traditional vision based place recognition methods on the dataset we have collected.**

I. INTRODUCTION

A visual place recognition system comprises three parts: image processing for the vision data (usually this would be image), the map and the belief generation. The belief generation part will take information from the vision data or both vision data and the motion data. It is responsible for updating the map and making the place recognition decision based on the vision input and the map which has been initialized. Most of the visual place recognition algorithms are not consistent and robust against extremely perceptual changes. Before recognizing a place, the system needs to build or incrementally build a map. There are three types map representations generally used in place recognition: pure image retrieval, topological map and topological-metric map. In reality, for a place recognition system, avoiding false positive matches is the first priority since detecting an incorrect loop closing will be destructive to the robot mapping and localization system. Apart from that, a place recognition system can also be evaluated base on how many matches are true positive within a certain distance.

The definition of place recognition defined in [17] is simple and straightforward; by giving an image that representing a place, can the robot tell whether this place has been visited or this is a new place. A human is very good at doing such a task and is able to identify a place even with the environment completely changed, for example, identifying a place in sunny day or dark night, and in summer or winter. But for a robot, it is very challenging to perform visual place recognition due to the appearance is changed drastically (see Figure 1), and what's more, different places might have similar appearance and visual features so called perceptual aliasing. Apart from that, when the robot revisits the same place, it could observe the scene from different view-point, in which the robot only receive partial information of the visited place.

Z. Hong and S. Wang are with Institute of Sensors, Signals, and Systems, Heriot-Watt University, Edinburgh, UK zh9@hw.ac.uk

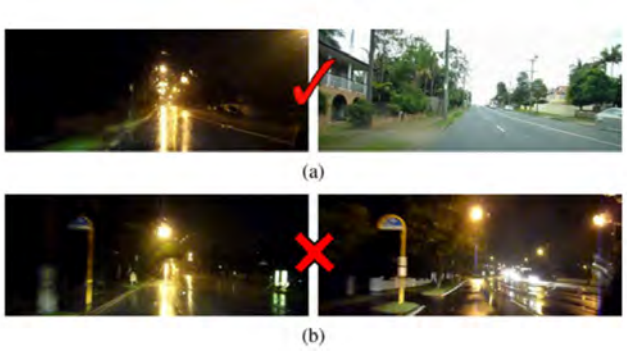


Fig. 1. Visual place recognition system needs to be able to deal with perceptual aliasing as well as perceptual changes due to the lighting and weather. [17]

II. RELATED WORK

Vision based place recognition algorithms are the major trend in the research community due to its nature of commonality and versatility, a low cost commodity RGB camera will satisfy the need of the vision based place recognition system. This chapter will mainly focus on vision based methods. Apart from visual place recognition, another category of place recognition algorithms are using range sensor like LiDAR and to build a global descriptor in a scene. The following content will mainly focus on vision based place recognition algorithms.

A. Vision based Place Recognition

Compared to 3D ranging sensor based place recognition methods, vision based methods have been explored by the research community for a long time since the evolution of camera. However this remains very challenging for an outdoor large-scale navigation. In large-scale outdoor environment , instead of using an accurate metric map, a topological map is more favorable in the context of place recognition.

1) Sparse Feature based Place Recognition: The development of local visual features like SIFT [16], SURF[3] and ORB[21] has provided accessibility to solve problem like structure from motion, camera motion estimation and image mosaic. Sparse feature based place recognition algorithms have shown great success and become dominant compared to other methods. In 2005, Newman and Ho [19] proposed a saliency-MSER-descriptor pipeline using SIFT to look for wide-baseline images matching. After that in 2008, inspired by the bag-of-words model for scene retrieval in [24], Cummins and Newman proposed a probabilistic localization and mapping model based on visual appearance which is named

as FAB-MAP[5]. In FAB-MAP, instead of trying to detect loop closure in a metric map, they adopt the topological map representation. In [1], Sunderhauf and Protze made use of BRIEF [4] descriptor and proposed a fast version of it named BRIEF-Gist. Hamming distance between two BRIEF-Gist descriptors is calculated for measuring the similarity of two images. Similar to FAB-MAP, [6] used bag of words model combining ORB features for fast place recognition. This method is the first one which makes use of binary descriptor, being one order of magnitude faster than all the approaches prior to this. To cope with large-scale and long-term navigation under the online processing time constrain, [10] split the most frequently and last seen keyframes in the working memory for a loop closure also based on bag-of-words approach . Different from FAB-MAP which counts the visual words in each images, Vector of Locally Aggregated Descriptors (VLAD) [9] computes the sum of the residuals between each visual word and the corresponding clustering center. In [1], a fast and incremental pipeline is implemented which relies on Bayesian filtering as a extension of bag-of-words method.

2) Sequence based Place Recognition: Instead of using sparse features and bag-of-words model, Milford and Wyeth [18] used the whole image as the descriptor for place recognition. It was the first attempt that tried to address the problem raised by extreme perceptual changes in the environment by exploiting the spatial and temporal information between frames. In SeqSLAM, it computes the intensity contrast vector between the query image and the template images, looking for the closest matching.

3) End To End Deep Learning based Place Recognition: In NetVLAD[2], a set of the local descriptors of the single image is learned by a convolutional neural network, the compact form of the local descriptors are computed similar to VLAD[9]. The query image is matched by searching the closest one in the images database. In [7], during training three-stream siamese network is deployed, both relevant and query image are used to compute the loss, additionally an irrelevant image is used to reinforce the network to produce similar representations for the relevant images. Whereas in [20], 3D models are constructed for each cluster of images and are used to guide for the image retrieval.

B. Scene text detection and recognition

Scene text detection and text recognition has a wide range of applications which will also facilitate robotics application. Inspired by SSD[14], [13] formulates the text region detection similar to object detection task, employing anchor boxes on multi-scales. Region proposal text detector like [8] directly gives the region where has highest probability on where the text appears. Both [12] and [27] are able to deal with oriented scene text. In [26], a set of multi-scale mid-level primitives are learned without using deep learning techniques. CRNN[22] is an end-to-end text recognition framework which predicts a sequence of characters by giving a crop of image in which text appears. FOTS [15] is unified text detection and text recognition framework which shares the features both in text detection and text recognition achieving 22.6 fps on dataset ICDAR 2015.

C. Text based Place Recognition and Localization

With deep learning techniques, text detection and text recognition become very robust and accurate. However text information has not been widely used for robotics application such as robot navigation or place recognition. The first attempt of putting text information in robot navigation is [25] where a conjunction text feature is used to encode text information, text is used as landmark for loop closure detection.

III. Text based place recognition system

The key ingredient in the proposed system is scene text in urban environment. The system contains two major part: the mapping stage and the place recognition stage, these two stages are processed separately. As shown in Figure 2, the input of the system are RGB images captured by a standard camera. Each frame will go through a deep learning network for detecting the text and recognizing text at each locations in the frame. A dictionary based text filtering will be applied to remove some false positive text detection and incorrect recognition of text both in training and testing stage. After that, each frame is represented by a histogram, each entity of the histogram is corresponding to a word which can be found in the predefined dictionary. The mapping module is responsible to initialize and update the map based on the forthcoming word histogram. In the place recognition stage, the text filtering part is the same as mapping stage, and then the place recognition system will perform a search in the map to find the image pair match. The detail information of the system will be explained in the following subsections.

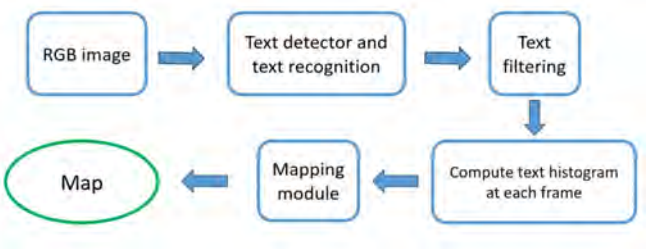


Fig. 2. Mapping stage

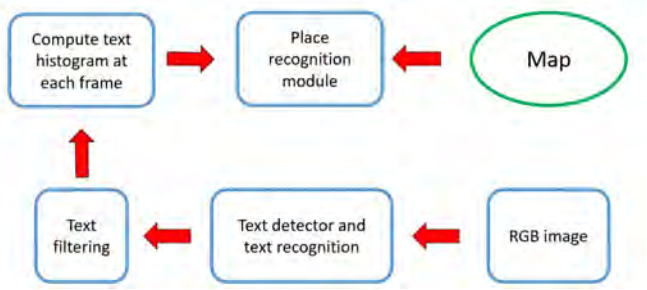


Fig. 3. Place recognition stage

A. Text Detection And Text Recognition

To spot the scene text in the wild, we select TextBoxes++ [12] a deep learning text detector to extract the bounding box of each word. TextBoxes++ is able to predict multiple words that appear in the image, what's more, the bounding boxes are in the form of quadrilaterals, which means the network is able detect arbitrary-oriented text from different view-points. Similar to SSD, they adopt the VGG-16 [23] architecture as the backbone, connecting 6 text prediction layers. At the end they applied a non-maximum suppression procedure.

After receiving the bounding boxes prediction from TextBoxes++, we extract the bounding boxes and crop them, then use another deep learning network named CRNN for text recognition proposed in [22]. Given an image filled with a word, CRNN will yield a sequence of characters following from left to right order.

B. Predefined Dictionary

Once we obtain text detection and text recognition across the whole image, we will apply text filtering to get rid of some false positive text detection and some meaningless string output from text recognition. But before applying text filtering, we need to predefined a dictionary. We build a dictionary containing all the shops' name and the words appear on the shops' sign around an area where we want to conduct our navigation. Why we do that is because those words appear on a shop sign will not be occluded or removed across daytime and night time. For example, if we are going to navigate in Princess Street at Edinburgh, we will manually put words like "starbucks", "coffee", "princes" and "street" into the dictionary.

C. Text Filtering

To filter the text recognition result in each frame, we measure the Levenshtein distance [11] between the predicted string and all the words in the predefined dictionary. Levenshtein distance is a string metric in information theory and linguistics. By definition, Levenshtein distance is the minimum operations including deletions, insertions and substitution to take for correcting string A to string B. For example, the Levenshtein distance between the string "sitting" and string "kitten" is 3. The following operations are operated to convert from "kitten" to "sitting":

1) replacing "k" with "s"
2) replacing "e" with "i"
3) inserting "g" after "n"

Even though TextBoxes++ is able to detect text bounding boxes from different view-point, the detection is not perfect so that sometimes the bounding box does not cover the whole area of the text region. Therefore the text recognition result is not complete and we are not able to find a perfect match in the dictionary. By introducing the measurement of Levenshtein distance, we will be able to filter those incorrect recognition and retrieve a complete string from partial observation of a string. We set up a threshold for rejecting the text recognition if the minimum Levenshtein distance across the whole dictionary is larger than the threshold distance. For example, if we set up the threshold distance to be 2, and we observe a text recognition being the string "starbuc" which is not a complete word in terms of the predefined dictionary, but since the distance is only 2, still no larger than the threshold, so we will eventually retrieve the string "starbucks". But for a string of recognition like "cofta", it will be discarded since the Levenshtein distance between "coffee" and "cofta" is 3.

D. Compute Word Histogram at Each Frame

After filtering the text information, we will build a word histogram for each frame. The number of bins of the histogram depends on how many words does the dictionary contain. Each bin represent a word, the occurrence for each word in each frame can only be 1 or 0 which means whether we detect a word from the dictionary or not in this frame. If the same word occur in multiple positions in a frame, we will only count 1.

E. Topological Mapping of Text Landmarks

In this project, we aim to deploy place recognition in a urban complex environment so that we argue that a topological mapping without metric would be appropriate and realistic for this goal. As mentioned above, each frame will be represented by a histogram of words coming from the dictionary. In the topological map(see Figure 4), each node contains an image and a histogram containing at least one word from the dictionary. If we do not detect a word in the frame we are checking, this frame will not be inserted into the map. Each frame is corresponding to a place, but we are allowing a place being represented by multiple consecutive frames. By storing all these consecutive frames and histograms in the map, we are able to preserve spatial-temporal information between all the text landmarks along the navigation.

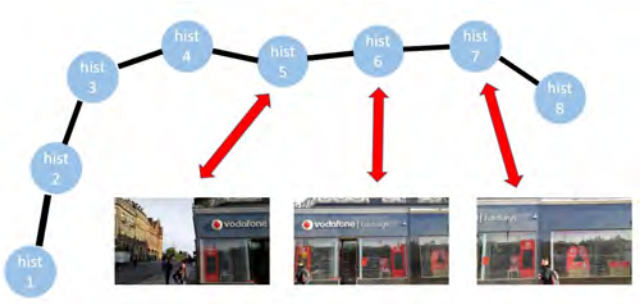


Fig. 4. Topological map of places

During the mapping stage, we might receive many consecutive frames which contain the same histograms, to avoid storing redundant frames, we need to set up a local frame buffering scheme instead of directly inserting a new frame(with text recognized) into the map. The global map will be incrementally built as we navigate through a route. The frame which has no text recognition will not be added to the global map.

F. Place Topological Matching

After building a topological map, the second part of the system is place recognition, finding image matching pairs. The place recognition module consists of two key components: the first one is called Global Searching, the second component is called Local Best Match. Similar to

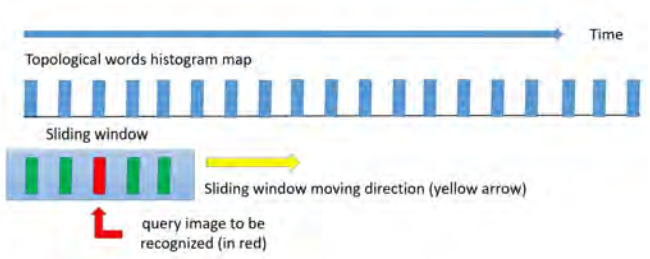


Fig. 5. Global searching scheme

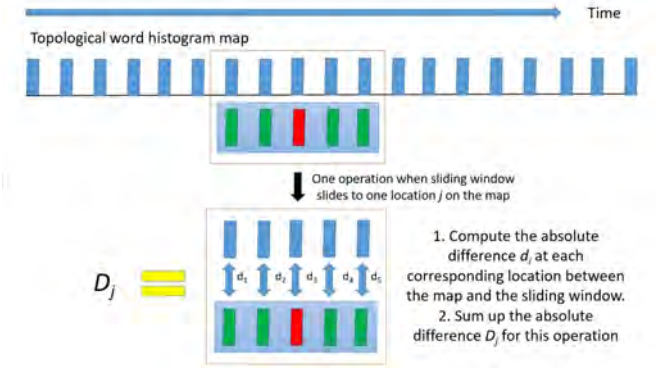


Fig. 6. One operation when the sliding window moves to location j

the mapping stage, each frame streaming from the camera will be applied text detection and text recognition, then the text filtering by comparing the recognized text string to the dictionary.

1) Global Searching: To recap what is mentioned above, the global map is a topological map, each node is consisting of a histogram and an image, nodes are connected without metric information but in a temporal order. Inspired by SeqSLAM [18], to exploit the spatial-temporal information between the text landmarks, we perform a global searching across the whole global map. A sliding window is created during run-time for place recognition testing, this sliding window will slide across the topological word histogram map at each test. The size of the sliding window is a user defined parameter which can be changed depends on the dataset. For each image that we want to conduct place recognition, we will push it to the center of the sliding window, so that its left and right neighborhoods will be the preceding words histogram and successive words histogram in a temporal sequential order. The assumption we made here is that during mapping and testing time, the spatial-temporal relation between the text landmark will not change according to lighting condition. So that we can find a close matching between the sliding window and the global map.

Let's denote the sliding window size to be w, M to be the number of nodes in global map, j to be the position where the center of the sliding window is pointing to, D_j to be sum of differences for the operation when the sliding window moves to the position of j. d_i is the absolute difference between the histogram h_i of the sliding window and its corresponding histogram H_k of the global map in one operation.

$$d_i = \mid h_i - H_k \mid \quad (1 \leqslant i \leqslant w) \tag{1}$$

where $k = j - (w-1)/2 + i - 1$, w is the size of sliding window.

$$D_j = \sum_{i=1}^{w} d_i \quad (1 \leqslant i \leqslant w) \tag{2}$$

After searching across the whole global map, we track down the position which provides the minimum sum of differences D_{min}.

$$D_{min} = \min_j D_j \quad (w-1)/2+1 \leqslant j \leqslant M-(w-1)/2+1) \tag{3}$$

where M is the number of nodes in the global map.

Once we find D_{min} in the global searching step, it means global searching has shrunk the search area for us and the correct matching will be within that area. Then we want to find the best match locally by simply comparing the histogram of the query image to all the candidate histograms within that area. We set up a threshold and calculate the minimum difference between the candidates, if the minimum difference is not less than the threshold, we consider the not finding a match between the query image and the global map. If the minimum is less than the threshold, it will be considered as finding a match.

2) Update The Sliding Window: Once a whole iteration of global searching and local best matching is finished, the front frame of the sliding window will be pop off, a new frame coming from the streaming will push back to the end of sliding window, again the center of the sliding window will be conducted with the same mechanism.

IV. Experiment and Result

We want to compare our proposed methodology with the state-of-the-art place recognition algorithms and evaluate the performance of our proposed method. Due to the nature of our proposed algorithm, we need to make use of scene text in a urban environment so that we collected a set of data between day time and night time at Edinburgh city center.

A. Evaluation Measurements

The standard performance measurement for place recognition is precision-recall curve. Precision is the portion of the true positive matches over the the sum of true positive and false positive. Recall is the portion of true positive matches over the sum of true positive and false negative.

The state of the art algorithms for large scale outdoor place recognition will be FAB-MAP and SeqSLAM. Here we will test them on all of the dataset we have collected.

B. Dataset Collection

The dataset collected for the experiment are only containing RGB images which are captured on an Android phone. We also equipped our phone on a camera stabilizer so that we can reduce the motion blur especially for night time imaging. The images are divided between day time and night time by going through the same routes. Some are captured with or without the stabilizer on a bus or by walking. The locations and the routes we choose for data collection are Princes Street , Edinburgh and South Bridge Street, Edinburgh where you could find many scene text along both the streets. We

collected 6 sequences in total, day_1 and day_2 are a pair, day_4 and night_4 are a pair, day_6 and night_6 are a pair.

1) Day Time Images Captured On A Bus Without A Stabilizer: In this set of images, they are captured without using the stabilizer on the bus during filming. The view point and the field of view is slightly different. Here we have day_1 set and day_2 captured in the morning and in the afternoon respectively.

2) Day Time Images Captured By Walking With A Stabilizer: In this set of images, they are captured with using a camera mounted on the stabilizer during day time while walking along the street. Here we name one set as day_4 which is captured at Princes Street, the other set as day_6 which is captured at South Bridge Street.

Fig. 7. day_4 images set

Fig. 8. day_6 images set

3) Night Time Images Capture By Walking With A Stabilizer: In this set of images, they are captured with using the stabilizer during night time and I was walking along the street. Here we name the one set as night_4 which is captured at Princes Street, the other set as night_6 which is captured at South Bridge Street.

Fig. 9. night_4 images set

C. Precision-recall curves comparison

Here we compare the proposed method with FAB-MAP and SeqSLAM by verifying the precision-recall curve.

Fig. 10. night_6 images set

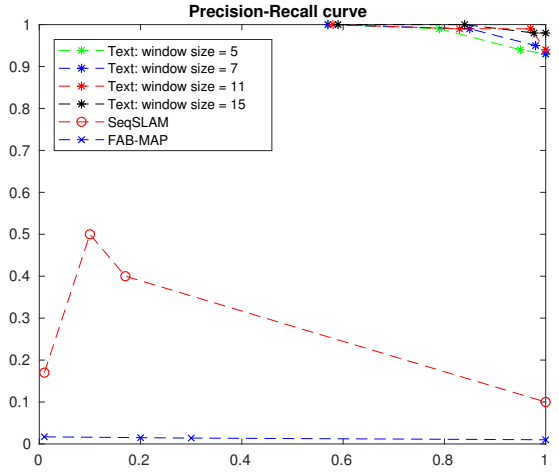


Fig. 11. Precision-Recall Comparison Between Three Methods on day_1 and day_2 image sequences

As we can in Figure 11, the performance of FAB-MAP is poor, SeqSLAM is better than FAB-MAP. Our method performs the best on the test between day_1 and day_2 dataset.

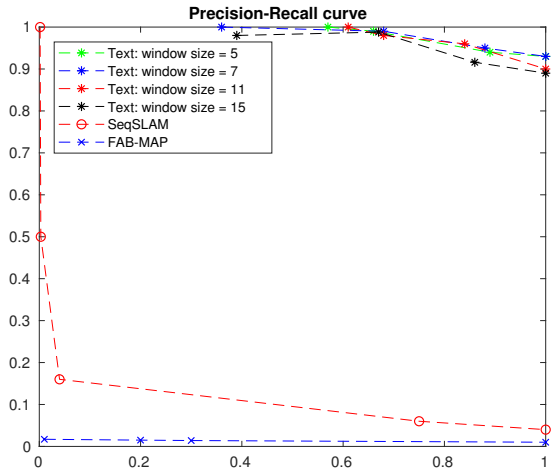


Fig. 12. Precision-Recall Comparison Between Three Methods day_4 and night_4 image sequences

In Figure 12, we can see that our methods still perform better than both FAB-MAP and SeqSLAM. And the precision is much higher than those two.

As it shows in Figure 13, once again our proposed method

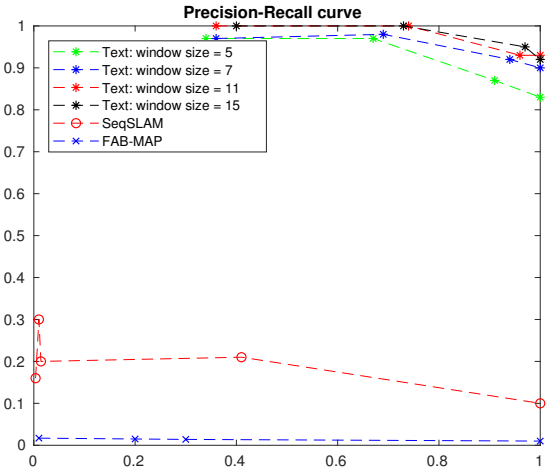


Fig. 13. Precision-Recall Comparison Between Three Methods day_6 and night_6

performs better than SeqSLAM and FAB-MAP.

V. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

To solve the vulnerability of existing vision based place recognition algorithms due to perceptual changes. We proposed and tested a novel place recognition system by making use of scene text information. The final remarks of our proposed system are listed below:

- We explored the feasibility of using scene text for describing a place or a scene and proved that high level feature can be used for place recognition.
- We designed a complete and novel system for place recognition. The topological map we proposed is very light-weight and highly scalable. It only takes few memory storage.
- We evaluated the performance of proposed system, the experiment result demonstrates that our proposed system is able to deal with extreme illumination changes and occlusions. It performs better than the traditional sparse feature based and sequential based place recognition algorithms.
- The matching between the map and the query image does not require any training, it can be scaled to any size of dataset.

B. Future Works

In the future work, we would like to incorporate the robot odometry from the wheel encoder to build a topological metric map which is similar to FAB-MAP. in such a topological map, each node is associated with metric information. Apart from that, we want to associate our topological map with GPS coordinate respect to the Google map so that it become a global localization system.

REFERENCES

[1] Adrien Angeli, David Filliat, Stéphane Doncieux, and Jean-Arcady Meyer. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, 24(5):1027–1037, 2008.

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016.

[3] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.

[4] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *European conference on computer vision*, pages 778–792. Springer, 2010.

[5] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

[6] Dorian Gálvez-López and Juan D Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012.

[7] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017.

[8] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[9] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010.

[10] Mathieu Labbe and Francois Michaud. Appearance-based loop closure detection for online large-scale and long-term operation. *IEEE Transactions on Robotics*, 29(3):734–745, 2013.

[11] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[12] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *arXiv preprint arXiv:1801.02765*, 2018.

[13] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *AAAI*, pages 4161–4167, 2017.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[15] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. *arXiv preprint arXiv:1801.01671*, 2018.

[16] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.

[17] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J. Leonard, David Cox, Peter Corke, and Michael J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2016.

[18] Michael J Milford and Gordon F Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1643–1649. IEEE, 2012.

[19] Paul Newman and Kin Ho. Slam-loop closing with visually salient features. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 635–642. IEEE, 2005.

[20] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European Conference on Computer Vision*, pages 3–20. Springer, 2016.

[21] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2564–2571, Washington, DC, USA, 2011. IEEE Computer Society.

[22] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE TPAMI*, 39(11):2298–2304, 2017.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[24] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *null*, page 1470. IEEE, 2003.

[25] Hsueh-Cheng Wang, Chelsea Finn, Liam Paull, Michael Kaess, Ruth Rosenholtz, Seth Teller, and John Leonard. Bridging text spotting and slam with junction features. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 3701–3708. IEEE, 2015.

[26] Cong Yao, Xiang Bai, Baoguang Shi, and Wenyu Liu. Strokelets: A learned multi-scale representation for scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4049, 2014.

[27] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. *arXiv preprint arXiv:1704.03155*, 2017.

3-D Semantic Localization with Graph Kernel Techniques

Yu Liu and Sen Wang

Abstract— **Place recognition lies at the heart of long-term mapping and localization for autonomous robots. The appearance-based approaches have proven remarkably successful but still face challenges under drastic appearance changes. On the other hand, recent advances in dense SLAM systems and semantic segmentation via Convolution Neural Networks (CNNs) enable reconstructed maps to encapsulate rich semantics and geometry. Inherently, descriptors based on these features should be minimally affected by appearance variations. In this paper. we explore leveraging the dense semantic features and their underlying geometry for place recognition and global localization. We exploit using an object graph to capture objects' topology for each place, and using a walking-based graph kernel method to match object graphs. Finally, we utilize point cloud registration method based on the Fast Point Feature Histogram (FPFH), followed by the iterative-closest-point (ICP) algorithm to align objects from the matched scenes. When semantic information are reliable, our global localization framework achieves accurate 6 Degree of Freedom (DoF) pose estimation, demonstrating the potential of addressing place recognition and global localization problems in the semantic object space.**

I. INTRODUCTION

Autonomous navigation is a crucial feature for a variety of mobile robotic platforms and applications. For this purpose, the robot platform must be able to reliably localize itself by interpreting its surrounding and be able to cope with unexpected changes in the environments. In addition, identifying a place that has already been visited, also referred as loop-closure detection, is the core technique of the Simultaneous Localization and Mapping (SLAM) problems to reduce estimation drift and reconstruct globally consistent maps of the environments.

With place recognition lying at the heart of the loop-closure and localization problems, relevant research, especially visual appearance-based methods, have been extensively investigated. One main spectrum of approaches is to utilize local invariant features, such as SURF[5], SIFT[20] and ORB[30] features extracted from images. This approach and its variants are adapted by many state-of-the-art SLAM methods [9][24][[25]][14][19]. Alternatively, successful results have also been acquired by using global image representations [43][3][23].

In spite of their impressive performance, in real-world applications appearance-based methods are typically affected by a number of challenging problems. For instance, seasonal or weather changes over long navigation, drastic viewpoint changes, natural or artificial illumination variations, as well as the emergence of new static objects or dynamic elements (e.g., vehicles) may drastically alter the appearances of the scene. Even though there exist a host of studies attempting to overcome such challenges, achieving robust place recognition using visual appearance remains a difficult task.

Today's RGB-D SLAM systems are capable of reconstructing dense 3-D maps [14][19][10], which provide rich texture and full structures of the environments. However, this rich amount of dense points are not directly applied to facilitate place recognition. On the other hand, the inclusion of rich semantic information within dense maps potentially permits a greater range of high-level tasks. As a specific example about localization, if the robot possesses a semantic and spatial understanding that a specific part of the environment corresponds to specific objects, similar to how humans query semantic information to their knowledge of the environment, the robot can localize itself when re-observing these objects.

From an intuitive standpoint, semantic features are inherently appearance and viewpoint invariant, e.g., a chair remains a chair regardless being captured under lighting or dark, viewed from right or left direction. With reliable semantic information in the map and the sensor's query, a semantic-based descriptor could potentially achieve absolute viewpoint and lighting invariance. Recent studies on place recognition have also been exploring the adoption of higher level visual features that have a closer relation to the semantic description of the environment [1][34][37][27][13][7][8][11]. In this paper, we explore leveraging rich 3-D semantics and geometry information on place recognition and localization problems. Our main contributions in this work is twofold. Firstly, our proposed framework utilizes 3-D dense object semantics and their underlying geometry and topology for place recognition and localization, fully operating in the object space. Secondly, via point cloud registration on objects, our method is able to estimate an accurate 6 DoF pose for the query scene for global localization.

II. RELATED WORK

In this section we briefly review the current state-of-the-art in appearance-based place recognition. This is followed by a number of recently proposed methods utilizing high-level features in relation to our proposed framework.

A. Visual based place recognition

Visual feature matching based on the visual Bag-of-Words (BoW) [35] techniques has been extensively used for place recognition and loop closure detection since the introduction of FABMAP (Fast Appearance-Based MAPping) [9]. Studies applying the BoW framework often utilize sets of

Y. Liu and S. Wang are with Institute of Sensors, Signals, and Systems, Heriot-Watt University, Edinburgh, UK smw2@hw.ac.uk

locally invariant features (such as SIFT[20], SURF[5], and ORB[30]) from the image space to represent each image associated with a location, and have shown success applying the TF-IDF (term frequency-inverse document frequency) [15] [2][22] and probabilistic scoring methods [9][18][38] for place recognition in the presence of viewpoint and moderate lighting changes. However. one obvious drawback of these methods is that they neglect the spatial arrangement among features, resulting in higher perceptual aliasing and incidence of false-positive. Studies have attempted incorporating spatial information to the recognition framework, such as modeling locations by visual landmarks and a distribution of 3-D distances provided by range finders or stereo cameras [28], extending the standard visual vocabulary with a spatial dictionary by dividing up images into regular grids[17], and modeling spatial relations between features by covisibility information [41][22].

B. Utilizing high-level features and objects

Appearance-based methods, in spite of their great success at place recognition and large-scale mapping, still suffer from the problem of changing environmental condition over longer periods. Recent studies have moved toward the adoption of higher level visual cues such as semantic cues for place recognition. For instance, Snderhauf et al. [42] use a landmark proposal system [48] to extract highly salient regions in the images as high level landmark, and employ features extracted from inner convolutional layer of a pretrained CNN network. Cascianell et al. [8] extend Snderhauf's work by adapting the covisibility graph [22] to model the environment as a structured collection of visual landmarks. Instead of RGB-based region proposals, Yu et al. [46] exploits clustering point cloud of similar surfaces as high-level landmarks to perform place recognition.

Alternatively, place recognition based on objects and their spatial relations have also been explored recently. In [26], the authors build an object graph for each place and perform place recognition by measuring node and edge differences. Other attempts such as [1] first detects objects by convolving pre-defined primitive kernel patches of basic shapes with the full 3-D dense maps generated from Kintinuous SLAM system [44]. The authors also represent a place as a sparse object graph. With the recent advances in learning-based semantic extraction methods, [13] proposes using CNN for semantic segmentation to obtain per-pixel semantics of the query images and represent their underlying topologies of objects. On the other hand, McCormat et al. [21] combine state-of-the-art SLAM system and CNN for semantic segmentation, building an geometrically consistent semantic map by probabalistically fusing multiple semantic prediction from observations of different viewpoints.

C. Graph matching

When it comes to representing a place as a constellation of visual objects, researcher often choose to utilize a graph structure to describe the object content and topology [1][13][46][26]. Similarity between places is hence reduced to measuring pair-wise similarities between graphs. One common solution is to formulate the graph matching problem as an assignment problem [46][26][1]. The turns matching into an optimization problem, with the objective of finding an exact correspondence between the nodes and edges from two graphs. However in the general case, solving graph matching in this manner is NP-hard [47].

An alternative spectrum of method to graph matching problems is inexact graph comparison, which does not explicitly solve for pair-wise correspondences. For example, [39] converts their graphs of visual landmarks into weighted graphs of known landmark categories. Similarity is then approximated by the normalized cross-correlation between the two associated adjacency matrices. On the other hand, [40] chooses to augment each node in the graph with a histogram describing the presence of connected nodes nearby. Graph similarity is then measured by standard dot product scheme. Others attempt to solve graph matching using graph kernels based on walks [12][13]. In [13], random walk descriptors are extracted for every node, each storing the base vertex label and the labels of visited nodes in sequence. Also employing a walk strategy, [12] inspects pair-wise similarity of walks' composing nodes and edges for scene modeling problems.

III. PROPOSED ALGORITHM

Our method leverages building object-based graphs from the incoming RGB images, depth maps and semantic classification. We then match graphs between the query and those built from the global map using a walking strategy. The final goal of our localization method is to estimate the 6 DoF pose of the camera with respect to the global map. Figure 1 below illustrates the architecture of our proposed method.

A. Reconstruction of dense semantic maps

We partially implement the Bayesian fusion method proposed by [21] to obtain dense semantic maps. Similar to their work, we employ the publicly available SLAM system - RTAB-Map (Real-Time Appearance-Based Mapping) [19] and the semantic segmentation CNN - SegNet[4]. After mapping, RTAB-Map generates a map of registered RGB point clouds, a series of camera poses and their associated keyframes in RGB. On the other hand, SegNet is a deep convolutional encoder-decoder architecture for pixel-wise semantic segmentation. Different variants of models and pretrained weights are publicly available; We choose to use the model trained with SUN RGB-D [36] indoor dataset for semantic labeling. This model takes raw RGB images and outputs semantic features over 37 indoor object classes for every pixel. To obtain the raw probability data in order to cope with McCormat's method, we minimally alter the model's output to extract full probability over all class.

To begin fusing semantic features, the global map is initialized with a uniform probability distribution. By using the camera pose associated to each RGB frame and the intrinsic matrix from the RGB-D camera, we project the semantic probability distribution derived from SegNet into

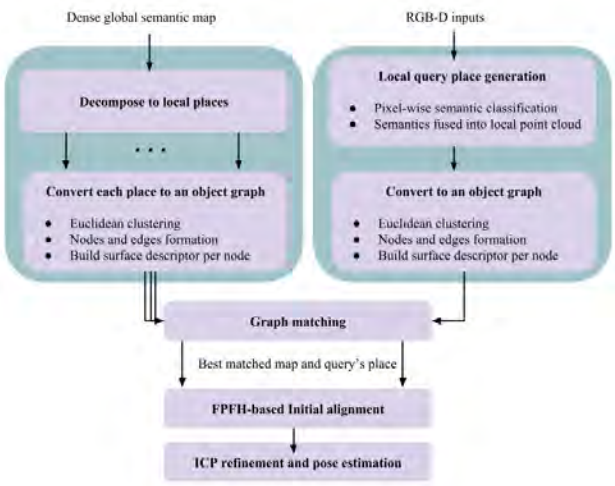


Fig. 1: Overview of our proposed global localization pipeline. The inputs from the sensor (e.g., RGB-D camera) are semantically segmented and fused into the local point cloud (query local place). The global semantic map is subdivided into map local place. Each local place is converted into its 3-D graph representation in terms of semantic objects as nodes and connections to each other as edges. Our method matches each query graph to the map's sub-graph. Finally, the point clouds associated to the matches are aligned for estimating the 6 DoF pose of the query place.

the map. In details, we apply frustum culling for each camera pose to only keep the points within the camera's field of view (excluding the rest of the points in the map). This is followed by the projection of these 3-D points within the field of view into the image space. Once the association between 3-D points and 2-D pixel coordinates are established, the semantic probability distribution at each pixel is back-projected to the associated 3-D point. This back-projected semantic distribution is then multiplied by the distribution a 3-D point currently has (and then normalized). This allows updating any 3-D point's semantic distribution from multiple observations. Subsequent observations of the same point will eventually converge and lead to a more confident labeling.

B. Local query place generation

We propose to register multiple incoming frames of point clouds to represent a query place. The motion between each consecutive frame is assumed small and known, which can be calculated using visual odometry [45]. In our work we treat each registered point cloud as a local place analogous to the key-frame based framework using an keyframe images.

Augmenting each local place with semantic features is done in the same way as we perform semantic mapping. The only difference is that in case of local place, the target point cloud is each registered point cloud instead of the global map.

C. Graph representation for each local place

Given a place of 3-D points with semantics, we represent the place with an object graph. We define an object as:

$$O \triangleq \{L_O, C_O, R_O, S_O\} \tag{1}$$

where L_O is the semantic label associated with the object, C_O is the center of the object, R_O is the radius of the bounding sphere which fully encloses every point of the object, and S_O is the surface descriptor which will be covered more in the following section. We choose to use radius of a sphere for our objects dimension as spheres hold the convenient property of being rotationally invariant.

To add object vertices to the object graph, we extract blobs of object by performing euclidean clustering [31] on points with the same semantic label. The dimension of the sphere is simply the distance of point furthest away from the cluster center.

We choose to form undirected edges between any two object vertices when they are within a proximity distance with each other. Nodes which are far apart will simply be unconnected. The distance is the relative offset between their center locations in the 3-D space. Figure 2 shows an example of a query location and its corresponding object graph.

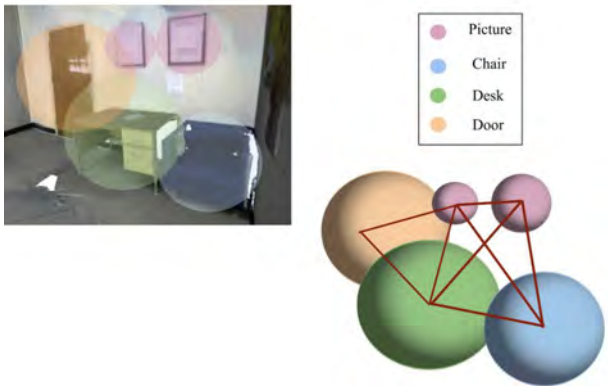


Fig. 2: Concept of object graph from the 3-D densely labelled semantic point cloud.

D. Shape descriptor

Each object vertex holds the shape property associated with the surface of that object. We implement the shape descriptor proposed by [46], which is built from the Fast Point Feature Histogram (FPFH) [32] of the object's point cloud.

The FPFH of a point is a histogram describing three types of angular relationship between the normal of that point and those of its neighbors within a sampling radius [32]. For a perfect plane model point cloud, the FPFH of every underlying point has an unique form of histogram distribution. The shape ratio descriptor proposed by [46] measures the variation of every points FPFH of the target point cloud from that of a plane model. In other words, we will distinguish one surface from another by assessing their degree of deformation from a plane.

To build the shape descriptor for each object, we first calculate the FPFH for every point in that object. This results in a 33-bin histogram FPFH signature for each point, where each angle property is divided into 11 divisions. We then calculate the Euclidean distance between every points FPFH and that of a plane element. We then assign each Eudlidean distance of point feature to a specific bin index b of a histogram. The calculation of the bin index is given by:

$$b = ceil\left[\frac{EDist_(P_r)}{EDist_{max}}D\right] \tag{2}$$

where D denotes the number of subdivisions determining the resolution classifying each angle point feature. The overall length of the histogram is thus D. Finally we normalize the resulting histogram between 0 and 1.

E. Graph matching

Our method is inspired by the intuition and related literature that semantic graphs hold reliable discrimination power, which can facilitate place recognition. However, explicitly solving for node and edge correspondence is an NP hard problem. Instead, we choose to perform an inexact matching by using the walk based graph kernel strategy employed by [12].

In [12], a walk of length 1 corresponds to an edge and the two nodes connected by that edge. Graph similarity is then defined as sum of similarity between walks of two graphs. To perform matching between two walks is equivalent to matching the similarity between individual pair of constituent nodes and edge of those walks.

1) Node comparison: Each node has a number of properties as described in section III-C. We choose to model node similarity using two of them: semantic label and shape descriptor. We use the same annotation as in [12]. Node comparison is formulated as follows:

$$k_{node}(r,s) = \delta_{rs}k_{shape}(r,s) \tag{3}$$

In the above equation, r and s each represents a node from a walk. δ_{rs} is a simple Kronecker delta function, which is 1 if the two nodes being compared have the same semantic label and 0 otherwise. We define $k_{shape}(r,s)$ as:

$$k_{shape}(r,s) = 1 - d \tag{4}$$

where d denotes the Euclidean distance between two shape histograms. Essentially, two nodes of different object types will not be matched even if their shape descriptor express high similarity.

2) Edge comparison: We define two edges' similarity based on both metric distance and semantic relations. Metric distance is simply the observed length of each edge. We believe such distances still encodes meaningful spatial information. However, metric distance is sensitive to viewpoint variations, as occlusions or objects being partially captured influence the center location of a node. Hence we try to incorporate semantic relations based on how two connected objects interact, in terms the degree of enclosure of their associated bounding sphere. We classify four semantic relations as illustrated in figure 3.

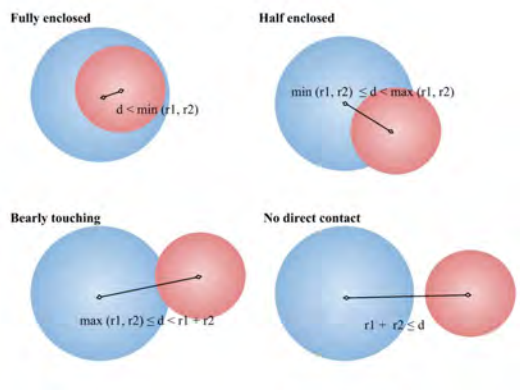


Fig. 3: Four semantic relations based on the degree of "enclosure" between the object pair. "d" denotes the Euclidean distance between two nodes' centers

3) Graph comparison: Given the node kernel and an edge kernel, the similarity between walks of length 1 is defined as the product of the two node scores and the edge score. In cases when multiple walks are present in each graph, we compute the similarity between all pairs of walks from each graph and sum the score. Finally, we normalize the score as equation 5 by dividing the result of a graph kernel by the maximum of the evaluation between each graph and itself:

$$normk_G(G_a,G_b) = \frac{k_G(G_a,G_b)}{max(k_G(G_a,G_a),k_G(G_b,G_b))} \tag{5}$$

F. Object alignment

During graph comparison, individual walks receiving high scores indicate potential good match between the associated objects and spatial relations from the two graphs. After conducting graph matching, for every query place and the best matched place found in the map, we only keep objects from both scenes which express high similarity. We then estimate the relative transformation between the two scenes by aligning the sets of objects from query place to map place.

To perform alignment, we once again rely on the FPFH features following the Sample Consensus initial alignment scheme (SAC-IA) as proposed by [32]. The estimated transformation gives us an initial alignment to localize the query place within the map. Finally, the transformation is further refined by ICP [6].

IV. EXPERIMENTAL RESULTS

We evaluate our global localization framework based on three considered dataset: two with groundtruth semantic labels, and the other one obtains semantics via SegNet. Our algorithm is implemented dominantly in C++ and partially in python under Robot Operating System (ROS) [29]. We also extensively make use of the Point Cloud Library (PCL) [33].

A. Experimental setup

1) Groundtruth semantics: We choose to use the publicly available ScenNN dataset [16]. This dataset consists of several sequences of indoor scans using an ASUS XTION Pro RGB-D camera. Each environment data provides the RGB, depth, and pixel-wise semantic classification frames for the full sequence, and two global maps (RGB and semantically labeled respectively). The complexity of an indoor scene and the number of different semantic classes vary. We use sequence 011 and 025 in our experiments.

We carry on our experiments as follows. As we model our algorithm based on the key-frame based method, we consider the original global map consisting of multiple key local places. Localization is then carried out by finding the best match query place to the best-match map place. To prepare for map's local places, we directly extract sets of local point clouds from the global map using frustum culling method and the provided camera poses.

The trajectory provided by SceneNN does not necessarily include loop closures, as the mapping environment is small and effect of camera drift is minimal. Hence, we model our localization experiment by using the same sequence but different frames for query places to imitate making a second travel in the map via a similar trajectory with the first one. Contrary to each map's local place which we obtain from the given global map, each query local place is generated by projecting the provided semantic images to the corresponding depth maps, and then register the resulting local point clouds. Probabilistic fusion is not required as we have the groundtruth segmentation. We also make sure frames used to build map's local places are not reused for generating query places.

2) CNN semantics: We apply SegNet and our Bayesian-fusion approach to acquire the global semantic map of our own collected data. According to our proposed pipeline, we would then collect a second sequence of the same environment to create query local places and try to match each of them against the global map.

In practice, in order to obtain the groundtruth trajectory for the query sequence without the use of additional sensor, we also perform semantic mapping on the second sequence and keep the initial camera pose as identical as possible to the first sequence. This will allow the second map to have the same map frame as that of the first one; hence the keyframe trajectory provided by RTAB-Map for the second sequence is used as groundtruths. Having the each keyframe pose and the semantic map of the query sequence, we simplify the generation of each local query place by directly extracting from the query semantic map by applying the frustum culling method again. This simplified way of generating a query local place is reasonable; our original proposal for generating a local query place is to utilize the Bayesian fusion technique and fuse semantic predictions of a scene observed from multiple viewpoints. Directly extracting a local place associated with a keyframe pose from the semantic map inherently has considered multiple frames of observations.

B. Localization on groudtruth semantics

Figure 6 (a-d) displays the mean localization error on both Sequence 11 and 25, in terms of position and orientation for each query local place. Indices with non-zero results indicate global localization was successfully carried out for the corresponding query places. Within our framework, a query place could fail to be localized if the SAC-IA alignment step could not find enough inliers to support its hypothesized transformation. This also suggests that incorrect matching of places would most certainly fail consequently in the localization step. On the other hand, due to the nature of our proposed framework using walks (two nodes linked by an edge) to measure graph similarity, any query place containing only one object or multiple unconnected objects would automatically not be taken into matching. This results in place recognition nor localization not being triggered. This second factor dominates in both sequences, as both environments are sparse in terms of the objects accommodated. In our framework, object misalignment would directly impact the localization accuracy. Figure 4 exhibits an example of a poor alignment in Sequence 11, reflecting a high error at the corresponding index.

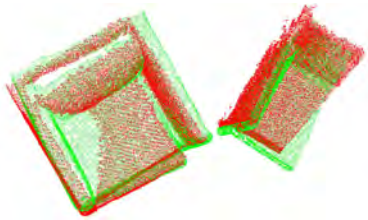


Fig. 4: Final alignment on query 29, Sequence 11.

Table I shows the mean localization error of each sequence. Overall, Sequence 25 exhibits a significantly larger error in roll and yaw. This is due to multiple query scenes, even though correctly matched to the right local map place, failed to completely align with the correspondences. In addition, during the computation of the mean error, we removed an outlier from Sequence 25 as the estimated pose failed completely (figure 5). To avoid incidence such as this one, a stricter inlier criteria can be set for the alignment step; however at the cost of possibly rejecting correct localization.

TABLE I: Localization mean error on groundtruth semantics (unit: cm and degree)

	X	Y	Z	Roll	Pitch	Yaw
Seq. 11	1.994	1.820	1.474	2.210	2.913	2.248
Seq. 25	1.6125	3.160	3.473	11.241	1.342	10.276

C. Localization on CNN semantics

As we do not hold any groundtruth semantics, we first qualitatively inspect the dense semantics for each generated local place. Figure 7 shows examples of local query places

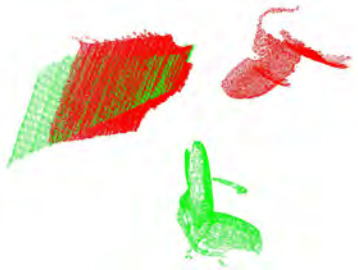


Fig. 5: Final alignment on query 32, Sequence 25.

and the associated RGB frames for reference. The semantic features are prone to a degree of error (mislabeling or oversegmentation) even after the Bayesian fusion; however, they are fairly consistent between the global and query maps.

Figure 6 (e-f) and table II show the localization error.

TABLE II: Localization mean error (unit: cm and degree)

	X	Y	Z	Roll	Pitch	Yaw
mean error	2.828	4.064	5.943	5.836	7.656	4.892

Overall, we observe an increase in localization error in terms of position, as compared to experiments with groundtruth semantics. This may be due to the nosier surfaces of the point cloud data from RTAB-Map, where the reconstructed map after loop closure detection still exhibit slight inconsistency. On the other hand, localization accuracy in orientation is better than that of Sequence 25, as it has multiple local places which do not perfectly align with their matches. This experiment demonstrates that the proposed framework's functionality even under nosier geometry and semantics.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

We have presented a global localization framework leveraging 3-D dense semantic features, its underlying surfaces and topology to match and estimate the 6 DoF pose of the camera associated with the given observation in the global semantic map.

Our approach was evaluated on two real-world data sequences where idealistic semantics are available, and one of our own collected indoor data utilizing semantic segmentation CNN to densely augment points in the map with semantic features. In both cases, our method was validated to be able to perform global localization, and finds the 6 DoF pose within a certain order of deviations from the groundtruths. This demonstrates the potential of using compact graph representation of semantics to capture the scene topology, and the estimation of the camera pose based on object alignments. The incorporation of the object alignment scheme did significantly raise the computation complexity due to the computation of each point' normal and FPFH features. Nevertheless, with reliable dense semantic features, the result of global localization proves accurate in terms of both position and orientation.

B. Future work

Even though our results demonstrate the potential of addressing the place recognition and global localization problems using dense semantic features, there remain many challenges and possible improvements to the proposed framework. Firstly, our method was experimented minimally. In experiments where the ideal semantic features are assumed, we have only experimented on two sequences with few objects, minimal repetition and simple arrangement which otherwise would have imposed great challenges. On the other hand when we took the experiment on our dataset without groundtruth semantics, we simplified experiment setups by extracting local places directly from a SLAM optimized global map, which would have granted us more information and advantages than generating local places from the sensor's incoming frames. Furthermore, we are motivated by the intuition that using objects as visual landmarks potentially copes well with drastic lighting and viewpoint variations, as objects remain the same properties under these changes. However relevant experiments have not yet been conducted during the write up of this thesis.

One possible improvement to the current framework is to incorporate a motion model and the probabilistic framework. Our proposed method was developed to handle the localization problem in a straightforward way; it determines which local place of the global map matches well with a query place, and then calculates the transformation between the underlying objects to align the two places. In reality, matching query scene in such discrete approach not only is prone to erroneous estimation and observation uncertainties, but also is computational costly as every local place of the map is compared with the query one. In the contrary, modeling the place recognition and localization problem probabilistically can facilitate identifying the correct match from a small number of match candidates based on prior observations.

Moreover, we believe that defining a local place by registering multiple frames and associated point clouds may only be an intermediate proposition. We started out in this approach due to our observations that single-frames capture limited views of objects and can be sensitive to occlusions. This would be challenging with our current framework which performs localization on discrete queries. On the other hand, incorporation of the probabilistic framework can potentially take advantage of multiple single-frame observations.

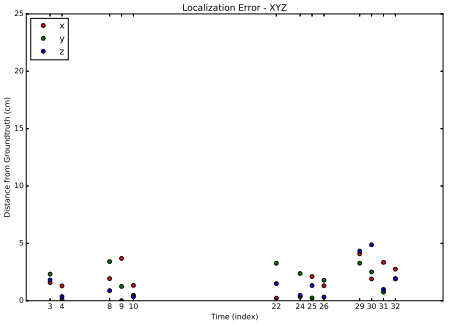
VI. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of reviewers' comments.

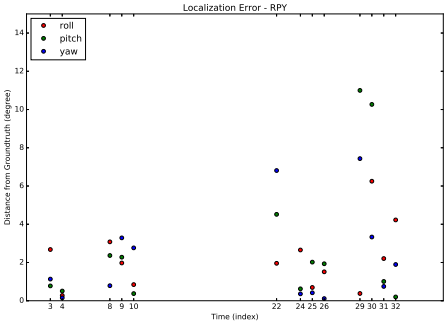
REFERENCES

[1] *Toward Object-based Place Recognition in Dense RGB-D Maps*, 05/2015 2015.

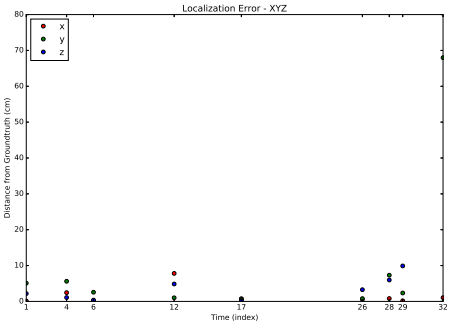
[2] Adrien Angeli, David Filliat, Stéphane Doncieux, and Jean-Arcady Meyer. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, 24(5):1027–1037, 2008.



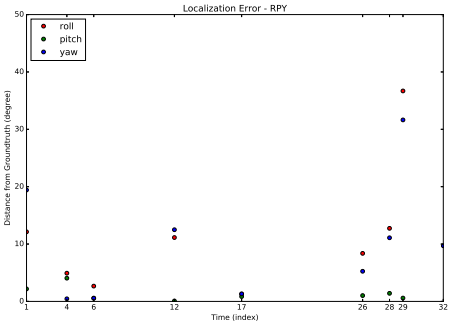
(a) Localization error - XYZ (Sequence 11)



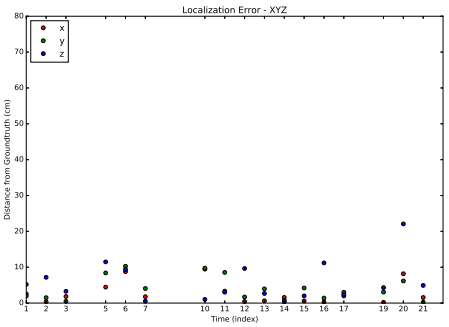
(b) Localization error - RPY (Sequence 11)



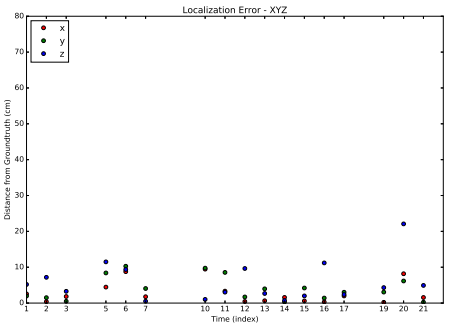
(c) Localization error - XYZ (Sequence 25)



(d) Localization error - RPY (Sequence 25)



(e) Localization error - XYZ (own collected data)



(f) Localization error - RPY (own collected data)

Fig. 6: Localization error in terms of position (XYZ) and orientation (RPY). Note the different scale in Y-axis.

Fig. 7: Generated query local places with CNN's semantics

[3] Hernán Badino, Daniel F. Huber, and Takeo Kanade. Real-time topometric localization. *2012 IEEE International Conference on Robotics and Automation*, pages 1635–1642, 2012.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561, 2015.

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.

[6] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*, volume 1611, pages 586–607. International Society for Optics and Photonics, 1992.

[7] Sean L Bowman, Nikolay Atanasov, Kostas Daniilidis, and George J Pappas. Probabilistic data association for semantic slam. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 1722–1729. IEEE, 2017.

[8] Silvia Cascianelli, Gabriele Costante, Enrico Bellocchio, Paolo Valigi, Mario L Fravolini, and Thomas A Ciarfuglia. Robust visual semi-semantic loop closure detection by a covisibility graph and cnn features. *Robotics and Autonomous Systems*, 92:53–65, 2017.

[9] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

[10] Felix Endres, Jürgen Hess, Nikolas Engelhard, Jürgen Sturm, Daniel

Cremers, and Wolfram Burgard. An evaluation of the rgb-d slam system. *2012 IEEE International Conference on Robotics and Automation*, pages 1691–1696, 2012.

[11] Ross Finman, Thomas Whelan, Liam Paull, and John J Leonard. Physical words for place recognition in dense rgb-d maps. In *ICRA workshop on visual place recognition in changing environments*, 2014.

[12] Matthew Fisher, Manolis Savva, and Pat Hanrahan. Characterizing structural relationships in scenes using graph kernels. *ACM Transactions on Graphics (TOG)*, 30(4):34, 2011.

[13] Abel Gawel, Carlo Del Don, Roland Siegwart, Juan I. Nieto, and Cesar Cadena. X-view: Graph-based semantic multi-view localization. *CoRR*, abs/1709.09905, 2017.

[14] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *Int. J. Rob. Res.*, 31(5):647–663, April 2012.

[15] Djoerd Hiemstra. A probabilistic justification for using tf× idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3(2):131–139, 2000.

[16] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 92–101. IEEE, 2016.

[17] Edward Johns and Guang-Zhong Yang. Feature co-occurrence maps: Appearance-based localisation throughout the day. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 3212–3218. IEEE, 2013.

[18] Edward Johns and Guang-Zhong Yang. Generative methods for long-term place recognition in dynamic scenes. *International Journal of Computer Vision*, 106(3):297–314, 2014.

[19] Mathieu Labbé and Franois Michaud. Online global loop closure detection for large-scale multi-session graph-based slam. *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2661–2666, 2014.

[20] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.

[21] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 4628–4635. IEEE, 2017.

[22] Christopher Mei, Gabe Sibley, and Paul Newman. Closing loops without places. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 3738–3744. IEEE, 2010.

[23] Michael Milford. Vision-based place recognition: how low can you go? *The International Journal of Robotics Research*, 32(7):766–789, 2013.

[24] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *CoRR*, abs/1502.00956, 2015.

[25] Raul Mur-Artal and Juan D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *CoRR*, abs/1610.06475, 2016.

[26] JH Oh, JD Jeon, and BH Lee. Place recognition for visual loop-closures using similarities of object graphs. *Electronics Letters*, 51(1):44–46, 2014.

[27] Parv Parkhiya, Rishabh Khawad, J Krishna Murthy, Brojeshwar Bhowmick, and K Madhava Krishna. Constructing category-specific models for monocular object-slam. *arXiv preprint arXiv:1802.09292*, 2018.

[28] Rohan Paul and Paul Newman. Fab-map 3d: Topological mapping with spatial and visual appearance. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 2649–2656. IEEE, 2010.

[29] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009.

[30] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2564–2571, Washington, DC, USA, 2011. IEEE Computer Society.

[31] Radu Bogdan Rusu. *Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments*. PhD thesis, Computer Science department, Technische Universitaet Muenchen, Germany, October 2009.

[32] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3212–3217. IEEE, 2009.

[33] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.

[34] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1352–1359. IEEE, 2013.

[35] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *null*, page 1470. IEEE, 2003.

[36] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, volume 5, page 6, 2015.

[37] Erik Stenborg, Carl Toft, and Lars Hammarstrand. Long-term visual localization using semantically segmented images. *arXiv preprint arXiv:1801.05269*, 2018.

[38] Elena Stumm, Christopher Mei, and Simon Lacroix. Probabilistic place recognition with covisibility maps. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 4158–4163. IEEE, 2013.

[39] Elena Stumm, Christopher Mei, Simon Lacroix, and Margarita Chli. Location graphs for visual place recognition. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5475–5480.

[40] Elena Stumm, Christopher Mei, Simon Lacroix, Juan Nieto, Marco Hutter, and Roland Siegwart. Robust visual place recognition with graph kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4535–4544, 2016.

[41] Elena S Stumm, Christopher Mei, and Simon Lacroix. Building location models for visual place recognition. *The International Journal of Robotics Research*, 35(4):334–356, 2016.

[42] Niko Sünderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. *Proceedings of Robotics: Science and Systems XII*, 2015.

[43] Niko Snderhauf and Peter Protzel. Brief-gist closing the loop by simple means. In *in Proc. of IEEE Intl. Conf. on Intelligent Robots and Systems (IROS*, 2011.

[44] Thomas Whelan, Michael Kaess, Maurice Fallon, Hordur Johannsson, John Leonard, and John McDonald. Kintinuous: Spatially extended kinectfusion. 2012.

[45] Khalid Yousif, Alireza Bab-Hadiashar, and Reza Hoseinnezhad. An overview to visual odometry and visual slam: Applications to mobile robotics. *Intelligent Industrial Systems*, 1(4):289–311, 2015.

[46] Hyejun Yu, Hee-Won Chae, and Jae-Bok Song. Place recognition based on surface graph for a mobile robot. In *Ubiquitous Robots and Ambient Intelligence (URAI), 2017 14th International Conference on*, pages 342–346. IEEE, 2017.

[47] Feng Zhou and Fernando De la Torre. Factorized graph matching. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 127–134. IEEE, 2012.

[48] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.

Compressed Low-Light Scene Reconstruction using Hyper-Spectral Single-Photon LiDAR

Anirudh Puligandla

***Abstract*— Recent introduction of the Single-Photon LiDAR technology has garnered wide interest to produce accurate 3D reconstructions from limited amounts of photons. Such devices have higher sensitivity over traditional laser scanning systems and are capable of imaging over multiple wavelengths from visible to near infra red simultaneously. With such advantages this technology has found many applications under low-light scenarios. Low-Light constraints may arise due to various factors such as short acquisition times (e.g, long range applications) or low light-flux of the emitted laser (e.g, bio-medical applications, where even low-powered lasers might damage the tissues). This work focuses on the reconstruction of Intensity images simultaneously, from measurements taken over multiple peak wavelengths, using a Poisson measurement model and Alternating Direction Method of Multipliers (ADMM) algorithm. Different sparsity promoting regularizers have been proposed under a Compressed Sensing framework. This work will also show that the proposed methods are capable of reconstruction with light flux as low as 1 or even 0.5 received photons per pixel over an image.**

I. INTRODUCTION

Photon counting techniques have been used for decades mainly to monitor the atmosphere [1]. For example, precipitation can be estimated by checking the amount of water present in the air [2], or, pollution can be estimated by checking the amount of aerosols present in the air [3]. Single-Photon Lidar (SPL) is an example of advanced photon counting device. Traditional lidar systems have been widely used in *geographical information systems* [4], [5], *underwater imaging* (mainly bathymetry) [6], [7], *forest monitoring* [8], [9], [10], *space imaging* [11], [12], and in molecular biology applications mainly for time-resolved fluorescence spectroscopy [13], [14], [15]. Time of Flight (ToF) SPL with an architecture to scan the scene in a raster scanning fashion will be used in this context [16]. As mentioned in [17], *Single-photon imaging is the detection of two-dimensional patterns of low-intensity light (i.e, when the number of detected photons in each pixel is lower than 10).*

Fig. 1 depicts an illustration of the photon detection process. Full circles in the images represent photon interaction while open circles indicate no interaction of photons. The image on the left illustrates photons interacting with the image sensor while the image on the right illustrates the subsequent electronic photo-charge detection and conversion process. The noise in the detection process arises due to two primary factors. Firstly, the variance of photon detection also includes the variance of the photon detection process. As a consequence, detection of photons becomes prone to errors (full circles on the left become open circles on the right, larger full circles on the right). Secondly, as the electronic noise is acting on all the pixels, photons may be reported even in locations where no photons are imaged onto the image sensor, known as ”dark counts” (shown with a cross symbol on the right hand side image).

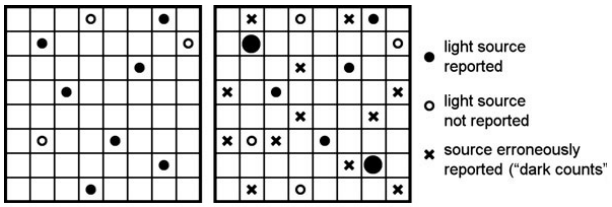
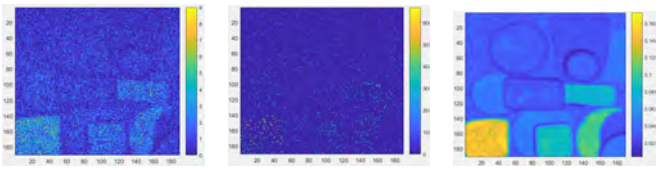


Fig. 1: An illustration showing the noisy photon detection process, *source: [17]*

A. Problem Definition



(a) (b) (c)

Fig. 2: Illustration showing observations and ground truth intensity image for one wavelength. (a) photons collected uniformly, (b) photons collected only on $1/16^{th}$ of the total number of pixels, (c) expected result.

Under low-light conditions, Poisson noise assumptions are appropriate to minimize noise. In the rest of the document, we will see how an optimization framework can be constructed to maximize the likelihood of photon detection. The architecture of single-pixel ToF cameras has shown to reduce the required size, complexity and cost of the photon detector array down to a single unit [18]. By combining the single-pixel SPL with random Compressed Sensing measurements, a trade-off between space and time can be achieved during image acquisition. We will also see how some regularizers (known as priors from a Bayesian point of view) can help in hyper-spectral intensity image reconstruction. Lastly, we will see how a minimization problem can be framed that can be optimized using a convex optimization algorithm such as the Alternating Direction Method of Multipliers (ADMM)

A short version of the thesis submitted for the degree of MSc. VIBOT

A. Puligandla is with the Institute of Sensors, Signals and Systems, School of Engineering and Physical Sciences, Heriot Watt University, EH14 4AS Riccarton, Edinburgh pv.anirudh@gmail.com

algorithm. Fig. 2 depicts an illustration of the problem stated above.

B. Prior Work

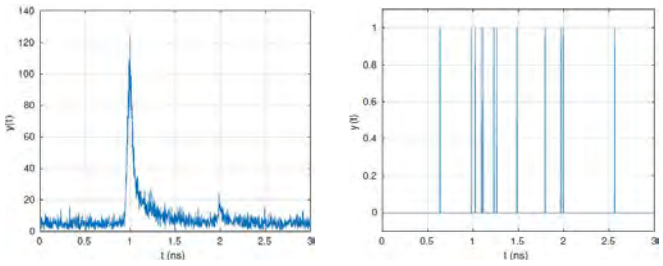
A recent methodology to obtain 3D reconstructions from SPL data was proposed by Shin et al. in [19], [20]. This method assumes that at a given pixel, the response signal can be the sum of responses from limited number of objects with different reflectivities at different depths. This method is shown to be able to model translucent objects with a better resolution at the pixels containing edges. But the union-of-subspaces model limits it only to a few pre-defined responses. However, this method works only for the cases with sufficient amount of photons at each pixel location while considering only single band. Recently, Altmann et al. had proposed an algorithm under the Bayesian context for simultaneous reconstruction and spectral unmixing of multi-spectral SPL data using the Markov Chain Monte Carlo (MCMC) method, [21]. This method works by estimating the abundances of each material in each pixel. Although the algorithm is efficient, it requires previous knowledge about the spectral response of all the materials present in the scene. In similar lines, they had also proposed another method to use convex optimization with an aim to minimize the estimated number of materials in each pixel while maintaining the data fidelity or the regularizer terms low, [22]. However, this method also requires prior knowledge about the spectral responses of all the materials present in the scene.

Shin et al. proposed another method in [23], [24], where a known baseline noise is assumed over all the measured pixels. This assumption is not adequate as the baseline noise is pixel dependent. Despite these assumptions, the method achieves good reconstructions for data with low photon counts because of the exploitation of spatial correlation among the pixels through the use of Total Variation (TV) regularization. From prior work it is evident that convex optimization techniques provide reconstructions with high accuracy under low-light (low photon reception) settings. In addition to the above mentioned methods, some comprehensive reviews of Poisson image restoration algorithms can be found in [25], [26]. Regularized variants of the classical Richardson-Lucy (RL) method have been proposed in [27], [28] that use Total Variation (TV) and Wavelet-based Regularization, respectively. Some multiscale approaches that handle image inverse problems can be found in [29], [30], [31], [32].

II. 2D Image Intensity Estimation

A. Observation Model

SPL works by emitting multiple time correlated light pulses and detecting the reflected photons. SPL uses the Time Correlated Single Photon Counting technique (TCSPC) to count the number of photons imaged onto the sensor [33]. TCSPC exploits the fact that for signals with high repetition rate, the light intensity is so low that the probability of detecting one photon in one signal period is far less than one [34]. The output signal is constructed by recording the photons, measuring their time of arrival in the signal period and building a histogram of the photon times. The probability of detecting more than one photon per period is close to zero [34]. Therefore, the probability of detection of photons for each signal period can be seen as a Bernoulli trial[1], as there can be either 1 or 0 photons per signal period. Hence, the detection of photons at each pixel is a sequence of consecutive Bernoulli trials. We know that the summation of consecutive Bernoulli trials (binary terms) leads to a binomial distribution [17]. The binomial distribution represents a Poisson distribution, as a special case, when the number of detected photons is very low, [35].



(a) High (b) Low

Fig. 3: Comparison between the output signal from SPL at each pixel for high or low number of detected photons, respectively

Now, if an object is present within the range of the SPL, the peak intensity value of the response signal depends on the reflectivity of the object at the wavelength of the emitted pulse while the position of the peak depends on the position of the object. This work requires some calibration data that contains response signals for an object with known reflectivity that is imaged at a known position w.r.t. the SPL device. This calibration data is collected for a predefined time of acquisition by imaging at all the wavelengths. Fig. 3 shows a comparison between the response signals at high and low number of observed photons. It can be seen that when the number of detected photons is very low ($<$10 photons per pixel (ppp)), the structure of the response signal becomes intractable. The estimation of the intensity images can be divided into two sequential steps, i.e, estimation of the background intensity (noise due to other light sources) followed by the estimation of true reflectivity (response intensity) of the scene. Two separate inverse problems that relate the underlying intensity images to the observations through a linear relationship. The baseline intensity can be estimated from the first few time bins of the response signal as these bins only represent background light. Fig. 4 shows an illustration of the measurement model. The red dotted lines in the images highlight the separation of the histogram bins into two parts to estimate the baseline (left of red line) and response (right of red line) intensities.

The baseline intensity can be related to the observations

[1]A Bernoulli trial is an experiment whose outcome is random and can only be of two possible outcomes, either success or failure.

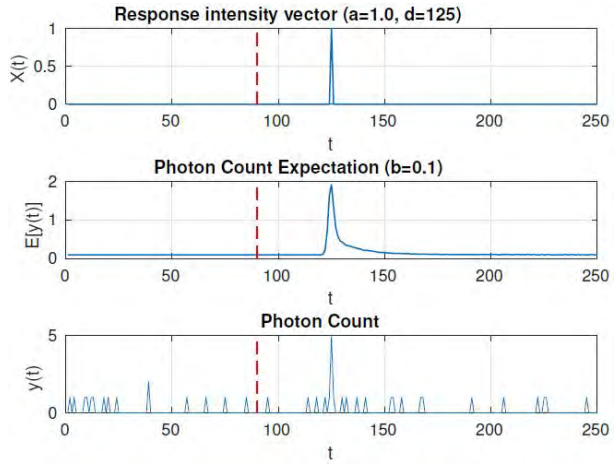


Fig. 4: An illustration of the measurement model. Top: example of true response intensity per depth. Middle: expected photon count for each time bin that depends on a,b and d. Bottom: simulated photon count y, at each pixel, using the Poisson model

by using the following equation

$$y_{p,l} \sim Pr(b_{p,l}) \tag{1}$$

Where, $Pr(\lambda)$ represents the Poisson Distribution with mean intensity λ, $y_{p,l}$ is the photon count at pixel p and wavelength l, summed over the first few time bins (upto t_b). Similarly, the response intensity can be related to the observations as

$$y_{p,l} \sim Pr(r_{p,l}F_l x_{p,l} + b_{p,l}) \tag{2}$$

Where, $y_{p,l}$ is again the photons counts summed over the remaining time bins x_p and b_p are the response and baseline intensities, respectively, F_l is each column of the concatenated calibration matrix $F = [F_1 ... F_l ... F_L]$, and $r_{p,l}$ is a pixel-wise multiplicative factor compensating for the wavelength vignetting effect that is the same as the reduction in brightness at the peripheries of the image often seen in photography or optics. $F \in \mathbb{R}^{D\times L}$ (where, D is the range of the SPL system in discrete steps and L is the number of wavelengths) is known as the calibration matrix. The data in F is usually collected for long a time (such as 100s) to minimize noise and this matrix has to be scaled according to the acquisition time of the observed data, using (3).

$$F = F_{calib}\left(\frac{\Delta t_0}{100}\right) \tag{3}$$

where, Δt_0 is the acquisition time in seconds. Now, using the negative log-likelihood of Poisson probability function [36], the observation model can be defined, supposing that under the compressed framework, we randomly select a set of pixels $S_{\alpha,l}$ from the P pixels for each wavelength l. The observations models for the baseline intensity is as follows

$$\mathcal{L}_{Y,\alpha}(B) = \sum_{l=1}^{L}\sum_{p\in S_{\alpha,l}} [-\log(b_{p,l})\sigma_{p,l} + T_b b_{p,l}] \tag{4}$$

Where, B represents the concatenation of L matrices $[b_1, ..., b_l, ..., b_L]$ and $B \in \mathbb{R}^{\mathbb{P}\times\mathbb{L}}$ (where, P is the total number of pixels in each wavelength l) and $\sigma_{p,l} = \sum_{t=1}^{T_b} y_{p,t,l}$ is the sum of the response signal for each pixel location. The observation model for response intensity is as follows

$$\mathcal{L}_{Y,\alpha}(A|\hat{B}) = \sum_{l=1}^{L}\sum_{p\in S_{\alpha,l}} [r_{p,l}\sigma_f a_{p,l} + T_a b_{p,l} - \sigma_{p,l}\log(r_{p,l}\sigma_f a_{p,l} + T_a b_{p,l})] \tag{5}$$

Where, $\hat{B}$ is the reconstructed solution matrix of the baseline intensity images over all the images, T_a is the total number of bins (time-steps) in the photon count histogram and again, $\sigma_{p,l} = \sum_{t=t_b+1}^{T_a} y_{p,t,l}$. Similarly to the previous case, A is considered to be a concatenation of L column-vector images, $A \in \mathbb{R}^{P\times L}$.

B. Minimization Problems

We will now see what regularizer terms are appropriate for image reconstruction from Hyper-spectral lidar data. The regularizer terms are some prior information about the scene that are imposed on the observation model as penalty terms. The prior information that is generally used under compressed sensing framework can be some sparsity coefficients in the gradient domain or some wavelet basis. The peak wavelengths at which the scene is imaged are close to one another. For example, for the data used in this work, the peak wavelengths range from 500nm to 820nm at steps of 10nm. Moreover, the surface responsible for generating response at each pixel location is the same over all the wavelengths. This implies that it is sensible to assume a high correlation between the intensity images over different wavelengths in addition to spatial correlation.

1) TV Regularization: The first efficient yet simple algorithms for minimization problems using Total Variation (TV) regularization were developed in the 2000s, [37], [38]. [39], contains a detailed overview on the working and convexity of TV regularization in a convex minimization problem. Total variation describes a phenomenon that is similar to the energy of a signal. Mathematically, it is the integral of the absolute gradient of the signal (or image in this context). We know that the gradient of an image describes the edges. Exactly in the same way Total Variation describes the detail present in an image. The TV regularization in this context will be performed individually over images at each wavelength. For simplicity, it will be called Spectral TV Regularization, STV, and it can be represented as follows

$$STV(X) = \sum_{l=1}^{L} TV(x_l) \tag{6}$$

where, X will correspond to B or A depending on whether we estimate the baseline intensity or the response intensity.

2) Nuclear Norm Regularization: Let us assume a data cube $X \in \mathbb{R}^{N\times L}$, where N is the total number of pixels in each image, imaged over L wavelengths. The nuclear norm over the data cube is represented by $||X||_*$, where $X =$

$U\Sigma V^T$ is the SVD decomposition of X. The nuclear norm works by estimating from the sum of absolute singular values (we have L absolute singular values in this case) over all the imaged wavelengths L. In other words, nuclear norm enhances the spectral correlation present in each row of our data cube X. NN regularization has been shown to be very efficient in the context of multi-spectral radio-interferometry imaging, [40], [41].

3) Joint Sparsity Regularization: A new regularization term that promotes joint sparsity and low-rankness on the data matrix was proposed recently in the context of radio astronomy, [40]. The term 'joint' refers to sparsity in both the spatial and spectral domains simultaneously. This model is based on the assumption that the photons received over the image can be seen to be reflected from a finite number of sources, ρ, each with a different spectral signature. Under this assumption, a linear mixture model $X = SH^\dagger$ can be adopted, where the columns of the matrix $S \in \mathbb{C}^{N\times\rho}$ represent the responsible sources present in the image and the columns of the matrix $H \in \mathbb{C}^{\rho\times L}$ are their corresponding spectral signatures. The rank of the matrix X is given by ρ, that implies low-rankness. Suppose if none of the sources are active at a given pixel location, then one whole row of X will be zero.

C. TVNN Model

This model is formed by combining the TV and Nuclear Norm regularization (hence, the name TVNN). The minimization problem to estimate the intensity images can be written as follows

$$\hat{X} = \arg\min_X [\mathcal{L}_{Y,\alpha}(X) + \tau_1 STV(X) + \tau_2||X||_* + i_{\mathbb{R}^+}(X)] \tag{7}$$

where, $X = B$ and $\mathcal{L}_{Y,\alpha}(X) = \mathcal{L}_{Y,\alpha}(B)$ for baseline intensity reconstruction and $X = A$ and $\mathcal{L}_{Y,\alpha}(X) = \mathcal{L}_{Y,\alpha}(A|\hat{B})$ for response intensity reconstruction, $i_{\mathbb{R}^+}(X)$ is an indicator function enforcing non-negative values on the solution X, and, τ_1 and τ_2 are small positive parameters.

D. Joint Sparsity Model

This minimization problem uses only the joint sparsity regularization. The regularization term can be defined by encapsulating joint-sparsity in some adequate basis Ψ along with an analysis prior based on the $l_{2,1}$ norm. The minimization problem can then be defined as

$$\hat{X} = \arg\min_X [\, \mathcal{L}_{Y,\alpha}(X) + \tau_1||\Psi^\dagger X||_{2,1} + i_{\mathbb{R}^+}(X)\,] \tag{8}$$

where, $||\Psi^\dagger X||_{2,1}$ stands for the component-wise $l_{2,1}$ norm. The notations are similar as in (7) and the likelihood term can be raplaced accordingly for estimating A or B. This regularization term promotes smoothness of the spectral lines along with joint-sparsity in basis Ψ. Ψ can be any wavelet basis in which the data can be assumed to be sparse.

III. Algorithmic Details

As calculating an exact solution of the formulated inverse problem is computationally impossible, optimization can be employed to approximate the solution iteratively. The ADMM algorithm belongs to the primal-dual class of convex optimization algorithms. The ADMM algorithm is not new but it is still widely used due to its accuracy and versatility. [42] has a detailed explanation on a variant of the ADMM algorithm that is able to model cost functions with more than two terms. Under the framework of the ADMM variant, a minimization problem can be written as follows

$$\min_{z\in\mathbb{R}^N} \sum_{j=1}^{J} g_j H^{(j)}(z) \tag{9}$$

where, $g_j(.)$ are closed, proper, convex functions and $H^{(j)} \in \mathbb{R}^{m\times n}$ are arbitrary matrices. This problem can be written in the form of the standard ADMM algorithm with the following considerations

$$f_1 = 0\,;\, G = \left[H^{(1)}\ \ldots\ H^{(J)}\right]^T \in \mathbb{R}^{m\times n} \tag{10}$$

where, $m = m_1 + \ldots + m_J$, and $f_2 : \mathbb{R}^{m\times n} \to \mathbb{R}$ given by,

$$f_2(u) = \sum_{j=1}^{J} g_j(u^{(j)}) \tag{11}$$

where, $u^{(j)} \in \mathbb{R}^{m_j}$ and $u = [(u^{(1)})^T, \ldots, (u^{(J)})^T]^T \in \mathbb{R}^m$. With this definition of the ADMM, the solution for each term in the minimization problem can be computed sequentially as

$$u_{k+1}^{(j)} \leftarrow \arg\min_{v\in\mathbb{R}^{m_j}} g_j(v) + \frac{\mu}{2}\left|\left|v - s_k^{(j)}\right|\right|_2^2 \tag{12}$$

for $j = 1, \ldots, J$, where, $s_k^{(j)} = H^{(j)} z_{k+1} - d_k^{(j)}$. The resulting algorithm for the JS model is shown in Fig. 5. This algorithm can be easily adapted for the TVNN model by replacing the joint sparsity terms with those corresponding to the TV and NN terms.

The solution to the minimization problems shown in the algorithm, such as, the ones at steps 13 and 15 is given by the so called "Moreau proximity operator". For example, the solution to step 15 is given by the soft-thresholding operation, defined for each row k as,

$$\left(\mathcal{S}_\alpha^{l_{2,1}}(Z)\right)_{k,:} \triangleq \begin{cases} \bar{z}\frac{||\bar{z}||_{l_2}-\alpha}{||\bar{z}||_{l_2}} & ||\bar{z}||_{l_2} > \alpha \\ 0 & ||\bar{z}||_{l_2} \le \alpha \end{cases} \quad \forall k \tag{13}$$

where, $\bar{z} = z_k$, is the row k of the input matrix Z and α can be chosen equal to $\frac{2\tau}{\mu}$.

IV. Results

The two methods proposed above were tested on real data consisting of 32 images, each of dimension 190×190 pixels, imaged at peak wavelengths of light ranging from 500nm to 810nm in steps 10nm. Figure 6 shows the setup that was used to acquire this real data. This data was acquired for $10ms$ for each pixel location at each wavelength step. As the reflectivity of the materials depends upon the

1) **Input:** M_α, μ, τ, Σ, σ_f, $\hat{B}$ (only if estimating A)
2) **Result:** $\hat{B}$ or $\hat{A}$
3) Initialize $u_0^{(1)}$, $u_0^{(2)}$, $u_0^{(3)}$, $d_0^{(1)}$, $d_0^{(2)}$, $d_0^{(3)}$,
4) Pre-compute $(M_\alpha^T M_\alpha + 2I)^{-1}$
5) Set $k = 0, \mu > 0, \tau > 0$
6) **while** $||z_k - z_{k-1}||_2/(min(||z_k||_2, ||z_{k-1}||_2) + \epsilon)$ AND $k \leq k_{max\_iter}$ **do**
7) $\quad \zeta_k^{(1)} \leftarrow u_k^{(1)} + d_k^{(1)}$
8) $\quad \zeta_k^{(2)} \leftarrow u_k^{(2)} + d_k^{(2)}$
9) $\quad \zeta_k^{(3)} \leftarrow u_k^{(3)} + d_k^{(3)}$
10) $\quad \gamma_k \leftarrow M_\alpha^T \zeta_k^{(1)} + P^T \zeta_k^{(2)} + \zeta_k^{(3)}$
11) $\quad z_{k+1} \leftarrow (M_\alpha^T M_\alpha + 2I)^{-1} \gamma_k$
12) $\quad \nu_k^{(1)} \leftarrow M_\alpha z_{k+1} - d_k^{(1)}$
13) $\quad u_{k+1}^{(1)} \leftarrow \arg\min_v \frac{\mu}{2}||v - \nu_k^{(1)}||_2^2 + \mathcal{L}_{Y,\alpha}(v)$
14) $\quad \nu_k^{(2)} \leftarrow P z_{k+1} - d_k^{(2)}$
15) $\quad u_{k+1}^{(2)} \leftarrow \arg\min_v \frac{\mu}{2}||v - \nu_k^{(2)}||_2^2 + \tau||\Psi^\dagger v||_{2,1}$
16) $\quad \nu_k^{(3)} \leftarrow z_{k+1} - d_k^{(3)}$
17) $\quad u_{k+1}^{(3)} \leftarrow max(0, \nu_k^{(3)})$
18) $\quad d_{k+1}^{(1)} \leftarrow d_k^{(1)} - (M_\alpha z_{k+1} - u_{k+1}^{(1)})$
19) $\quad d_{k+1}^{(2)} \leftarrow d_k^{(2)} - (P z_{k+1} - u_{k+1}^{(2)})$
20) $\quad d_{k+1}^{(3)} \leftarrow d_k^{(3)} - (z_{k+1} - u_{k+1}^{(3)})$
21) $\quad k \leftarrow k + 1$
22) **end**
23) $\hat{B} \leftarrow z_k$ or $\hat{A} \leftarrow z_k$

Fig. 5: Variant of the PIDAL-FA algorithm, [42]

wavelength of the light pulse, the average photon counts of the measurements over the data varies from $\sim$ 230 to $\sim$ 1800 photons per pixel depending on the wavelength used for the image. To evaluate the results quantitatively, the reconstructions, from data at full measurements, obtained using the TVNN method were used as the ground truth images. These reconstructions were obtained with very low values for the two regularization parameters (see (7)) to keep their impact low. The Signal to Noise Ratio (SNR) metric was used to quantitatively evalute the reconstructions that can be given as

$$SNR = 10 \log \left(\frac{||\hat{x}||_2^2}{||\hat{x} - x||_2^2} \right) \tag{14}$$

where, $\hat{x}$ is the ground truth image matrix ($\hat{x} \in \mathbb{R}_+^{N \times L}$) and x is the estimated intensity image matrix of same dimensions, with N as the number of pixels in each image and L as the total number of wavelengths, and, $||.||_2^2$ represents the squared l_2 norm.

Table I shows the correspondences between different average photon counts over the data and the corresponding sub-sampling ratios. The relation is that when the average photon counts per pixel over each image is doubled, the number of considered pixels in the image is halved. The joint sparsity model was tested using Discrete Cosine Transform (DCT) and Daubechies Wavelet Transform (DWT) as two separate sparsity bases. The three models will hence, be called TVNN, JS-DCT and JS-DWT for simplicity.

Fig. 6: The setup that was repeatedly imaged over different peak wavelengths

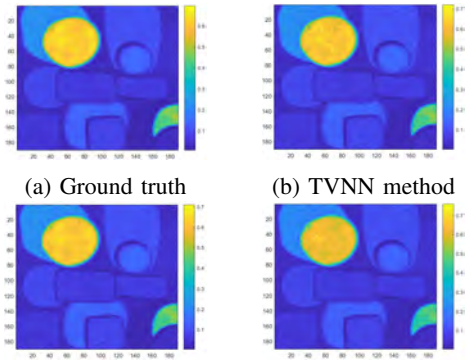
TABLE I: Table showing the correspondence of sub-sampling ratio (α) with the mean photons per pixel (ppp) per each band of wavelength. 1^{st} and 3^{rd} columns show the average number of photons per pixel

before sampling	**s**ub-sampling ratio	**a**fter sub-sampling
0.5, 1, 2, 4, 8	1, 1/2, 1/4, 1/8, 1/16	0.5
1, 2, 4, 8, 16	1, 1/2, 1/4, 1/8, 1/16	1
10, 20, 40, 80, 160	1, 1/2, 1/4, 1/8, 1/16	10
50, 100, 200	1, 1/2, 1/4	50
100, 200	1, 1/2	100

Table II shows the SNR values for the reconstructions from the data with different average photon counts per pixel at different levels of sub-sampling. The table is divided into cells of three rows each corresponding to the three proposed methods for some fixed average number of photons per pixel (first column). The best reconstruction for each cell is highlighted as the bold SNR value. It can be observed from the highlighted values that JS-DWT method has the best reconstructions out-performing the TVNN method by a margin, specially, in the case low measurements ($\leq$ 10 ppp). The improvement with compressed sensing can be seen at low observations (shown with an underline highlighting the best value of that row). It can be observed that in some

TABLE II: Table showing the SNR (dB) values for the reconstructions obtained using the proposed methods for different average values of photons per pixel (rows) against different sub-sampling ratios (columns).

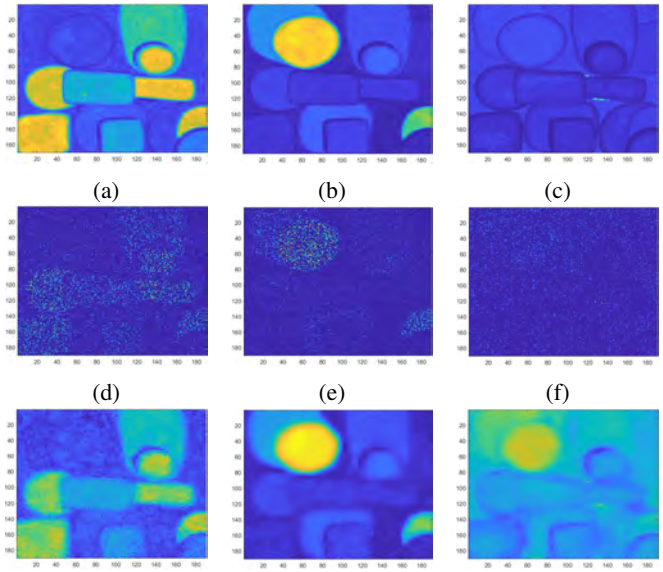
ppp \ α		1	1/2	1/4	1/8	1/16
100	TVNN	53.08	50.48			
	JS-DCT	53.59	49.89			
	JS-DWT	**54.59**	51.16			
50	TVNN	46.74	48.69	46.33		
	JS-DCT	47.95	47.27	42.68		
	JS-DWT	**48.99**	47.97	43.79		
10	TVNN	34.38	41.33	42.79	39.75	35.69
	JS-DCT	37.58	42.12	40.22	36.03	31.69
	JS-DWT	39.29	**43.14**	41.43	37.34	27.83
1	TVNN	22.65	27.77	34.42	**35.89**	32.73
	JS-DCT	28.25	30.64	31.06	30.34	28.01
	JS-DWT	27.22	27.31	27.41	27.15	21.87
0.5	TVNN	20.91	21.32	22.33	22.53	23.09
	JS-DCT	26.68	26.95	26.89	26.78	26.16
	JS-DWT	26.70	28.75	**29.67**	28.99	24.64



(a) Ground truth (b) TVNN method

(c) JS-DWT method (d) JS-DCT method

Fig. 7: Response Intensity image reconstruction results using the three methods at high measurements (~ 100 photons per pixel per band) at the 15^{th} band



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 8: Response Intensity image reconstruction results. The columns correspond to the $6^{th}, 15^{th}$ and 32^{nd} wavelength bands respectively. Row 1 ((a)-(c)) shows the ground truth images, Row 2 ((d)-(f)) shows the measurements corresponding to an average of ~ 0.5 photons per pixel per band collected only on $1/4^{th}$ of the total number of pixel locations, and, Row 3 ((g)-(i)) shows the reconstructions using the JS-DWT method.

cases, it is better to consider fewer samples (upto $1/8^{th}$ of the total number of pixels) than to measure data at all the pixel locations.

Figure 7 shows the reconstruction of the response intensity images from the data with an average of 100 photons per pixel per band. The varied effect of smoothness from the three different methods can be seen from those images. It can also be noted that the three methods have equally appealing results, visually. However, the difference between the three methods is more evident at low observations.

It has been observed that data with ~ 0.5 photon per pixel per band is the limit to obtain visually satisfactory results. The methods start to collapse when the measurements go lower than that. Figure 8 shows that the joint sparsity method performs very well at all the wavelengths even from data with extremely low measurements.

V. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

Firstly, an in-painting method has been successfully implemented for a Poisson noise model by minimizing the negative log-likelihood of the observed data. The intensity image estimation problem was efficiently modelled by separating the estimation process into two sequential steps of estimating the baseline intensity and response intensity. The proposed minimization problems have shown to be able to model the data appropriately, specially, when the data has observations consisting of an average of <10 photons per pixel. The methods have shown to perform extremely well while reconstructing intensity images at all the considered wavelengths, simultaneously. Lastly, it was shown that the proposed observation model and the minimization problems support the compressed sensing framework.

B. Future Works

This work did not make any contributions towards the estimation of the depth profile of the scene. So, further research can be conducted in the direction of estimating the depth images of the hyper-spectral SPL data, while keeping in mind that estimating the depth at each pixel location poses a non-convex problem. Although, depth estimation is non-convex, the reconstructed response intensity images should facilitate depth reconstruction. As this work obtains good reconstructions of the response intensity images, future research can focus on material classification using the reconstructed response intensity images. By looking at the quality of the reconstructed response intensity images and the contrast between reflectivity of different objects present in the scene, material classification can be easily achieved even at low-light observations.

REFERENCES

[1] C. Weitkamp, *Lidar: range-resolved optical remote sensing of the atmosphere*. Springer Science & Business, 2006, vol. 102.

[2] J. Goldsmith, F. H. Blair, S. E. Bisson, and D. D. Turner, "Turn-key raman lidar for profiling atmospheric water vapor, clouds, and aerosols," *Applied Optics*, vol. 37, no. 21, pp. 4979–4990, 1998.

[3] A. Ansmann, M. Riebesell, and C. Weitkamp, "Measurement of atmospheric aerosol extinction profiles with a raman lidar," *Optics letters*, vol. 15, no. 13, pp. 746–748, 1990.

[4] G. Zhou, C. Song, J. Simmers, and P. Cheng, "Urban 3d gis from lidar and digital aerial images," *Computers & Geosciences*, vol. 30, no. 4, pp. 345–353, 2004.

[5] K. Zhang, J. Yan, and S.-C. Chen, "Automatic construction of building footprints from airborne lidar data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 9, pp. 2523–2533, 2006.

[6] A. Maccarone, A. McCarthy, X. Ren, R. E. Warburton, A. M. Wallace, J. Moffat, Y. Petillot, and G. S. Buller, "Underwater depth imaging using time-correlated single-photon counting," *Optics express*, vol. 23, no. 26, pp. 33 911–33 926, 2015.

[7] N. Cadalli, P. J. Shargo, D. C. Munson, and A. C. Singer, "Three-dimensional tomographic imaging of ocean mines from real and simulated lidar returns," in *Ocean Optics: Remote Sensing and Underwater Imaging*, vol. 4488. International Society for Optics and Photonics, 2002, pp. 155–167.

[8] M. A. Lefsky, W. B. Cohen, G. G. Parker, and D. J. Harding, "Lidar remote sensing for ecosystem studies: Lidar, an emerging remote sensing technology that directly measures the three-dimensional distribution of plant canopies, can accurately estimate vegetation structural attributes and should be of particular interest to forest, landscape, and global ecologists," *AIBS Bulletin*, vol. 52, no. 1, pp. 19–30, 2002.

[9] M. A. Wulder, C. W. Bater, N. C. Coops, T. Hilker, and J. C. White, "The role of lidar in sustainable forest management," *The Forestry Chronicle*, vol. 84, no. 6, pp. 807–826, 2008.

[10] D. Martinez-Ramirez, G. Buller, A. McCarthy, X. Ren, A. W. S. Morak, C. Nichol, and I. Woodhouse, "Developing hyperspectral lidar for structural and biochemical analysis of forest data," in *Proc. EARSEL Conf. Adv. Geosci.*, 2012, pp. 1–11.

[11] J. De Bruijne, R. Kohley, and T. Prusti, "Gaia: 1,000 million stars with 100 ccd detectors," in *Space Telescopes and Instrumentation 2010: Optical, Infrared, and Millimeter Wave*, vol. 7731. International Society for Optics and Photonics, 2010, p. 77311C.

[12] G. Finger, R. J. Dorn, S. Eschbaumer, D. Ives, L. Mehrgan, M. Meyer, and J. Stegmeier, "Infrared detector systems at eso," in *Workshop Detectors for Astronomy*, 2009.

[13] K. J. Moore, S. Turconi, S. Ashman, M. Ruediger, U. Haupts, V. Emerick, and A. J. Pope, "Single molecule detection technologies in miniaturized high throughput screening: fluorescence correlation spectroscopy," *Journal of biomolecular screening*, vol. 4, no. 6, pp. 335–353, 1999.

[14] A. Esposito, T. Oggier, H. Gerritsen, F. Lustenberger, and F. Wouters, "All-solid-state lock-in imaging for wide-field fluorescence lifetime sensing," *Optics express*, vol. 13, no. 24, pp. 9812–9821, 2005.

[15] F. Müller and C. Fattinger, "Exploiting molecular biology by time-resolved fluorescence imaging," in *Single-Photon Imaging*. Springer, 2011, pp. 329–344.

[16] G. Buller and A. Wallace, "Ranging and three-dimensional imaging using time-correlated single-photon counting and point-by-point acquisition," *IEEE Journal of selected topics in quantum electronics*, vol. 13, no. 4, pp. 1006–1015, 2007.

[17] P. Seitz and A. J. Theuwissen, *Single-photon imaging*. Springer Science & Business Media, 2011, vol. 160.

[18] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 83–91, 2008.

[19] D. Shin, J. H. Shapiro, and V. K. Goyal, "Single-photon depth imaging using a union-of-subspaces model," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2254–2258, 2015.

[20] J. H. Shin, Dongeek & Shapiro and V. K. Goyal, "Computational single-photon depth imaging without transverse regularization," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 973–977.

[21] Y. Altmann, A. Maccarone, A. McCarthy, G. Newstadt, G. Buller, S. McLaughlin, and A. Hero, "Robust spectral unmixing of sparse multispectral lidar waveforms using gamma markov random fields," *IEEE Transactions on Computational Imaging*, 2017.

[22] Y. Altmann, A. Maccarone, A. Halimi, A. McCarthy, G. Buller, and S. McLaughlin, "Efficient range estimation and material quantification from multispectral lidar waveforms," in *Sensor Signal Processing for Defence (SSPD), 2016*. IEEE, 2016, pp. 1–5.

[23] D. Shin, A. Kirmani, V. K. Goyal, and J. H. Shapiro, "Computational 3d and reflectivity imaging with high photon efficiency," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 46–50.

[24] A. Shin, Dongeek & Kirmani, V. K. Goyal, and J. H. Shapiro, "Photon-efficient computational 3-d and reflectivity imaging with single-photon detectors," *IEEE Transactions on Computational Imaging*, vol. 1, no. 2, pp. 112–125, 2015.

[25] P. Sarder and A. Nehorai, "Deconvolution methods for 3-d fluorescence microscopy images," *IEEE Signal Processing Magazine*, vol. 23, no. 3, pp. 32–45, 2006.

[26] J.-L. Starck and F. Murtagh, *Astronomical image and data analysis*. Springer Science & Business Media, 2007.

[27] N. Dey, L. Blanc-Feraud, C. Zimmer, P. Roux, Z. Kam, J.-C. Olivo-Marin, and J. Zerubia, "Richardson–lucy algorithm with total variation regularization for 3d confocal microscope deconvolution," *Microscopy research and technique*, vol. 69, no. 4, pp. 260–266, 2006.

[28] J.-L. Starck, F. Murtagii, and A. Bijaoui, "Multiresolution support applied to image filtering and restoration," *Graphical models and image processing*, vol. 57, no. 5, pp. 420–431, 1995.

[29] R. Nowak and E. D. Kolaczyk, "A bayesian multiscale framework for poisson inverse problems," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 3. IEEE, 1999, pp. 1741–1744.

[30] K. E. Timmermann and R. D. Nowak, "Multiscale modeling and estimation of poisson processes with application to photon-limited imaging," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 846–862, 1999.

[31] R. M. Willett and R. D. Nowak, "Platelets: a multiscale approach for recovering edges and surfaces in photon-limited medical imaging," *IEEE Transactions on Medical Imaging*, vol. 22, no. 3, pp. 332–350, 2003.

[32] R. D. Willett, Rebecca M & Nowak, "Fast multiresolution photon-limited image reconstruction," in *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*. IEEE, 2004, pp. 1192–1195.

[33] G. S. Buller, R. D. Harkins, A. McCarthy, P. A. Hiskett, G. R. MacKinnon, G. R. Smith, R. Sung, A. M. Wallace, R. A. Lamb, K. D. Ridley *et al.*, "Multiple wavelength time-of-flight sensor based on time-correlated single-photon counting," *Review of Scientific Instruments*, vol. 76, no. 8, p. 083112, 2005.

[34] W. Becker, *Advanced time-correlated single photon counting techniques*. Springer Science & Business Media, 2005, vol. 81.

[35] T. Yamane, "Statistics: An introductory analysis," 1973.

[36] F. A. Haight, "Handbook of the poisson distribution," 1967.

[37] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical imaging and vision*, vol. 20, no. 1-2, pp. 89–97, 2004.

[38] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2419–2434, 2009.

[39] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, "An introduction to total variation for image analysis," *Theoretical foundations and numerical methods for sparse recovery*, vol. 9, no. 263-340, p. 227, 2010.

[40] A. Abdulaziz, A. Dabbech, A. Onose, and Y. Wiaux, "A low-rank and joint-sparsity model for hyper-spectral radio-interferometric imaging," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 388–392.

[41] M. Golbabaee and P. Vandergheynst, "Compressed sensing of simultaneous low-rank and joint-sparse matrices," *arXiv preprint arXiv:1211.5058*, 2012.

[42] M. A. Figueiredo and J. M. Bioucas-Dias, "Restoration of poissonian images using alternating direction optimization," *IEEE transactions on Image Processing*, vol. 19, no. 12, pp. 3133–3145, 2010.

Robot-Assisted Game Incorporating Emotional Postures for Children with ASD

D. Stoeva
VIBOT,
Heriot Watt University,
University of Burgundy,
University of Girona
darja.stoeva@gmail.com

G. Rajendran
Psychology,
Heriot Watt University,
Edinburgh, Scotland, UK
t.rajendran@hw.ac.uk

M. Dragone
Engineering
and Physical Sciences,
Heriot Watt University,
Edinburgh, Scotland, UK
m.dragone@hw.ac.uk

E. M. Suphi
Engineering
and Physical Sciences,
Heriot Watt University,
Edinburgh, Scotland, UK
m.s.erden@hw.ac.uk

***Abstract*— Joint attention has been shown to have an important impact on the development of the social cognition in children. Children with autism spectrum disorder (ASD) are especially affected by the lack of joint attention skills. Few studies have shown that early interventions and improvement of these skills in the early child development can improve the social skills in the long term. As a consequence research in using robots as therapy tools providing aid in the development of social skills for children with ASD has increased over the past few years. The goal of this study is to design and implement an interactive game with the humanoid robot Pepper, which incorporates the two most basic emotional postures, happiness and sadness, as an engaging and motivating factor for the children. Furthermore, the game is designed to improve social skills, more specifically the initiation of joint attention in pre-school children with ASD. In total twenty volunteers participated in this study to evaluate the proposed method, 13 played the game and 7 watched a promotional video of the game. The obtained results from the questionnaires indicated that emotional postures in interactive game are beneficial for increasing engagement and motivation in users. However, it was difficult to conclude whether the designed game could promote the behaviour of initiating joint attention in children with ASD.**

I. INTRODUCTION

Socially assistive robotics (SAR) aims at developing robots that provide assistance to the user through social interaction rather than physical contact [1]. These robots have a wide variety of application domains and so far they have been used for physical disabilities and rehabilitation such as post-stroke [2, 3] and children with cerebral palsy [4] in cognitive and behavioural disorders as interactive therapy [5, 6], in elderly to reduce stress and depression [7] and as educational tools for students [8]. SAR has become increasingly popular research area over the past decade mostly because it has a great potential to encourage independence and improve the quality of life for people with disabilities [9].

A. Autism Spectrum Disorder

From the cognitive and behavioural disorders, the most researched area is therapy in the autism spectrum disorder (ASD). ASD is a term defining a group of few neurodevelopmental disorders such as autism and Asperger's syndrome, that are characterized by deficits in social interaction, problems with verbal and non-verbal communication, and restricted, repetitive behaviours [10]. The exact cause of the disorder is still not known, although the most recent research indicates that a possible cause is the formation of too many synapses due to the RNF8 gene [11]. Consequently, a medical test has not been established, thus the diagnosis is based on the specialist's observations of child's development and behaviour, more specifically social skills such as facial expressions, body postures and gestures, and eye contact [10]. Children with ASD usually fail to identify other's emotional states, fail to make eye contact and have difficulties to follow social norms [12]. Often but not necessarily, children with ASD have to some degree a language deficit which can be in the context of verbal communication absence or intense talking without giving a chance to the other speaker. It is difficult to say at what age individuals get diagnosed since it differs from one person to another. The aforementioned symptoms can be indicators that can be noticed by the parents in the first 18 months the earlies [13]. Early diagnosis and treatment have shown to improve the quality life and independence in individuals with ASD [12].

1) Joint Attention and ASD: Recently a lot of research has been focused on the impairment in the development of joint attention (JA) is children with ASD, as the earliest indicator of the disorder [14]. JA refers to the person's ability to share attention with a social partner on the same object or event [15, 16]. In order to do so the individual exploits social behaviours such as pointing, eye gazing, articulation, and gestures to express interest in sharing an information with another individual. There are two categories of JA skills: initiating joint attention (IJA) - the ability to want to share interest with others, and responding to joint attention (RJA) - ability to follow other's gaze and gestures. For example, when someone wants to direct the attention of another person to a specific object, in order to coordinate the attention of the other, they will point towards the object, IJA, and the other person will attend the object by focusing their gaze to the object, RJA. So both of them will focus their attention both on the object and between each other at the same time.

The behaviours of JA facilitate social learning which makes JA essential or more like a prerequisite for the development of language skills such as intentional communication, social cognition and theory of mind [17]. Typically developing infants exhibit behaviours of JA in the first 4 to 18 months [14], but children with ASD have impairments in JA skills, more specifically deficits in the development of the

IJA behaviour rather then the RJA [16]. If ASD is diagnosed early, the research on JA intervention demonstrates that early intervention can improve the JA skills which as a result will also improve the development of social and communication skills [18, 19].

2) Social Motivation and ASD: As a consequence of the impairments in social interaction which are considered to be the core problem in ASD, many different theories on the cause of social deficits have been developed. The social motivation theory of autism states that the social deficits in individuals with ASD are in close relation to issues with their social motivation [20]. Social motivation refers to the people's drive to engage in social interaction and usually individuals who lack social motivation may not make eye contact with other people, they may not interact or learn from other. All of these behavioural indicators fall under the characterized symptoms of individuals diagnosed with ASD. The social motivation theory explains that children with ASD experience decrease in their motivation for social interactions where the main cause is deficits in the brain's reward pathway [21]. To be more precise, they have irregularities in the circuit transmitting reward signals which is responsible for the dysfunctional 'wanting' circuit [21]. Social reward is one of the socio-biological mechanism that belongs to the social motivation framework. There are two main composite elements in social reward: 'wanting' referring to striving or having motivation for something regardless of the reward, and 'liking' referring to the feeling of pleasure caused from the reward. Children with ASD have dysfunctional 'wanting' circuit as opposed to the typically developing children [21]. They exhibit 'liking' behaviours, but they lack motivation to manifest 'wanting' behaviours.

B. Robots in ASD Therapy

When it comes to social interaction and behaviour, interacting with humans represents the best type for learning social skills. However, children with ASD show more interest in interacting with objects rather than humans, possibly because the social actions of humans are unpredictable while children with ASD prefer routines and repetitive behaviours which make them feel comfortable and calm [22]. Interestingly, many studies [23–28] have shown that children with ASD react positively when interacting with humanoid robots, triggering a boost in engagement and attention. Therefore the emerging technology of social robotics has a great potential to deliver social behaviours and interactions that will allow and encourage children with ASD to comfortably identify and improve targeted skills. Currently in the field of robots in ASD therapy there are methods that have been developed and evaluated by testing, but there are also studies where a method has been just proposed as a potential therapy on the basis of existing treatment techniques. In addition this field is still being researched in order to reach the full-potential of SAR and to provide the necessary therapy.

Stanton et al. examined whether the robotic dog AIBO could provide help in the development of social skills in children with ASD [29]. Their method consisted of comparison of how children interacted with the robotic dog and a toy dog. In this study 11 children participated with age range from 5 to 8 and it was reported that the children showed more engagement and increased interaction with the robotic dog. On the other hand, as part of the AuRoRa project focused on robots as tools for therapy in ASD, Robins at el. investigated whether robots can be used as therapeutic or educational toys using the humanoid-toy robot Roberta [23] . In a long-term study, they aimed to decrease anxiety in children with ASD and encourage interaction skills such as eye-contact, imitation, turn-taking, while slowly increasing the unpredictability of the robot's actions. Four children with ASD from 5 to 10 years old participated in this study and the results showed that the children perceived the robot as salient object which resulted in increased interaction.

Duquette et al. studied how can the robot Tito facilitate interaction such as imitative play in children with ASD [30]. One group of children were interacting with a human and another with the robot. The imitation play was focused on facial expressions, body movements, and familiar actions with and without objects (such as pointing at objects or waving hello). The experimental phase was performed with four children with ASD and it was reported that the children interacting with the human showed higher imitation of body movements and familiar actions, but the children interacting with the robot showed increased shared focused attention in all types of imitation play especially in imitating facial expressions. This observation is very important for the future development of robots in ASD therapy, since children with ASD usually avoid eye contact and have difficulties understanding facial expressions when interacting with humans. In [24] they developed a similar system like the one designed in this study. Aiming to promote JA skills, they proposed an interactive robot-assisted system using the HOAP-3 robot, for children with ASD using eye gazing as social behaviour, where in this study pointing was used as a behaviour to IJA. Regarding the JA skills their method consisted of two parts, first the robot tried to attract the child by introducing itself, dancing etc., and once the child was interested, the robot continued the interaction with a JA task. The JA task was designed in such a way that the robot would point out an object and detect the child's eye gaze. If the child managed to successfully achieve JA with the robot, the robot responded with joyful motion. They have tested their system with 5 children (10-11 years old) and reported that majority of time, the children interacted with the robot with attention, either focusing on the robot's face when static or focusing on its movements when in motion. Finally they have concluded that their approach showed that robots are capable of providing motivation and a socially learning environment for children with ASD.

From the literature reported, it is clear that children with ASD show increased motivation and interaction when in contact with robots. Accordingly, it can be observed that robots definitely have a great potetnial to aid children with ASD in the development of their social and communication skills. In addition the data from studies that are focused on

JA and how can robots deliver JA behaviours shows that robots are capable of of providing JA.

This study is focused on designing and implementing an interactive game with a humanoid robot that conveys emotional postures aiming to improve social skills in children with ASD. More specifically, engage children with ASD in game to aid them in developing and improving the skills of IJA while also incorporating emotional postures as a social reward as part of the social motivation framework. This rises the following questions: What are the possible benefits of emotional postures in an interactive game with a humanoid robot? How can social robots promote the initiation of joint attention in children with ASD through an interactive game? Can emotional postures be considered as a social motivation in an interactive game that aims to promote initiating joint attention?

II. METHODS

To investigate the proposed research questions, two methods were developed. To test the impact of the emotional postures on the game, whether they provide motivation to the user, the human-robot interaction and user's experience, the designed game was tested with participants who played the same game once with emotional postures and once without. To evaluate the designed system in terms of whether it will encourage JA in children with ASD and whether the emotional postures can be perceived as social reward, a 'promotional' video of the game was made.

A. Pepper Robot

For the physical appearance of the robots used in autism research, humanoid robots are reported to be more favourable, as Scassellati et al. explain, the resemblance of a human might help the child with ASD to not only recognize the social cues that the robot is projecting but also to identify them in social environments [31]. Pepper, the robot used in this research, is a humanoid robot produced by the Aldebaran Robotics Company (see Figure 1), which is 1.21 m tall and 0.48 m wide, has 20 degrees of freedom in its head, arms and back. It has two HD cameras on the forehead used for perception and one 3D camera behind the eyes used for localization and navigation, and a tablet attached to his chest. The operating system of the robot is NAOqi OS and the available APIs are the NAOqi API, SDK Python, C++, and JavaScript for the tablet. In this research the NAOqi API was used with the graphical user interface Choregraphe Suite.

B. Game Desgin

When designing the game the following aspects from the ASD symptomatology were taken into account: inability to socially interact, avoidance of eye contact, language difficulties and difficulties understanding emotions and associating them contextually. The game consists of a quiz-like structure, where Pepper asks a general knowledge question and waits for the user to point to the correct answer. The questions were inspired from several websites which offer example of questions appropriate for 2 to 5 years old children [32–34].

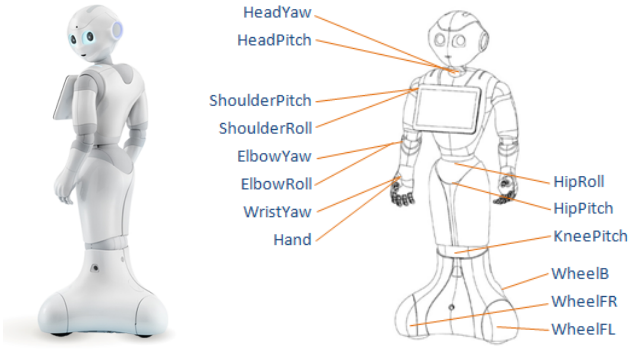


Fig. 1: The Pepper robot and the joint angles

The speech recognition feature of the robot was avoided due to speech and language difficulties present in many children with ASD, so instead the tablet was used to start and stop the game. When the robot has finished introducing itself and explaining the instructions of the game, a click-to-start image is shown on the tablet. Before the robot starts with the questions it explains to the user that the game can be stopped at any time by pressing on the tablet which is showing a click-to-stop image. Then the robot begins with asking the pre-formulated questions. Each time a question has been asked, two images appear on the monitors, one correct and one incorrect. The images were taken from the internet and were resized to fit the screen and also to show written word describing the object shown. For each question the user has to answer by pointing to (IJA) one of the screens. Once pointing has been detected, Pepper looks at the screen to which the user has pointed at (RJA) and according to the answer performs a corresponding posture and then says whether the answer is correct or not. When the answer is incorrect Pepper shows a sad posture, and when the answer is right a happy posture. When the game has finished, the robot does a verbal expression of gratitude by thanking the user for playing and goes to sleep.

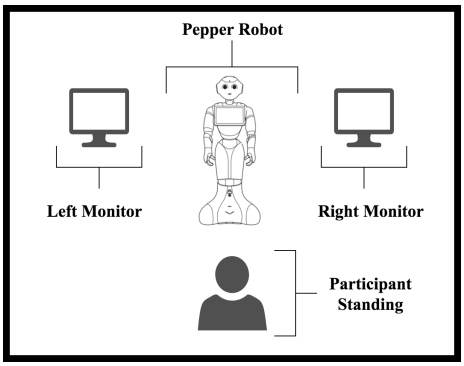


Fig. 2: Framework of the game setting showing the positions of Pepper, the monitors and the participant

C. Emotional Postures

Postures were incorporated in the game as an engaging factor for the children with ASD, but also as a social

motivation. Having a quiz-like structure where only one answer has to be chosen from two offered leaves room only for two emotional behaviours. Accordingly the two considered emotions were happiness and sadness. Due to a lack of emotional postures implemented for the Pepper robot, the postures used in this study were adapted from tested emotional postures for NAO. Both, NAO and Pepper, have the same range of the joint angles for the head and arms, the only difference is their legs. Pepper has one hip and knee with total of 3 DOF (see Figure 1) while NAO has two legs and each has hip, knee and ankle with 11 DOF in total and higher range. When the posture adapted from NAO had different values for each leg, the average of the value was used for the hip and knee values in Pepper. Seven different postures were implemented for both sadness and happiness taken from three different research studies [35–37].

D. Participants

Twenty individuals participated in this study for both, playing the game and watching the video of the game. Thirteen participants, 12 male and 1 female (mean age = 25.23, std = 3.72) volunteered to take part in this research and play the game. All participants were recruited from the Heriot Watt University and were either master, PhD or postdoc students. The majority of them had either a background or some knowledge in Robotics, from which two have worked on Social Robotics but none has worked on therapy in cognitive and behaviour disorder. The experimental procedure was approved by the ethics committee at Heriot Watt University and prior to the game testing, the participants were provided with written consent form. Seven participants volunteered to watch the 'promotional' video of the game and fill a questionnaire. Three of these participants were parents, three special educators and one research student. All of them have worked or been in contact with children with ASD in a range from 2 to 16 years.

E. Procedure

For the first part, where the participants were playing the game, the testing took place in the Earl Mountbatten building at Heriot Watt University, Edinburgh, in the EM1.50 room. Once the participants have read the information sheet and signed the consent form they started playing the interactive game. To avoid any bias towards the role of the emotional postures on the game, two versions of the game were played and the questions asked by the robot were the same for both versions. The two versions of the game were: one without emotional postures, and one with emotional postures used as a social reward as part of the social motivation framework. Both of the games were played in a continues flow where once the first version of the game has finished, the robot informs the participant that in 5 seconds they will start playing the second version of the game. Six of the participants first played the version with emotional postures and then the version without, and seven of them vice versa. There were 15 general knowledge questions in total per version, and the total participation time was approximately 10 minutes. The participants were told that the answers were not scored and they were encouraged to also give wrong answers if they wished. The participants filled a questionnaire with questions regarding the effect of the emotional postures on the whole experience.

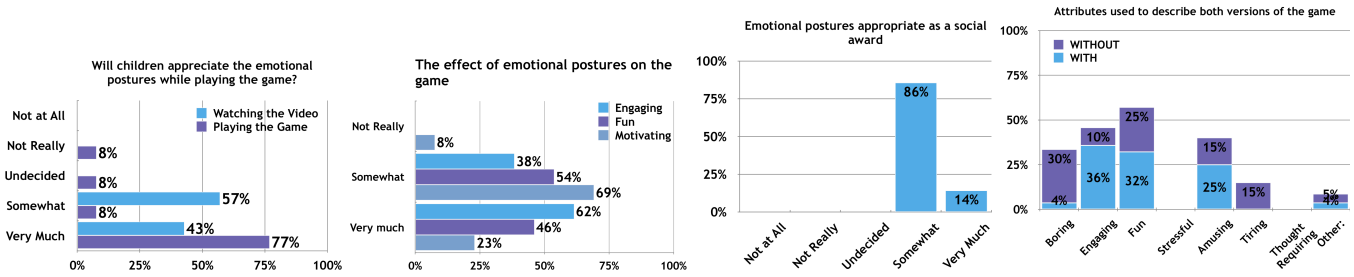
For the second part, several ASD organisations and schools for children with ASD were contacted and asked to participate in the study and it was explained that the study is open for anyone who has been in any relation with children with ASD (e.g. parent, teacher, researcher, etc.). They were send a link containing a short description of the aim of the game, the video and the questionnaire divided in 3 sections: ASD, Joint Attention and Emotional Postures each with regard to the game. The questionnaire investigated two aspects, namely whether emotional postures can be beneficial for children with ASD and if so in which way. Two societies, Lothian Autistic Society and Autism Berkshire, responded positively and shared the study on their Facebook page.

III. RESULTS AND DISCUSSION

The questionnaires use for evaluation contained multiple choice, Likert scale and open questions. For each of the Likert 5-scale questions in the questionnaires these bar charts a Chi-Square test of significance was performed and the p–value was determined (with significance level $p = 0.05$). The p–values for the questionnaire data from the participant that played the game were all found to be statistically significant. On the other hand there were three questions from the video questionnaire which have a p–value greater than 0.05. For multiple choice questions where 2 or 3 options were given and the open questions are reported without visual representation. Three of the open answer questions from the questionnaire for the participants watching the video are reported in Section IV-B, since they were regarding possible improvements of the system and possible problems that might happen if a child with ASD would have played the game.

A. Benefits of The Emotional Postures

The participants playing the game were asked questions regarding the emotional postures including specific questions on the effects of the emotional postures on the overall experience (see Figure 3b), to describe the game with and without emotional postures with already offered attributes (see Figure 3d) and lastly, a question on which version did they prefer and why. Figure 3b shows that all of the participants thought that the emotional postures make the game somewhat more engaging, fun and motivating or entirely more engaging, fun and motivating. Only one participant found the emotional postures not so motivating. The p–value for the three questions is $p << 0.05$, making the obtained answers statistically significant. Taking the average for each question indicates that the participants found the emotional postures very useful to make the game more engaging, more fun and motivating. When asked to describe both versions of the game, the one without emotional postures was described with the attribute "Boring" with highest percentage 46.3%. Meanwhile the



(a) (b) (c) (d)

Fig. 3: Answers from the questionnaire of the participants playing the game, asking whether they found the emotional postures engaging, fun and motivating

version with emotional postures was described the most with the attribute "Engaging", more than half of the participants (76.9%). Both games were described as "Fun" but the version with emotional postures exceeds with 69.2%. These observations can relate back to the first two questions evaluated, which confirms the benefit of using emotional posture to make the game more engaging and fun. Another thing to note regarding the comparison of the two versions, is that few participants (23.1%) described the version without emotional postures as "Tiring" while more than half described the version as "Amusing". These two are the only attributes that were uniquely assigned to each version of the game that emphasize the quality of emotional postures in terms of bringing fun and engagement in the game. It was noticed that during the experiments, the participants were getting bored and tired from the game since they were answering the same questions for both version. However, only the version without emotional postures was described as "Tiring". Almost all of the participants (92.3%) preferred the version with emotional postures over the one without, except one participant who did not show any preference. This question was followed by an open answer question asking the participants to explain why did they prefer this particular version. The answers can be summarized as follows: the interaction was more fun and amusing, the feedback (emotional postures) brings more excitement to the user, the game was more engaging, and the robot seemed more engaged.

Participants that watched the video including both versions of the game, were asked questions regarding the usage of emotional postures to improve JA skills, or as a social reward in the motivational framework, and whether children will appreciate them while playing the game. The data regarding JA was found to be statistically insignificant and therefore it si not discussed. However, when asked whether the emotional postures are appropriate as a social award 85.7% answered with somewhat and 14.3% with very much ($p << 0.05$). Since these answers are statistically significant, the average indicates that the participants found the emotional postures somewhat appropriate as a social award.

B. Promoting IJA in Children with ASD

When both groups of participants were asked whether children would appreciate the emotional postures while playing the game, both groups either somewhat or very much. However, one participant that played the game answered as undecided and another as not really. The answers for both groups were statistically significant meaning that both groups agree that the children will very much appreciate the emotional postures.

When JA behaviour is provided by a robot it is very important for the robot to be seen as a social partner. More than half of the participants (61.5%) that played the game found the game very encouraging to point in order to show an answer ($p << 0.05$) and 53.8% thought they would not experience the game in the same way if the robot did not look at the answer they were showing. These findings suggest that people experienced IJA especially because when asked whether they felt like playing alone or with a social partner, 61.5% answered with a social partner.

On the other hand, the participants watching the video had to answer more specific questions regarding the JA involved in the game and for children with ASD. It was assumed that these participants would have a great knowledge and experience of at least 2 year in the fields of ASD, especially children. The participants were asked with an open answer question for the age at which this game would be appropriate. When asked for which age is the game appropriate, the answers gave a range from 3 to 6 but also one participant answered any depending on the ability. The answers for whether the participants thought the game would improve IJA in children with ASD were found to be statistically insignificant. The follow-up open answer question that asked the participants how will this game improve IJA, almost all of them answered "I don't know" or "Can't be sure", confirming the insignificance of the answers. The answers regarding to the possible benefits that the game might provide for the children with ASD did not yield as statistically significant ($p > 0.05$). This question was followed by an open answer question, asking the participants to explain how and in what ways would it help the children with ASD. Even though the participants were not sure about the game being beneficial for children with ASD, they had several interesting points when answering how it might help. Some of these answers were: the usage of pointing as a pre-speech skill, children with ASD would enjoy the robot celebrating the correct answer, improving social interactions and develop social skills, and

again imitating of the movements was mentioned by the same participant.

IV. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

The ultimate objective of this study was to investigate three aspects: the effect of emotional postures in an interactive game, the usage of social robots to promote IJA through interactive game, the possibility of using emotional postures as social reward. In order to do so a game incorporating emotional postures was designed and tested with young adults. To test the user experience of the game with emotional postures, two version of the game were tested, with and without emotional postures.Additionally, the game was focused on improving a specific social skill in children with ASD, JA which is very important in the development of social and communication skills. More specifically, the targeted skill was IJA, and pointing was taken as a social behaviour that people use to express the concept of sharing. A brief video was made including both version of the game and it was sent to few autism societies along with a questionnaire. For this part, the participants were evaluating the game in terms of whether and how would the game be beneficial for children with ASD and whether the game promotes IJA.

From the observation made in the discussion it is clear that there is a difference between the versions with and without emotional postures, where in both methods participants preferred the game with emotional postures and found it more engaging, fun and appropriate as a social motivation. Furthermore, in both methods participants thought that children will appreciate the emotional postures while playing the game.

If we combine the data observed from both methods together it can be concluded that the game proposed in this study is suitable for pre-school children with ASD and the emotional postures are beneficial in terms of encouraging and motivating the user to play the game. Further research is needed to evaluate how and whether the designed game can improve social skills in children with ASD, especially whether it promotes IJA.

B. Future Works

For future work, the most apparent approach would be to test the game with pre-school children with ASD, however there are several improvements that have to be made before proceeding with testing. The questionnaire formulated for the participants that watched the video had open answer questions regarding the possible problems and improvements of the system.

Regarding the language used in the game, majority of the participants noted that most children with ASD have difficulty processing spoken language and in order for them to understand the instructions and questions given by the robot, they have suggested to make bigger pauses between two sentences. Additionally, they have also suggested simplifying the questions or including someone in the game such as caregiver, parent, teacher etc. who will be able to aid the children when needed. Furthermore, few participants suggested adding an option for the robot to repeat the question and also if the child's response is incorrect the robot should encourage the child to try to answer the question again by repeating it. Regarding the appearance of the robot, several participants suggested that the robot should have more humanistic look and bigger eyes that blink (the robot used was programmed to blinked but maybe it was not clearly visible in the video). Although children with ASD avoid eye contact with people, when interacting with robots they spent most of the time looking at the robot's face. Few of the participants also noted that the movements of the robot take too long and could be more fluid. Further, by one parent was noted that the child might touch the tablet repeatedly.

Taking into consideration the aforementioned suggestions, the game can be improved in the following way: simplifying the language used by the robot, adding an option to repeat the question, changing the answer the robot gives when an answer is incorrect by giving a chance to the user to answer it again. Once these changes have been made the game could be tested with pre-school children with ASD. The evaluation of the game could be done in several sessions, where the experimental group would play the game and the control group would be playing the same game but without emotional postures and the robot would not turn to look where the child has pointed. The child's JA capabilities should be observed before the trial, during the trials and after the final session to see if there has been any improvement in both groups.

REFERENCES

[1] D. Feil-Seifer and M. J. Mataric, "Defining socially assistive robotics", in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.*, Jun. 2005, pp. 465–468.

[2] M. J. Matarić, J. Eriksson, D. J. Feil-Seifer, and C. J. Winstein, "Socially assistive robotics for post-stroke rehabilitation", *Journal of NeuroEngineering and Rehabilitation*, vol. 4, p. 5, Feb. 19, 2007.

[3] A. C. Lo, P. D. Guarino, L. G. Richards, J. K. Haselkorn, G. F. Wittenberg, D. G. Federman, R. J. Ringer, T. H. Wagner, H. I. Krebs, B. T. Volpe, C. T. J. Bever, D. M. Bravata, P. W. Duncan, B. H. Corn, A. D. Maffucci, S. E. Nadeau, S. S. Conroy, J. M. Powell, G. D. Huang, and P. Peduzzi, "Robot-assisted therapy for long-term upper-limb impairment after stroke", *New England Journal of Medicine*, vol. 362, no. 19, pp. 1772–1783, May 13, 2010.

[4] N. A. Malik, F. A. Hanapiah, R. A. A. Rahman, and H. Yussof, "Emergence of socially assistive robotics in rehabilitation for children with cerebral palsy: A review", *International Journal of Advanced Robotic Systems*, vol. 13, no. 3, p. 135, 2016.

[5] D. Feil-seifer, U. Viterbi, *et al.*, "Development of socially assistive robots for children with autism spectrum disorders", 2009.

[6] S. Tariq, S. Baber, A. Ashfaq, Y. Ayaz, M. Naveed, and S. Mohsin, "Interactive therapy approach through collaborative physical play between a socially assistive humanoid robot and children with autism spectrum disorder", in *Social Robotics*, Springer, Cham, Nov. 1, 2016, pp. 561–570. (visited on 02/09/2018).

[7] K. Wada, T. Shibata, T. Saito, and K. Tanie, "Analysis of factors that bring mental effects to elderly people in robot assisted activity", in *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*, Ieee, vol. 2, 2002, pp. 1152–1157.

[8] M. Saerbeck, T. Schut, C. Bartneck, and M. D. Janse, "Expressive robots in education: Varying the degree of social supportive behavior of a robotic tutor", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2010, pp. 1613–1622.

[9] A. Tapus, M. Mataric, and B. Scassellati, *The grand challenges in socially assistive robotics. robotics and automation magazine, 14 (1), 1-7*, 2007.

[10] E. J. Jones, T. Gliga, R. Bedford, T. Charman, and M. H. Johnson, "Developmental pathways to autism: A review of prospective studies of infants at risk", *Neuroscience & Biobehavioral Reviews*, vol. 39, pp. 1–33, 2014.

[11] P. Valnegri, J. Huang, T. Yamada, Y. Yang, L. A. Mejia, H. Y. Cho, A. Oldenborg, and A. Bonni, "Rnf8/ubc13 ubiquitin signaling suppresses synapse formation in the mammalian brain", *Nature communications*, vol. 8, no. 1, p. 1271, 2017.

[12] D. J. Ricks and M. B. Colton, "Trends and considerations in robot-assisted autism therapy", IEEE, May 2010, pp. 4354–4359.

[13] I. Cohen, R. Looije, and M. A. Neerincx, "Child's perception of robot's emotions: Effects of platform, context and experience", *International Journal of Social Robotics*, vol. 6, no. 4, pp. 507–518, 2014.

[14] T. Charman, "Why is joint attention a pivotal skill in autism?", *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 358, no. 1430, pp. 315–324, 2003.

[15] P. Mundy and L. Newell, "Attention, joint attention, and social cognition", *Current directions in psychological science*, vol. 16, no. 5, pp. 269–274, 2007.

[16] G. Little, L. Bonnar, S. Kelly, K. S. Lohan, and G. Rajendran, "Gaze contingent joint attention with an avatar in children with and without asd", in *Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2016 Joint IEEE International Conference on*, IEEE, 2016, pp. 15–20.

[17] P. Mundy, M. Sigman, and C. Kasari, "A longitudinal study of joint attention and language development in autistic children", *Journal of Autism and developmental Disorders*, vol. 20, no. 1, pp. 115–128, 1990.

[18] K. K. Poon, L. R. Watson, G. T. Baranek, and M. D. Poe, "To what extent do joint attention, imitation, and object play behaviors in infancy predict later communication and intellectual functioning in asd?", *Journal of autism and developmental disorders*, vol. 42, no. 6, pp. 1064–1074, 2012.

[19] C. Kasari, S. Freeman, and T. Paparella, "Joint attention and symbolic play in young children with autism: A randomized controlled intervention study", *Journal of Child Psychology and Psychiatry*, vol. 47, no. 6, pp. 611–620, 2006.

[20] C. Chevallier, G. Kohls, V. Troiani, E. S. Brodkin, and R. T. Schultz, "The social motivation theory of autism", *Trends in cognitive sciences*, vol. 16, no. 4, pp. 231–239, 2012.

[21] G. Kohls, C. Chevallier, V. Troiani, and R. T. Schultz, "Social 'wanting'dysfunction in autism: Neurobiological underpinnings and treatment implications", *Journal of Neurodevelopmental Disorders*, vol. 4, no. 1, p. 10, 2012.

[22] B. Scassellati, "How social robots will help us to diagnose, treat, and understand autism", in *Robotics research*, Springer, 2007, pp. 552–563.

[23] B. Robins, P. Dickerson, P. Stribling, and K. Dautenhahn, "Robot-mediated joint attention in children with autism: A case study in robot-human interaction", *Interaction Studies*, vol. 5, no. 2, pp. 161–198, 2004.

[24] R. S. De Silva, K. Tadano, M. Higashi, A. Saito, and S. G. Lambacher, "Therapeutic-assisted robot for children with autism", in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, IEEE, 2009, pp. 3561–3567.

[25] H. Yussof, L. Ismail, S. Shamsuddin, F. Hanapiah, S. Mohamed, H. Piah, and N. Zahari, "Human-robot interaction intervention therapy procedure for initial response of autism children with humanoid robot", in *Joint International Symposium on System-Integrated Intelligence*, 2012.

[26] Z. E. Warren, Z. Zheng, A. R. Swanson, E. Bekele, L. Zhang, J. A. Crittendon, A. F. Weitlauf, and N. Sarkar, "Can robotic interaction improve joint attention skills?", *Journal of autism and developmental disorders*, vol. 45, no. 11, pp. 3726–3734, 2015.

[27] H. Kozima, C. Nakagawa, and Y. Yasuda, "Children–robot interaction: A pilot study in autism therapy", *Progress in Brain Research*, vol. 164, pp. 385–400, 2007.

[28] B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard, "Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills?", *Universal Access in the Information Society*, vol. 4, no. 2, pp. 105–120, 2005.

[29] C. M. Stanton, P. H. Kahn Jr, R. L. Severson, J. H. Ruckert, and B. T. Gill, "Robotic animals might aid in the social development of children with autism", in *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, ACM, 2008, pp. 271–278.

[30] A. Duquette, F. Michaud, and H. Mercier, "Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism", *Autonomous Robots*, vol. 24, no. 2, pp. 147–157, 2008.

[31] B. Scassellati, H. Admoni, and M. Matarić, "Robots for use in autism research", *Annual review of biomedical engineering*, vol. 14, pp. 275–294, 2012.

[32] MrChewypoo, *Quiz for kids - preschool (ages 2-5)*. [Online]. Available: https://www.sporcle.com/games/MrChewypoo/quiz-for-kids-preschool-ages-2-5.

[33] S. Jana, *105 basic gk questions and answers for kids*, Feb. 2018. [Online]. Available: http://www.momjunction.com/articles/general-knowledge-questions-for-kids_00439953/#gref.

[34] *Trivia quiz questions for children and teenagers*. [Online]. Available: http://www.free-for-kids.com/trivia-quizzes-for-children.shtml.

[35] A. Beck, A. Hiolle, A. Mazel, and L. Cañamero, "Interpretation of emotional body language displayed by robots", in *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, ACM, 2010, pp. 37–42.

[36] I. Cohen, R. Looije, and M. A. Neerincx, "Child's perception of robot's emotions: Effects of platform, context and experience", *International Journal of Social Robotics*, vol. 6, no. 4, pp. 507–518, 2014.

[37] M. S. Erden, "Emotional postures for the humanoid-robot nao", *International Journal of Social Robotics*, vol. 5, no. 4, pp. 441–456, 2013.

Learning Underwater Motion and 3D Reconstruction from Optical and Acoustic Sensors

Shubham Wagh, Sen Wang and Yvan Petillot

***Abstract*— Autonomous Underwater Vehicles (AUVs) are in increasing demand to perform different underwater tasks such as coral reef monitoring, harbours security, mine countermeasure missions etc. However, its autonomy still presents some challenges in terms of navigation and localization. One of the key features still missing from the current system is the capability to automatically gain knowledge and improve performance through end-to-end learning. Based on the recent works, we propose two unsupervised learning framework for 3D reconstruction and motion estimation in underwater scenarios using raw optical and sonar image frames individually to solve the challenging problem of localization and navigation in underwater environments. The results of the experiments demonstrate appreciable results and showed a great potential in improving the performance if supplemented with large number of good datasets both for sonar and optical images.**

I. INTRODUCTION

In recent years, there has been an increasing need for autonomous vehicles that operate underwater. AUVs have many applications as they perform dangerous and monotonous tasks like ship hull inspection, coral reef monitoring, surveying etc. AUVs should be able to localize itself accurately in the environment in order to carry out these operations efficiently and autonomously. However, its autonomy presents challenges on multiple levels in terms of localization and navigation.

The existing geometric feature based methods make use of vision sensors (optical camera or sonar sensor) for 3D scene reconstruction and ego-motion estimation and have demonstrated an incredible performance in terms of accuracy. However, one of the key features still missing from the current system is the capability to automatically gain knowledge and improve performance through end-to-end learning [22]. The current systems heavily rely on manual troubleshooting to analyze failure cases and refine localization results.

In this work, we propose two end-to-end unsupervised learning framework for 3D reconstruction and motion estimation in underwater scenarios using optical images and sonar images (images from 2D Forward Looking Sonar) individually:

1) depth map and pose prediction network using unstructured sequences of optical images.
2) elevation map and pose prediction network using unstructured sequences of sonar images in polar form, also known as *SonarNet*.

We are particularly inspired by the recent work in [25], where unsupervised learning is adopted for air images to estimate 3D scene depth and pose. They perform appreciatively compared to established SLAM systems under similar input settings. Also, unsupervised learning removes the need for separate supervisory signals or ground truth. To the best of our knowledge, no previous systems exist that learns motion estimates and 3D information (i.e. depth map from optical images and elevation map from sonar images) in an unsupervised manner from raw image sequences in underwater environments.

S. Wagh, S. Wang and Y. Petillot are with Institute of Sensors, Signals, and Systems, Heriot-Watt University, Edinburgh, UK smw2@hw.ac.uk

II. RELATED WORK

In this section, we discuss some key state-of-the-art methods for 3D reconstruction and pose estimation in underwater scenarios solely based on optical or sonar images. We also deliberate about deep learning approaches for images on land as no learning method exist for underwater scenarios.

A. Underwater 3D reconstruction

1) Using Optical Sensors: Sedlazeck *et al.* in [21] reconstructs a real 3D scenario using a HD color camera with the help of features are based on image gradients using a corner detector. Nicosevici *et al.* [20] use SIFT features in a robotics approach, with an average error of 11 mm. Beall *et al.* [4] use a wide baseline stereo rig to extract SURF features from left and right image pairs. After a smoothing and mapping (SAM) step, they recover structure of the environment by tracking these features. Negre *et al.* [6], [5] use a SLAM approach in a micro AUV equipped with two stereo rigs to perform 3D reconstruction of underwater environments.

The above methods need correspondence matching between two views to accomplish 3D reconstruction. Factors like low visibility and low lighting conditions make the task of finding correspondences between two underwater optical views more challenging.

2) Using Sonar Sensors: Negahdaripour *et al.* in [17] utilize radiometric information in a sonar image to detect object and shadow regions, and various geometric cues and constraints to determine sizes, shapes and spatial distributions of 3-D objects resting on the sea floor. Furthermore, in [18], an SfM approach from a set of images taken from an imaging sonar is used to recover 3D data. Aykin *et al.* in [2] estimate the lost elevation angles by first determining the 3-D information at object boundaries based on cast shadow cues and then employ the proposed image formation model to obtain the remaining elevation values through each object. Its main requirement is that the shadow is distinguishable and that it lays on a known flat surface. In [3], Aykin et al. propose a novel method of space carving framework for

the 3-D reconstruction of targets from multiple forward scan sonar images captured at known sonar poses.

Thee methods manually select the features and perform matching only for the sparse features. Thereby the resultant 3d structure is not dense.

B. Underwater Motion Estimation

Apart from the general Structure from Motion (SfM) approach, we present some direct methods used to estimate the motion.

1) Using Optical Sensors: Xu *et al.* in [24] propose a direct method for accurate 3-D motion estimation from the brightness variations in an optical sea bed image sequence. In [7], Garcia et al. present a mosaicing algorithm, based on this it is able to provide positional coordinates to an underwater vehicle. Futhermore, in [8] propose a new method to improve image matching in underwater image sequences for estimating the motion of an underwater robot. However, it has serious limitations when all the correspondence matches are coplanar.

Based on the above methods, it can be seen that image registration is a pre-requisite to accomplish motion estimation which is computationally expensive.

2) Using Sonar Sensors: Negahdaripour *et al.* in [16] investigate the model of sonar image flow in 2D FLS by tracking the 2-D images of scene features to enable automated dense correspondences and 3D motion estimation. Furthermore in [19], the author present the mathematical models of sonar image flow for 3-D objects and their cast shadows and utilize them in devising various 3-D sonar motion estimation solution.

The above methods are not capable enough to automatically gain knowledge and improve performance of the motion estimation system.

C. Deep Learning approaches for 3D reconstruction & motion estiation

Garg *et al.* in [9], present first unsupervised single view depth prediction method by using the left-right photometric constraint of stereo image pairs. This method was further improved in [10] by wrapping the left and right images across each other which improve the accuracy of depth prediction. Li *et al.* in [13] propose an UnDeepVO architecture to estimate the 6-DoF pose of a monocular camera and the depth of its view in an unsupervised fashion by incorporating spatial and temporal dense information in the loss function for training the network. However, a number of challenges such as robustness to image blurs, camera parameters, or illumination are not addressed. Recently, Zhou *et al.* in [25], propose a novel SfMLearner architecture for the task of monocular depth and camera motion estimation from unstructured video sequences based on the task of view synthesis as supevisory signals. This setup is most aligned with our work as we similarly learn depth/elevation and ego-motion from optical/sonar image in an unsupervised setting for underwater scenarios.

III. PROPOSED METHODOLOGY

A. Depth Map and Ego-Motion from Optical Images

In this section, we discuss the proposed framework in terms of geometry of the problem, network architecture and then describe the individual loss functions used in the framework.

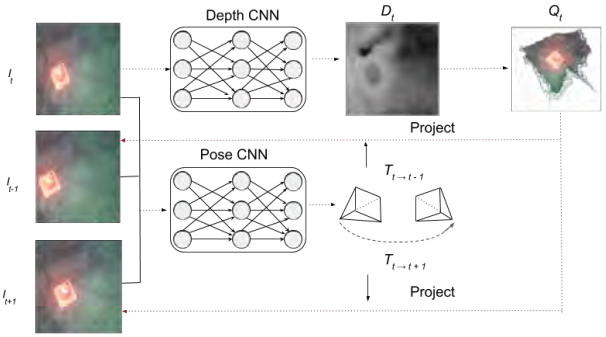


Fig. 1: Overview of the supervision pipeline based on view synthesis. The depth network takes only I_t as input, and outputs a per-pixel depth map D_t. The pose network takes both the target view (I_t) and the nearby/source views (e.g., I_{t-1} and I_{t+1}) as input, and outputs the relative camera poses ($T_{t\to t-1}, T_{t\to t+1}$).

1) Problem Geometry:

i Given an input image I_t , we estimate depth D_t and with the help of nearby source views i.e images I_{t-1} and I_{t+1}, ego-motions $T_{t\to t-1}$ and $T_{t\to t+1}$ are estimated. The depth and ego-motion estimations are done using depth and pose network which is described in the next section.

ii Using the pixel coordinates (i, j) of image I_t and estimated depth D_t^{ij}, it is then projected into a structured 3D point cloud Q_t.

$$Q_t^{ij} = D_t^{ij} K^{-1}[i, j, 1]^T \tag{1}$$

where K is intrinsic matrix of the camera (using homogeneous coordinates)

iii Given the estimates for camera movement $T_{t\to t-1}$ and $T_{t\to t+1}$, Q_t can be transformed to get the estimates for 3D point clouds $\hat{Q}_{t-1}$ and $\hat{Q}_{t+1}$.

$$\hat{Q}_{t-1} = T_{t\to t-1} Q_t \tag{2}$$

$$\hat{Q}_{t+1} = T_{t\to t+1} Q_t \tag{3}$$

iv $\hat{Q}_{t-1}$ and $\hat{Q}_{t+1}$ can then be projected onto the nearby camera frames respectively as $K\hat{Q}_{t-1}$ and $K\hat{Q}_{t+1}$.

v Combining this transformation and projection, we obtain I_t image's projected pixel coordinates onto the source views $\hat{I}_{t-1}$ and $\hat{I}_{t+1}$ at time $t-1$ and $t+1$ respectively.

$$\hat{p}_{t-1} = K T_{t\to t-1} D_t(p_t) K^{-1} p_t \tag{4}$$

$$\hat{p}_{t+1} = K T_{t\to t+1} D_t(p_t) K^{-1} p_t \tag{5}$$

where,

p_t is pixel coordinate of input image at time t

$\hat{p}_{t-1}$ is pixel coordinates of $\hat{I}_{t-1}$
$\hat{p}_{t+1}$ is pixel coordinates of $\hat{I}_{t+1}$
$D_t(p_t)$ is depth value at p_t pixel coordinate

The above formulation allows us to reconstruct the image frame $\hat{I}_{t-1}$ and $\hat{I}_{t+1}$ by warping input image I_t based on D_t and ego-motions $T_{t\to t-1}$ and $T_{t\to t+1}$. As the projected pixel coordinates are continuous values we follow the procedure given in [25] wherein a soft sampling is performed from the four pixels in I_t whose pixel coordinates overlap with $\hat{p}_{t-1}$ to get $\hat{I}_{t-1}$ or $\hat{p}_{t+1}$ to get $\hat{I}_{t+1}$ (figure 1).

2) Network Architecture:

i **Single-view depth network** : We adopt the SfMLearner architecture [25] which is in turn based on DispNet [15]. It is mainly an encoder-decoder or autoencoders design with skip connections and multiscale predictions. Given a single image as input it produces a dense depth estimate mapping each pixel of the input to a metric depth value. The network is fully convolutional. All *conv* layers are followed by a ReLU activation function except for the prediction layers where we use, $1/(\alpha * sigmoid(x) + \beta)$ with $\alpha = 10$ and $\beta = 0.01$ for positive depth value constraints.

ii **Pose network** : The input to the pose estimation network is a sequence of three images (e.g. I_{t-1}, I_t and I_{t+1}). We concatenate the target image I_t with nearby source images (I_{t-1} and I_{t+1}) along the color channels. The outputs of the pose network are the relative poses between the target image and each of the source views (i.e $T_{t\to t-1}$ and $T_{t\to t+1}$). These ego motion predictions are represented by six numbers corresponding to relative 3D rotation (Euler angles) and metric translations between the two frames.

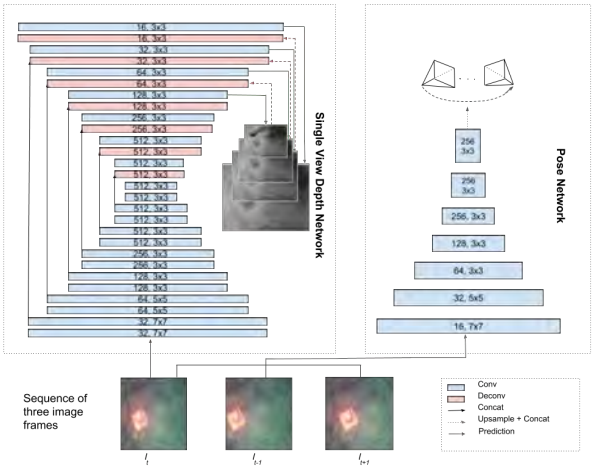


Fig. 2: Network architecture for our depth/pose prediction modules.

3) Loss Functions:

i **Image Reconstruction loss** : We use view synthesis supervision to form the image reconstruction loss. Let $< I_1, I_2, ...I_N >$ be training image sequence with one of the frames being target image and the rest being the source view $I_s (1 \leq s \leq N, s \neq t)$. The image reconstruction loss is formulated as:

$$L_{rec} = \sum_{ij} ||I_t^{ij} - \hat{I}_s^{ij}|| \tag{6}$$

where, $\hat{I}_s$ is the source view I_s warped to the target coordinate frame based on the predicted depth D_t, ego-motion $T_{t\to s}$ and the input source view I_s as discussed in the problem geometry.

ii **Structured Similarity loss** : This is a metric to evaluate the quality of image predictions. It measures the similarity between two images patches x and y and is defined as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x + \sigma_y + c_2)} \tag{7}$$

where μ_x, σ_x are the local means and variances [23]. μ and σ are computed by simple pooling operation, $c_1 = 0.01^2$ and $c_2 = 0.03^2$. Since SSIM is upper bounded to 1 and needs to be maximized, we instead minimize:

$$L_{SSIM} = \sum_{ij} [1 - SSIM(\hat{I}_s^{ij}, I_t^{ij})] \tag{8}$$

iii **Depth Smoothness loss** : This incorporates sharp changes in depth at pixel coordinates where there are sharp changes in the image. It is a refinement of the depth smoothness loss defined in Zhou *et al.* [25].

$$L_{smooth} = \sum_{ij} ||\partial_x D^{ij}||e^{-||\partial_x I^{ij}||} + ||\partial_y D^{ij}||e^{-||\partial_y I^{ij}||} \tag{9}$$

All the aforementioned loss functions are applied at all the scales s, ranging from the models input resolution, to an image that is $\frac{1}{8}$ in width and height. The total loss is defined as:

$$L_{total} = \sum_s \alpha L_{rec}^s + \beta L_{SSIM}^s + \gamma L_{smooth}^s \tag{10}$$

where α, β and γ are hyper-parameters.

B. Elevation Map and Ego-Motion from Sonar Images

In Forward Looking Sonar (FLS) image formation process, the elevation angle ϕ is lost through projection from the 3D space (R, Θ, Φ) onto the 2D (r, θ) or (x_s, y_s) sonar image domain [17]. For accurate 3D scene reconstruction from 2D FLS sonar imagery, we need to compute the unknown elevation angle from the image data with high precision. However, an approximate estimation of ϕ is more than enough for 3D sonar ego-motion estimation. We propose an analogous unsupervised learning framework known as "**SonarNet**" to estimate elevation map and ego-motion using sonar video stream.

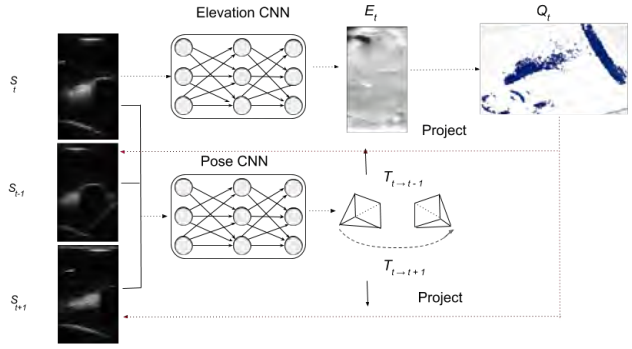


Fig. 3: Overview pipeline for SonarNet model

1) Problem Geometry:

i Given an input sonar image S_t, we estimate elevation map E_t and with the help of nearby source views i.e images S_{t-1} and S_{t+1}, ego-motions $T_{t\to t-1}$ and $T_{t\to t+1}$ are estimated. The elevation map and ego-motion estimations are done using SonarNet architecture which is described in the next section. Note that the input sonar image is in polar form (r, θ) as that is the raw image obtained from the FLS sonar sensor.

ii Sonar being an acoustic sensor, the concept of intrinsic matrix is not valid. So the pixel coordinates (r, θ) are linearly scaled to actual (R, Θ) measurements depending on the window length and FOV of the sonar sensor, where R is in meters and Θ in degrees.

iii Elevation map is defined as for every (R, Θ) real world coordinate of an object in the scene, there is a corresponding elevation angle Φ (in degrees) for that object. The elevation angle range for mostly every sonar sensor is within -7 deg to $+7$ deg. The predicted elevation map (E_t) from the network have values in the scale ranging from 0 to 1 (i.e. linear mapping from [-7 deg - $+7$ deg] scale to range [0 to 1]) which is further linearly mapped to gray scale intensity range $0 - 255$ for visualization purpose.

$$E_t(r, \theta) = \phi \tag{11}$$

where (r, θ) are pixel coordinates of elevation map and ϕ is the intensity value of elevation map at that coordinate.

iv A structured 3D point cloud in spherical coordinate system is formed by linearly scaling the pixel coordinates (r, θ) and its corresponding intensity (ϕ) of elevation map E_t to get real world coordinates (R, Θ, Φ). The azimuth (Θ) and elevation (Φ) angles are converted to radians for further computations.

v The structured 3D point cloud is converted to Cartesian coordinate system to get (X, Y, Z) real world coordinates [17], denoted by Q_t.

vi Given the estimates for the sonar sensor movement $T_{t\to t-1}$ and $T_{t\to t+1}$, Q_t can be transformed to get the estimates for 3D point clouds $\hat{Q}_{t-1}$ and $\hat{Q}_{t+1}$.

$$\hat{Q}_{t-1} = T_{t\to t-1} Q_t \tag{12}$$

$$\hat{Q}_{t+1} = T_{t\to t+1} Q_t \tag{13}$$

vii $\hat{Q}_{t-1}$ and $\hat{Q}_{t+1}$ can then be projected onto the sonar image frame at $t-1$ and $t+1$ respectively to get (R, Θ) and then by appropriate linear scaling to get (r, θ) pixel coordinates.

viii Combining this transformation and projection, we obtain S_t image's projected pixel coordinates onto the source views $\hat{S}_{t-1}$ and $\hat{S}_{t+1}$.

The above formulation allows us to reconstruct the sonar image frame $\hat{S}_{t-1}$ and $\hat{S}_{t+1}$ by warping input sonar image S_t based on E_t and ego-motions $T_{t\to t-1}$ and $T_{t\to t+1}$ (figure 3).

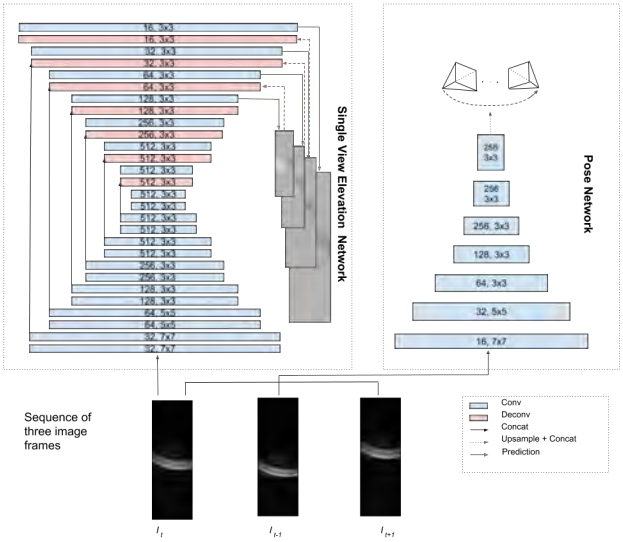


Fig. 4: Network architecture for the SonarNet mode

2) Network Architecture:

i **Elevation Network** : It is very much similar to the single view depth network architecture except here we don't have inverse prediction.

ii **Pose Network** : It is similar to the pose network for optical images except here we input sequence of three sonar images in polar form.

3) Loss Functions: Here we only make use of image reconstruction loss and elevation smoothness loss as sonar images do not have color, texture etc. features. For smoothness, we minimize the L1 norm of the second-order gradients for the predicted elevation map.

The learning setup is analogous to setup for optical images except here $\beta = 0$.

IV. EXPERIMENTS AND RESULTS

A. Training Details

1) Depth map and Ego-motion from Optical Images: We implemented the proposed system using the publicly available Tensorflow framework [1]. For all the experiments, as per the learning setup in section 4.2.4, we set the value of $\alpha = 0.15$, $\beta = 0.85$ and $\gamma = 0.1$ for all the scales. During training, we used batch normalization [11] for all the layers

except for the output layers, and the Adam [12] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate of 0.0001 and mini-batch size of 16. We jointly train the system on modern GPU TITAN Xp for a maximum iteration number of $200,000$.

We train the system using *Universitat de Girona's* publicly available data set collected with an AUV testbed in the unstructured environment of an underwater cave complex [14]. It contains a total of 9774 raw image frames. having a dimension of 288×384. Both the depth and pose networks can be run fully-convolutionally for images of arbitrary size at test time.

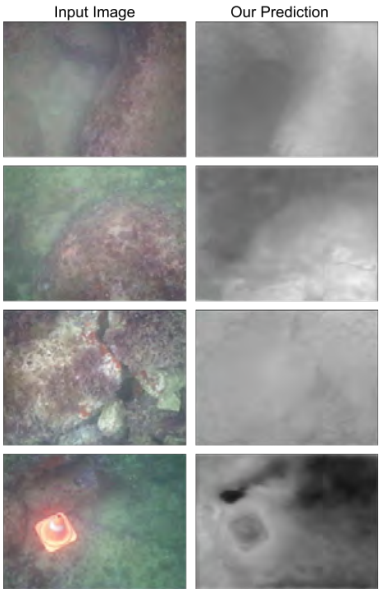


Fig. 5: Our sample depth predictions on the Gironas underwater dataset using the model trained on the same datase

2) Elevation map and Ego-motion from Sonar Images: Similarly, we jointly train the SonarNet model. For all the experiments, as per the learning setup, we set value of $\alpha = 1$, $\beta = 0$ and $\gamma = 0.25$ for all the scales. We set learning rate of 0.00001 and maximum iteration upto $300,000$. Apart from this, all the details are similar to previous one.

The model is trained using the available underwater sonar dataset collected in a water tank at Ocean Systems Laboratory, Heriot-Watt University. It contains a total of $21,348$ sonar image frames, having a dimension of 1352×128. We resize this image dataset to 467×128 as an appropriate image size for CNN's input.

B. Results

In figure 5 we show our single view depth predictions. As currently we do not have ground truth for the available dataset, we validate our depth map by reconstructing the point cloud as shown in figure 7. We see that the depth predictions are appreciable. As the dataset is limited (9772 sequences of 3) and has artificial illumination in the dataset, our unsupervised framework has not been able to learn complete underwater features, thereby the pose predictions are affected and may not be correct at this point.

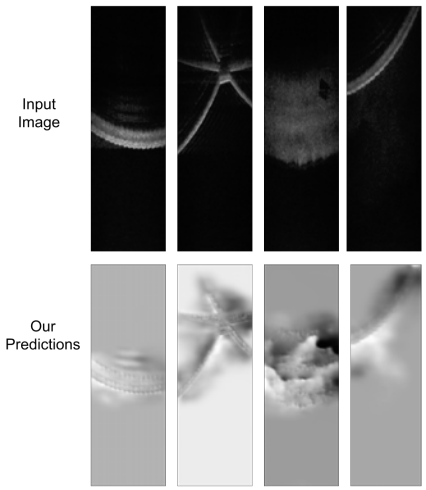


Fig. 6: Elevation map predictions

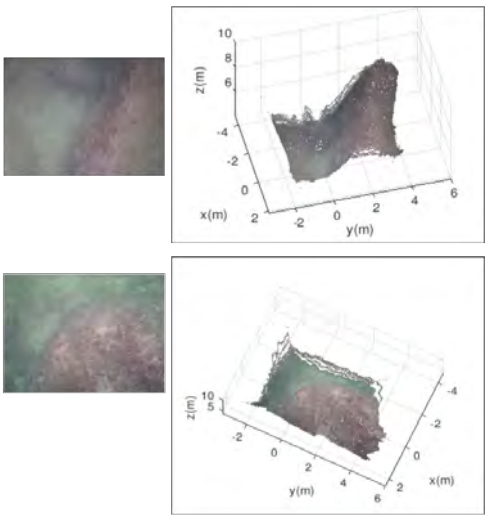


Fig. 7: Reconstructed 3d point cloud of an input image

In figure 6, we show elevation map prediction of SonarNet model. As we do not have ground truth, we again verify this by reconstructing 3D point cloud as shown in figure 8 for first image in figure 6. We get appreciable and logical results at this point. We also plot the pose trajectory for an example data-set as shown in figure 9. The example dataset is in cartesian form for easy interpretation of the motion. We get approximately good estimation of the motion with little drift as in the example the sonar is rotating about $+z$ axis with some motion along $+y$ axis. Most of the sonar images in the dataset are sparse, hence not significant learning was possible. This can be significantly improved with more number of dataset which is required for unsupervised learning.

V. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

In this paper, we presented two unsupervised learning framework for 3d reconstruction and motion estimation us-

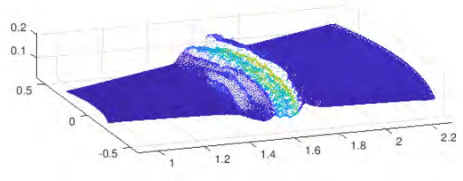


Fig. 8: Reconstructed 3d point cloud of sonar image in fig. 6

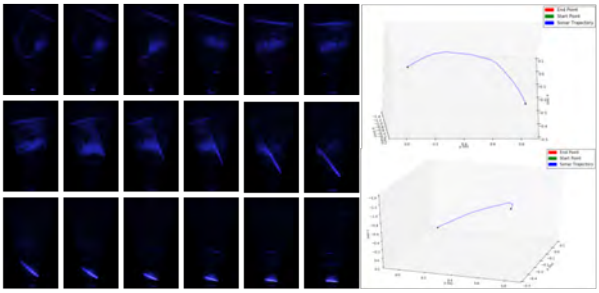


Fig. 9: Sonar pose trajectory of the sequential sonar frames

ing: 1) Optical and 2) Sonar images respectively. Both the systems make use of view synthesis as its supervisory signal. The results of the experiments demonstrate logical results and showed a great potential in improving the performance if supplemented with large number of good datasets both for sonar and optical images.

B. Future Works

In general, unsupervised learning methods have the potential to improve their performance with the increasing size of training datasets. In the next step, we will investigate how to train the proposed networks with large amount of datasets to improve its performance, such as robustness to illumination changes for optical images and sparsity for sonar images. In the future, we also plan to extend our system to a visual SLAM system to reduce the drift.

VI. ACKNOWLEDGMENTS

The authors gratefully acknowledge the Girona Underwater Vision and Robotics group, Spain and Ocean Systems Lab, UK for providing the dataset.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Murat D Aykin and Shahriar Negahdaripour. Forward-look 2-d sonar image formation and 3-d reconstruction. In *Oceans-San Diego, 2013*, pages 1–10. IEEE, 2013.

[3] Murat D Aykin and Shahriar Negahdaripour. On 3-d target reconstruction from multiple 2-d forward-scan sonar views. In *OCEANS 2015-Genova*, pages 1–10. IEEE, 2015.

[4] Chris Beall, Brian J Lawrence, Viorela Ila, and Frank Dellaert. 3d reconstruction of underwater structures. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 4418–4423. IEEE, 2010.

[5] Francisco Bonin-Font, Aleksandar Cosic, Pep Lluis Negre, Markus Solbach, and Gabriel Oliver. Stereo slam for robust dense 3d reconstruction of underwater environments. In *OCEANS 2015-Genova*, pages 1–6. IEEE, 2015.

[6] Pep Lluis Negre Carrasco, Francisco Bonin-Font, and Gabriel Oliver Codina. Stereo graph-slam for autonomous underwater vehicles. In *Intelligent Autonomous Systems 13*, pages 351–360. Springer, 2016.

[7] Rafael Garcia, J Batlle, Xavier Cufi, and Josep Amat. Positioning an underwater vehicle through image mosaicking. In *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, volume 3, pages 2779–2784. IEEE, 2001.

[8] Rafael Garcia, Xevi Cufi, and Marc Carreras. Estimating the motion of an underwater robot from a monocular image sequence. In *Intelligent Robots and Systems, 2001. Proceedings. 2001 IEEE/RSJ International Conference on*, volume 3, pages 1682–1687. IEEE, 2001.

[9] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016.

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, page 7, 2017.

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[12] D Kinga and J Ba Adam. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[13] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. *arXiv preprint arXiv:1709.06841*, 2017.

[14] Angelos Mallios, Eduard Vidal, Ricard Campos, and Marc Carreras. Underwater caves sonar data set. *The International Journal of Robotics Research*, 36(12):1247–1251, 2017.

[15] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.

[16] Shahriar Negaharipour. On 3-d motion estimation from 2-d sonar image flow. In *Oceans, 2012*, pages 1–7. IEEE, 2012.

[17] Shahriar Negaharipour. On 3-d scene interpretation from fs sonar imagery. In *Oceans, 2012*, pages 1–9. IEEE, 2012.

[18] S Negahdaripour. On 3-d reconstruction from stereo fs sonar imaging. In *OCEANS 2010*, pages 1–6. IEEE, 2010.

[19] Shahriar Negahdaripour. On 3-d motion estimation from feature tracks in 2-d fs sonar video. *IEEE Transactions on Robotics*, 29(4):1016–1030, 2013.

[20] Tudor Nicosevici, Nuno Gracias, Shahriar Negahdaripour, and Rafael Garcia. Efficient three-dimensional scene modeling and mosaicing. *Journal of Field Robotics*, 26(10):759–788, 2009.

[21] Anne Sedlazeck, Kevin Koser, and Reinhard Koch. 3d reconstruction based on underwater video from rov kiel 6000 considering underwater imaging conditions. In *OCEANS 2009-EUROPE*, pages 1–10. IEEE, 2009.

[22] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 2043–2050. IEEE, 2017.

[23] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[24] Xun Xu and Shahriar Negahdaripour. Vision-based motion sensing for underwater navigation and mosaicing of ocean floor images. In *OCEANS'97. MTS/IEEE Conference Proceedings*, volume 2, pages 1412–1417. IEEE, 1997.

[25] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, page 7, 2017.