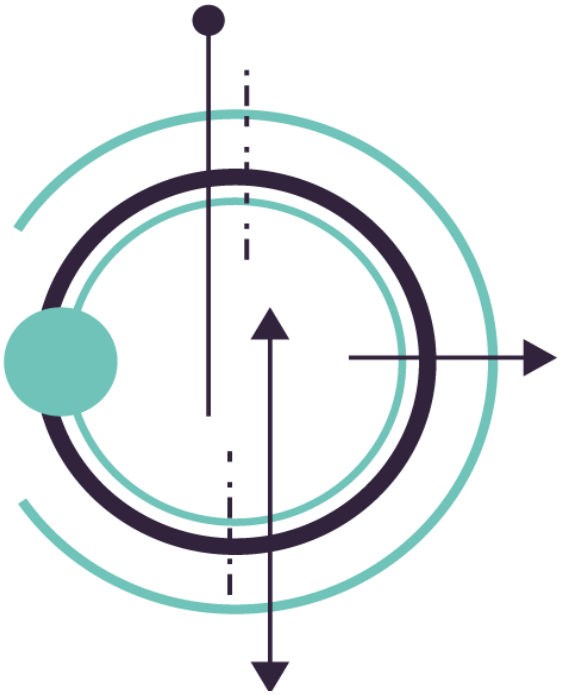
IMVIP 2022

Irish Pattern Recognition and Classification Society



Conference Proceedings

24th Irish Machine Vision and Image Processing Conference

31st August – 2nd September, Queen's University, Belfast

Sponsored by:

Queen's University Belfast | Faculty of Engineering and Physical Sciences

bcs The Chartered Institute for IT

McC fpmccann

Published by the Irish Pattern Recognition & Classification Society

iprcs.org

ISBN 978-0-9934207-7-1

Welcome

The 24th Irish Machine Vision and Image Processing Conference (IMVIP 2022) is hosted in person at The Bridge Lecture Theatre and Student Hub at Queen's University, Belfast.
The IMVIP Conference is Ireland's primary meeting for those researching in the fields of machine vision and image processing. This year marks the 25th year since the first IMVIP conference which was held in Magee Campus at Ulster University. To mark this special milestone, Prof Bryan Scotney (Ulster University) will present during a Special Session on this anniversary. The conference continues to be a platform for researchers across Ireland and beyond to showcase their novel work and share ideas. This single-track conference will see work presented through 19 oral presentations and 13 poster presentations.
As organiser's, we have been delighted to have several sponsors support this year's conference including the Faculty of Engineering and Physical Sciences within QUB, FP McCann, and the British Computer Society (BCS) Northern Ireland. Furthermore, the Irish Pattern Recognition and Classification Society (IPRCS) have supported a small number of national and international travel bursaries along with sponsoring the Best Paper Award. We are delighted to additionally award the BCS Northern Ireland Best Oral and Poster Presentation prizes. The support of sponsors is integral to help us put on an enjoyable and thought-provoking conference whilst keeping registration costs manageable particularly for student delegates. We are pleased that IMVIP continues to be a welcoming, supportive, and inspiring environment for presenters and delegates.
This year's conference welcomes three insightful and inspiring talks from Dr Mairéad Grogan, Prof Amos Storkey, and Dr Sandra Scott-Hayward. Having completed her PhD and post-doctoral research in Trinity College Dublin, Dr Grogran returns to IMVIP now as a Researcher Engineer at Foundry to provide insights to Image Processing/AI research in industry and application to the arts. Prof Storkey shares some of his expertise on developing image analysis in the medical domain and the importance of developing robust models for such critical applications. Extending on this theme and cognisant of the increasing adoption of AI and computer vision in the real-world, Dr Scott-Hayward looks at how we should be looking beyond performance metrics when building machine vision and image processing models. We know these talks are of great interest to all attending the conference but particularly important for early career researchers to learn about as they start their research journey.
We would like to extend our thanks to all those who have helped facilitate and promote the conference, for their time, resources, expertise, assistance and indeed patience! We are delighted to welcome all attendees to Belfast and to Queen's University.

Richard Gault
Organiser/Editor
Queen's University, Belfast
August 2022

Programme Chair

Richard Gault, Queen's University, Belfast

Organising Committee

Dr. Muhammad Fahim, Queen's University, Belfast
Dr. Darragh Lydon, Queen's University, Belfast
Victoria Porter, Queen's University, Belfast
Kristopher McCombe, Queen's University, Belfast
Dr. Ross McWhirter, Queen's University, Belfast and FP McCann Ltd

Programme Committee

Baharak Ahmaderaghi, Queen's University, Belfast
Vincent Andrearczyk, HES-SO
Donald Bailey, Massey University
Francesco Bianconi, Università degli Studi di Perugia
Sonya Coleman, Ulster University
Joan Condell, Ulster University
Jane Courtney, Technological University Dublin
Kathy Clawson, University of Sunderland
Kathleen Curran, University College Dublin
Rozenn Dahyot, National University of Ireland, Maynooth
Kenneth Dawson-Howe, Trinity College Dublin
Catherine Deegan, Technological University Dublin
Soumyabrata Dev, University College Dublin
Cem Direkoglu, Middle East Technical University - Northern Cyprus Campus
Ciaran Eising, University of Limerick
Robert Fisher, The University of Edinburgh
Bryan Gardiner, Ulster University
Jonathan Horgan, Valeo Vision Systems
Ihsan Ullah, National University of Ireland, Galway
Dermot Kerr, Ulster University
Anil Kokaram, Trinity College Dublin
Vladimir Krylov, Dublin City University
Suzanne Little, Dublin City University
Charles Markham, National University of Ireland, Maynooth
Jesus Martinez-Del-Rincon, Queen's University, Belfast
Sally McClean, Ulster University
John McDonald, National University of Ireland, Maynooth
Kevin McGuinness, Dublin City University
Paul McKevitt, Ulster University
Niall McLaughlin, Queen's University, Belfast
Paul Miller, Queen's University, Belfast
Derek Molloy, Dublin City University
George Moore, Ulster University
Sean Mullery, Institute of Technology Sligo
Omar Nibouche, Ulster University
Noel E. O Connor, Dublin City University
Robert Ross, Technological University Dublin

 ISBN 978-0-9934207-7-1

Keynote Speaker: Mairéad Grogan

Mairéad Grogan is a Research Engineer at Foundry working as part of the AI Research Team. She received an MSc and PhD in Computer Science from Trinity College Dublin in 2013 and 2016, with her PhD dedicated to 3D shape and image processing. She was a postdoctoral researcher in the V-SENSE group for three years, working in various areas including light fields, colour processing and image compositing. In 2020, she joined Foundry as a Research Engineer specialising in Machine Learning algorithms for 3D scene reconstruction and image processing tasks.

Unleashing the power of the artist with Machine Learning

It's no secret that Machine Learning has been on the rise in visual effects. Over the past few years, we've seen the impact it's had on the VFX industry and the technology that has emerged because of it. Join Foundry Research Engineer Mairéad Grogan as we take a look at how the company is implementing cutting-edge Machine Learning technology to put into the hands of the VFX community. Hear how this work inspired the development of the CopyCat node in Foundry's compositing software Nuke, giving artists full control over the types of effects they can generate using Machine Learning.

Keynote Speaker: Amos Storkey

Amos Storkey is Professor of Machine Learning and AI at the School of Informatics, University of Edinburgh. He leads the Bayesian and Neural Systems Research Group and is Director of the EPSRC Centre for Doctoral Training in Data Science. On the methodological side, he is known for his contributions to meta-learning and few shot learning, efficient neural network design, reinforcement learning, dataset shift, and transactional mechanisms for machine learning. His focus is machine learning for images and video; as part of that he has a long history of developments in medical imaging - historically in brain MRI and diffusion MRI, and more recently in brain CT and retinal imaging.

On Robust Machine Learning for Natural and Medical Computer Vision

Robustness is a catch-all term used to describe the expectation that our computer vision methods should be broadly applicable in general circumstances and not break in the context of some minor domain shift. However it has long been known that neural networks can produce fragile methods that easily break under dataset shift. This is an issue throughout computer vision, but is particularly an issue in medical imaging, where methods developed need to be robust to changes in demographic, changes in imaging equipment, variations in pathology, differences in prior medical health, and differences in choices or setting a radiographer or imager use. Failure in robustness can result in both poor performance and the introduction of serious bias. In this talk I will demonstrate, with examples, reasons why neural networks can be non-robust, and characterise solutions to robustness into three types of approaches: structural, invariant and equivariant measures. I will illustrate a toolkit of approaches we can apply to neural architectures and learning to help improve model robustness. I will also discuss the place of generative models and adversarial methods in robustness for computer vision.

Keynote Speaker: Sandra Scott-Hayward

Sandra Scott-Hayward is an Associate Professor at Queen's University Belfast. She began her career in industry and became a Chartered Engineer in 2006. Since joining academia, she has contributed security designs and solutions for softwarized networks based on her research on network security architectures and security functions for emerging networks. She is Director of the QUB Academic Centre of Excellence in Cyber Security Education (ACE-CSE), co-lead of the QUB Leverhulme Interdisciplinary Network on Algorithmic Solutions (LINAS) doctoral training programme, and a Polymath Fellow of the Global Fellowship Initiative at the Geneva Centre for Security Policy (GCSP) from 2021 to 2023. With LINAS and GCSP, she explores the impact of ML and AI technologies on security and society.

99.99% accurate - what's the problem?

Innovations in systems and services based on advances in artificial intelligence (AI)/machine learning (ML) are increasing. The main metric for demonstrating the value of such solutions is, generally, performance with a reported accuracy/precision/recall. Given the reported vulnerabilities associated with the use of AI/ML in a broad range of areas, is accuracy the only aspect that we should evaluate? In this talk, we illustrate some of the vulnerabilities associated with AI-based systems and discuss the range of considerations (including technical, economic, legal, social, ethical, and environmental) that should be made in their design and development.

Table of Contents

 ISBN 978-0-9934207-7-1

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/ZVHF9195

Towards Temporal Stability in Automatic Video Colourisation

Rory Ward and John Breslin

Data Science Institute,
National University of Ireland, Galway
Galway, Ireland

Abstract

Much research has been carried out into the automatic restoration of archival images. This research ranges from colourisation, to damage restoration, and super-resolution. Conversely, video restoration has remained largely unexplored. Most efforts to date have involved extending a concept from image restoration to video, in a frame-by-frame manner. These methods result in poor temporal consistency between frames. This manifests itself as temporal instability or flicker. The purpose of this work is to improve upon this limitation. This improvement will be achieved by employing a hybrid approach of deep-learning and exemplar based colourisation. Thus, informing current frame colourisation about its neighbouring frame's colourisations and therefore alleviating the inter-frame discrepancy issues. This paper has two main contributions. Firstly, a novel end-to-end automatic video colourisation technique with enhanced flicker reduction capabilities is proposed. Secondly, six automatic exemplar acquisition algorithms are compared. The combination of these algorithms and techniques allow for an 8.5% increase in non-referenced image quality over the previous state of the art.

Keywords: Colourisation, Machine Vision, Deep Learning, Vision for graphics, Video

1 Introduction

Video colourisation is the process of applying colour to monochrome videos [Liu et al., 2021]. It has a broad range of applications, from image and video restoration [Luo et al., 2021], to self-supervised pre-training [Larsson et al., 2016]. Previous to the invention of colour video recorders, black-and-white video recorders were the standard. This has resulted in large amounts of videos which are unavailable in colour. As a result of this, a whole industry has relied on manually colouring these old videos to appeal to a modern audience who expect a more colourful cinematic experience. The process of manually colouring these videos frame-by-frame is very slow and expensive, thus leading to research into automatic video colourisation [Reinhard et al., 2001, Welsh et al., 2002, Hertzmann et al., 2001]. One of the main issues with these types of applications is temporal instability or more commonly referred to as flicker. Reducing this flicker will be the main focus of

Figure 1: Three consecutive frames from Ours (top) compared to DeOldify (bottom). As well as the frames themselves, some of their details are also displayed, these are the frame's saturation and average colour.

 ISBN 978-0-9934207-7-1

this paper. Flicker reduction is the process of removing temporal inconsistency between consecutive frames of a video [Naranjo and Albiol, 2000, Delon, 2006]. This paper has two main contributions. Firstly, a novel end-to-end automatic video colourisation technique with enhanced flicker reduction capabilities. This is achieved by combining the benefits of two standard colourisation methods while minimising their respective limitations. Secondly, six automatic exemplar acquisition algorithms are compared. These algorithms are the connection that allow for the interface between the deep learning and the exemplar based colourisation approaches. The combination of these algorithms and techniques allow for an 8.5% increase in non-referenced image quality over the previous state of the art.

The rest of this journal has the following structure. Section 2 explores the state of the art. Section 3 describes our implementation. Experimental results are shown and discussed in Section 4 and conclusions are drawn in Section 5. This is followed by future work recommendations in Section 6.

2 State of the Art

Three different methods of automatic video colourisation have been developed, these are Scribble [Levin et al., 2004], Exemplar [Zhang et al., 2019], and Deep-Learning [Zhang et al., 2016, Isola et al., 2016, Nazeri and Ng, 2018, Antic, 2019] based approaches, (See Table 1). These methods will be examined in more detail in this section.

Method	User Input Required	Temporal Consistency
Scribble	Yes	No
Exemplar	Yes	Yes
Learning	No	No

Table 1: Comparison of the various automatic video colourisation techniques. The criteria for comparison is whether the technique requires user input and if the technique ensures temporal consistency.

Some of the earliest solutions to this problem involved scribble based methods. These approaches generally worked by the user inputting scribbles into certain areas of an image, and the algorithm then colourised the image based on where it believed these scribbles were applicable to the image. These methods worked reasonably well, but did require user input and their performance depended on how well the user knew which colours were required [Heu et al., 2009].

Exemplar based methods provide the network with the reference image for it to base its colourisation on. This process tends to work well, but is highly dependent on the user's skill in finding a suitable exemplar to match the video being colourised. This generally involves a lot of competent human interaction or a strong database retrieval algorithm with a large selection of images [Liu et al., 2008].

The final method to be examined is that of deep-learning based approaches. These methods work by initially training a model on large datasets of related media. This model is then used to colourise the inputted video. This approach uses deep neural networks to learn complex relationships in media, whether they be temporal or spatial and apply them to new unseen data. These models have achieved state-of-the-art performance in this area. However, one main issue regarding learning based approaches is that of flicker or temporal inconsistency [Lai et al., 2018].

From our analysis of the existing approaches, we have observed that each one either requires user input or does not allow for temporal consistency. Our approach does not require user input and allows for temporal consistency.

3 Implementation

Our Implementation can be segmented into four main stages that follow on sequentially from each other. These stages are video collection, video pre-processing, flicker reduction and finally performance evaluation. Each of these steps will now be explained in greater detail.

3.1 Video Collection

Initially a suitable video to test our methods on needed to be sourced. Bold Emmet was selected, and then obtained through Trinity College Dublin (TCD). This video was selected because it provided a temporarily inconsistent deep-learning based colourisation which flicker reduction could be applied to. This temporarily inconsistent deep-learning based colourisation is typical of colourisation applied to older and low resolution videos. Therefore, this makes Bold Emmet a good representative of this category of videos.

3.2 Video Pre-Processing

Before flicker reduction could be performed, some pre-processing was required. Firstly the video needed to be segmented into scenes. Our flicker reduction technique works on the principle that colours within a scene are relatively consistent to the point where knowing a good representation of these colours in one frame is beneficial to the colourisation of the whole scene [Welsh et al., 2002]. Once the scene had been obtained it was necessary to colourise it using a state-of-the-art colouriser [Antic, 2019].

3.3 Flicker Reduction

Flicker reduction is a two-step process. The first step is the acquisition of a suitable exemplar image. The next step is to use this exemplar to colourise the monochrome clip. Usually obtaining a "good" exemplar image is difficult, as explained in Section 2. One of the novel parts of this paper is that this is no longer the case. One of the major challenges of this project was understanding how to describe the term "good" in relation to an exemplar image. Two metrics were investigated to determine their correlation to "good". These metrics were average colour per frame and saturation per frame.

3.3.1 Average Colour Per Frame

Flicker can be defined as a sudden change in colour between consecutive frames. As such, the average colour per frame of the video was calculated, (See Figure 2). From this the minimum and maximum values could be obtained, and then the corresponding frames compared, (See Figure 3).

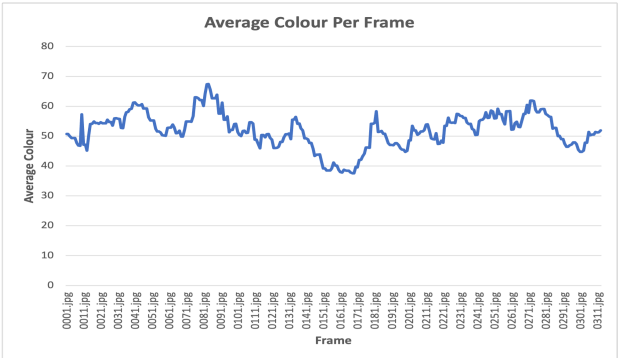


Figure 2: Plotting the average colour per frame of the colourised clip.

Figure 3: Comparing the frames with the minimum (left) and the maximum (right) average colour.

3.3.2 Saturation Per Frame

In addition to average colour per frame, saturation per frame was also calculated, (See Figure 4). Saturation per frame was chosen to be examined because it appeared that some of the colourised frames were being over-saturated during colourisation. The index of the frame with the maximum and minimum values were identified and the corresponding frames compared, (See Figure 5).

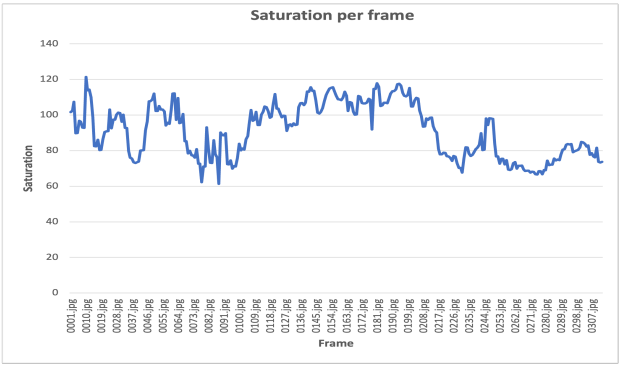


Figure 4: Plotting the saturation per frame of the colourised clip.

Comparing the relative graphs and the relative frames, (See Table 2), it can be seen that saturation has greater discriminatory power in terms of suitability of exemplar than average colour, as it has a larger standard deviation. Through visual inspection it can also be seen that saturation correlates more closely with the trend of the flicker in the video. Moreover, the sharpness of the graph suggests that oversaturation happens regularly and is not just an isolated incidence.

Method	Maximum	Minimum	Average	Std Dev
Average Colour	67.41	37.55	51.89	5.93
Saturation	121.44	61.35	91.15	15.06

Table 2: Comparing the average colour of the clip to the saturation of the clip in terms of their relative maximum value, minimum value, average value and standard deviation.

Following on from these findings, two classes of exemplar selection techniques were employed. The first class of exemplar criteria is based on the saturation of the image. Since flicker is highly correlated with change in saturation, it was used as an exemplar selection mechanism. The minimum (Min_SEA), average (Avg_SEA) and maximum (Max_SEA) saturated frames were used as an exemplar and their resultant colourisation performances were recorded. The second class of exemplar selection techniques is based on non-referenced image quality analysis, specifically the frame with the lowest NIQE (NIQEd) and BRISQUE (BRISQUEd).

Figure 5: Comparing the frame with the maximum saturation (left) to the frame with the minimum saturation (right).

The various saturation criteria was chosen to optimise the clip's saturation and therefore image quality. The blind image quality exemplar selection criteria were chosen to emphasise the best quality frames in the clip. This exemplar then provides the reference to facilitate the colourisation of the rest of the scene. The original monochrome clip was then re-colourised using the process proposed in [Zhang et al., 2019] and the selected exemplars. The results of each of the criteria were analysed and the best clips were combined using a ranking system (Ours).

3.4 Performance Evaluation

Performance can be evaluated qualitatively by visual inspection of the final video. However, performance must also be assessed quantitatively to determine its success in a more objective way. In order to achieve this characterisation of the process, six metrics will be evaluated, four using referenced analysis and two using non-referenced analysis. The four referenced metrics are Power Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM) [Horé and Ziou, 2010], Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al., 2018] and Fréchet Inception Distance (FID) [Heusel et al., 2017]. The two non-referenced metrics are are NIQE [Mittal et al., 2013], and BRISQUE [Mittal et al., 2012]. In order to compare our methods using the non-referenced techniques, we needed a standard dataset to colourise. The REDS [Nah et al., 2019] dataset was chosen as it would then allow for like-for-like comparisons with other techniques [Wan et al., 2022]. The results of the comparisons will be presented and discussed in the following section.

4 Results and Discussions

Following the methodology described in the implementation section, the evaluation could begin. Temporal consistency needed to be compared for each of the exemplar selection techniques, this would be achieved through the use of frame quality as a proxy metric. This would allow for the comparison of our approach to existing techniques. Furthermore, each of the exemplar acquisition algorithms will also be compared.

The results of our method can primarily be seen qualitatively in the finished colourised video, (See Figure 1). This figure compares two image sequences, the top being an image sequence taken from our clip and the bottom image sequence is the same image sequence but taken from the DeOldify clip. As well as the sequences themselves, some of their associated details are also included.

We will analyse our clip first, it can be seen that the frames are more consistent and natural looking. This is reflected in the associated saturation and average colour details. In contrast, looking at the DeOldify clip, it can be seen that it is less consistent and natural looking. This can particularly be seen in the third frame, where the soldier's hat takes on a very over-saturated yellow colour. This is also reflected in the associated saturation and average colour details, these values have a large spread compared to our clip. These differences can also

 ISBN 978-0-9934207-7-1

be seen in the quantitative analysis, both in the referenced and non-referenced comparisons. The results of this quantitative analysis will now be reported upon.

4.1 Referenced comparisons

Table 3 examines the referenced comparisons. This table was created by initially colourising the validation set of the REDS dataset (as outlined in the "Performance Evaluation" subsection of the "Our Implementation" section) using each of the exemplar acquisition methods. The outputs of this were then analysed using the PSNR, SSIM, LPIPS and FID metrics.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DeOldify	**30.35**	0.89	0.1	86.50
Min_SEA	30.12	0.89	0.1	87.14
Avg_SEA	30.07	0.89	0.1	85.47
Max_SEA	30.11	0.89	0.1	85.24
BRISQUEd	30.06	0.89	0.1	84.15
NIQEd	30.13	0.89	0.1	**77.23**
Ours	30.14	0.89	0.1	79.00

Table 3: Quantitative referenced comparisons on the REDS dataset.

Looking at the table, it is evident that there is no substantial difference in terms of SSIM or LPIPS between the algorithms. Perhaps, these metrics are not able to detect subtle changes in colour. NIQEd achieved the best FID score out of the exemplar selection techniques, followed closely by our technique. PSNR is highest in the DeOldify clip, although just marginally ahead of our model. Summarising the whole table, there appears to be a narrow spectrum of results. The issue with each of these metrics is that they are referenced, they are essentially measuring the distance between the colourised clip and the ground truth. However, this is generally not the main focus in colourisation as it is an ill-posed problem with multiple plausible solutions. Non-referenced metrics solve this problem by considering the naturalness of an image as opposed to its distance from the ground truth.

4.2 Non-referenced comparisons

Table 4 examines the non-referenced comparisons. This table was created by colourising Bold Emmet using each of the exemplar acquisition methods. The outputs of this were then analysed using the BRISQUE and NIQE metrics.

Method	NIQE ↓	BRISQUE ↓
DeOldify	16.23	65.43
Min_SEA	16.12	66.31
Avg_SEA	16.15	66.32
Max_SEA	16.16	66.33
BRISQUEd	16.14	66.33
NIQEd	16.15	66.33
Ours	**15.01**	**60.02**

Table 4: Quantitative non-referenced comparisons on real old films.

Looking at the table, our method outperforms DeOldify by about 8% in terms of the NIQE score and 9% in terms of the BRISQUE score. To summarise, out method achieves an 8.5% performance gain on the previous state of the art in terms of non-referenced image quality analysis.

 ISBN 978-0-9934207-7-1

5 Conclusion

We have seen that by using exemplar based colourisation after deep-learning based colourisation we can reduce flicker in a video. An appropriate exemplar must be chosen from the deep-learning based colourised video. Saturation and blind image quality evaluation were found to be useful indicators of suitability of an image to be the exemplar. We have seen that different criteria are applicable to different scenes. We learned that an optimal combination of deep-learning and exemplar colourisation techniques with automatic exemplar selection can outperform any singular colourisation technique. A hybrid approach is better than any individual exemplar, scribble or deep-learning based approach. We have contributed a novel end-to-end automatic video colourisation technique with enhanced flicker reduction capabilities. Six automatic exemplar acquisition algorithms were also compared. The combination of these algorithms and techniques allowed for an 8.5% increase in non-referenced image quality over the previous state of the art.

6 Future Work

An area of possible future work would be to expand the criteria for exemplar selection and assess their usefulness. Contrast could be investigated and assessed to evaluate its suitability as an exemplar selection criteria. This could result in a further increase in performance. Another interesting area to investigate would be using optical flow. Optical flow works on the idea of calculating motion vectors between successive frames. This could allow for more precise colourisation of particular objects within a video sequence as they travel through time.

Acknowledgments

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No.18/CRT/6223, and also by Grant Nos.16/RC/3918, 12/RC/2289_P2 and 16/RC/3835. We would like to thank the Irish Centre for High End Computing (ICHEC) for their computing resources. We would also like to thank Trinity College Dublin (TCD) for allowing us to use their videos.

References

[Antic, 2019] Antic, J. (2019). Deoldify. `https://github.com/jantic/DeOldify`.

[Delon, 2006] Delon, J. (2006). Movie and video scale-time equalization application to flicker reduction. *IEEE Transactions on Image Processing*, 15(1):241–248.

[Hertzmann et al., 2001] Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B., and Salesin, D. H. (2001). Image analogies. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, page 327–340, New York, NY, USA. Association for Computing Machinery.

[Heu et al., 2009] Heu, J.-H., Hyun, D.-Y., Kim, C.-S., and Lee, S.-U. (2009). Image and video colorization based on prioritized source propagation. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 465–468.

[Heusel et al., 2017] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium.

[Horé and Ziou, 2010] Horé, A. and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369.

[Isola et al., 2016] Isola, P., Zhu, J., Zhou, T., and Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004.

[Lai et al., 2018] Lai, W.-S., Huang, J.-B., Wang, O., Shechtman, E., Yumer, E., and Yang, M.-H. (2018). Learning blind video temporal consistency.

[Larsson et al., 2016] Larsson, G., Maire, M., and Shakhnarovich, G. (2016). Learning representations for automatic colorization. In *European Conference on Computer Vision (ECCV)*.

[Levin et al., 2004] Levin, A., Lischinski, D., and Weiss, Y. (2004). Colorization using optimization. *ACM Transactions on Graphics*, 23.

[Liu et al., 2008] Liu, X., Wan, L., Qu, Y., Wong, T.-T., Lin, S., Leung, C.-S., and Heng, P.-A. (2008). Intrinsic colorization. *ACM Trans. Graph.*, 27(5).

[Liu et al., 2021] Liu, Y., Zhao, H., Chan, K. C. K., Wang, X., Loy, C. C., Qiao, Y., and Dong, C. (2021). Temporally consistent video colorization with deep feature propagation and self-regularization learning. *CoRR*, abs/2110.04562.

[Luo et al., 2021] Luo, X., Zhang, X. C., Yoo, P., Martin-Brualla, R., Lawrence, J., and Seitz, S. M. (2021). Time-travel rephotography. *ACM Transactions on Graphics*, 40(6):1–12.

[Mittal et al., 2012] Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708.

[Mittal et al., 2013] Mittal, A., Soundararajan, R., and Bovik, A. C. (2013). Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212.

[Nah et al., 2019] Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., and Lee, K. M. (2019). Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*.

[Naranjo and Albiol, 2000] Naranjo, V. and Albiol, A. (2000). Flicker reduction in old films. In *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, volume 2, pages 657–659 vol.2.

[Nazeri and Ng, 2018] Nazeri, K. and Ng, E. (2018). Image colorization with generative adversarial networks. *CoRR*, abs/1803.05400.

[Reinhard et al., 2001] Reinhard, E., Adhikhmin, M., Gooch, B., and Shirley, P. (2001). Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41.

[Wan et al., 2022] Wan, Z., Zhang, B., Chen, D., and Liao, J. (2022). Bringing old films back to life. *CVPR*.

[Welsh et al., 2002] Welsh, T., Ashikhmin, M., and Mueller, K. (2002). Transferring color to greyscale images. *ACM Trans. Graph.*, 21(3):277–280.

[Zhang et al., 2019] Zhang, B., He, M., Liao, J., Sander, P. V., Yuan, L., Bermak, A., and Chen, D. (2019). Deep exemplar-based video colorization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8044–8053.

[Zhang et al., 2016] Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful image colorization. *CoRR*, abs/1603.08511.

[Zhang et al., 2018] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595.

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/SUHE3553

Fast and Efficient Scene Categorization for Autonomous Driving using VAEs

Saravanabalagi Ramachandran [1], Jonathan Horgan[2], Ganesh Sistu[2], and John McDonald [1]

[1]*Department of Computer Science, Maynooth University, Ireland*
[2]*Valeo Vision Systems, Ireland*

Abstract

Scene categorization is a useful precursor task that provides prior knowledge for many advanced computer vision tasks with a broad range of applications in content-based image indexing and retrieval systems. Despite the success of data driven approaches in the field of computer vision such as object detection, semantic segmentation, etc., their application in learning high-level features for scene recognition has not achieved the same level of success. We propose to generate a fast and efficient intermediate interpretable generalized global descriptor that captures coarse features from the image and use a classification head to map the descriptors to 3 scene categories: Rural, Urban and Suburban. We train a Variational Autoencoder in an unsupervised manner and map images to a constrained multi-dimensional latent space and use the latent vectors as compact embeddings that serve as global descriptors for images. The experimental results evidence that the VAE latent vectors capture coarse information from the image, supporting their usage as global descriptors. The proposed global descriptor is very compact with an embedding length of 128, significantly faster to compute, and is robust to seasonal and illumination changes, while capturing sufficient scene information required for scene categorization.

Keywords: Scene Categorization, Image Embeddings, Coarse Features, Variational Autoencoders

1 Introduction

Scene categorization is a precursor task with a broad range of applications in content-based image indexing and retrieval systems. Content Based Image Retrieval (CBIR) uses the visual content of a given query image to find the closest match in a large image database [Aliajni and Rahtu, 2020]. The retrieval accuracy of CBIR depends on both the feature representation and the similarity metric. The retrieval process can be accelerated by selectively searching based on certain scene categories e.g. given a query image with multiple high-rise buildings, searching the rural regions would not be beneficial and can be skipped. The knowledge about the scene category can also assist in context-aware object detection, action recognition, and scene understanding and provides prior knowledge for other advanced computer vision tasks [Khan et al., 2016, Xiao et al., 2010].

Figure 1: Images from the Scene Categorization Dataset. Top Left: Rural (Utah), Top Right: Urban (Toronto), Bottom Left: Rural (Stockport to Buxton), Bottom Right: Suburban (Melbourne)

[1]{saravanabalagi.ramachandran, john.mcdonald}@mu.ie. [2]{jonathan.horgan, ganesh.sistu}@valeo.com. This research was supported by Science Foundation Ireland grant 13/RC/2094 to Lero - the Irish Software Research Centre and grant 16/RI/3399.

In autonomous driving scenarios, location context provides an important prior for parameterising autonomous behaviour. Generally GPS data is used to determine if the vehicle has entered the city limits, where additional caution is required e.g. to set the pedestrian detection threshold to watch out for pedestrians in populated regions. However, such an approach requires apriori labelling of the environment and due to rapid development of regions around the cities and suburbs, it has become increasingly hard to distinguish such regions of interest only using GPS coordinates. A more scalable and lower cost approach would be to automatically determine the scene type at the edge using locally sensed data.

We present a deep learning based unsupervised holistic approach that directly encodes coarse information in the multi-dimensional latent space without explicitly recognizing objects, their semantics or capturing fine details. Models equipped with intermediate representations train faster, achieve higher task performance, and generalize better to previously unseen environments [Zhou et al., 2019]. To this end, rather than directly mapping the input image to the required scene categories as with classic data-driven classification solutions, we propose to generate an intermediate generalized global descriptor that captures coarse features from the image and use a separate classification head to map the descriptors to scene categories. More specifically, we use an unsupervised convolutional Variational Autoencoder (VAE) to map images to a multi-dimensional latent space. We propose to use the latent vectors directly as global descriptors, which are then mapped to 3 scene categories: Rural, Urban and Suburban, using a supervised classification head that takes in these descriptors as input.

2 Background

The success of deep learning in the field of computer vision over the past decade has resulted in dramatic improvements in performance in areas such as object recognition, detection, segmentation, etc. However, the performance of scene recognition is still not sufficient to some extent because of complex configurations [Xie et al., 2020]. Early work on scene categorization includes [Oliva and Torralba, 2001] where the authors proposed a computational model of the recognition of real world scenes that bypasses the segmentation and the processing of individual objects or regions. Notable early global image descriptor approaches include aggregation of local keypoint descriptors through Bag of Words (BoW) [Csurka et al., 2004], Fisher Vectors (FV) [Perronnin et al., 2010, Sanchez et al., 2013] and Vector of Locally Aggregated Descriptors (VLAD) [Jégou et al., 2010]. More recently, researchers have also used Histogram of Oriented Gradients (HOG) and its extensions such as Pyramid HOG (PHOG) for mapping and localization [Garcia-Fidalgo and Ortiz, 2017]. Although these approaches have shown strong performance in constrained settings, they lack the repeatability and robustness required to deal with the challenging variability that occurs in natural scenes caused due to different times of the day, weather, lighting and seasons [Ramachandran and McDonald, 2019].

To overcome these issues recent research has focussed on the use of learned global descriptors. Probably the most notable here is NetVLAD which reformulated VLAD through the use of a deep learning architecture [Arandjelovic et al., 2016] resulting in a CNN based feature extractor using weak supervision to learn a distance metric based on the triplet loss.

Variatonal Autoencoder (VAE), introduced by [Kingma and Welling, 2013], maps images to a multi-dimensional standard normal latent space. Although since the introduction of the CelebA dataset [Liu et al., 2015] multiple implementations of VAEs have shown success in generating human faces, VAEs often produce blurry and less saturated reconstructions and have been shown to lack the ability to generalize and generate high-resolution images for domains that exhibit multiple complex variations e.g. realistic natural landscape images. Besides their use as generative models, VAEs have also been used to infer one or more scalar variables from images in the context of Autonomous Driving such as for vehicle control [Amini et al., 2018].

A number of researchers have developed datasets to accelerate progress in general scene recognition. Examples include MIT Indoor67 [Quattoni and Torralba, 2009], SUN [Xiao et al., 2010], and Places 365 [Zhou et al., 2017]. Whilst these datasets capture a very wide variety of scenes they lack suitability when developing scene categorisation techniques that are specific to autonomous driving. Given this, in our research we choose to use images from public driving datasets such as Oxford Robotcar [Maddern et al., 2017] in an unsupervised manner and curate our own evaluation dataset targeted at our domain of interest.

In this paper, we propose to use a trained Variational Autoencoder to map images to a multi-dimensional latent space and use the latent vectors as compact embeddings that serve directly as global descriptors for images. To the best of our knowledge, this is the first time VAE latent vectors are used as global image descriptors. In detail, we train a convolutional Variational Autoencoder in an unsupervised manner with images from Oxford Robotcar dataset [Maddern et al., 2017] that exhibit strong visual changes caused by seasons, weather, time of the day, etc. and use the latent vectors inferred using the encoder as global descriptors. We show that VAE encoder captures coarse features of the image and produces a mapping in multi-dimensional standard normal latent space. We then use a simple two-layer linear classification perceptron head to map the global descriptors to required scene categories: Rural, Urban and Suburban.

3 Methodology

In our method, we train a Variational Autoencoder from scratch on images from publicly available Oxford Robotcar dataset [Maddern et al., 2017] in an unsupervised manner. We then obtain latent vectors during inference and use them as compact embeddings that serve as global descriptors. The global descriptors are then used as input to a simple linear classifier that maps them to 3 scene categories: Rural, Urban and Suburban. Figure 2 provides a high-level view of our overall architecture.

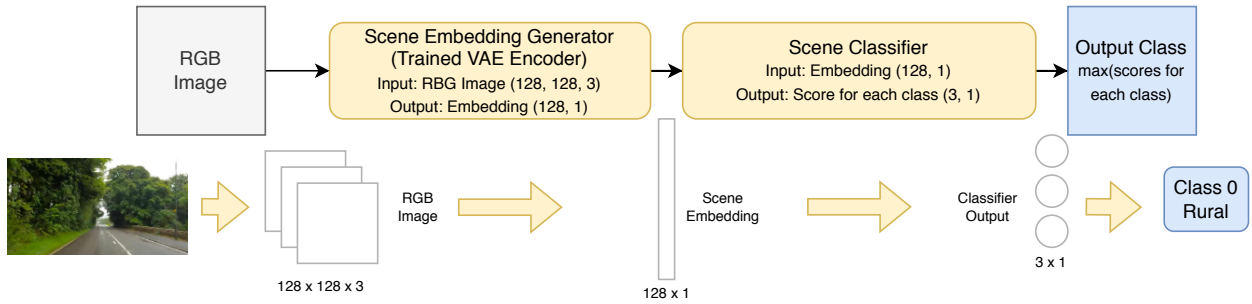


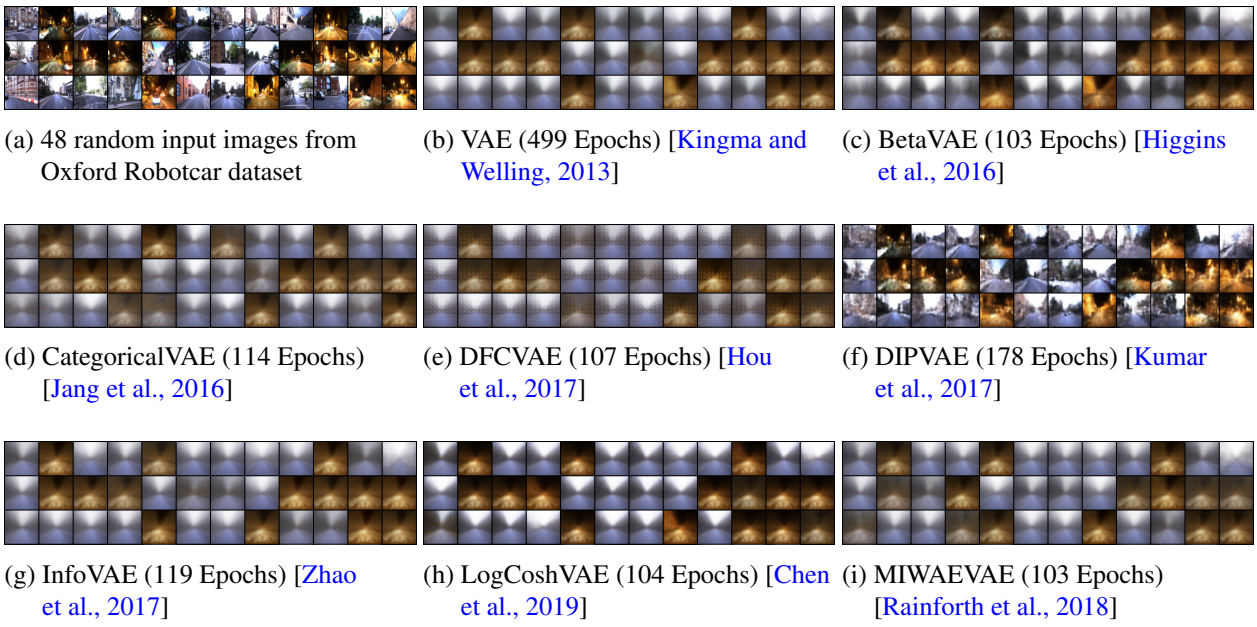
Figure 2: Overall system architecture for scene categorization. Input and output for each module is shown in the bottom row including an example.

3.1 Scene Embedding

In order to evaluate the performance of different VAEs for scene embedding, we train multiple variants on 3 Oxford Robotcar traversals *2014-12-09-13-21-02* (Winter Day), *2014-12-10-18-10-50* (Winter Night), and *2015-05-19-14-06-38* (Summer Day). These traversals exhibit changes due to seasons and time of the day. We sub-sample each traversal using the sequential adaptive sampling strategy reported in [Ramachandran and McDonald, 2021] with parameters $\tau_{d_{\text{acc}}} = 5m$ and $\tau_{\theta_{\text{acc}}} = 15°$ to obtain 1787, 1879 and 1825 visually dissimilar images respectively. The following variants of VAE are trained:

- BetaVAE [Higgins et al., 2016]
- CategoricalVAE [Jang et al., 2016]
- DFCVAE [Hou et al., 2017]
- DIPVAE [Kumar et al., 2017]
- InfoVAE [Zhao et al., 2017]
- LogCoshVAE [Chen et al., 2019]
- MIWAE [Rainforth et al., 2018]
- VAE (Vanilla) [Kingma and Welling, 2013]

For all VAEs, we use the standard convolutional encoder and decoder with LeakyReLU [Maas et al., 2013] activated strided convolutional and transposed convolutional blocks with BatchNorm [Ioffe and Szegedy, 2015], respectively. This architectural setting allows various implementational advantages where convolutional accelerators and other ASICs can be used to speed up inference to achieve lower-latency and near-realtime performance. We resize the images to 64x64 and all VAEs use reconstruction as the primary task, where the loss



(a) 48 random input images from Oxford Robotcar dataset

(b) VAE (499 Epochs) [Kingma and Welling, 2013]

(c) BetaVAE (103 Epochs) [Higgins et al., 2016]

(d) CategoricalVAE (114 Epochs) [Jang et al., 2016]

(e) DFCVAE (107 Epochs) [Hou et al., 2017]

(f) DIPVAE (178 Epochs) [Kumar et al., 2017]

(g) InfoVAE (119 Epochs) [Zhao et al., 2017]

(h) LogCoshVAE (104 Epochs) [Chen et al., 2019]

(i) MIWAEVAE (103 Epochs) [Rainforth et al., 2018]

Figure 3: Reconstructions of different variants of VAE

function is given as:

$$\mathrm{L}(x;\phi,\theta) = -\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] + \mathrm{D}_{\mathrm{KL}}(q_\phi(z|x)||p_\theta(z)) \tag{1}$$

where x is the input image, z is the latent vector, ϕ and θ denote parameters of the encoder and decoder, respectively, and D_{KL} is the Kullback-Leibler divergence. Note that for each VAE variant, the corresponding loss will also contain additional specialized loss terms. The reader is referred to the associated paper for each variant for further details.

We train all VAEs at a constant learning rate of 0.005 and no weight decay up to a maximum of 500 epochs with an early stopping criteria set on validation loss with a patience of 100 epochs, i.e, if there is no improvement for the past 100 epochs, the training ends. Once the training is complete, we check the reconstructed images manually as shown in Figure 3. DIPVAE [Kumar et al., 2017] produces reasonably good reconstructions among all other VAEs, and the reconstructions evidence that it captures coarse information necessary to understand the scene. We hypothesize that DIPVAE reconstructions are less blurry as disentanglement is encouraged by introducing a regularizer over the induced inferred prior. β-VAE [Higgins et al., 2016] also encourages disentanglement, but with DIPVAE there is no extra conflict introduced between disentanglement of the latents and the observed data likelihood. Further, we additionally trained a DIPVAE on 128x128 images, which yielded similar reconstruction results.

Once trained, we use the VAE encoder to infer latent vectors for input images to use as compact global descriptor embeddings. As such, the decoder module of the VAE contributes to loss during training and is not used during inference. Nonetheless, the decoder may still be used to reconstruct images from latent vectors which facilitates interpreting and visualizing the global descriptors. The encoder is confined to standard normal distribution and hence, this makes it easier to tweak latent vectors and to visualize their corresponding reconstructions. This makes it possible to understand the feature or variation encoded in required dimensions using the decoded image. Additionally, DIPVAE encourages disentangling of features in the latent space, which results in less overlap of variations across the dimensions, producing more meaningful and interpretable intermediate representations to use as global descriptors. Ultimately, such configurations producing intermediate representations instead of directly mapping pixels to actions, are known to achieve higher task performance, and generalize better to previously unseen environments [Zhou et al., 2019].

 ISBN 978-0-9934207-7-1

3.2 Scene Classification

We train a linear classifier on top of the frozen base VAE network on the train split of the dataset for 100 epochs. A two-layer input-output perceptron without any activation function is used as the linear classifier.

We curate our own dataset[1] for training and testing the classifier, by manually selecting screenshots at different timestamps from driving videos on YouTube. This dataset includes 3 scene categories: Rural, Urban and Suburban, each category includes images captured in or around a diverse set of cities and regions as shown in Table 1. Some examples are shown in Figure 1 and as can be seen the dataset covers a variety of landscapes including desert and mountainous landscapes and exhibit mild to moderate illumination and seasonal changes such as fallen leaves, different time of the day, etc. The train split is made by randomly selecting two-thirds of the images from each route yielding 314 images and the linear classifier is trained.

Rural		**Suburban**		**Urban**	
Jarrahdale Perth	33	Hawaii	8	Indianapolis	30
Missouri Ozarks	22	Howth	17	Nashville	21
Southern Illinois	43	Melbourne	33	Paris	24
Stockport Buxton	25	Stockport Buxton	17	St Louis	52
Utah	75	Wimbledon	21	Toronto	33
Rural	198	Suburban	96	Urban	160
Total					**454**

Table 1: Information about our Scene Categorization dataset

4 Experiments

To verify the suitability of the embeddings for scene categorization, we use the widely used evaluation procedure employed to test embeddings for classification tasks [Deng et al., 2009, He et al., 2020]. Our proposed architecture already uses an intermediate global descriptor representation, which is then input to the linear classifier for scene categorization. Hence, we evaluate the resultant output without adding any additional layers.

4.1 Evaluation

The output of the linear classifier is tested on the test split of the dataset containing the remaining 140 images and the test accuracy is used as a proxy for representation quality of the embeddings used. Evaluation was done on an Intel i9-9900K (8 cores @3.60 GHz) and Nvidia RTX 2080 Ti and all images were resized to 128x128. We compare the results with benchmark learned and handcrafted holistic image descriptors: (1) NetVLAD[2] [Arandjelovic et al., 2016], a weakly supervised CNN with generalized VLAD (Vector of Locally Aggregated Descriptors) layer. (2) PHOG[3] [Bosch et al., 2007], Pyramid Histogram of Gradients. For the evaluation we consider the following candidates:

- NetVLAD 4096 dimensions: Supervised, pretrained on Pittsburgh dataset[4]
- NetVLAD 128 dimensions: Supervised, pretrained, same as above, cropped to 128 dimensions and L2-normalized from NetVLAD 4096 embedding
- PHOG 1260 dimensions: Handcrafted, 60 bins and 3 levels [Garcia-Fidalgo and Ortiz, 2017]
- DIPVAE[5] 128 dimensions: Unsupervised, pretrained on 128x128 Oxford Robotcar dataset images

The experimental results are shown in Table 3. As expected, the supervised techniques scores higher than the unsupervised and handcraft techniques. NetVLAD 4096 tops the evaluation with 99.29% accuracy, followed by the NetVLAD 128 with 94.29% accuracy. The high accuracy is the result of (1) the technique's use

[1] Dataset available to download from https://gist.github.com/saravanabalagi/1cda6ae06c4cf722fd2227e83eadc792
[2] MATLAB implementation provided by authors at https://github.com/Relja/netvlad is used
[3] Code extracted from C++ implementation provided by authors at https://github.com/emiliofidalgo/htmap is used
[4] Off-the-shelf VGG16+NetVLAD+whitening model provided at https://www.di.ens.fr/willow/research/netvlad/
[5] Our own implementation in Python 3.8 and PyTorch 1.11 (CUDA 11.3) is used

 ISBN 978-0-9934207-7-1

Descriptor	Type	Dimensions ↓	Accuracy (%) ↑	Compute Time (μs) ↓
Random	Trivial	4096	34.29	71.3 ± 0.0
Random	Trivial	128	28.57	2.7 ± 0.0
NetVLAD	Supervised	4096	**99.29**	27560.0 ± 230.2
NetVLAD Cropped	Supervised	**128**	94.29	27563.9 ± 230.4
PHOG	Hand-crafted	1260	84.29	123.6 ± 3.9
DIPVAE (Ours)	Unsupervised	**128**	82.86	**60.4 ± 3.0**

Table 3: Test Accuracy reported on Scene Categorization Dataset for different descriptors mentioned in Section 3.2. Random descriptor, constructed trivially by sampling numbers from normal distribution, is shown in top provide a baseline for trivial descriptors that do not capture any relevant information from the images. Compute time is the time taken to obtain the descriptor from a decoded image loaded in memory.

of supervised learning, (2) the embedding length of 4096 allows capturing more information about the scene, and, (3) NetVLAD 128 is computed from NetVLAD 4096 by cropping and normalizing. DIPVAE R128 (128 dimensions) achieves 82% accuracy while only using 10% embedding size as that of PHOG (1260 dimensions) and 3.1% embeddings size as that of NetVLAD (4096 dimensions). We note that both NetVLAD and DIPVAE uses GPU acceleration, while PHOG uses CPU optimizations and multi-threading. DIPVAE is computed more than twice as fast as PHOG and several orders of magnitude faster than NetVLAD.

We further evaluate the linear classifiers on a second video based dataset[1]. Here, we utilised frames from a variety of extended driving videos collected from YouTube as shown in Table 2. Each video was labelled as a single scene category, where collectively this resulted in a total of over 185K images. On each route, we remove the first 900 frames (30 seconds at 30 fps) to avoid encountering intro text, crossfades and other effects, and use first 20% of the frames (~40K) for training and the rest 80% (~145K) for evaluation. We note that some portios of the sequences may exhibit ambiguous scene types and hence there will be a consequent noise in the results e.g. some images from urban sequences driven may resemble suburbs or rural regions. However given the length of each sequence, we estimate all Rural (Wicklow and Redwood) and City (Dublin and Vancouver) videos contain at least three quarters of images are unambiguously mapped to *Rural* and *Urban* labels, respectively. Therefore, we consider a model to be demonstrating good performance if it scores above 75%. Table 2 shows the accuracy of the descriptors on videos of each region or city. As such, DIPVAE performs consistently well and shows similar performance to that of NetVLAD and PHOG while having much smaller embedding dimensionality and significantly faster compute time.

Route	Dublin	Vancouver	Wicklow	Redwood
Total Images	14677	64672	60509	45253
Test Images	11022	51018	47688	35483
NetVLAD 4096	93.37	96.99	99.99	99.86
NetVLAD 128	78.47	98.84	99.98	99.87
PHOG 1260	91.25	98.50	88.46	86.60
DIPVAE 128	95.70	83.72	99.44	95.60

Table 2: Accuracy (in %) on the Scene Categorization Video Dataset

5 Conclusion and Future Work

Our proposed solution to scene categorization uses an architecture made up of an unsupervised VAE-based embedding generator and a supervised light linear classifier head, and produces meaningful and interpretable intermediate representations as opposed to an end-to-end pixel-to-class approach with no explicit intermediate representations. The experimental results evidence that the DIPVAE latent vectors capture coarse information from the image, supporting their usage as global descriptors. These global descriptors exist in the multi-dimensional standard normal manifold, allowing easier comparison and interpretation compared to unbounded embedding hyperspaces. The proposed global descriptor is very efficient with a compact embedding length of

128, significantly faster to compute, and is robust to seasonal and illuminational changes, while capturing sufficient scene information required for scene categorization. Further, the VAE backbone's architecture made up of standard convolutional blocks allows more efficient, fast, low latency and near-realtime inference using hardware convolutional accelerators, substantiating their use in autonomous vehicles to quickly determine location context as precursor task. We note that there is a potential to further improve this performance using supervised and weakly supervised techniques. If and when labels are available, the VAE backbone can be set to learn with a small learning rate (e.g. one-tenth relative to that of the head) in an end-to-end manner. Additionally, further available information, such as GPS, together with recent predictions, could also be used to make more temporally consistent decisions about the scene category (e.g. avoiding categorising the environment as rural when driving along a tree-lined route in a city). Finally, we indicate that the proposed global descriptors, being intermediate representations, are useful for other tasks and actions that only require coarse features including scene information present in the image.

In our future work we intend to explore the potential of adding further categories such as motorways, tunnels, car-parks, etc. that are useful and provide more context for various autonomous driving tasks. Given the successful results, we further intend to integrate this approach in a hierarchical place recognition pipeline, where these compact global representations are used to aggregate images to facilitate faster image retrieval.

References

[Aliajni and Rahtu, 2020] Aliajni, F. and Rahtu, E. (2020). Deep Learning Off-the-shelf Holistic Feature Descriptors for Visual Place Recognition in Challenging Conditions. In *2020 IEEE 22nd MMSP*, pages 1–6.

[Amini et al., 2018] Amini, A. et al. (2018). Variational autoencoder for end-to-end control of autonomous driving with novelty detection and training de-biasing. In *2018 IEEE/RSJ IROS*, pages 568–575. IEEE.

[Arandjelovic et al., 2016] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE CVPR 2016*, pages 5297–5307.

[Bosch et al., 2007] Bosch, A., Zisserman, A., and Munoz, X. (2007). Representing Shape with a Spatial Pyramid Kernel. In *Proceedings of the 6th ACM CIVR 2007*, page 401–408, New York, NY, USA. ACM.

[Chen et al., 2019] Chen, P., Chen, G., and Zhang, S. (2019). Log Hyperbolic Cosine Loss Improves Variational Auto-Encoder.

[Csurka et al., 2004] Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*. Prague.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE CVPR*, pages 248–255.

[Garcia-Fidalgo and Ortiz, 2017] Garcia-Fidalgo, E. and Ortiz, A. (2017). Hierarchical Place Recognition for Topological Mapping. *IEEE Transactions on Robotics*, 33(5):1061–1074.

[He et al., 2020] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF CVPR*, pages 9726–9735.

[Higgins et al., 2016] Higgins, I. et al. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR 2017*.

[Hou et al., 2017] Hou, X., Shen, L., Sun, K., and Qiu, G. (2017). Deep feature consistent variational autoencoder. In *2017 IEEE WACV*, pages 1133–1141. IEEE.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456. PMLR.

[Jang et al., 2016] Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

[Jégou et al., 2010] Jégou, H., Douze, M., Schmid, C., and Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *2010 IEEE CVPR*, pages 3304–3311. IEEE.

[Khan et al., 2016] Khan, S. H., Hayat, M., Bennamoun, M., Togneri, R., and Sohel, F. A. (2016). A discriminative representation of convolutional features for indoor scene recognition. *IEEE TIP*, 25(7):3372–3383.

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Kumar et al., 2017] Kumar, A., Sattigeri, P., and Balakrishnan, A. (2017). Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*.

[Liu et al., 2015] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738.

[Maas et al., 2013] Maas, A. L., Hannun, A. Y., Ng, A. Y., et al. (2013). Rectifier nonlinearities improve neural network acoustic models. In *ICML*, volume 30, page 3. Citeseer.

[Maddern et al., 2017] Maddern, W., Pascoe, G., Linegar, C., and Newman, P. (2017). 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research*, 36(1):3–15.

[Oliva and Torralba, 2001] Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175.

[Perronnin et al., 2010] Perronnin, F., Sánchez, J., and Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer.

[Quattoni and Torralba, 2009] Quattoni, A. and Torralba, A. (2009). Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE.

[Rainforth et al., 2018] Rainforth, T., Kosiorek, A., Le, T. A., Maddison, C., Igl, M., Wood, F., and Teh, Y. W. (2018). Tighter variational bounds are not necessarily better. In *ICML*, pages 4277–4285. PMLR.

[Ramachandran and McDonald, 2019] Ramachandran, S. and McDonald, J. (2019). Place Recognition in Challenging Conditions. In *Irish Machine Vision and Image Processing Conference*.

[Ramachandran and McDonald, 2021] Ramachandran, S. and McDonald, J. (2021). OdoViz: A 3D Odometry Visualization and Processing Tool. In *2021 IEEE ITSC*, pages 1391–1398.

[Sanchez et al., 2013] Sanchez, J., Perronnin, F., Mensink, T., and Verbeek, J. (2013). Image Classification with the Fisher Vector: Theory and Practice. *International Journal of Computer Vision*, 105(3):222–245.

[Xiao et al., 2010] Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE CVPR*, pages 3485–3492. IEEE.

[Xie et al., 2020] Xie, L., Lee, F., Liu, L., Kotani, K., and Chen, Q. (2020). Scene recognition: A comprehensive survey. *Pattern Recognition*, 102:107205.

[Zhao et al., 2017] Zhao, S., Song, J., and Ermon, S. (2017). InfoVAE: Information Maximizing Variational Autoencoders. *CoRR*, abs/1706.02262.

[Zhou et al., 2019] Zhou, B., Krähenbühl, P., and Koltun, V. (2019). Does computer vision matter for action? *Science Robotics*, 4(30):eaaw6661.

[Zhou et al., 2017] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on PAMI*, 40(6):1452–1464.

 ISBN 978-0-9934207-7-1

Detection and Isolation of 3D Objects in Unstructured Environments

Dylan Do Couto[1], Joseph Butterfield[1], Adrian Murphy[1], Karen Rafferty[1], Joseph Coleman[2]

[1]*Queen's University, Belfast,* [2]*Combilift*

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/AFMZ9460

Abstract

3D machine vision is a growing trend in the filed of automation for Object Of Interest (OOI) interactions. This is most notable in sectors such as unorganised bin picking for manufacturing and the integration of Autonomous Guided Vehicles (AGVs) in logistics. In the literature, there is a key focus on advancing this area of research through methods of OOI recognition and isolation to simplify more established OOI analysis operations. The main constraint in current OOI isolation methods is the loss of important data and a long process duration which extends the overall run-time of 3D machine vision operations. In this paper we propose a new method of OOI isolation that utilises a combination of classical image processing techniques to reduce OOI data loss and improve run-time efficiency. Results show a high level of data retention with comparable faster run-times to previous research. This paper also hopes to present a series of run-time data points to set a standard for future process run-time comparisons.

Keywords: 3D vision, Image Processing, Object Detection.

1 Introduction

3D vision is a field of automation engineering that, despite existing for over 20 years, is still in the early stages of development. A current research interest is in the isolation of Objects of Interest (OOIs) in 3D data from its surroundings for 3D object analysis, through the use of object recognition [Qi et al., 2021]. Key methods presented in the state-of-the-art literature have been developed using a Machine Learning (ML) approach [Sun et al., 2021], this allows for a system to be trained on a dataset of comparable objects to identify and separate OOIs from complex scenes. Despite continual optimisation in this approach two problems persist; data loss and long run times as a result of a dependence on neural network techniques [Aloni and Yitzhaky, 2015, Brazil and Liu, 2019, Li et al., 2022, Huang et al., 2020, Sindagi et al., 2019]. This is more prevalent in real world environments where the scene is populated with data that is not related to OOI, also known as unstructured environments [Guastella and Muscato, 2021]. A solution to this problem is to preform pre-processing operations that reduce the number of objects in the environment that are considered by the ML process, [Li et al., 2020a], however these methods still have a reliance on ML techniques that are both computationally intensive and subject to false positive feature matches. In this study we present an accurate and efficient method of isolating OOIs from complex unstructured environments in contrast what can be found in state-of-the-art ML object recognition systems. The current state-of-the-art methods of object isolation, most notably ML methods, do not provided a direct metric for the accuracy of object isolation but rather a confidence interval of the correct object being isolated [Avraham and Yitzhaky, 2021, Aloni and Yitzhaky, 2015, Williams et al., 2021]. In this study a direct accuracy metric is established based on both the percentage of pixels correctly isolated related to the OOI and the percentage of excess pixels isolated that do not relate to the OOI. The method presented in this study incorporates a dual data approach to classical image processing of 2D depth map and colour image datasets. This approach allows for the use of Morphological technique's [Dougherty, 2018, Maragos, 1996] to isolate the OOI data from the depth map data, while simultaneously recording the 3D capture devices properties and capture conditions to convert the isolated data into a standard point cloud format [Chmelar et al., 2016, Perra et al., 2016, Chen et al., 2017]. Acknowledged in this study is the limitation of this approach for cases of

clustered OOIs (unorganised bin picking), and fringe cases of obscured objects. However, the ultimate use case for this method is intended to aid in large item autonomous logistics, such as but not limited to, Autonomous Guided Forklifts/Vehicles (AGFs/AGVs). To summarise, presented in this study is a new method of OOI isolation from complex scenes without the use of ML technique's. The ultimate goal of this study is to retrieve an isolated point cloud of an OOI while simultaneously reducing operation run-time and data loss. To achieve this goal an in-depth study and implementation of classical image processing techniques is required. This is presented in the methodology section. Lastly to verify this method extensive and varying datasets are captured and presented in detail in the results section, with further discussion presented in the conclusion.

2 Methodology

The goal of this study is to develop an efficient and precise method of isolating 3D data of an OOI from an environment where the properties of the OOI and it's position in 3D space is not explicitly known. The method presented in ths study isolates an OOI based on its approximate position in 3D space, in contrast to isolation methods that operate using keypoints, convoluted neural networks and other forms of template matching [He et al., 2020, Li et al., 2020b, Barabanau et al., 2019, Feng et al., 2019]. To achieve this goal, a morphological approach was taken that relies on research conducted for unstructured 2D image processing [Dougherty, 2018, Serra and Soille, 2012, Codaro et al., 2002]. Before exploring the methods involved in isolated the object in 2D data, it is important to first discuss the hardware used and how it can best utilize in this use case. In this study a Time Of Flight (TOF) 3D camera (Basler Blaze) was used due to several properties of the system, such as a relatively high capture rate of 30 frames per second and a moderate resolution of 640 by 480 pixels. However, the main properties of this technology that makes it suitable for this study is the method of capture. TOF cameras operate by pulsing ultraviolet light and recording the delay from its light source to the capture camera as the light reflects off of objects in the scene. This returns 3D data in the form of a 2D depth map represented as a gray-scale image. The camera will also record the raw image captured of the reflected light, presented as an intensity image, also captured in gray-scale. The depth map data provides useful context to the position of objects captured in 3D space, while the intensity data returned provides context to the visual properties of the objects captured such as the objects texture and colour in the form of gray-scale data (Figure 1). These two distinct types of data captured allow for the application of classical image process techniques in novel ways. To explore this, we will begin with a description of what these classical techniques are followed by how they can be applied to the data captured in this study.

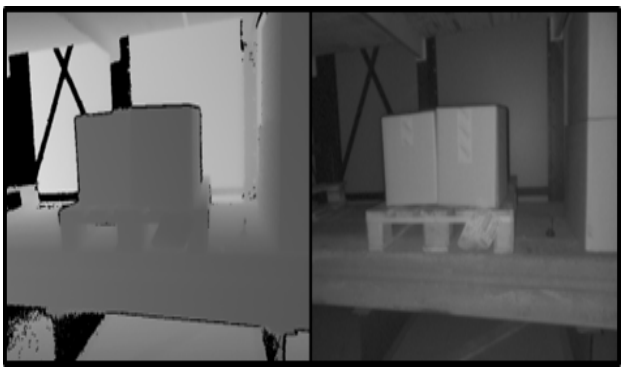


Figure 1: Depth map of scene (left), intensity image of scene (right).

2.1 Classical Image Processing

In classical image processing, the approach is to analyse the structure and composition of pixels to determine patterns that constitute the OOI. In colour images the simplest form of this is determining the colour of the desired OOI and isolating this colour in the image data. In Figure 2 [a] a 2d colour-based dataset is presented. The OOIs in this case is simply the cardboard boxes placed on a pallet. The colour of these boxes is easily distinguishable from other objects in the scene such as the yellow support beams, the white walls, etc. Using this distinct brown colour, we can create a mask that only includes data within a certain range of this colour, Figure 2 [b], and use this mask to remove all other data from the source image, presenting only the OOIs, Figure 2 [c].

 ISBN 978-0-9934207-7-1

Figure 2: Source image [a], isolated data in desired colour range [mask] [b], mask applied to source data [c].

2.1.1 Intensity Data

Observing the intensity data provided by the TOF vision system, the same colour-based data is presented, however the data in this case is in gray-scale format. The same principle of classical image processing can be applied to this data, within a broader isolation range due to the data being single channel (gray-scale) rather than three channel RGB (Red, Green, Blue). To separate the OOIs in intensity data, as shown in Figure 3 [a], a gray-scale value rather than a RGB range is used to threshold hold the image. This is an effective tool for isolating objects based on the material and colour properties. Applying a simple threshold operation on the source data, Figure 3 [a], we can isolate objects in the scene that are a lighter colour, such as the brown of cardboard boxes when captured in grayscale data, and objects with a relatively smooth surface, in contrast to the rough texture of back wall, Figure 3 [b]. Using this data as a mask it can be applied to the source data to isolate only the OOI data, Figure 3[c].

Figure 3: Colour-based isolation method applied to TOF intensity data. Source data [a], threshold data [mask][b], mask applied to source [isolated] data [c].

2.1.2 Depth Data

In contrast to the intensity data, depth data is presented in a depth map, a grayscale image where the value of each pixel relates to a scaled distance from the capture device. Without knowing the precise position of an OOI in a scene, we can use an approximation of its position to isolate the OOI within a certain range. observing Figure 4, the isolation process is presented as it was for the previous sections. It is important to note that in contrast to colour-based isolation explored with the intensity data, Figure 3, a larger range of data is isolated to encompass both the cardboard boxes and the full view of the pallet, resulting in more unwanted data captured.

 ISBN 978-0-9934207-7-1

Figure 4: Isolation method applied to range data [depth map]. Source image [a], data within desired depth range [mask][b], isolated depth data using mask [c].

2.2 Combination of Methods

The methodology thus far has presented techniques for isolating OOIs in each dataset, however two problems persist. The intensity data can be easily and effectively isolated to present only the OOIs, the carboard boxes and the front of the pallets, however the data holds no useful information concerning the objects position in 3D space. The same isolation process can be applied to the depth data, which can precisely convey an objects 3D position within +- 5mm, however the classical methods applied here result in excessive amounts of unwanted data and poor OOI isolation. However, there is a property of TOF systems that can be exploited here. Both datasets are captured from the same optical sensor, allowing for a direct overlay of the datasets and isolation masks. Using a combination of the two processes, the isolation process of the intensity data and the range data separately, can allow for the isolation only the OOIs that match a general visual description and are in an approximate position in 3D space we're interest in. This allows for two stages of separation when combined through a bitwise "AND" operation, isolating only the range data of an OOI within a desired general position in 3D space, Figure 5. As a result, the process allows for the separation of only the OOIs in the scene cleanly with relatively lower amounts of noise, with excess data collected typically below 25% of all isolated data in comparison to excess values of over 40% when using only one of the isolation techniques.

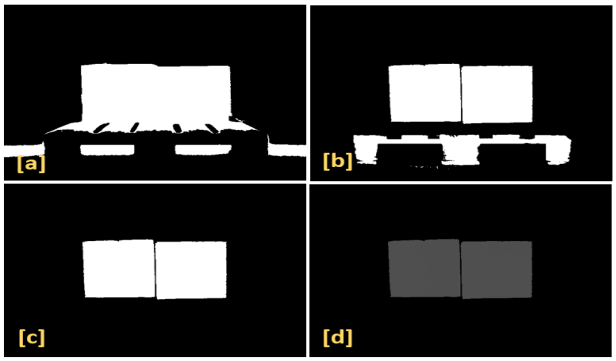


Figure 5: Range based isolation [a], Colour based Isolation [b], Combined isolation (Bitwise-and of masks)[c], Final isolation of depth map [d].

3 Results

To further explore this process 10 OOI scenes were captured as shown in Figure 6, with the intensity image, isolated range map and resulting point cloud displayed left to right. Each scene consists of increasingly complex arrangements of OOIs, such as multiple objects that match the visual description of the OOI with a varying depth distribution, i.e. scene 5, and partial obstructed views of the OOI producing low levels of occlusion, such as scene 8 and 9. To evaluate the isolation process, two approaches were taken. The first approach is to assess the accuracy of the isolation process by comparing the algorithm isolated point cloud data to point cloud data that was hand curated to isolate only the OOI. The second approach is to determine the efficiency of the isolation process by recording the runtime of isolation for each data and evaluating the mean, standard deviation and the 99th percentile of runtime data.

 ISBN 978-0-9934207-7-1

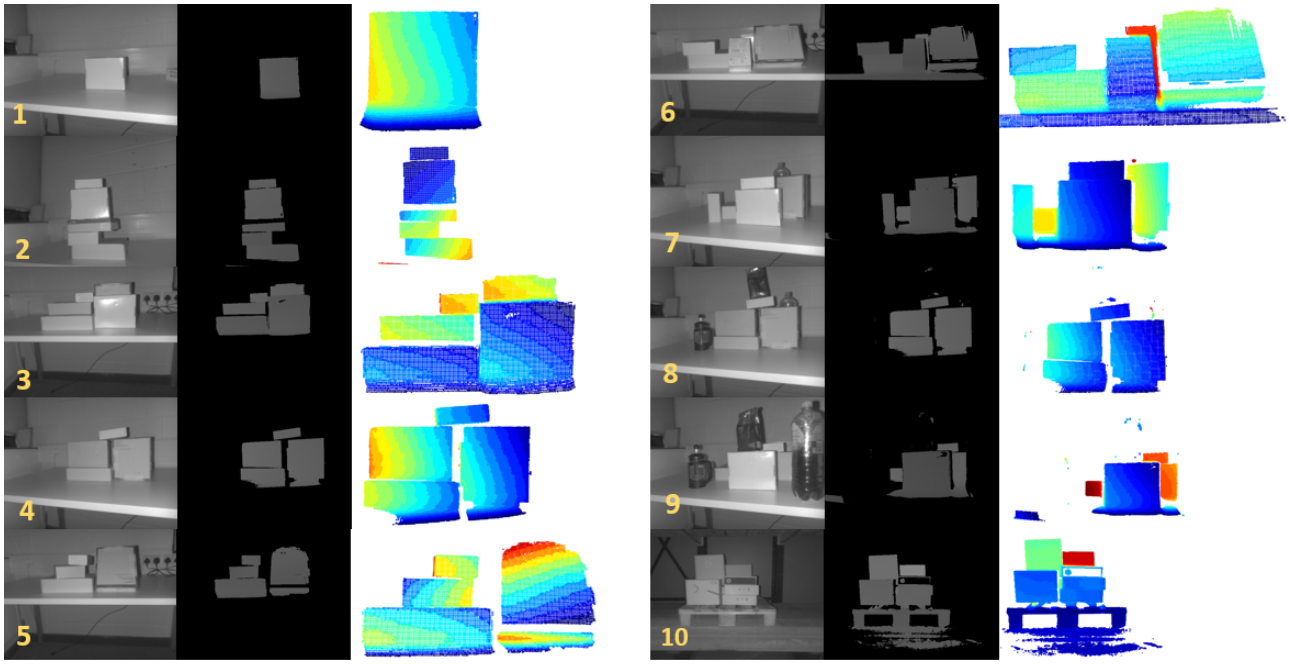


Figure 6: Images of experimental scenes 1 through 10.

3.1 Isolation Accuracy

The accuracy of each of the isolated data varies greatly due to several factors. The first factor is the positioning of multiple objects relative to each other, objects that are in close proximity to each other are more difficult to segment based on their 3D position or distance from the capture device. A further factor that greatly impacts the isolation process is the range of object colours. For the intensity data, the colour and texture of the object is a key property used to distinguish them. As a result, objects that are near the OOIs in 3D space and share a comparable colour are harder to distinguish and as a result are not removed from the isolation process. This constraint can be reduced through using more advanced 3D imaging systems that also collect RGB colour data that is overlaid on the depth data, resulting in a larger range of data that can be used for colour-based isolation, as shown in Figure 2. With these constraints in mind, the isolation process was evaluated against hand curated isolated data to determine the accuracy of the method presented as shown in Figure 7.

Figure 7: Intensity image of load (left), curated depth map (centre), point cloud of curated load (right).

Using this ideal data, a bit wise “NOT” operation is preformed to remove the ideal data from the algorithm output. The remaining data can therefore be classified as excess data that is not related to the OOI. The number of points present in this excess data gives an indication of points that were incorrectly isolated. In addition to this, a bitwise “AND” operation can be performed to determine points in common between the isolated data

and the hand curated data. The number of points present after this operation indicate how many points were correctly isolated in the isolation process, here after referred to as inliers. To achieve this, the depth maps of the algorithmic isolated process and the hand curated process were used. This allows for an easier bitwise operation as the data is in image format rather than point cloud format. The number of excess points detected relative to the total number of points present in the isolated data was used to determine the excess percentage of points isolated. The number of points present in the inlier data relative to total number of points present in the hand curated data was used to determine the percentage of correctly isolated points. Performing this calculation on each scenes, 1 through 10, the following table was produced to present the accuracy of the isolation process. For scenes that consist of simple OOI structures and clear colour space differences between the OOI and the background, a higher percentage of correctly isolated points can be seen, with a simultaneous low excess point percentage. The opposite can be seen for more complex scenes that have a similar colour space to the background data in addition to the 3D position of the OOI coinciding with other objects of a similar colour space as shown in scene 6, Figure 6. Calculating the average of these values, the process can be determines to have an average excess percentage of 21.94% and inlier percentage of 90.09%., with a 99th percentile value of 34.23% and 82.24% respectively.

scene	Excess %	Inliers %	scene	Excess %	Inliers %
1	20.14%	99.64%	6	34.54%	88.55%
2	15.44%	87.76%	7	29.92%	87.87%
3	6.81%	93.85%	8	17.21%	90.18%
4	16.16%	90.48%	9	23.47%	94.12%
5	24.66%	86.68%	10	31.06%	81.80%

Table 1: Isolation efficiency results.

3.2 Isolation Efficiency

For each of these datasets collected, the isolation and point cloud generation algorithm as detailed in the methodology section was run with the precise run-time of the program recorded, as shown below in table 2. The average run-time value was calculated as 1.1211 seconds, as well as the 99th percentile value calculated as 1.3248. The results show that the isolation process typically takes around 1 second to produce the isolated point cloud with simpler OOIs taking less than a second of run-time. Review of the literature has shown no precedent for value-based efficiency, however a goal of this study is to present these values so that they may be used in future studies as a metric for comparison for similar isolation processes or other 3D data operations. Despite no precedent for comparison, we are confident that these results of operation run-time, on a modest personal computer, demonstrate the high-level efficiency of this operation in comparison to more complex isolation procedures that rely on neural network processing.

scene	Run Time (seconds)	scene	Run Time (seconds)
1	0.8858	6	1.3280
2	0.9890	7	1.1458
3	1.0580	8	1.2922
4	1.0900	9	1.1400
5	1.0070	10	1.2752

Table 2: Isolation accuracy results.

 ISBN 978-0-9934207-7-1

4 Conclusion

In this study a method was proposed that allows for the isolation of OOI data from a clustered scene without explicit knowledge of the object, relying only on knowledge of a vague position in 3D space and the general colour of the object. While the isolation process does involve manual intervention, the setting of threshold values, due to the nature of the dual isolation approach these conditions can be set to broad values. Once these threshold values have been determined for a representative scene, a typical distance from the object and an object that is a typical colour and texture of other OOIs, the same values determined can be applied to other comparable scenes with little effect on the outcome of the isolated data. This can be seen in the results with the threshold values for isolation determined in scene 1 and applied to remaining scenes 2-9. The final scene, scene 10, required recalculating the correct depth threshold value due to a greater distance from the camera, however the colour space was still adequate for isolating the OOI effectively. The ultimate goal of this project was to isolate OOIs in unstructured scenes while maintaining low OOI data loss and process runtimes. The results presented indicate that the isolation process maintains a 90% average capture rate of OOI data with an average runtime of below 1.3 seconds, outperforming what can be found in the state-of-the-art literature. The isolation process does exhibit a high degree of excess data isolated with an average of 22%, however with further work on this process and the integration of RGB data used instead of grayscale data, this method can a see substantial drop in excess data captured.

References

[Aloni and Yitzhaky, 2015] Aloni, D. and Yitzhaky, Y. (2015). Automatic 3D object localization and isolation using computational integral imaging. *Applied Optics*, 54(22):6717–6724. Publisher: Optica Publishing Group.

[Avraham and Yitzhaky, 2021] Avraham, D. and Yitzhaky, Y. (2021). Effects of Depth-Based Object Isolation in Simulated Retinal Prosthetic Vision. *Symmetry*, 13(10):1763. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[Barabanau et al., 2019] Barabanau, I., Artemov, A., Burnaev, E., and Murashkin, V. (2019). Monocular 3D Object Detection via Geometric Reasoning on Keypoints. Number: arXiv:1905.05618 arXiv:1905.05618 [cs].

[Brazil and Liu, 2019] Brazil, G. and Liu, X. (2019). M3D-RPN: Monocular 3D Region Proposal Network for Object Detection. pages 9287–9296.

[Chen et al., 2017] Chen, L., He, Y., Chen, J., Li, Q., and Zou, Q. (2017). Transforming a 3-D LiDAR Point Cloud Into a 2-D Dense Depth Map Through a Parameter Self-Adaptive Framework. *IEEE Transactions on Intelligent Transportation Systems*, 18(1):165–176. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[Chmelar et al., 2016] Chmelar, P., Beran, L., and Rejfek, L. (2016). The Depth Map Construction from a 3D Point Cloud. *MATEC Web of Conferences*, 75:03005. Publisher: EDP Sciences.

[Codaro et al., 2002] Codaro, E. N., Nakazato, R. Z., Horovistiz, A. L., Ribeiro, L. M. F., Ribeiro, R. B., and Hein, L. R. O. (2002). An image processing method for morphology characterization and pitting corrosion evaluation. *Materials Science and Engineering: A*, 334(1):298–306.

[Dougherty, 2018] Dougherty, E. (2018). *Mathematical Morphology in Image Processing*. CRC Press. Google-Books-ID: TVCQDwAAQBAJ.

[Feng et al., 2019] Feng, M., Hu, S., Ang, M. H., and Lee, G. H. (2019). 2D3D-Matchnet: Learning To Match Keypoints Across 2D Image And 3D Point Cloud. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4790–4796. ISSN: 2577-087X.

[Guastella and Muscato, 2021] Guastella, D. C. and Muscato, G. (2021). Learning-Based Methods of Perception and Navigation for Ground Vehicles in Unstructured Environments: A Review. *Sensors*, 21(1):73. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[He et al., 2020] He, Y., Sun, W., Huang, H., Liu, J., Fan, H., and Sun, J. (2020). PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation. pages 11632–11641.

[Huang et al., 2020] Huang, T., Liu, Z., Chen, X., and Bai, X. (2020). EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 35–52, Cham. Springer International Publishing.

[Li et al., 2020a] Li, D., Wang, H., Liu, N., Wang, X., and Xu, J. (2020a). 3D Object Recognition and Pose Estimation From Point Cloud Using Stably Observed Point Pair Feature. *IEEE Access*, 8:44335–44345. Conference Name: IEEE Access.

[Li et al., 2020b] Li, P., Zhao, H., Liu, P., and Cao, F. (2020b). RTM3D: Real-Time Monocular 3D Detection from Object Keypoints for Autonomous Driving. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 644–660, Cham. Springer International Publishing.

[Li et al., 2022] Li, Y., Yu, A. W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., Le, Q. V., Yuille, A., and Tan, M. (2022). DeepFusion: Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection. pages 17182–17191.

[Maragos, 1996] Maragos, P. (1996). Differential morphology and image processing. *IEEE Transactions on Image Processing*, 5(6):922–937. Conference Name: IEEE Transactions on Image Processing.

[Perra et al., 2016] Perra, C., Murgia, F., and Giusto, D. (2016). An analysis of 3D point cloud reconstruction from light field images. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. ISSN: 2154-512X.

[Qi et al., 2021] Qi, S., Ning, X., Yang, G., Zhang, L., Long, P., Cai, W., and Li, W. (2021). Review of multi-view 3D object recognition methods based on deep learning. *Displays*, 69:102053.

[Serra and Soille, 2012] Serra, J. and Soille, P. (2012). *Mathematical Morphology and Its Applications to Image Processing*. Springer Science & Business Media. Google-Books-ID: T82qCAAAQBAJ.

[Sindagi et al., 2019] Sindagi, V. A., Zhou, Y., and Tuzel, O. (2019). MVX-Net: Multimodal VoxelNet for 3D Object Detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282. ISSN: 2577-087X.

[Sun et al., 2021] Sun, K., Zhang, J., Liu, J., Yu, R., and Song, Z. (2021). DRCNN: Dynamic Routing Convolutional Neural Network for Multi-View 3D Object Recognition. *IEEE Transactions on Image Processing*, 30:868–877. Conference Name: IEEE Transactions on Image Processing.

[Williams et al., 2021] Williams, D. P., España, A., Kargl, S. G., and Williams, K. L. (2021). A family of algorithms for the automatic detection, isolation, and fusion of object responses in sonar data. *Proceedings of Meetings on Acoustics*, 44(1):070022. Publisher: Acoustical Society of America.

View Sub-sampling and Reconstruction for Efficient Light Field Compression

Yang Chen, Martin Alain, and Aljosa Smolic

V-SENSE project
Graphics Vision and Visualisation group (GV2)
Trinity College Dublin

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/SWCC3620

Abstract

Compression is an important task for many practical applications of light fields. Although previous work has proposed numerous methods for efficient light field compression, the effect of view selection on this task is not well exploited. In this work, we study different sub-sampling and reconstruction strategies for light field compression. We apply various sub-sampling and corresponding reconstruction strategies before and after light field compression. Then, fully reconstructed light fields are assessed to evaluate the performance of different methods. Our evaluation is performed on both real-world and synthetic datasets, and optimal strategies are devised from our experimental results. We hope this study would be beneficial for future research such as light field streaming, storage, and transmission.

Keywords: Light Field View Synthesis, Light Field Compression

1 Introduction

A 4D light field is described as a collection of light rays passing through a 3D volume with specific intensity and direction, which can be represented as the interaction between each ray and two parallel planes: image plane and camera plane. The original concept of 4D light fields was firstly introduced by Levoy et. al. in 1996 [Levoy and Hanrahan, 1996]. After years of active research, light fields were applied in various immersive visual applications such as VR, 3DTV and holographic systems. The capturing process of the light field usually produces a huge amount of data, which requires plenty of storage space and transmission bandwidth. To reduce these, efficient compression of light field data is crucial for practical applications. One potential way to compress light field data is to sub-sample sub-aperture views for encoding, and to reconstruct the missing views after decoding, which is enabled through advances in light field view synthesis. In this paper, we study different sub-sampling strategies in combination with a recent view synthesis method based on deep learning.

2 Related Work

Light field compression is a crucial research topic, due to the huge amount of data needed for light field imaging. Conti et. al. provided a thorough review about recent light field coding techniques [Conti et al., 2020]. One common approach is utilizing light field reconstruction methods to complete view-subsampled light fields, i.e. compressing only a subset of the original views [Chen et al., 2017, Zhao and Chen, 2017, Viola et al., 2018, Jiang et al., 2017b, Jiang et al., 2017a]. Various methods use a video codec such as HEVC as coding component, while JPEG launched the JPEG Pleno initiative for standardizing compression of plenoptic data including light fields [Ebrahimi et al., 2016].

Recently, deep learning techniques were introduced to light field compression. Chen et, al. investigate the impact of subsampling and reconstruction on the light field view synthesis [Chen et al., 2020b]. Chen et. al.

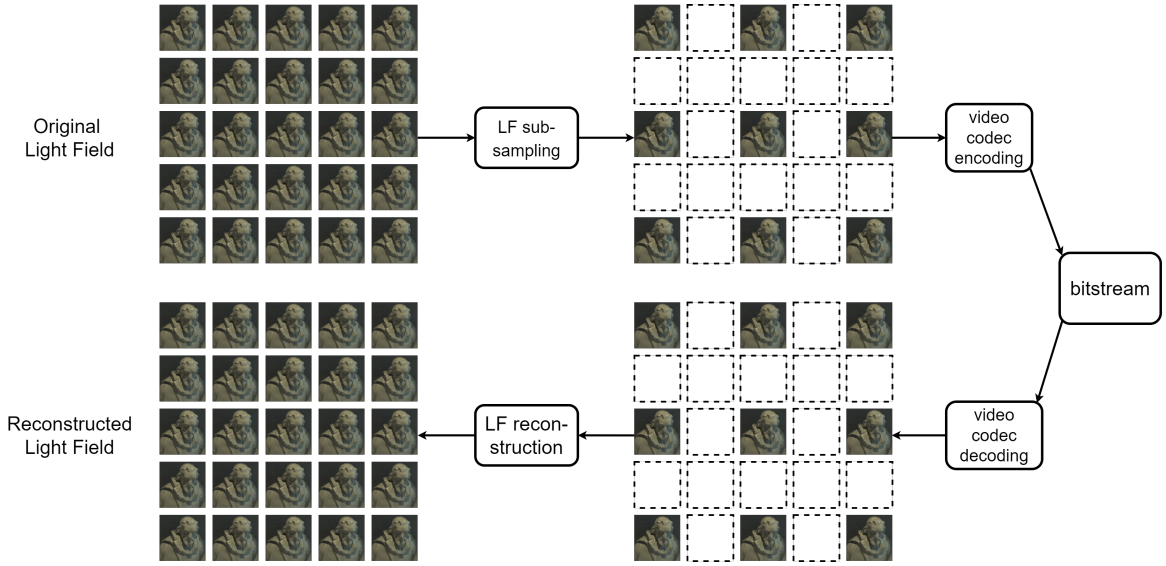


Figure 1: Light field compression pipeline with sub-sampling and reconstruction strategies.

proposed a self-supervised learning method to synthesize novel views of light field [Chen et al., 2020a]. Zhao et. al. presented a learning-based method which combines view enhancement and view synthesis to reconstruct complete light fields from decoded views [Zhao et al., 2018]. Wafa et. al. proposed a deep recursive residual network to synthesize intermediate views after the sparse views are decoded [Wafa et al., 2021]. Singh et. al. introduced an end-to-end disparity-aware 3D-CNN for light field compression, which utilizes the disparity information between views and the middle view [Singh and Rameshan, 2021]. Other works target light field compression using adversarial learning [Jia et al., 2018, Bakir et al., 2020, Liu et al., 2021].

All aforementioned approaches perform light field compression with one certain pattern of sub-aperture views, but we argue that the impact of different sub-sampling patterns on compression performance would be worth investigating. Thus, in this paper, we focus on evaluating view selection strategies for light field compression.

3 Subsampling and reconstruction for efficient light field compression

In this section, we present different sub-sampling and reconstruction strategies for light field compression. We first select various sub-sampling strategies, then the sub-sampled views are encoded, and finally the light field is completed by view synthesis.

3.1 Sub-sampling and Encoding

With the classical two-plane representation of light fields, three basic view sub-sampling strategies are introduced, row, column and corners, as shown in Fig. 2. Each one of them will have their corresponding reconstruction process, as discussed in Section 3.2. We further investigate different sub-sampling densities (2x) and (4x). The remaining views after sub-sampling are scanned in snake order Fig. 3a and encoded with a video codec.

3.2 Decoding and Reconstruction

After decoding at the receiver, we perform reconstruction by inverting the corresponding sub-sampling strategies. We take corners_4x of a 9×9 light field as an example, as shown in Fig. 3b. The complete reconstruction process is a cascade of one row wise and one column wise view synthesis, while (4x) reconstruction is a cascade of two (2x) interpolations. For row and column based sub-sampling, it is simple to apply row wise and column wise interpolation to complete the light field, accordingly.

 ISBN 978-0-9934207-7-1

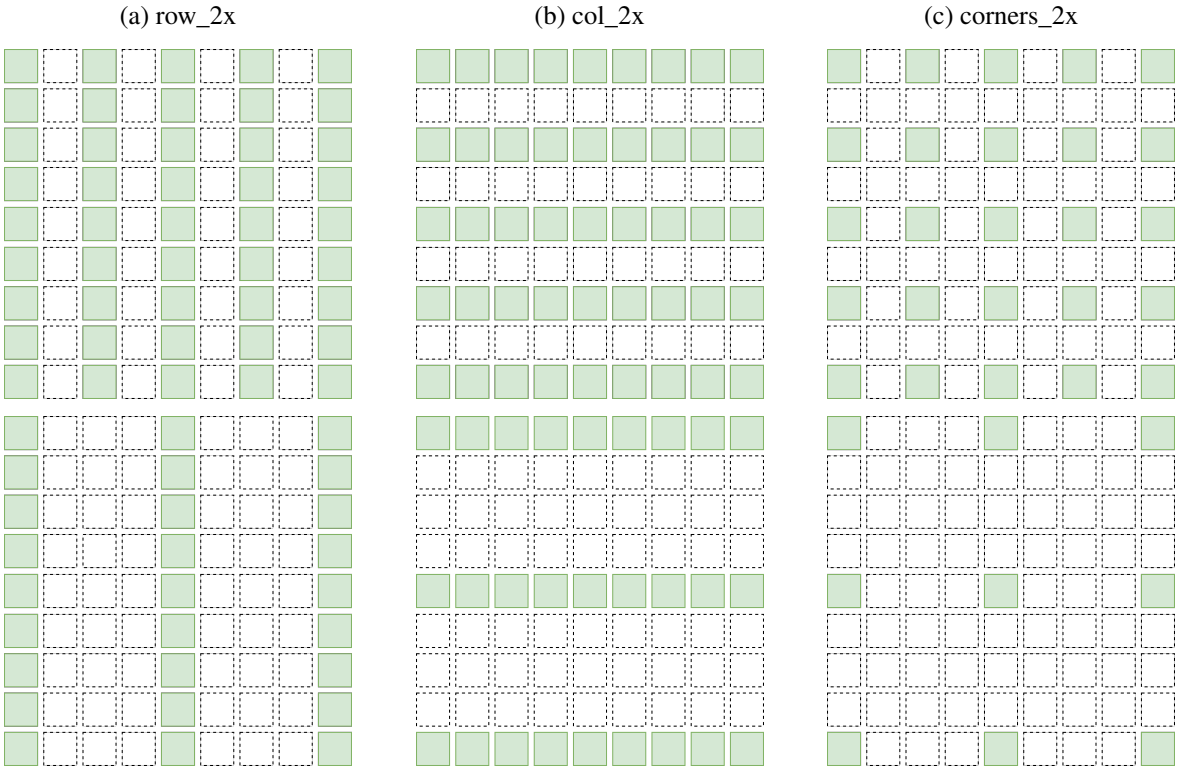


Figure 2: Six strategies to sub-sample views from a 9×9 light field. Sub-sampled views are shown in green. The columns show three different type of strategies (row, column and corners). Top (2x) and bottom (4x) rows illustrate different sampling densities.

4 Experiments

In this section, we present the details of our experiments. All computations were performed on an Intel Core i7-6700k 4.0GHz CPU. To implement our pipeline, sub-sampled RGB views were converted into YUV 420 video using the open-source FFMPEG software [FFM,]. Then, these video files were encoded by an efficient video codec, HM (HEVC) 16.22 [HEV,]. We used typical quantization parameters to vary bitrate and quality (QP: 20, 25, 30, 35, 40, 45). As baseline comparison we also present the results of encoding the full light field without any sub-sampling and reconstruction, which is indicated as "anchor" in all the figures.

We performed all experiments with synthetic light fields, Lytro images, and gantry-robot captured data to investigate different properties. The **Bedroom** light field is from the widely used synthetic HCI dataset. **Bee_2** is extracted from the Lytro Illum dataset using a Lytro enhancement pipeline [Matysiak et al., 2018]. **LegoKnight** is from the Stanford dataset and pre-cropped to 512×512 to have similar spatial resolution to other light fields. This is a challenging light field due to its large disparity compared to others and extended textureless areas. From all light fields we used 9×9 views.

Figure 4 shows the PSNR results after reconstruction as heatmaps, for 6 different strategies. These light fields were reconstructed by the state-of-the-art synthesis method CycleLF [Chen et al., 2020a] with $QP = 30$. We can recognize the sub-sampling patterns in these results, as directly decoded views have higher PSNR than interpolated views. These quality fluctuations are analyzed in more detail in Figure 7, where we show the standard deviations of PSNR results over the whole range of bitrates. We find that anchor encoding has the lowest fluctuation as expected. Fluctuations increase with the sub-sampling ratio as can also be expected, i.e. (4x) versions exhibiting highest fluctuations. While overall fluctuations are quite low for HCI and Lytro, they are significant for the relatively sparse Lego Knights. Thus, quality fluctuations have to be considered when applying view sub-sampling for light field compression.

 ISBN 978-0-9934207-7-1

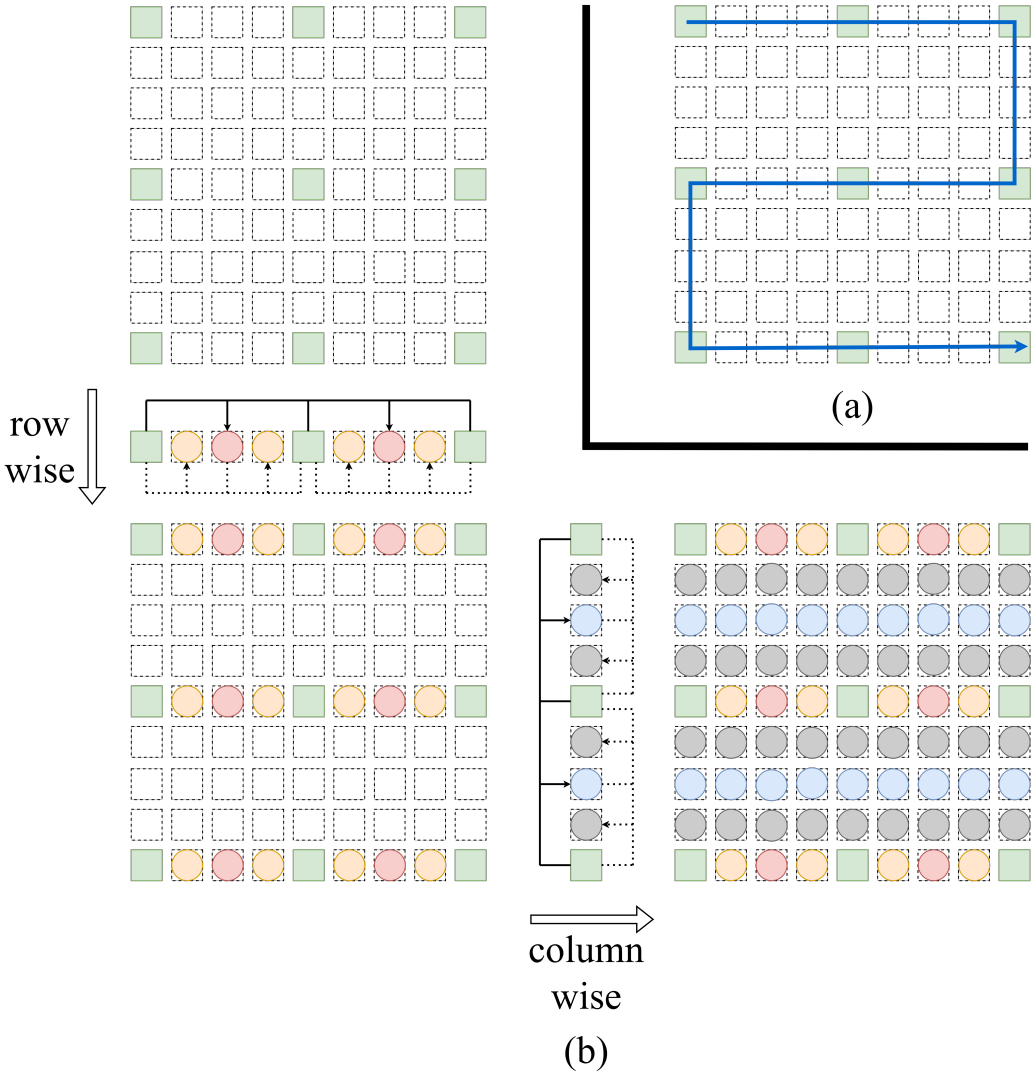


Figure 3: (a) Snake order for compression and (b) multi-step reconstruction of a light field. Different colors indicate views reconstructed in different stages of a hierarchical process, i.e. (red, blue) circles mean first step and (yellow, grey) mean second step.

Complete Rate-distortion (RD) results for PSNR and SSIM over bitrate are shown in Figure 5 and Figure 6, respectively. All strategies with CycleLF outperform bilinear on all three datasets. All strategies with CycleLF reconstruction outperform anchor encoding on **Bedroom** and show equivalent performance on **Bee_2**. Strategies with CycleLF reconstruction fall behind the anchor for **LegoKnight**, because of the large baseline of this light field affecting the performance of the reconstruction method. Meanwhile, regarding SSIM, CycleLF-based strategies consistently outperform the anchor.

Bjontegaard metrics (BD-PSNR and BD-Rate) [Wien, 2015] are shown in Table 1 and Table 2 including a number of additional light fields. Compression with sub-sampling and CycleLF reconstruction consistently outperforms the anchor. Especially on the HCI dataset, the CycleLF-based method achieves an average BD-DSNR gain of 1.63dB and an average BD-DSNR bitrate saving of -53.6% over anchor compression with the "corner_4x" pattern. Please note that we can't show BD-scores for the Stanford dataset as the large differences between the involved RD curves make this metric unsuitable for this case [Wien, 2015].

 ISBN 978-0-9934207-7-1

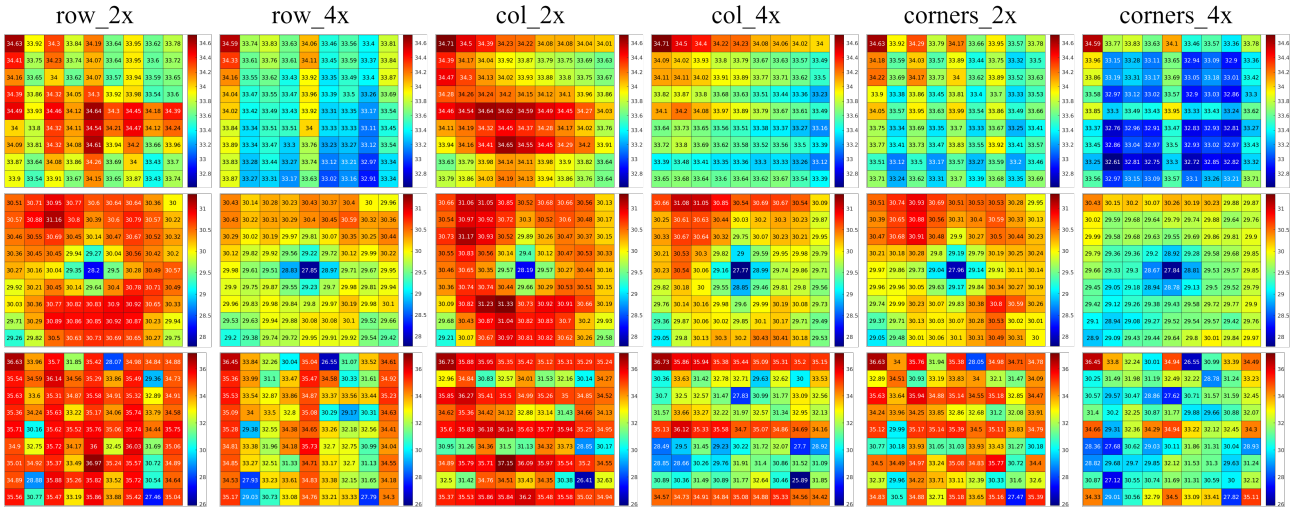


Figure 4: PSNR scores after light field reconstruction with $QP = 30$. Top, middle and bottom rows show results from HCI (**Bedroom**), Lytro (**Bee_2**) and Stanford (**Lego Knights**) datasets using CycleLF [Chen et al., 2020a].

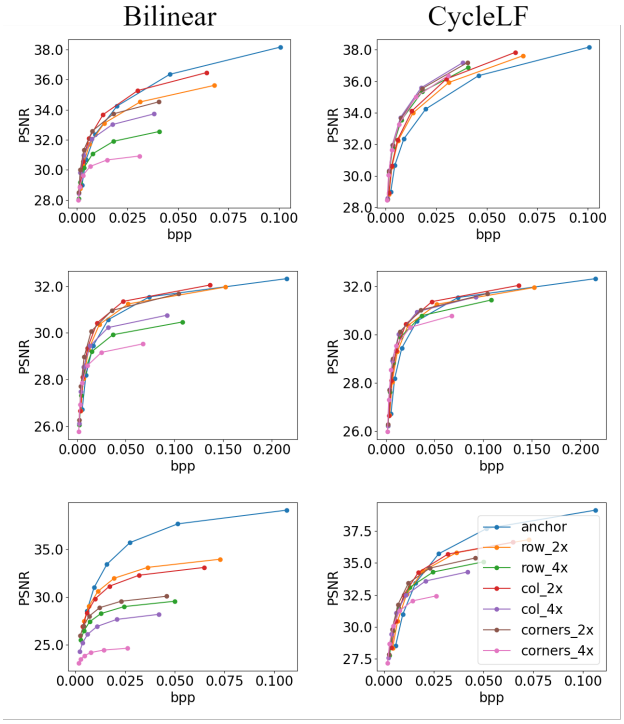


Figure 5: Rate-distortion results as PSNR over bit per pixel (bpp). Top row, middle row and bottom row are results on HCI (**Bedroom**), Lytro (**Bee_2**) and Stanford (**Lego Knights**), respectively.

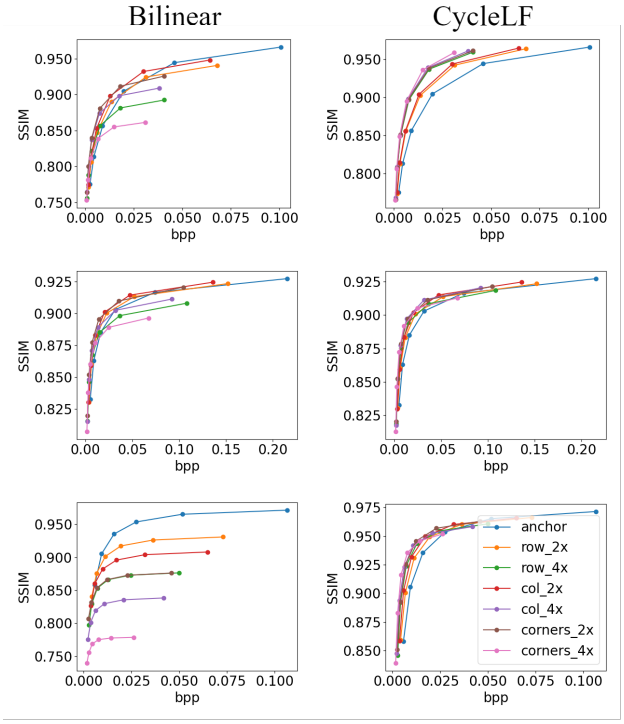


Figure 6: Rate-distortion results as Structural Similarity Index (SSIM) over bit per pixel (bpp). Top row, middle row and bottom row are results on HCI (**Bedroom**), Lytro (**Bee_2**) and Stanford (**Lego Knights**), respectively.

5 Conclusion

In this paper we presented a comprehensive investigation about the influence of different view selection strategies on the light field compression task. To achieve this goal, we tested our complete pipeline including subsampling, encoding, decoding, and reconstruction with various strategies. Our results show that sub-sampling can improve compression efficiency, especially for dense light fields. Higher sub-samling can give more gain

 ISBN 978-0-9934207-7-1

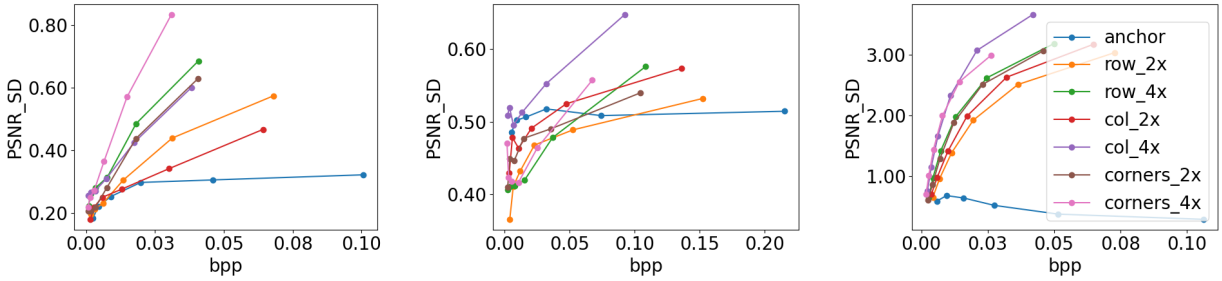


Figure 7: Per-view standard deviations of PSNR results over bitrate. Left, middle and right for HCI (**Bedroom**), Lytro (**Bee_2**) and Stanford (**Lego Knights**), respectively.

		Bed	**Bic**	**Her**	**Ori**	avg
row_2x	BD-PSNR	0.68	0.60	0.51	0.47	0.57
	BD-Rate	-24.8	-22.6	-21.4	-22.3	-22.8
row_4x	BD-PSNR	1.47	1.27	0.89	1.08	1.18
	BD-Rate	-48.0	-43.8	-36.9	-44.6	-43.4
col_2x	BD-PSNR	0.91	0.96	0.78	0.78	0.86
	BD-Rate	-31.5	-32.7	-30.0	-33.0	-31.8
col_4x	BD-PSNR	1.79	1.72	1.26	1.54	1.58
	BD-Rate	-53.8	-51.6	-46.1	-55.0	-51.6
cors_2x	BD-PSNR	1.66	1.56	1.24	1.26	1.43
	BD-Rate	-51.6	-48.4	-44.8	-49.3	-48.5
cors_4x	BD-PSNR	1.62	1.89	1.28	1.76	1.63
	BD-Rate	-52.7	-55.3	-48.2	-58.2	-53.6

Table 1: BD-scores of CycleLF reconstruction on the HCI dataset, compared to baseline compression.

		Bee	**Bik**	**Che**	**Des**	avg
row_2x	BD-PSNR	0.28	0.47	0.58	0.30	0.49
	BD-Rate	-18.1	-20.8	-28.9	-19.7	-24.8
row_4x	BD-PSNR	0.36	0.89	1.11	0.59	0.63
	BD-Rate	-29.4	-39.8	-52.7	-42.2	-36.4
col_2x	BD-PSNR	0.50	0.61	0.59	0.23	0.55
	BD-Rate	-29.3	-25.7	-29.5	-17.1	-27.9
col_4x	BD-PSNR	0.68	0.89	1.10	0.34	0.48
	BD-Rate	-41.6	-41.3	-52.0	-36.6	-36.6
cors_2x	BD-PSNR	0.63	1.05	1.18	0.52	0.97
	BD-Rate	-39.1	-43.7	-54.1	-40.7	-47.3
cors_4x	BD-PSNR	0.45	0.90	1.17	0.24	0.40
	BD-Rate	-40.6	-43.2	-52.8	-37.8	-37.3

Table 2: BD-scores of CycleLF reconstruction on the Lytro dataset, compared to baseline compression.

in these cases. However, fluctuations of output view quality have to be considered, which increase with the sub-sampling ratio.

Acknowledgments

All authors are from the Trinity College Dublin, College Green, Ireland. Contact cheny5@tcd.ie for further questions about this work. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. 978-1-7281-9320-5/20/$31.00 2020 European Union

References

[FFM,] FFMPEG: A complete, cross-platform solution to record, convert and stream audio and video. `https://www.ffmpeg.org/`. accessed: 16-01-2022.

[HEV,] HM reference software for high efficiency video coding (HEVC). `https://vcgit.hhi.fraunhofer.de/jvet/HM/`. accessed: 16-01-2022.

[Bakir et al., 2020] Bakir, N., Hamidouche, W., Fezza, S. A., Samrouth, K., and Déforges, O. (2020). Light field image coding using dual discriminator generative adversarial network and vvc temporal scalability. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

[Chen et al., 2017] Chen, J., Hou, J., and Chau, L.-P. (2017). Light field compression with disparity-guided sparse coding based on structural key views. *IEEE Transactions on Image Processing*, 27(1):314–324.

[Chen et al., 2020a] Chen, Y., Alain, M., and Smolic, A. (2020a). Self-supervised light field view synthesis using cycle consistency. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE.

[Chen et al., 2020b] Chen, Y., Alain, M., and Smolic, A. (2020b). A study of efficient light field subsampling and reconstruction strategies. *arXiv preprint arXiv:2008.04694*.

[Conti et al., 2020] Conti, C., Soares, L. D., and Nunes, P. (2020). Dense light field coding: A survey. *IEEE Access*, 8:49244–49284.

[Ebrahimi et al., 2016] Ebrahimi, T., Foessel, S., Pereira, F., and Schelkens, P. (2016). Jpeg pleno: Toward an efficient representation of visual reality. *Ieee Multimedia*, 23(4):14–20.

[Jia et al., 2018] Jia, C., Zhang, X., Wang, S., Wang, S., and Ma, S. (2018). Light field image compression using generative adversarial network-based view synthesis. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(1):177–189.

[Jiang et al., 2017a] Jiang, X., Le Pendu, M., Farrugia, R. A., and Guillemot, C. (2017a). Light field compression with homography-based low-rank approximation. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):1132–1145.

[Jiang et al., 2017b] Jiang, X., Le Pendu, M., and Guillemot, C. (2017b). Light field compression using depth image based view synthesis. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 19–24. IEEE.

[Levoy and Hanrahan, 1996] Levoy, M. and Hanrahan, P. (1996). Light field rendering. In *Proc. ACM SIGGRAPH*, pages 31–42.

[Liu et al., 2021] Liu, D., Huang, X., Zhan, W., Ai, L., Zheng, X., and Cheng, S. (2021). View synthesis-based light field image compression using a generative adversarial network. *Information Sciences*, 545:118–131.

[Matysiak et al., 2018] Matysiak, P., Grogan, M., Pendu, M. L., Alain, M., and Smolic, A. (2018). A pipeline for lenslet light field quality enhancement. In *Proc. IEEE ICIP*, pages 639–643.

[Singh and Rameshan, 2021] Singh, M. and Rameshan, R. M. (2021). Learning-based practical light field image compression using a disparity-aware model. *arXiv preprint arXiv:2106.11558*.

[Viola et al., 2018] Viola, I., Maretic, H. P., Frossard, P., and Ebrahimi, T. (2018). A graph learning approach for light field image compression. In *Applications of Digital Image Processing XLI*, volume 10752, page 107520E. International Society for Optics and Photonics.

[Wafa et al., 2021] Wafa, A., Pourazad, M. T., and Nasiopoulos, P. (2021). Learning-based light field view synthesis for efficient transmission and storage. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 354–358. IEEE.

[Wien, 2015] Wien, M. (2015). High efficiency video coding. *Coding Tools and specification*, 24.

[Zhao and Chen, 2017] Zhao, S. and Chen, Z. (2017). Light field image coding via linear approximation prior. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4562–4566. IEEE.

[Zhao et al., 2018] Zhao, Z., Wang, S., Jia, C., Zhang, X., Ma, S., and Yang, J. (2018). Light field image compression based on deep learning. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

A Comparative Study of Traditional Light Field Methods and NeRF

Pierre Matysiak, Susana Ruano, Martin Alain, Aljosa Smolic

V-SENSE Project, School of Computer Science and Statistics, Trinity College Dublin

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/IQKC6774

Abstract

Neural Radiance Fields (NeRF) is a recent technology which had a large impact in computer vision, promising to generate high quality novel views and corresponding disparity map, all using a fairly small number of input images. In effect, they are a new way to represent a light field. In this paper, we compare NeRF with traditional light field methods for novel view synthesis and depth estimation, in an attempt to quantify the advantages brought by NeRF, and to put these results in perspective with the way both paradigms are used practically. We provide qualitative and quantitative comparisons, discuss them and highlight some aspects of working with NeRF depending on the type of light field data used.

1 Introduction

NeRF is a recent deep-learning based method [Mildenhall et al., 2020] that created a small revolution in how light fields are thought of, constructed, and processed. While most previous light field data is precisely defined, with specific parameters, fixed baseline between views, and a number of constraints, NeRF instead works from using only a relatively small number of images pointing at the same scene from different angles. After some training of the underlying network, it builds a 3D representation of the scene, from which additional information can be extracted, such as novel views, or disparity maps. This is based on the concept of unstructured light fields. While this new technology seemed almost miraculous when it was first presented, it also seems to suffer, in a way not dissimilar to traditional light field methods, from some drawbacks in how it can be used, and in particular the type of data it can be used with. While traditional light field methods struggle to generalise to data containing a very wide baseline or very high resolution, NeRF on the other hand requires the principal point of the images to be centred. This is not the case with data captured with plenoptic cameras, or digitally synthesised images, and those require some minor additional processing before being used by NeRF. This paper aims at looking in more detail and comparing the respective output and failure cases of both traditional light field methods and NeRF with regards to two applications: view synthesis and depth estimation.

2 Related work

In this section we briefly describe the current state of the art regarding traditional light field view synthesis as well as depth estimation, and relevant papers using NeRF for those same applications.

Light Field view synthesis. We first have a brief look at classical computer vision methods. After observing that the sparsity is much greater in the continuous Fourier spectrum than the discrete spectrum, [Shi et al., 2015] proposed an approach to reconstructing views optimised for sparsity in the continuous Fourier spectrum, to reduce sampling requirements and improve quality. In [Chen et al., 2017] consistent disparity maps are produced using the combination of a feature flow method and a spatio-temporal edge-aware filter. The sparse representation of Epipolar Plane Images (EPI) in the shearlet transform domain is used by [Vagharshakyan et al., 2018] and they take advantage of the straight line characteristic of EPIs for reconstruction. In particular their method handles semi-transparent objects in a scene with a much higher degree of precision.

 ISBN 978-0-9934207-7-1

In the work by [Kalantari et al., 2016], they were among the first to use machine learning to mitigate the usual trade-off between spatial and angular resolution of plenoptic cameras. They break down the view synthesis process into disparity and colour estimation components trained simultaneously to obtain high quality reconstruction. A 4DCNN network combining convolutions on stacked EPIs, and detail-restoration 3DCNNs to effectively synthesise 4D light fields from a sparse selection of views are presented in [Wang et al., 2018]. A coarse to fine scheme to extrapolate high-dimensional spatio-angular features in a two-step method first generating intermediate coarse novel views which are later refined using guided residual learning and 4D convolutions is used in [Yeung et al., 2018]. More recently, in [Chen et al., 2020], they look at the data collection drawback of other learning-based approaches, and propose a self-supervised framework. They first train their network on natural videos, and use that prior knowledge combined with a cycle consistency constraint to build a bidirectional mapping and generate input-consistent views.

Predating NeRF in concept, a new technique was developed by [Zhou et al., 2018], called *multi-plane images* (MPI) and generated some interest as a new representation of light fields. MPIs approximate a light field by generating a stack of semi-transparent coloured layers organised at various depth levels, which allows for real-time synthesis of novel views. Early work was constructing MPIs from dense sets of views, but this was soon generalised to sparser sets of real-life images [Flynn et al., 2019, Srinivasan et al., 2019, Mildenhall et al., 2019].

Light Field depth estimation. Depth estimation on light fields is an extremely rich and still active field of research. Starting with classical computer vision approaches, in [Yu et al., 2013] they analysed the geometric structure of 3D lines in a light field image and obtained depth maps by matching those lines between sub-aperture images (SAI). In [Tosic and Berkner, 2014] they formulated a method to construct light field scale-depth spaces, indicating regions of constant depth, before solving the finer depth estimation in each space separately. This allowed to obtain good results in both highly textured and uniform regions. In [Zhang et al., 2016] they provided a solution to deal with occlusion artefacts, by implementing a spinning parallelogram operator to divide EPIs into regions and locating depth lines by maximising distribution distances of those regions.

After the advent of deep learning, several new methods were developed. Based on the EPI or epipolar geometry property, in [Luo et al., 2017] they proposed to formulate the depth estimation as a classification problem, in which a standard CNN-architecture is employed on horizontal and vertical EPI patches. Since a shallow CNN is inadequate to guarantee proper accuracy, a global optimisation with traditional approach is utilised. A similar approach is presented by [Feng et al., 2018] in which a shallower CNN is considered and the output of the fully-connected layer is more than one pixel. In [Jiang et al., 2019] they proposed to estimate initial depths by a fine-tuned flow-based network and then refine these initial results using a multi-view stereo refinement network. Epinet is presented by [Shin et al., 2018], an end-to-end network to predict depth, which takes as input the horizontal, vertical, left diagonal and right diagonal camera views, instead of EPI patches. With richer information of light fields, Epinet achieves a better accuracy. In [Khan et al., 2021] they used the idea that depth edges are more sensitive than texture edges to local constraints, and tell the two apart using a bidirectional diffusion process. Some of the most recent work is looking at using attention-based models, providing a better selection of features even is complex and texture-rich scenes, and leading to more accurate disparity ([Tsai et al., 2020]) or depth maps ([Chen et al., 2021]).

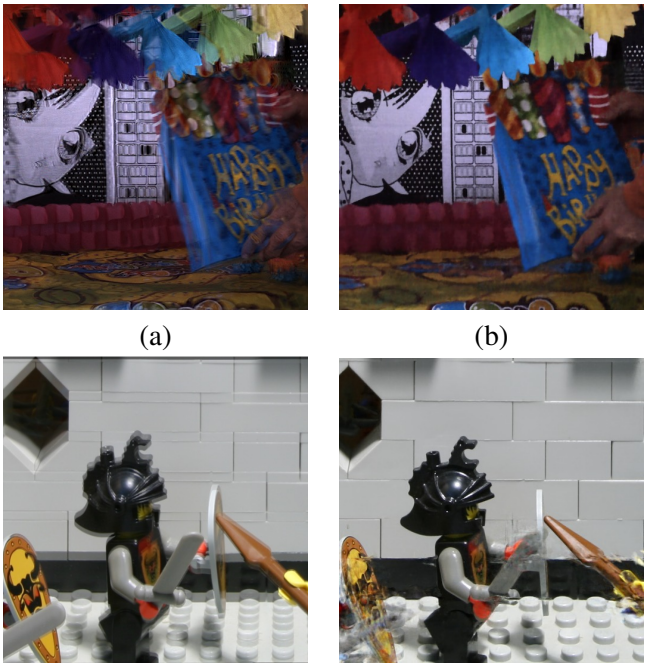
NeRF. The seminal paper [Mildenhall et al., 2020] is intended to be a novel view synthesis method, and does so by rendering and optimising a continuous volumetric scene using a sparse set of input views. The input to their fully-connected non-convolutional network is a single continuous 5D coordinate (spatial location and viewing direction), which output the volume density and view-dependent emitted radiance at that location. While intended for novel view synthesis, as the network performs a dense 3D reconstruction of the scene, it can also be used for accurate high-resolution depth estimation, generated concurrently with novel views. Building on this foundation, and looking more specifically at the problem of depth estimation on indoor scenes, in [Wei et al., 2021] they combine structure-from-motion (SfM) and learning-based priors and plug them into a NeRF network to obtain high-resolution depth estimation. The sparse SfM reconstruction is fine-tuned using a monocular depth network, and use those priors to fix the inherent shape-radiance ambiguity of NeRF. Finally they further improve the results by using a per-pixel confidence map.

 ISBN 978-0-9934207-7-1

3 Comparing novel view synthesis

We first describe the data used in this section. We selected three images from the HCI synthetic dataset (*boardgames*, *rosemary*, *table*) [Honauer et al., 2016], and five images from Lytro datasets: INRIA (*fruits*) [Pendu et al., 2018], EPFL (*bikes*) [Rerabek and Ebrahimi, 2016], and our own (*guinness*, *frog*, *cards*). The former three were processed using the pipeline of [Matysiak et al., 2020] and the latter two contain full camera calibration data [Aenchbacher and Smolic, 2022]. These two types of images pose a challenge to NeRF because the principal point of these images is not centered, which is a prerequisite of the method. To counter this issue we modify the original SAIs to shift the focal plane to infinity. While this puts the principal point in the centre of the views by simulating the images being taken by a single camera, it comes at the price of a small loss of resolution due to cropping. In addition we use one image from the Stanford gantry dataset (*lego knights*) [Stanford, 2021], and four high resolution images from the Technicolor (*birthday*, *painter*) [Sabater et al., 2017] and SAUCE datasets (*cellist*, *fire_dancer*) [Herfet et al., 2018, Trottnow et al., 2019]. These high resolution images, with a wider baseline, pose some difficulty for traditional light field methods which are not designed for such sets. The gantry and high resolution set, having been captured by a single camera, do not suffer from the aforementioned issue, and NeRF can handle them directly.

Methods. There are many traditional methods for light field novel view synthesis, who were initially all depth-based, but some novel methods were proposed to increase the accuracy of the reconstruction. For example, in [Vagharshakyan et al., 2018] they use the EPI representation of light fields, and perform inpainting in the shearlet transform domain to generate novel views between two existing views. Those methods were soon replaced with machine learning approaches, the main drawback of which is the need for large amounts of labelled data to train any network. In [Chen et al., 2020] they bypass this issue and instead first train their network on labelled video data, more widely available, and use in turn a self-supervised network guided by a cycle consistency constraint, used to build bidirectional mapping and enforce the generated views to be consistent with input views. We use this method for comparison in the rest of this section.



(a) (b) (c) (d)

Figure 1: View synthesis on images *birthday* (detail) and *lego knights* obtained using the method of [Chen et al., 2020] (a, c) and NeRF [Mildenhall et al., 2020] output (b, d).

On the other hand, NeRF works by approximating a continuous 5D scene representation with an MLP network (for details see [Mildenhall et al., 2020], whose input in a 3D location (x, y, z) and a 2D viewing direction (θ, ϕ). Its output is an emitted colour (r, g, b) and volume density σ. The weights obtained encode the volume of the underlying scene by mapping each input 5D coordinate to its volume density and emitted colour. This model is view dependent, which allows it to handle non-Lambertian effects and realistic specularity while rendering novel views.

Visual results. Looking first at large baseline images, NeRF seems to offer a higher accuracy of reconstruction, see Figure 1 (top). Several points must be noted. First of all, NeRF can work on the full resolution image, while most traditional light field methods work better with square (cropped) views, particularly deep learning

 ISBN 978-0-9934207-7-1

based methods, and as a result there is necessarily loss of information in the second case. On top of that, the difference in quality between the reconstructed views is pretty obvious, NeRF comes out with high quality and high resolution novel views, capturing most of the minute details of the scene, even in complex ones like in Figure 1 ($birthday$), despite the high number of occlusions present. Traditional light field methods however come out with noisier results, as if motion was present.

Comparisons on smaller baseline images show both approaches seem to have their issues, and it is more difficult to determine which is preferable, see Figure 1 (bottom). While traditional light field methods seem to work fairly well, on some depth levels it is still possible to see some reconstruction artefacts, as if motion occurred, however those issues are not generalised: notice how the shield in the corner has high detail, but the rest of the image suffers from artefacts akin to motion blur. NeRF on the other hand seems to handle some parts of the image fairly well (helmet, spear, wall), but suffers from reconstruction artefacts in many places which detracts from the details in the rest of the image. In particular those artefacts seem to occur on the edges of the image, however it is not limited to that (sword in the centre). One thing to note is that both methods can only render novel views within the angular space defined by the input views.

Objective comparison. For objective comparison we use two classical metrics, PSNR and SSIM. They are properly representative as both traditional light field methods and NeRF generate views that are directly aligned with existing views, used as ground truth. As we can see in Table 1, NeRF performs better on all datasets. This is not surprising as it renders the scene in a continuous underlying 3D model, which is then used to generate novel views and thus does not have to approximate parts of the scene. Both methods seem to fare better with smaller baseline data, however when looking at wider baseline there is a large discrepancy between the images used. We posit that since the $birthday$ image is rife with minuscule details, it is more difficult to generate novel views that fool the metric well, even though it fools the eye, whereas the $cellist$ image contains a large uniform area and limited number of elements in the scene.

	$P-LF$	$P-NF$	$S-LF$	$S-NF$
boardg	34.83	43.13	0.912	0.993
rosemary	34.23	41.28	0.904	0.983
table	33.91	39.19	0.895	0.954
E_bikes	32.36	32.45	0.862	0.963
I_fruits	31.63	30.29	0.856	0.948
V_guinn	32.92	33.21	0.848	0.940
V_cards	32.47	33.17	0.861	0.957
V_frog	35.69	41.63	0.873	0.982
legoK	21.61	24.67	0.711	0.849
birthday	19.29	23.69	0.542	0.750
painter	23.24	28.08	0.563	0.786
cellist	27.18	35.70	0.632	0.970
fire_danc	27.25	30.82	0.625	0.972
Mean	29.74	33.64	0.714	0.927

Table 1: Metric comparison results (PSNR (denoted P) and SSIM (denoted S)) on novel views, comparing the method of [Chen et al., 2020] (denoted LF) and NeRF [Mildenhall et al., 2020] (denoted NF).

4 Comparing depth estimation

While both traditional light field methods and NeRF allow to obtain accurate results in their preferred environment, neither method truly generalises to all types of data. For this comparison, since we need ground truth depth estimation to properly use the selected metrics, the only data for which we have objective comparison are synthetic images. For other data types, visual comparison will be used instead.

Methods. When it comes to traditional light fields methods we used here the one by [Chen et al., 2017]; their process follows a three-step approach. First they extract a 3D volume of the light field by selecting views along a single angular dimension. Second they perform an optical flow estimation to obtain disparity estimates between the selected views. Finally the last aggregation step allows to obtain depth maps from the multiple disparity map estimates. This process is relatively fast and runs in about 20 seconds for a whole row or column of the light field.

NeRF on the other hand is providing with 'direct' disparity estimation, which can then be converted to depth estimation, as the network first trains to obtain a 3D representation of the scene, from which each new view is rendered, as well as the corresponding depth. As a result the accuracy of the latter is very high, at the

 ISBN 978-0-9934207-7-1

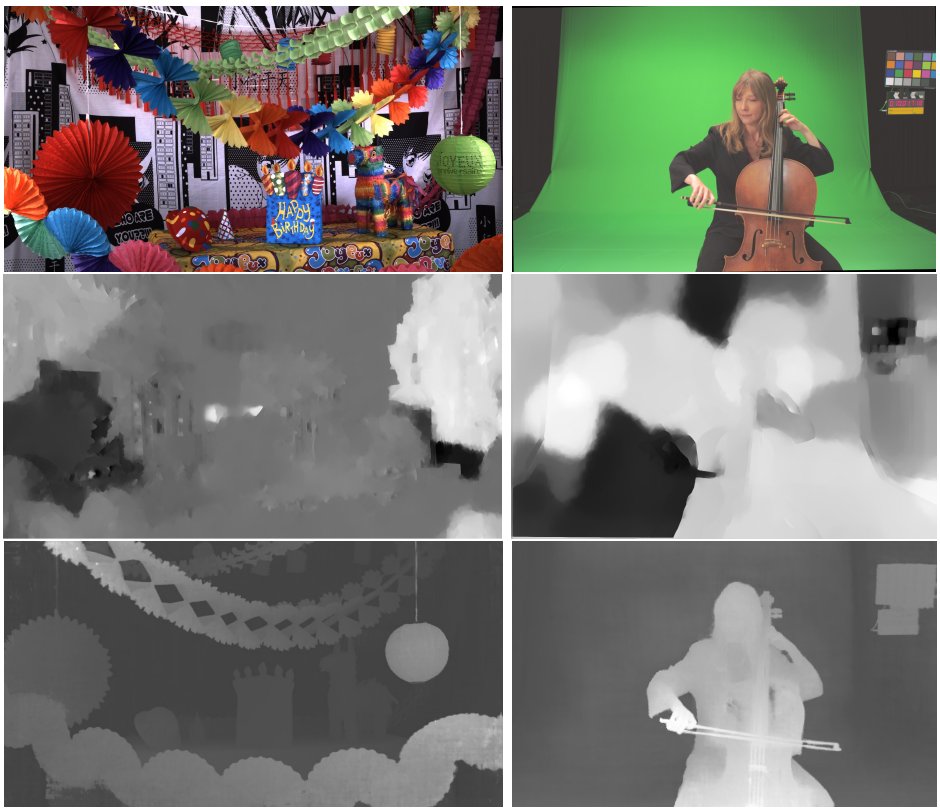


Figure 2: Depth maps obtained using method of [Chen et al., 2017] (middle) and NeRF [Mildenhall et al., 2020] output (bottom) on the $birthday$ (left) and $cellist$ (right) images.

cost of higher computational time. For example, generating a single novel view of resolution 512x512 takes an average of 18 seconds, while the same for a view of resolution 2048x1080 takes an average of 3 minutes.

Visual comparison. From the images selected the differences are clear between both approaches. On high quality, high baseline images (see Figure 2) NeRF has a clear advantage. By generating a comprehensive 3D render of the scene, it has access to fine features and details from the scene. Considering that representation, it also has detailed information regarding camera position and its distance to every single point of the scene, which, at the same time as it allows the generation of high quality novel views, also helps generating a corresponding high quality disparity map. Some artefacts can be visible on the edges of the image, which we posit could be explained by the lower number of views in which those parts are present, which lowers the quality of the rendering in these parts. Traditional light field methods on the other hand fail to accurately obtain a proper depth estimation, in part due to the higher baseline, the absence of camera parameters, and the fact that the input views are not aligned on a perfect grid - as the data we used was unstructured - which is one of the prerequisites of those methods.

When it comes to smaller baseline images, the advantage is still on the side of NeRF, however there are counter-examples, see Figure 3. We posit this image in particular is tricky for NeRF as only parts of small objects come into view on the sides of the images and potentially make a 3D representation of the scene, this already led to artefacts visible in Figure 1.

Objective comparison. To compare depth estimation output, we use again two metrics, MSE and Bad Pixel Count (2.0), see Table 2. Comparison for synthetic images is fairly straightforward as ground truth exists, and in this case we notice that NeRF does perform much better compared to traditional methods. However when it comes to comparing natural images, whether Lytro or otherwise, ground truth depth maps do not exist, making objective comparisons less relevant. We therefore only use visual comparison for these images.

 ISBN 978-0-9934207-7-1

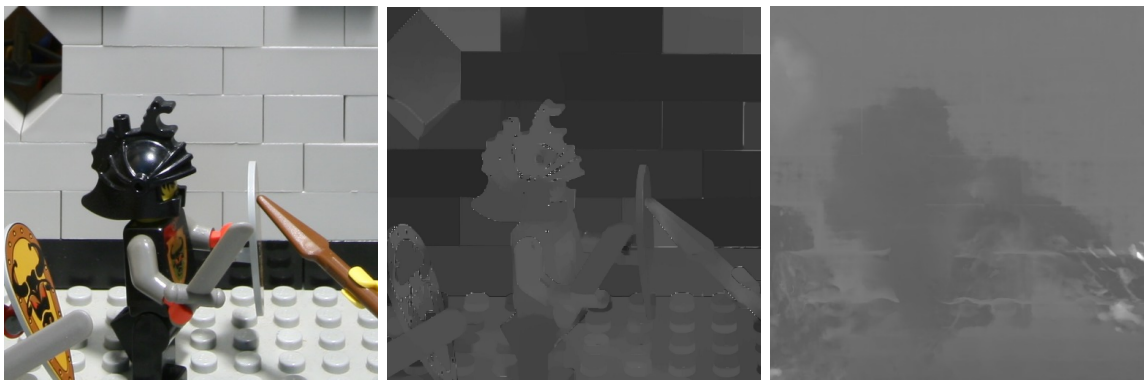


Figure 3: Depth maps obtained using method of [Chen et al., 2017] (centre) and NeRF [Mildenhall et al., 2020] output (right) on *Lego Knights* image.

	$M-LF$	$M-NF$	$B2-LF$	$B2-NF$
boardgames	4.513	0.602	15.41	3.57
rosemary	7.135	1.916	17.84	5.12
table	6.205	1.931	17.12	5.86
Mean	5.951	1.483	16.79	4.85

Table 2: Metric comparison results (MSE (denoted M) + Bad Pixel Count (denoted B2)) on depth estimation performed on synthetic images, between the method of [Chen et al., 2017] (denoted LF) and NeRF [Mildenhall et al., 2020] (denoted NF).

5 Runtime

On that front traditional light field methods show a relative advantage, as for both view synthesis and depth estimation, the runtime is between 6 and 8 minutes depending on the light field image. Whereas for NeRF, on a single NVidia GeForce GTX 1080 GPU, the model training runtime is in the order of hours to days depending on input resolution. On top of that, generating a single novel high resolution (2048x1088) view using a trained NeRF model takes on average ~3 minutes. Knowing that a light field can contain hundreds of views, these high processing times could be a deterrent to some when image quality is not absolutely essential. It is worthy of note that some recent work has been looking specifically at reducing NeRF computational time, while retaining high quality output [Lindell et al., 2021]. These could be used as an alternative.

6 Conclusion and future works

We have presented a comparative study between the newly developed NeRF with regards to view synthesis and depth estimation, and traditional light field methods, which it aimed to replace. While impressive, it has shown some limitations especially in the data it can process directly, and tricks need to be used to make it adaptable to either synthetic or Lytro image. We have however detailed the necessary pre-processing required to use NeRF on Lytro or synthetic data. In general, we can say that both schemes have their advantage for the specific type of data they target, but fail to generalise to any type of light field out of the box, which leaves traditional light field research some opportunities to keep growing. Potential future work includes adapting some editing methods that use traditional light fields to using NeRF instead, and analysing the quality of those results.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776.

References

[Aenchbacher and Smolic, 2022] Aenchbacher, W. and Smolic, A. (2022). The object-side light field, natural light field reference planes, and the importance of entrance and exit pupils for unfocused plenoptic cameras.

[Chen et al., 2021] Chen, J., Zhang, S., and Lin, Y. (2021). Attention-based multi-level fusion network for light field depth estimation. *Proc. AAAI Conf. on AI*, 35(2):1009–1017.

[Chen et al., 2017] Chen, Y., Alain, M., and Smolic, A. (2017). Fast and accurate optical flow based depth map estimation from light fields. In *Proceedings of the Irish Machine Vision and Image Processing Conference*.

[Chen et al., 2020] Chen, Y., Alain, M., and Smolic, A. (2020). Self-supervised light field view synthesis using cycle consistency. In *IEEE Int. Workshop on MMMSP*, pages 1–6.

[Feng et al., 2018] Feng, M., Wang, Y., Liu, J., Zhang, L., Zaki, H., and Mian, A. (2018). Benchmark data set and method for depth estimation from light field images. *IEEE TIP*, 27(7):3586–3598.

[Flynn et al., 2019] Flynn, J., Broxton, M., Debevec, P., DuVall, M., Fyffe, G., Overbeck, R., Snavely, N., and Tucker, R. (2019). Deepview: View synthesis with learned gradient descent. In *IEEE/CVF CVPR*, pages 2362–2371.

[Herfet et al., 2018] Herfet, T., Lange, T., and Hariharan, H. P. (2018). Enabling multiview- and light field-video for veridical visual experiences. In *IEEE ICCC*, pages 1705–1709.

[Honauer et al., 2016] Honauer, K., Johannsen, O., Kondermann, D., and Goldluecke, B. (2016). A dataset and evaluation methodology for depth estimation on 4d light fields. In *ACCV*. Springer.

[Jiang et al., 2019] Jiang, X., Shi, J., and Guillemot, C. (2019). A learning based depth estimation framework for 4d densely and sparsely sampled light fields. In *IEEE ICASSP*, pages 2257–2261.

[Kalantari et al., 2016] Kalantari, N., Wang, T.-C., and Ramamoorthi, R. (2016). Learning-based view synthesis for light field cameras. *ACM Trans. on Graphics*, 35(6).

[Khan et al., 2021] Khan, N., Kim, M., and Tompkin, J. (2021). Edge-aware bidirectional diffusion for dense depth estimation from light fields.

[Lindell et al., 2021] Lindell, D., Martel, J., and Wetzstein, G. (2021). Autoint: Automatic integration for fast neural volume rendering. In *IEEE/CVF CVPR*.

[Luo et al., 2017] Luo, Y., Zhou, W., Fang, J., Liang, L., Zhang, H., and Dai, G. (2017). Epi-patch based convolutional neural network for depth estimation on 4d light field. In *Neural Information Processing*, pages 642–652, Cham.

[Matysiak et al., 2020] Matysiak, P., Grogan, M., Pendu, M. L., Alain, M., Zerman, E., and Smolic, A. (2020). High quality light field extraction and post-processing for raw plenoptic data. *IEEE TIP*, pages 1–1.

[Mildenhall et al., 2019] Mildenhall, B., Srinivasan, P., Ortiz-Cayon, R., Kalantari, N., Ramamoorthi, R., Ng, R., and Kar, A. (2019). Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Trans. Graph.*, 38(4).

[Mildenhall et al., 2020] Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J., Ramamoorthi, R., and Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*.

[Pendu et al., 2018] Pendu, M. L., Jiang, X., and Guillemot, C. (2018). Light field inpainting propagation via low rank matrix completion. *IEEE TIP*, 27(4):1981–1993.

 ISBN 978-0-9934207-7-1

[Rerabek and Ebrahimi, 2016] Rerabek, M. and Ebrahimi, T. (2016). New light field image dataset. In *Proceedings of the International Conference on Quality of Multimedia Experience*.

[Sabater et al., 2017] Sabater, N., Boisson, G., Vandame, B., Kerbiriou, P., Babon, F., Hog, M., Langlois, T., Gendrot, R., Bureller, O., Schubert, A., and Allie, V. (2017). Dataset and pipeline for multi-view light-field video. In *CVPR Workshops*.

[Shi et al., 2015] Shi, L., Hassanieh, H., Davis, A., Katabi, D., and Durand, F. (2015). Light field reconstruction using sparsity in the continuous fourier domain. *ACM Trans. Graph.*, 34(1).

[Shin et al., 2018] Shin, C., Jeon, H.-G., Yoon, Y., Kweon, I., and Kim, S. (2018). Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *IEEE/CVF CVPR*, pages 4748–4757.

[Srinivasan et al., 2019] Srinivasan, P., Tucker, R., Barron, J., Ramamoorthi, R., Ng, R., and Snavely, N. (2019). Pushing the boundaries of view extrapolation with multiplane images. In *IEEE/CVF CVPR*, pages 175–184.

[Stanford, 2021] Stanford (2021). The stanford light field archive. `http://lightfield.stanford.edu/lfs.html`. accessed: 27-12-2021.

[Tosic and Berkner, 2014] Tosic, I. and Berkner, K. (2014). Light field scale-depth space transform for dense depth estimation. In *IEEE CVPR Workshops*, pages 441–448.

[Trottnow et al., 2019] Trottnow, J., Spielmann, S., Herfet, T., Lange, T., Chelli, K., Solony, M., Smrz, P., Zemcik, P., Aenchbacher, W., Grogan, M., Alain, M., Smolic, A., Canham, T., Vu-Thanh, O., Vázquez-Corral, J., and Bertalmío, M. (2019). The potential of light fields in media productions. In *SIGGRAPH Asia Tech. Briefs*, SA '19, page 71–74.

[Tsai et al., 2020] Tsai, Y.-J., Liu, Y.-L., Ouhyoung, M., and Chuang, Y.-Y. (2020). Attention-based view selection networks for light-field disparity estimation. *Proc. AAAI Conf. on AI*, 34(07):12095–12103.

[Vagharshakyan et al., 2018] Vagharshakyan, S., Bregovic, R., and Gotchev, A. (2018). Light field reconstruction using shearlet transform. *IEEE Trans. PAMI*, 40(1):133–147.

[Wang et al., 2018] Wang, Y., Liu, F., Wang, Z., Hou, G., Sun, Z., and Tan, T. (2018). End-to-end view synthesis for light field imaging with pseudo 4dcnn. In *Computer Vision – ECCV 2018*, pages 340–355, Cham.

[Wei et al., 2021] Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., and Zhou, J. (2021). Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *ICCV*.

[Yeung et al., 2018] Yeung, H., Hou, J., Chen, J., Chung, Y., and Chen, X. (2018). Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. In *Computer Vision – ECCV 2018*, pages 138–154, Cham.

[Yu et al., 2013] Yu, Z., Guo, X., Lin, H., Lumsdaine, A., and Yu, J. (2013). Line assisted light field triangulation and stereo matching. In *IEEE ICCV*.

[Zhang et al., 2016] Zhang, S., Sheng, H., Li, C., Zhang, J., and Xiong, Z. (2016). Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145:148 – 159.

[Zhou et al., 2018] Zhou, T., Tucker, R., Flynn, J., Fyffe, G., and Snavely, N. (2018). Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4).

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/KPPV3732

Diversity Issues in Skin Lesion Datasets

Neda Alipour*, Ted Burke†, Jane Courtney‡

School of Electrical and Electronic Engineering, Technological University Dublin

Abstract

Melanoma is one of the most threatening skin cancers in the world, which may spread to other parts of the body if it has not been detected at an early stage. Thus, researchers have put extra efforts into using computer-aided methods to help dermatologists to recognise this kind of cancer. There are many methods for solving this issue, many based on deep learning models. In order to train these models and have high accuracy, datasets which are large enough to cover gender, race, and skin type diversity are required. Although there is a large body of data on melanoma and skin lesions, most do not cover a broad diversity of skin types, which can affect the accuracy of models trained on them. To understand the issue, first the diversity of each database must be assessed and then, based on the existing shortcomings, such as minority skin types, a suitable method must be developed to solve any diversity issues. This article summarizes the problem of the lack of diversity in gender, race and skin type in skin lesion datasets and takes a brief look at potential solutions to this problem, especially the lesser discussed colour-based methods.

Keywords: Skin Lesions, Data Augmentation, Deep Learning, Racial Bias, Gender bias.

1 Introduction

Melanoma is a type of skin cancer that can be more threatening than non-melanoma skin cancer, because it is more likely to spread to other organs of the body. Therefore, to decrease the death risk, it is important to diagnose this type of cancer in early rudimentary stages. The most prominent sign of melanoma is the appearance of moles on the body, as shown in Figure 1. On the other hand, manual examination of skin lesion images for skin cancer detection is time-consuming for dermatologists. Due to recent advances in computer vision, and most especially the outstanding performance of deep learning models in analysis of medical images, the focus on computer-aided systems increased dramatically in recent years. Various deep learning and machine learning techniques have emerged to be applied practically in the diagnosis of this disease. For example, El-khatib et al. [El-Khatib et al., 2020] proposed a system to diagnose skin lesions based on deep learning and feature-based methods. The models they employed were GoogLeNet, ResNet-101 and NASNet-Large. Also, Zhang et al. [Zhang et al., 2019] utilized a deep supervised multi-scale network (DSM network) for automatic segmentation of skin lesions in dermoscopy images. Although the results of these methods have been impressive, there remains inequality in these systems due to the lack of diversity in skin lesion datasets [Wen et al., 2022]. There have been many attempts to employ data augmentation methods to address the problem of diversity in datasets [Pham et al., 2018, Sayed et al., 2021, Andrade et al., 2020], however, none of these report on the problem of gender and race bias or skin type diversity.

Figure 1: Signs of melanoma [Kate, 2021]

* D21124371@mytudublin.ie; † ted.burke@tudublin.ie; ‡ jane.courtney@tudublin.ie

 ISBN 978-0-9934207-7-1

In this paper, the diversity reporting of 23 skin datasets is investigated. Data augmentation is commonly used to solve the problem of diversity in datasets, so the most common data augmentation methods used in skin lesion datasets are also investigated here. From this research, we found that in most of the available datasets, diversity reporting is limited, with skin type the least common metadata. Despite this, colour-based methods are less common in data augmentation than methods based on geometry, suggesting that this skin colour diversity issue is not being sufficiently reported or addressed.

2 Methodology

Although applications such as skin detection algorithms boast high classification accuracy (over 90%), these outcomes are not universal. A growing body of research exposes divergent error rates across demographic groups, with the poorest accuracy consistently found in dark-skinned female subjects [Gloster Jr and Neal, 2006, Mehrabi et al., 2021, Hardesty, 2018, Raji and Buolamwini, 2019, Rosenberg et al., 2004, Najibi, 2020]. Researchers are already concerned that skin cancer detection algorithms, many of which are trained primarily on light-skinned individuals, do worse at detecting skin cancer affecting darker skin [Adamson and Smith, 2018, Agbai et al., 2014, Goel et al., 2020, Goyal et al., 2020]. This paper investigates this issue by examining the metadata contained in datasets. It was found that skin type information is the least reported characteristic. Secondly, the types of augmentation methods used to diversify the data in each skin dataset were examined. Here, it was found that colour augmentation is the least common form of augmentation, meaning that the skin colour diversity issue is not being sufficiently addressed. A more detailed description of these two steps follows.

2.1 Dataset

In paper [Wen et al., 2022], 30 datasets were investigated for diversity. Here, a further 23 datasets have been included, totalling 53 publicly available skin lesion datasets and were investigated in research databases like Google Scholar, IEEE Xplore, and SpringerLink. Note that many works use common datasets such as PH2 [Mendonça et al., 2013] and ISIC [Rotemberg et al., 2021]. Although a few of these datasets include skin type information, they do not include all types of skin. Although ethnicity is sometimes reported, this is different from skin type, since ethnicity mostly reflects the geographical area to which each datasets belongs. In many ethnicities, a variety of skin colours are observed, as illustrated in Figure 2, which shows different types of skin colour found in Iran. As shown in Table **1** and Figure 3, in skin lesions datasets, skin type is the least reported metadata, which means that skin colour diversity is not reported or captured.

Figure 2: Different skin types in an Asian country (Iran)

The goal in addressing this diversity issue is to create a database with as much variety as possible in terms of gender, age, race, skin type and any other characteristics. As shown in Figure 3, skin type is reported less in comparison with other features and even ethnicity, which is more commonly reported, is included in just 40% of the datasets. Before training models on datasets, the diversity of the dataset should be known. Without reporting this information, the results of using these datasets in deep learning methods cannot be trusted.

Dataset	Number of images	Number of classes	Metadata			
			Gender	Age	Ethnicity	Skin type
AtlasDerm [Argenziano et al, 2000, Liao et al, 2016, Zhang et al., 2021]	9,503	534	-	-	✓	-
DanDerm [Liao et al, 2016]	1,110	91	-	-	✓	-
Derm7pt [Kawahara et al., 2018]	> 2000	20	-	-	-	-
Derm101 [Boer and Nischal, 2007]	107,656	541	-	-	✓	-
Dermatology Dataset [Güvenir et al., 1998]	336	34	-	✓	-	-
Dermnet [Liao et al, 2016]	19,500	626	-	-	✓	-
DermNet NZ [Zhang et al, 2021]	246	6	-	-	-	-
Dermoscopic Atlas [Argenziano et al, 2000]	872	3	-	-	-	-
ISBI 2018 [Marchetti et al., 2018]	10,015	7	-	-	-	-
Light Field Image [de Faria et al., 2019]	250	8	✓	✓	-	-
MoleMap [Gu et al, 2019, Mikołajczyk and Grochowski, 2018]	102,451	25	-	-	-	-
Normal [Han et al., 2020]	48,271	2	✓	✓	✓	-
OLE [Liao, 2016]	1,300	19	-	-	-	-
SD-128 [Sun et al., 2016]	5,619	128	-	-	✓	-
SD-198 [Sun et al., 2016, Yang et al., 2018]	6,584	198	✓	✓	✓	✓
Skin-10 [He et al., 2019]	10,218	10	-	-	-	-
Skin-100 [He et al., 2019]	19,807	100	-	-	-	-
Skin Cancer' Malignant vs. Benign [Fanconi, 2019, Ashim et al., 2021]	6,594	2	-	-	-	-
SMARTSKINS [Vasconcelos et al., 2014]	-	106	✓	✓	-	-
The Cancer Genome Atlas [Argenziano et al, 2000]	2,860	-	-	-	-	-
The Dermoscopy Skin Lesion Multispectral Image Database [Lézoray et al., 2014]	30	-	-	-	-	-
The Interactive Atlas of Dermoscopy (IAD) [Argenziano et al., 2000]	> 2, 000	2	-	-	-	-
Web [Han et al., 2020]	51,459	174	✓	✓	✓	-

Table 1: Various skin lesion datasets and their metadata

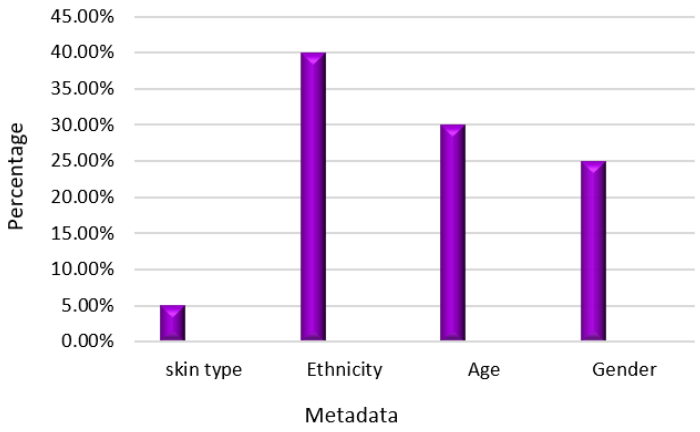


Figure 3: Percentage of reporting of each type of metadata in the 23 datasets

 ISBN 978-0-9934207-7-1

2.2 Augmentation Methods

Achieving skin type diversity in a dataset requires having a sufficient number of subjects with different skin colours. Ideally, diversity would be achieved by simply gathering more data from a more diverse set of subjects. This is not always feasible due to a number of reasons, including lack of available subjects, lower prevalence of melanoma in darker skin, poor quality images due to poor quality of care, racial bias in camera technology and reduced access to screening in some countries and demographics. In the absence of such an ideal dataset, augmentation methods are one way of addressing this problem [Abdelhalim et al., 2021, Pham et al., 2018]. Augmentation methods are routinely used to diversify datasets (in a much more general sense, not specifically related to skin types) before feeding them to classifiers. Data augmentation can be described as increasing the number of samples of training during training phase to not only improve the overall prediction accuracy but also allow a network to better generalize. Although there are various types of augmentation methods, two categories of them are widely used for this purpose. The first group is based on geometric transformations such as translations, rotations, cropping, flipping, scaling [Hameed et al., 2021, Mishra et al., 2021, Sae-Lim et al., 2019] and the second one includes using deep learning-based models, like GAN, CNN, Alex-net and etc [Ahmad et al., 2021, Gu et al, 2019, Ren et al., 2021]. Augmentation for datasets was investigated through overview of publications that were acquired using keywords including “skin lesion augmentation”, “skin lesion classification with deep learning”, “augmentation race bias” in research databases such as Google Scholar, IEEE Xplore, and SpringerLink. Colour is a useful feature to describe and analyse skin texture, skin diversity and different races and ethnicities of people around the world. It was found that most of the augmentation methods in skin lesion articles are based on geometric-based methods rather than colour-based ones. Therefore, they do not address the skin type issue. Figure 4 shows that the percentage using colour-based augmentation methods is just 10.9% in compared with other methods that are mostly based on geometric information.

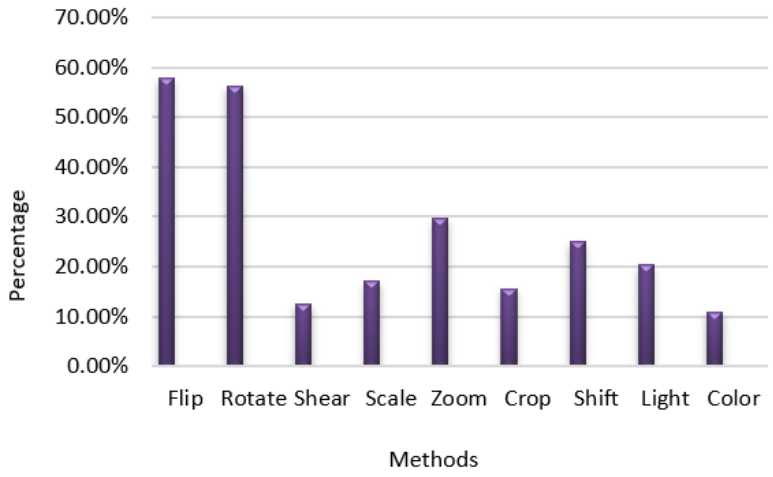


Figure 4: Percentage of most common augmentation methods to diversify skin lesion dataset from 2017 to 2022

3 Conclusions

A significant problem in available datasets used to diagnose and classify skin lesions is that most do not provide information on skin type diversity. Lack of colour diversity in the database can lead to racial bias in the models trained on that data. For example, just 5.0% of datasets provide information on skin types, and even those do not include all skin types. On the other hand, most data augmentation methods on skin lesion datasets are based on geometric information, not colour information, which relates to skin type. The problem of colour diversity in the resulting datasets are not examined and do not state for which type of skin they could achieve the highest accuracy using the augmentation method. Only 10.9% of augmentation methods even diversify colour in the datasets. And after applying data augmentation methods, no criteria for the success of the methods for solving the colour diversity problem were provided.

One measure that can help test for and report on diversity is the Fitzpatrick skin type scale [Fitzpatrick, 1988], which is a measure for classification of skin type based on its reaction to exposure to sunlight due to the amount of melanin pigment in the skin [Zhang et al, 2021]. For example, classifiers' performance can be assessed on each type of skin, and on the other hand, if skin lesion datasets have Fitzpatrick skin type label, not only sources of bias and underrepresented groups can be detected easily, but also more accurate models are developed. In future work, this scale will be used to test the diversity of datasets and the success of augmentation methods. Also, quantifying and assessing the datasets using Fitzpatrick (or possibly another method) will be investigated in our future work.

References

[Abdelhalim et al., 2021] Abdelhalim, I.S.A., Mohamed, M.F. and Mahdy, Y.B. (2021). *Data augmentation for skin lesion using self-attention based progressive generative adversarial network*. Expert Systems with Applications, 165, p.113922.

[Adamson and Smith, 2018] Adamson, A. S., and Smith, A. (2018). *Machine learning and health care disparities in dermatology*. *154*(11), 1247-1248.

[Agbai et al., 2014] Agbai, O.N., Buster, K., Sanchez, M., Hernandez, C., Kundu, R.V., Chiu, M., Roberts, W.E., Draelos, Z.D., Bhushan, R., Taylor, S.C. and Lim, H.W. (2014). *Skin cancer and photoprotection in people of color: a review and recommendations for physicians and the public*. Journal of the American Academy of Dermatology, *70*(4), 748-762.

[Ahmad et al., 2021] Ahmad, B., Jun, S., Palade, V., You, Q., Mao, L. and Zhongjie, M. (2021). *Improving Skin Cancer Classification Using Heavy-Tailed Student T-Distribution in Generative Adversarial Networks (TED-GAN)*. Diagnostics, 11(11), p.2147.

[Andrade et al., 2020] Andrade, C., Teixeira, L. F., Vasconcelos, M. J. M., and Rosado, L. (2020). *Data Augmentation Using Adversarial Image-to-Image Translation for the Segmentation of Mobile-Acquired Dermatological Images*. Journal of Imaging, *7*(1), 2.

[Argenziano et al., 2000] Argenziano, G., Soyer, H. P., De Giorgio, V., Piccolo, D., Carli, P ., Delfino, M., Ferrari, A., Hofmann-Wellenhof, R., Massi, D., Mazzocchetti, G, Scalvenz, M., and Wolf, I. H. (2000). Interactive Atlas of Dermoscopy. Milan, Italy: Edra Medical Publishing & New Media.

[Ashim et al., 2021] Ashim, L.K., Suresh, N. and Prasannakumar, C.V. (2021). *A Comparative Analysis of Various Transfer Learning Approaches Skin Cancer Detection*. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1379-1385). IEEE

[Boer and Nischal, 2007] Boer, A., and Nischal, K. C. (2007). www. derm101. com: *A growing online resource for learning dermatology and dermatopathology*. Indian J Dermatol Venereol Leprol 2007;73:138-140

[de Faria et al., 2019] de Faria, S.M., Filipe, J.N., Pereira, P.M., Tavora, L.M., Assuncao, P.A., Santos, M.O., Fonseca-Pinto, R., Santiago, F., Dominguez, V. and Henrique, M. (2019). *Light field image dataset of skin lesions.* In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 3905-3908). IEEE.

[El-Khatib et al., 2020] El-Khatib, H., Popescu, D. and Ichim, L. (2020). *Deep learning–based methods for automatic diagnosis of skin lesions*. Sensor , *20*(6), p.1753.

[Fanconi, 2019] Fanconi, C. (2019) Retrieved from https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign

[Fitzpatrick, 1988] Fitzpatrick, T. B. (1988). *The validity and practicality of sun-reactive skin types I through VI*. Archives of dermatology, 124(6), pp.869-871.

[Gloster Jr and Neal, 2006] Gloster Jr, H. M., and Neal, K. (2006). *Skin cancer in skin of color*. Journal of the American Academy of Dermatology. *55*(5), 741-760.

[Goel et al., 2020] Goel, K., Gu, A., Li, Y. and Ré, C. (2020). *Model patching: Closing the subgroup performance gap with data augmentation*. arXiv preprint arXiv:2008.06775.

[Goyal et al., 2020] Goyal, M., Knackstedt, T., Yan, S. and Hassanpour, S. (2020). *Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities*. Computers in Biology and Medicine, 127, p.104065.

[Gu et al, 2019] Gu, Y., Ge, Z., Bonnington, C.P. and Zhou, J. (2019). *Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification*. IEEE journal of biomedical and health informatics, 24(5), pp.1379-1393.

[Güvenir et al., 1998] Güvenir, H.A., Demiröz, G. and Ilter, N. (1998). *Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals*. Artificial intelligence in medicine, 13(3), pp.147-165.

[Hameed et al., 2021] Hameed, A., Umer, M., Hafeez, U., Mustafa, H., Sohaib, A., Siddique, M.A. and Madni, H.A. (2021). *Skin lesion classification in dermoscopic images using stacked Convolutional Neural Network*. Journal of Ambient Intelligence and Humanized Computing, pp.1-15.

[Han et al., 2020] Han, S.S., Park, I., Chang, S.E., Lim, W., Kim, M.S., Park, G.H., Chae, J.B., Huh, C.H. and Na, J.I. (2020). *Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders*. Journal of Investigative Dermatology, 140(9), pp.1753-1761.

[Hardesty, 2018] Hardesty, L. (2018). *Study finds gender and skin-type bias in commercial artificial-intelligence systems*. MIT News. Retrieved from https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 on 2 June 2022

[He et al., 2019] He, X., Wang, S., Shi, S., Tang, Z., Wang, Y., Zhao, Z., Dai, J., Ni, R., Zhang, X., Liu, X. and Wu, Z. (2019). *Computer-Aided Clinical Skin Disease Diagnosis Using CNN and Object Detection Models.* In 2019 IEEE International Conference on Big Data (Big Data) (pp. 4839-4844). IEEE.

[Kate R., 2021] Kate Robinson, TechnologyNetworks Cancer Research (2021). Using Artificial Intelligence To Detect Melanoma. Retrieved from https://www.technologynetworks.com/cancer-research/blog/using-artificial-intelligence-to-detect-melanoma-355242

[Kawahara et al., 2018] Kawahara, J., Daneshvar, S., Argenziano, G., and Hamarneh, G. (2018). *Seven-point checklist and skin lesion classification using multitask multimodal neural nets*. IEEE journal of biomedical and health informatics, 23(2), pp.538-546

[Lézoray et al., 2014] Lézoray, O., Revenu, M., and Desvignes, M. (2014). *Graph-based skin lesion segmentation of multispectral dermoscopic images.* In 2014 IEEE International Conference on Image Processing (ICIP) (pp. 897-901). IEEE

[Liao et al., 2016] Liao, H., Li, Y., and Luo, J. (2016). *Skin disease classification versus skin lesion characterization: Achieving robust diagnosis using multi-label deep neural networks.* In 2016 23rd International Conference on Pattern Recognition (ICPR) (pp. 355-360). IEEE.

[Liao, 2016] Liao, H. (2016). *A deep learning approach to universal skin disease classification.* University of Rochester Department of Computer Science, CSC.

[Marchetti et al., 2018] Marchetti, M.A., Codella, N.C., Dusza, S.W., Gutman, D.A., Helba, B., Kalloo, A., Mishra, N., Carrera, C., Celebi, M.E., DeFazio, J.L. and Jaimes, N. (2018). *Results of the 2016 international skin imaging collaboration isbi challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images*. Journal of the American Academy of Dermatology, 78(2), p.270.

[Mehrabi et al., 2021] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021). *A survey on bias and fairness in machine learning*. ACM Computing Surveys (CSUR), *54*(6), pp.1-35.

[Mendonça et al., 2013] Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R. and Rozeira, J. (2013). *PH 2-A dermoscopic image database for research and benchmarking.* In 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC) (pp. 5437-5440). IEEE.

[Mikołajczyk and Grochowski, 2018] Mikołajczyk, A. and Grochowski, M. (2018). *Data augmentation for improving deep learning in image classification problem*. In 2018 international interdisciplinary PhD workshop (IIPhDW) (pp. 117-122). IEEE

[Mishra et al., 2021] Mishra, V., A. Kumar, and M. Arora. (2021). *Deep convolution neural network based automatic multi-class classification of skin cancer from dermoscopic images*. in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). 2021. IEEE.

[Najibi, 2020] Najibi, A. (2020). *Racial Discrimination in Face Recognition Technology*. In Science in the News. Harvard University Graduate School of Arts and Sciences. Retrieved from https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/ on 2 June 2022

[Pham et al., 2018] Pham, T.C., Luong, C.M., Visani, M. and Hoang, V.D. (2018). *Deep CNN and data augmentation for skin lesion classification.* In Asian Conference on Intelligent Information and Database Systems (pp. 573-582). Springer, Cham.

[Raji and Buolamwini, 2019] Raji, I. D., and Buolamwini, J. (2019). *Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products.* In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 429-435).

[Ren et al., 2021] Ren, Z., Guo, Y., Stella, X.Y. and Whitney, D. (2021). December. *Improve Image-based Skin Cancer Diagnosis with Generative Self-Supervised Learning*. In 2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) (pp. 23-34). IEEE.

[Rosenberg et al., 2004] Rosenberg, C.A., Greenland, P., Khandekar, J., Loar, A., Ascensao, J. and Lopez, A.M. (2004). *Association of nonmelanoma skin cancer with second malignancy: the women's health initiative observational study*. Cancer: Interdisciplinary International Journal of the American Cancer Society, 100(1), pp.130-138.

[Rotemberg et al., 2021] Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D. and Halpern, A. (2021). *A patient-centric dataset of images and metadata for identifying melanomas using clinical context*. Scientific data, 8(1), pp.1-8.

[Sae-Lim et al., 2019] Sae-Lim, W., Wettayaprasit, W. and Aiyarak, P. (2019). *Convolutional neural networks using MobileNet for skin lesion classification*. In 2019 16th international joint conference on computer science and software engineering (JCSSE) (pp. 242-247). IEEE.

[Sayed et al., 2021] Sayed, G.I., Soliman, M.M. and Hassanien, A.E. (2021). *A novel melanoma prediction model for imbalanced data using optimized SqueezeNet by bald eagle search optimization*. Computers in Biology and Medicine, 136, p.104712.

[Sun et al., 2016] Sun, X., Yang, J., Sun, M. and Wang, K. (2016). *A benchmark for automatic visual classification of clinical skin disease images.* In European Conference on Computer Vision (pp. 206-222). Springer, Cham.

[Vasconcelos et al., 2014] Vasconcelos, M.J.M., Rosado, L. and Ferreira, M. (2014). *Principal axes-based asymmetry assessment methodology for skin lesion image analysis.* In International symposium on visual computing (pp. 21-31). Springer, Cham.

[Wen et al., 2022] Wen, D., Khan, S.M., Xu, A.J., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A.K., Liu, X. and Matin, R.N. (2022) *Characteristics of publicly available skin cancer image datasets*. training, 11, p.12.

[Yang et al., 2018] Yang, J., Sun, X., Liang, J. and Rosin, P.L. (2018). *Clinical skin lesion diagnosis using representations inspired by dermatologist criteria.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1258-1266).

[Zhang et al., 2019] Zhang, G., Shen, X., Chen, S., Liang, L., Luo, Y., Yu, J. and Lu, J. (2019). *DSM: A deep supervised multi-scale network learning for skin cancer segmentation*. IEEE Access, 7, pp.140936-140945.

[Zhang et al., 2021] Zhang, L., Mishra, S., Zhang, T., Zhang, Y., Zhang, D., Lv, Y., Lv, M., Guan, N., Hu, S., Chen, D.Z. and Han, X (2021). *Design and Assessment of Convolutional Neural Network Based Methods for Vitiligo Diagnosis*. Frontiers in medicine, p.1901.

Pre- and Post-Operative Analysis of Planar Radiographs in Total Hip Replacement

Oscar Denton[1], Christopher Madden-McKee[2], Janet Hill[2], David Beverland[2], Nicholas Dunne[3], Alex Lennon[1]

[1]Queens University Belfast, School of Mechanical & Aerospace Engineering, Belfast, UK.

[2]Musgrave Park Hospital, Primary Joint Unit, Belfast, UK.

[3]School of Mechanical and Manufacturing Engineering, Dublin City University, Dublin, Ireland.

Abstract

Computed-Tomography scans represent the gold standard for accuracy when preoperatively templating and postoperatively assessing the hip. However, planar radiographs are used as standard, sacrificing accuracy. In this work, a method is proposed to more accurately assess femoral offset and neck-shaft angle from two planar radiographs (frontal and lateral), allowing more reliable templating of a modular stem. A second method is proposed to accurately assess postoperative stem version from planar frontal radiographs.

Keywords: Imaging, Image Processing, Medical Imaging, Hip Arthroplasty.

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/EXJL3727

1 Introduction

In total hip replacement (THR), medical images are used to define the native joint preoperatively and assess the accuracy of reconstruction postoperatively. Preoperative templating forewarns likely surgical complications and helps identify a modular stem to best replicate the native joint, whilst postoperative assessment defines the accuracy of joint reconstruction versus that targeted. The gold standard for accurate assessment of joint geometry is 3D computed tomography (CT), which produces a 3D model of the joint. However, cost of image acquisition and patient radiation exposure prevent routine use. Radiographs have a relatively low acquisition cost and exposes the patient

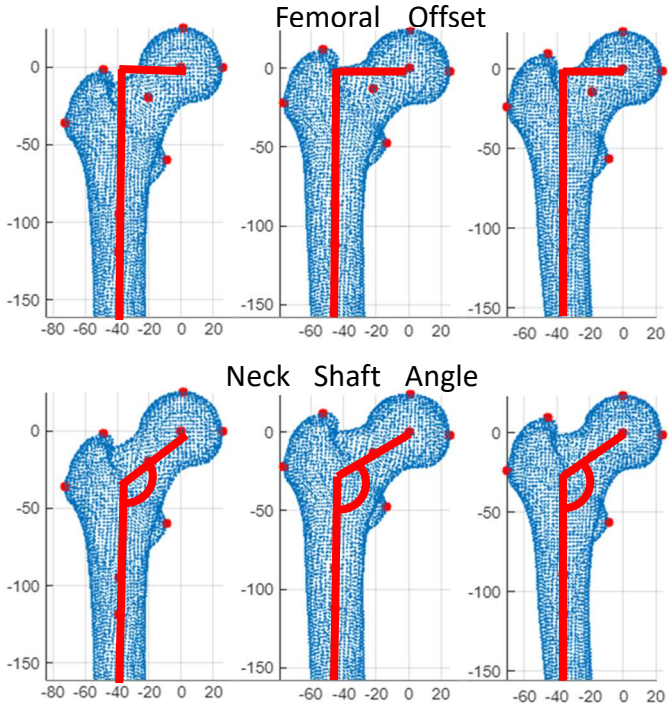


Figure 1 - Top, graphic demonstrating FO on varying femoral geometries. Bottom, graphic demonstrating NSA on varying femoral geometries.

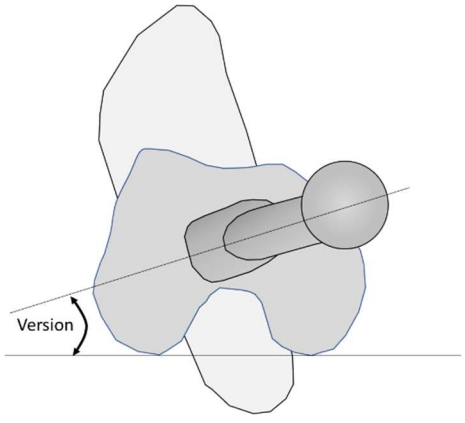


Figure 2 - Stem version. View of femur and foot from above. Version is angle relative to line joining rearmost points of the distal femur.

 ISBN 978-0-9934207-7-1

to a lesser radiation dose, making them the standard imaging technique both preoperatively and postoperatively. During radiograph acquisition, the 3D joint geometry is projected to a 2D geometry. This process introduces magnification and projection error. While magnification error can largely be controlled through magnification markers, projection error leads to distortion of the geometry and can misrepresent key features.

Two key variables used to define a modular stem are femoral offset (FO); distance from femoral head centre to long-axis of the bone) and neck-shaft angle (NSA) (Figure 1). Projection error leads to FO being under reported in anterior-posterior (AP; frontal view) radiographs [O'Connor *et al*, 2018, Lechler *et al,* 2014]; this exceeds 5 mm in 35% of AP radiographs [Weber *et al,* 2014]. Similarly, projected NSA has been shown to depend on orientation of the femur about one or more of its anatomical axes [O'Connor *et al*, 2018]. Mismeasurement of these features can lead to poor choice in modular stem components and inaccurate joint reconstruction. Accurate templating of FO and NSA is thus key to the pre-operative planning of THR.

Accurate restoration of stem version (Figure 2) is considered a prerequisite for a successful THR procedure and has been reported to influence risk of impingement [Malik *et* al, 2007], range of motion [D'Lima *et al*, 2000], implant wear [Patel *et al*, 2010], and loosening [Kiernan *et al*, 2013]. It has also been suggested that excessive version impacts biomechanics of the joint through alterations to muscle lines of action and lever arms [Scorcelletti *et al*, 2020]. Restoration of native femoral version is targeted as standard [Patel et al, 2010, Belzunce *et al,* 2020]. Accurate assessment of stem version postoperatively is therefore key to studying its impact on variable outcomes in postoperative patient cohorts.

2 State of the Art

2.1 Pre-operative Radiographic Templating for FO and NSA

Standard templating based on AP radiographs involves overlaying components to determine a best fit [DePuy Synthes, 2017]. This process doesn't attempt to correct for projection error, instead assuming an idealised radiograph. State of the art templating uses CT scans, which produce a 3D reconstruction of the femur, from which FO and NSA can be measured. As discussed earlier, CT is unsuitable for routine use. Low-dose digital stereo-radiography, using an EOS imaging system (EOS Imaging SA, Paris, France), has been validated as a method for reconstructing 3D proximal femoral geometries [Le Bras *et al,* 2004*,* Illés and Somoskeöy, 2012]. To achieve this, simultaneous biplanar X-ray images are captured by slot scanning in an upright, physiological load-bearing position, using ultra-low radiation doses. Despite its accuracy, its application as standard is limited by availability and the necessity that the patient be standing. The National Institute for Health and Care Excellence doesn't recommend EOS for routine use in the NHS [National Institute for Health and Care Excellence, 2011].

2D to 3D reconstructions of the proximal femoral geometry from standard radiographs has been investigated as a cost-effective means to obtain 3D patient geometries whilst avoiding significant radiation doses; once reconstructed, the 3D geometry can be processed to obtain features. These methods often utilise a parametric representation of shape, via a statistical shape model (SSM) [Zheng and Schumann, 2009, Schumann *et al* 2010], or shape and intensity, via a statistical shape and intensity model (SSIM) [Whitmarsh *et al*, 2010, Whitmarsh *et al* , 2011, Zheng, 2015]. Standard practice involves AP and lateral radiographs taken sequentially, meaning that perfectly orthogonal or defined conditions are unlikely. Despite asynchronous AP and lateral radiographs being common in practice, many techniques assume numerous radiographs at well-defined orientations or positioning, limiting their application with standard radiographic techniques. Reconstructing, or directly templating, the femoral geometry based on clinical radiographs, approximately orthogonal and only capturing the proximal femur, is thus an area of research interest.

 ISBN 978-0-9934207-7-1

2.2 Post-operative Stem Version Assessment

Numerous techniques have been proposed for postoperative assessment of stem version. Cross-table radiographs have been shown to be inaccurate [Kanazawa *et al*, 2016], despite their applicability to assess cup version [Nunley *et al*, 2011, Seo *et al,* 2017]. Lee *et al* proposed the modified Budin radiograph, a *seated* rear-front radiograph [Lee *et al*, 2013]. The technique requires a nonstandard radiograph, specifically targeting stem version prediction, and is inapplicable to patients with poor hip mobility. Low-dose digital stereo-radiography has been utilised to asses stem version, via a 3D reconstruction, by Guenoun *et al* [Guenoun *et al*, 2015]. This process requires access to the EOS imaging system, which is not widely available. Radio-stereometric analysis has been used to accurately measure changes in prosthetic orientation [Alfaro-Adrián *et al*, 2001], but is considered impractical for routine clinical practice [Kärrholm *et al*, 2006]. For most applications, markers are attached to the prosthesis prior to implantation, increasing manufacturing costs, and potentially changing prosthesis behaviour *in vivo* [Kärrholm *et al*, 2006]. In a study that avoided non-standard follow-up data, Weber *et al* [Weber *et al*, 2015] proposed a method to assess stem version from standard AP radiographs. The technique input the projected neck-shaft angle ($NSA_{Projected}$) and known implant neck-shaft angle (NSA_{True}) to a trigonometric equation (Equation 1).

$$\text{Stem Version} = cos^{-1}\left[tan\left(NSA_{Projected}\right) / tan\left(NSA_{True}\right)\right] \quad (1)$$

Femoral flexion was found to correlate with version prediction error, suggesting femoral flexion interferes with the methodology. Ha *et al* proposed a similar technique, in which the projected NSA in both an AP and trans-lateral radiograph are processed through a trigonometric equation to predict stem version [Ha *et al*, 2021]. The modified Budin method and Weber method have been shown to achieve comparable accuracy in an independent study [Woerner *et al*, 2016], with the Budin method found to underestimate angles <10° and overestimate angles >10° (trend to the middle), whilst the Weber technique produced a number of outliers where stem version was significantly overestimated. This was attributed to uncontrolled femoral rotation, rather than unquantified radiographic stem tilt.

3 Methods

3.1 Pre-Operative Radiographic Templating of FO and NSA

3.2.1 Optimisation

Forty-two corresponding female femoral geometries were used to construct an SSM, by defining each geometry as a deformation about a mean case and performing principal component analysis (PCA) on the deformation fields. The SSM is a parametric representation of shape variation within femoral geometries. The target contours are defined as a series of Cartesian coordinates representing 500 points evenly spaced along the contour of the radiographic femur. It is assumed that to replicate the contours observed in a given pair of radiographs, both the geometry, parametrised by the SSM, and the orientation, distinct in each radiograph and defined in degrees about the head centre, must match.

A fitness function was defined in which an SSM sample, orientation in AP, and orientation in lateral are inputs. The geometry is sampled from the SSM, and rotations applied based on the input variables. Ray casting is then used to replicate radiograph acquisition. The boundary of the resultant 2D point cloud is extracted and the boundary is resampled with the start point defined as the distal lateral corner. The similarity to the target contours is evaluated using Equation 2. Particle swarm optimisation is used to evaluate a set of inputs that adequately match the target case (Figure 3).

$$Health = sum\big(abs(Target[x,y] - Evaluated[x,y])\big) \quad (2)$$

 ISBN 978-0-9934207-7-1

Initially the orientation in the lateral radiograph is optimised in isolation; the output is used to bound the search space in a subsequent optimisation in which the SSM and both orientations are considered. The optimised SSM is reconstructed; FO and NSA are automatically measured from the reconstructed geometry. Optimisations were carried out using five SSM modes, accounting for 96% of the shape variance of the 10 modes used to generate targets. Eighteen instances were run and the error in FO and NSA compared to direct radiographic measurement.

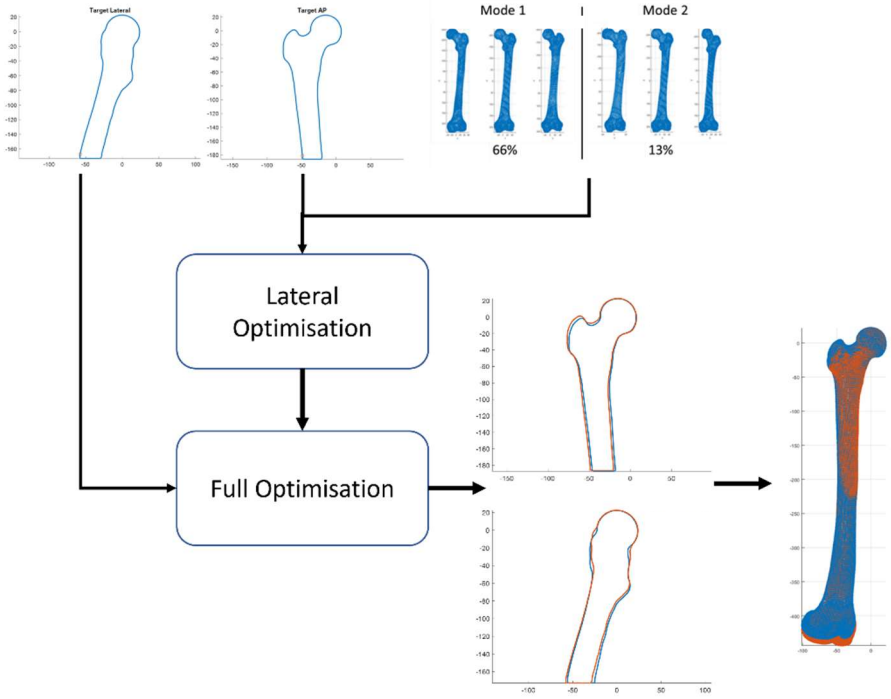


Figure 3 - Optimisation process starting with target contours and an SSM, through to optimised simulated radiographs and reconstructed 3D geometry.

3.2.2 Computational Validation

Target radiographs were generated by sampling from the first 10 modes of the SSM and ray-casting from a point source to generate AP and lateral radiographs of the femurs. Orientations were randomly sampled from a uniform distribution between ±25° flexion, ±5° abduction and ±30° rotation (Figure 4); rotation of the lateral geometry was constrained to within 75° and 105° of the AP, where 90° would represent ideal, orthogonal radiographs. Contours were extracted and used as inputs to the optimisation.

3.2 Post-operative Stem Version Assessment

3.2.1 Model Building

The surface mesh of a size 11 standard offset collared stem (KA) (CORAIL, DePuy Synthes Inc., USA,) was imported to MATLAB (MathWorks Inc., USA) and aligned with the Cartesian axes, with head centre positioned at the origin. Radiographic stem versions and tilts of 0–50° in 1° increments were applied to the stem. At each instance a stem tilt (mean 0, SD 3), and translation along each of the Cartesian axes (medial-lateral: mean 0, SD 15, superior-inferior: mean 0, SD 15, anterior-posterior: mean 0, SD 5) were randomly sampled from a normal distribution and applied to the stem. An AP radiograph, with focus to film distance of 1.15 m and an object to detector distance of 0.15 m (before the applied random translation) was simulated. The measurement of radiographic features, FO, NSA, stem-length (SL), neck-length (NL), and tip-to-head length (TH) was automated from the simulated X-ray (Figure 5). A random error sampled from a uniform distribution between ±1° or ±1 mm was applied to each measured

 ISBN 978-0-9934207-7-1

feature to replicate variability in manual measurement of features in a clinical context. The NSA, with applied error, was then processed using the Weber method (Equation 1), to give a rough stem version, SV.

SV was used as a sixth radiographic feature for model training. The data generation was repeated three times for a total of 7803 datapoints. This resulted in a dataset of radiographic features at known radiographic stem versions, which were used to train a Gaussian process regression (GPR) model using the MATLAB Statistics and Machine Learning Toolbox.

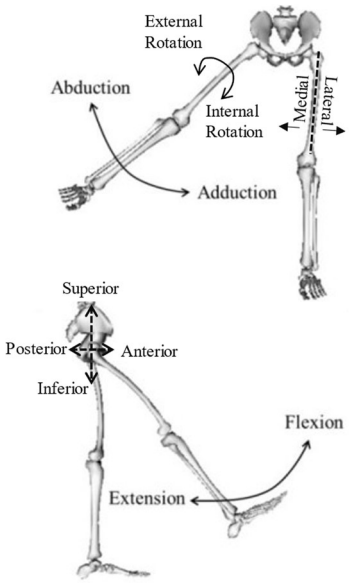


Figure 4 – Translational and rotational degrees of freedom of the hip

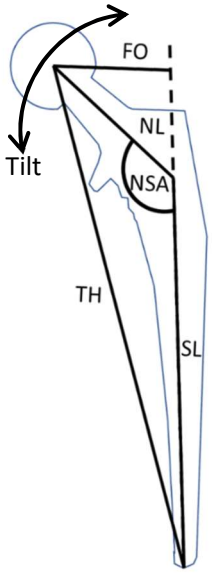


Figure 5 - Radiographic features generated and used as inputs to the GPR model.

3.2.2 *In Vitro* Validation

A size 11 standard offset collared stem (KA) and 28mm head (CORAIL, DePuy Synthes Inc., USA,) was mounted on a tripod ball head gimbal (BEIKE, China; 82 model – BK-03) allowing orientation to be manipulated. A dual axis inclinometer was used to orient the stem to within ±1°, verified by two independent authors (Figure 6). Radiographs were obtained at 5° increments between 0° and 40° of both radiographic stem version and radiographic stem tilt. Extreme instances of high combined rotations could not be replicated with the gimbal setup, resulting in a measured range of 0–40° version up to 15° tilt, 0-30° version at 20° tilt, 0–25° version at 25° tilt, 0–20° version at 30° tilt, 0–15° version at 35° tilt and 0° version at 40° tilt. AP radiographs of the implant were taken with a focus-to-film distance of 1.15 m. Head centre was maintained at 90 mm lateral of the radiographic source and 150 mm from the detector in all radiographs.

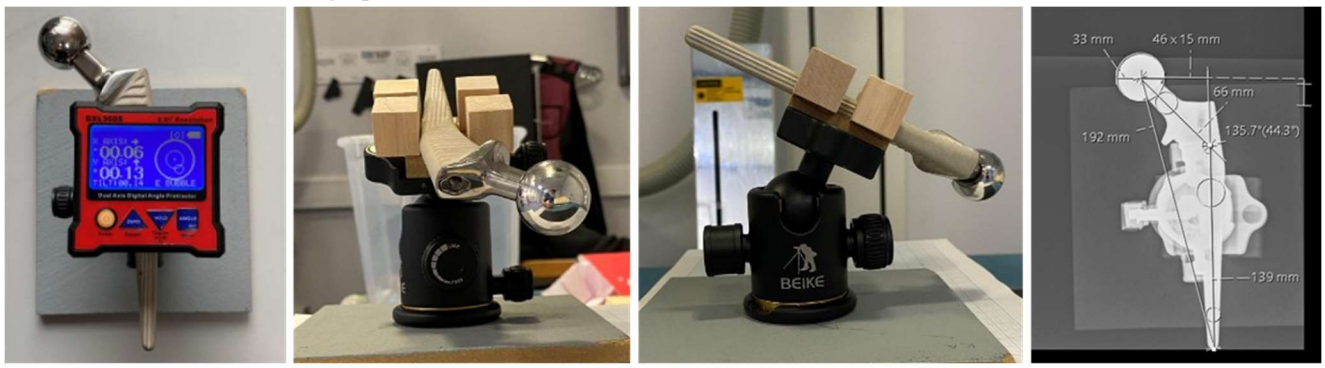


Figure 6 – Left, Gimbal setup. CORAIL stem mounted on a gimbal to allow orientation and a dual-axis inclinometer to accurately control orientation. Right, radiographic features measured from an in vivo radiograph using NIPACS.

 ISBN 978-0-9934207-7-1

Features were manually extracted from the *in vitro* radiographs utilising the Northern Ireland Picture Archiving and Communication System (NIPACS, NI Dept Health), as would be the case with patient radiographs, replicating an error source within clinical practice (Figure 6). The radiographic features were used to assess the accuracy of the proposed GPR version predictor and compared to the Weber technique. Mean error, RMSE, percentage within 10° of true, and coefficient of determination, R^2, values were produced for GPR and Weber predictions.

4 Results

4.1 Pre-operative Radiographic Templating of FO and NSA

Standard radiographic FO trended low of true by a mean of 10.77 mm (Table 1), whilst radiographic NSA trended high of true by a mean of 9.39°. The mean error was reduced when optimising both variables using the SSM-based optimisation, by a factor of ~3.6 for FO and ~2.2 for NSA (Figure 7). Optimising FO resulted in a reduction in *maximum* error by a factor of ~3.6, whilst maximum NSA was reduced by a factor of ~1.8 when optimised. The correlation coefficient between optimised and true FO was 0.1 greater than radiographic and true. The correlation coefficient was approximately equal between optimised and true NSA and between radiographic and true NSA (Table 1).

	Femoral Offset Error (mm)		Neck-Shaft Angle Error (°)	
	Radiographic	Optimised	Radiographic	Optimised
Mean	10.77	3.02	9.39	4.22
Max	20.35	5.65	18.57	10.27
R	0.84	0.94	0.87	0.85

Table 1 – Mean and maximum error, and correlation for radiographic and optimised measures of FO and NSA.

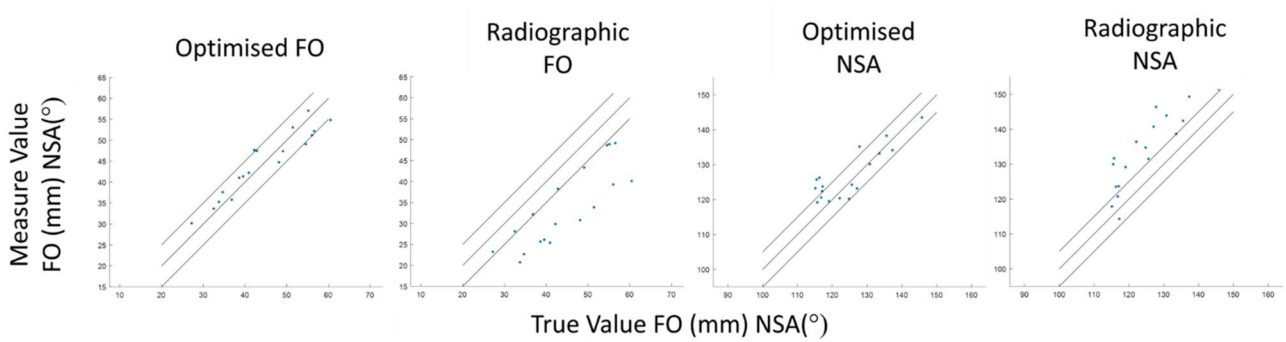


Figure 7 – Comparison between FO and NSA measure directly from radiographs and using the proposed technique versus anatomic values.

4.2 Post-operative Stem Version Assessment

The mean radiographic stem version error in the experimental dataset utilising the GPR model was 1.9° (overestimation) versus 13.8° (overestimation) using the Weber technique (Figure 8). The RMSE using the GPR model was 2.64° versus 14.97° using the Weber technique. 100% of GPR predictions were within 10° of true, versus only 20% using the Weber technique. The GPR predictor produced R^2=0.98 ($p<0.001$) compared with R^2=0.89 ($p<0.001$) using the Weber technique. Fitting a linear polynomial to each predictor, the GPR predictor has a slope of 0.95, meaning error is lower when higher radiographic stem versions are being predicted. The inverse is the case with the Weber technique, slope 1.08, which is more accurate when lower radiographic stem versions are being predicted.

 ISBN 978-0-9934207-7-1

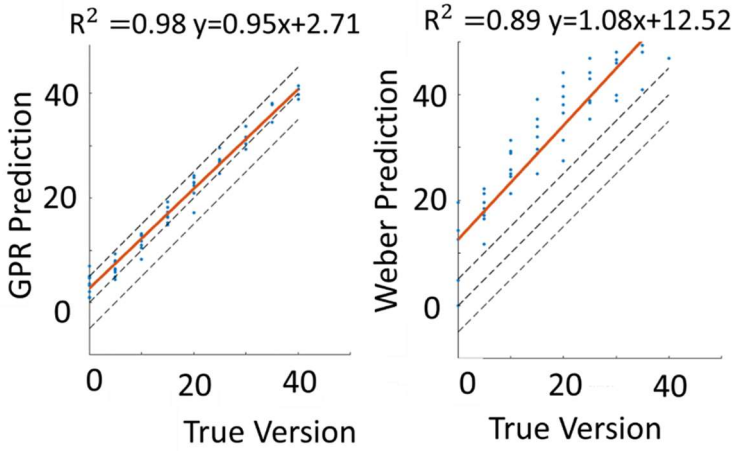


Figure 8 - Experimental validation. Predicted radiographic stem version of an invitro dataset using GPR and Weber technique predictors versus known radiographic stem version. Blue points represent in vitro radiographs and associated version prediction. Dashed lines represent +5°, 0° and -5° error. Red lines are a linear polynomial fitted to each in vitro dataset.

7 Conclusions

Computational testing indicates that SSM-enabled 2D-3D reconstruction is capable of pre-operative templating both FO and NSA, from asynchronous AP and lateral radiographs, more accurately than the standard technique of direct radiographic measurement. Future work will focus on automating contour extraction directly from radiographs, and validating the technique based on real radiographs.

Additionally, a GPR predictor of post-operative stem version validated against in vitro radiographs was shown to be more accurate and more robust to variable radiographic tilt than existing planar radiograph-based techniques. However, the proposed technique is currently only validated for a single modular stem combination; distinct models are needed for application to different stems. The proposed technique also requires 5 features to be extracted with an associated time cost. Future work will involve automating the feature extraction process from radiographs.

8 References

[O'Connor *et al*, 2018] J. D. O'Connor, M. Rutherford, J. C. Hill, D. E. Beverland, N. J. Dunne, and A. B. Lennon, "Effect of combined flexion and external rotation on measurements of the proximal femur from anteroposterior pelvic radiographs," *Orthop. Traumatol. Surg. Res.*, vol. 104, pp. 449–454, 2018.

[Lechler *et al, 2014*] P. Lechler *et al.*, "The influence of hip rotation on femoral offset in plain radiographs," *Acta Orthop.*, vol. 85, no. 4, pp. 389–395, Aug. 2014, doi: 10.3109/17453674.2014.931196.

[Weber et al, 2014] M. Weber, M. L. Woerner, H. R. Springorum, A. Hapfelmeier, J. Grifka, and T. F. Renkawitz, "Plain Radiographs fail to reflect femoral offset in total Hip Arthroplasty," *J. Arthroplasty*, vol. 29, no. 8, pp. 1661–1665, 2014, doi: 10.1016/j.arth.2014.03.023.

[Malik et al, 2007] A. Malik, A. Maheshwari, and L. D. Dorr, "Impingement with total hip replacement," *J. Bone Jt. Surg. - Ser. A*, vol. 89, no. 8, pp. 1832–1842, 2007, doi: 10.2106/JBJS.F.01313.

[D'Lima et al, 2000] D. D. D'Lima, A. G. Urquhart, K. O. Buehler, R. H. Walker, and C. W. Colwell Jr, "The Effect of the Orientation of the Acetabular and Femoral Components on the Range of Motion of the Hip at Different Head-Neck Ratios," *J. Bone Jt. Surg.*, vol. 82a, no. 3, 2000.

[Patel et al, 2010] A. B. Patel, R. R. Wagle, M. M. Usrey, M. T. Thompson, S. J. Incavo, and P. C. Noble, "Guidelines for Implant Placement to Minimize Impingement During Activities of Daily Living After Total Hip Arthroplasty," *J. Arthroplasty*, vol. 25, no. 8, pp. 1275-1281.e1, 2010, doi: 10.1016/j.arth.2009.10.007.

[Kiernan et al, 2013] S. Kiernan, K. L. Hermann, P. Wagner, L. Ryd, and G. Flivik, "The importance of adequate stem anteversion for rotational stability in cemented total hip replacement," *Bone Joint J.*, vol. 95-B, no. 1, pp. 23–30, Jan. 2013, doi: 10.1302/0301-620X.95B1.30055.

[Scorcelletti et al, 2020] M. Scorcelletti, N. D. Reeves, J. Rittweger, and A. Ireland, "Femoral anteversion: significance and measurement," *J. Anat.*, vol. 237, no. 5, pp. 811–826, Nov. 2020, doi: 10.1111/JOA.13249.

[Belzunce et al, 2020] M. A. Belzunce, J. Henckel, A. Di Laura, and A. Hart, "Uncemented femoral stem orientation and position in total hip arthroplasty: A CT study," *J. Orthop. Res.*, vol. 38, no. 7, pp. 1486–1496, Jul. 2020, doi: 10.1002/jor.24627.

[DePuy Synthes, 2017] DePuy Synthes, "DePuy Synthes CORAIL Hip System Product Rationale and Surgical Technique," 2017.

[Le Bras et al, 2004] A. Le Bras *et al.*, "3D reconstruction of the proximal femur with low-dose digital stereoradiography," *Comput. Aided Surg.*, vol. 9, no. 3, pp. 51–57, Jan. 2004, doi: 10.3109/10929080400018122.

[Illés and Somoskeöy, 2012] T. Illés and S. Somoskeöy, "The EOS™ imaging system and its uses in daily orthopaedic practice," *Int. Orthop.*, vol. 36, no. 7, p. 1325, Jul. 2012, doi: 10.1007/S00264-012-1512-Y.

[National Institute for Health and Care Excellence, 2011] National Institute for Health and Care Excellence, "The EOS 2D/3D imaging system," 2011.

[Zheng and Schumann, 2009] G. Zheng and S. Schumann, "3D reconstruction of a patient-specific surface model of the proximal femur from calibrated x-ray radiographs: A validation study," *Med. Phys.*, vol. 36, no. 4, pp. 1155–1166, 2009, doi: 10.1118/1.3089423.

[Schumann et al, 2010] S. Schumann, M. Tannast, L. P. Nolte, and G. Zheng, "Validation of statistical shape model based reconstruction of the proximal femur-A morphology study," *Med. Eng. Phys.*, vol. 32, no. 6, pp. 638–644, Jul. 2010, doi: 10.1016/j.medengphy.2010.03.010.

[Whitmarsh et al, 2010] T. Whitmarsh *et al.*, "3D Bone Mineral Density Distribution and Shape Reconstruction of the Proximal Femur from a Single Simulated DXA Image: An In Vitro Study," 2010.

[Whitmarsh et al, 2011] T. Whitmarsh, L. Humbert, M. De Craene, L. M. D. R. Barquero, and A. F. Frangi, "Reconstructing the 3D Shape and Bone Mineral Density Distribution of the Proximal Femur From Dual-Energy X-Ray Absorptiometry," *IEEE Trans. Med. Imaging*, vol. 30, no. 12, 2011, Accessed: Mar. 17, 2021. [Online]. Available: https://ieeexplore-ieee-org.queens.ezp1.qub.ac.uk/stamp/stamp.jsp?tp=&arnumber=5962359.

[Zheng, 2015] G. Zheng, "Personalized X-Ray Reconstruction of the Proximal Femur via Intensity-Based Non-rigid 2D-3D Registration," *Med. Imaging 2015 Image Process.*, vol. 9413, p. 94133B, 2015, doi: 10.1117/12.2082339.

[Kanazawa et al, 2016] M. Kanazawa, Y. Nakashima, S. Hamai, M. Hirata, and Y. Iwamoto, "Is a Stem Version on the Crosstable Lateral Radiograph Accurate in Total Hip Arthroplasty?," *J. Arthroplasty*, vol. 31, no. 6, pp. 1356–1360, Jun. 2016, doi: 10.1016/j.arth.2015.12.022.

[Nunley et al, 2011] R. M. Nunley, J. A. Keeney, J. Zhu, J. C. Clohisy, and R. L. Barrack, "The reliability and variation of acetabular component anteversion measurements from cross-table lateral radiographs," *J. Arthroplasty*, vol. 26, no. SUPPL. 6, pp. 84–87, Sep. 2011, doi: 10.1016/j.arth.2011.03.039.

[Seo et al, 2017] H. Seo *et al.*, "New cross-table lateral radiography method for measuring acetabular component anteversion in total hip arthroplasty: A prospective study of 93 primary THA," *HIP Int.*, vol. 27, no. 3, pp. 293–298, May 2017, doi: 10.5301/hipint.5000456.

[Lee et al, 2013] Y. K. Lee, T. Y. Kim, Y. C. Ha, B. J. Kang, and K. H. Koo, "Radiological measurement of femoral stem version using a modified Budin method," *Bone Jt. J.*, vol. 95 B, no. 7, pp. 877–880, 2013, doi: 10.1302/0301-620X.95B7.31195.

[Guenoun et al, 2015] B. Guenoun, F. El Hajj, D. Biau, P. Anract, and J. P. Courpied, "Reliability of a new method for evaluating femoral stem positioning after total hip arthroplasty based on stereoradiographic 3d reconstruction," *J. Arthroplasty*, vol. 30, no. 1, pp. 141–144, Jan. 2015, doi: 10.1016/j.arth.2014.07.033.

[Alfaro-Adrián et al, 2001] J. Alfaro-Adrián, H. S. Gill, and D. W. Murray, "Should total hip arthroplasty femoral components be designed to subside? A radiostereometric analysis study of the Charnley Elite and Exeter stems," *J. Arthroplasty*, vol. 16, no. 5, pp. 598–606, 2001, doi: 10.1054/arth.2001.23576.

[Kärrholm et al, 2006] J. Kärrholm, R. H. S. Gill, and E. R. Valstar, "The history and future of radiostereometric analysis," *Clin. Orthop. Relat. Res.*, vol. 448, no. 448, pp. 10–21, 2006, doi: 10.1097/01.blo.0000224001.95141.fe.

[Weber et al, 2015] M. Weber *et al.*, "The validity of a novel radiological method for measuring femoral stem version on anteroposterior radiographs of the hip after total hip arthroplasty," *Bone Jt. J.*, vol. 97-B, no. 3, pp. 306–311, Mar. 2015, doi: 10.1302/0301-620X.97B3.34618.

[Ha et al, 2021] Y. C. Ha, J. Il Yoo, J. M. Ahn, Y. K. Lee, Y. Kang, and K. H. Koo, "Trans-lateral decubitus radiograph of the hip: A new view to measure the anteversion of the femoral stem," *Asian J. Surg.*, vol. 44, no. 1, pp. 99–104, Jan. 2021, doi: 10.1016/j.asjsur.2020.03.016.

[Woerner et al, 2016] M. L. Woerner, M. Weber, B. S. Craiovan, H. R. Springorum, J. Grifka, and T. F. Renkawitz, "Radiographic Assessment of Femoral Stem Torsion in Total Hip Arthroplasty-A Comparison of a Caput-Collum-Diaphyseal Angle-Based Technique With the Budin View," *J. Arthroplasty*, vol. 31, no. 5, pp. 1117–1122, May 2016, doi: 10.1016/j.arth.2015.11.013.

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/XBAV3102

A Data Augmentation and Pre-processing Technique for Sign Language Fingerspelling Recognition

Frank Fowley[1,2], Ellen Rushe[1], and Anthony Ventresque[1]

[1]*School of Computer Science, University College Dublin & SFI Lero*
[2]*SFI Centre for Research Training in Digitally-Enhanced Reality (D-REAL)*

Abstract

The reliance of deep learning algorithms on large scale datasets is a significant challenge for sign language recognition (SLR). The shortage of data resources for training SLR models inevitably leads to poor generalisation, especially for low-resource languages. We propose novel data augmentation and pre-processing techniques based on synthetic data generation to overcome these generalisation difficulties. Using these methods, our models achieved a top-1 accuracy of 86.7% and a top-2 accuracy of 95.5% when evaluated against an unseen corpus of Irish Sign Language (ISL) fingerspelling video recordings. We believe that this constitutes a state-of-the-art performance baseline for an Irish Sign Language recognition model when tested on an unseen dataset.

Keywords: Sign Language recognition, data augmentation, pose estimation, convolutional neural networks

1 Introduction

Approximately 45,000 people use Irish Sign Language (ISL) including 5,000 Deaf people who use it as their first language[1]. Despite recent research, it remains extremely difficult to develop a practical automated solution for sign language recognition [Bragg et al., 2019, Rastgoo et al., 2021]. For a Deaf person, the potential benefits of such assistive technology include a heightened sense of autonomy and personal well-being along with greater confidence in education and employment scenarios [Dyzel et al., 2020].

Fingerspelling in ISL represents a specific challenge in sign language recognition. The manual elements in ISL include handshape (single-handed or two-handed), orientation, movement and location. Of the 81 fundamental handshapes in ISL, 26 are used for the fingerspelling alphabet. The use of fingerspelling in ISL is not restricted to the signing of abbreviations and proper nouns, such as place names, where there is a one-to-one correspondence between the hand signs and the English alphabetic letters. It is also used in compound words where only one part of a word is fingerspelled, as well as for emphasis and during meta-linguistic discourse [Leeson et al., 2020, Petitta et al., 2016]. The ISL lexicon is replete with signs which are formed based on a fingerspelled letter that is contained in the equivalent English spelling, such as the 'M' sign in "mother and 'F in "coffee" [Cormier et al., 2008, Matthews, 1996]. Lexical frequency studies of sign languages have shown rates of fingerspelling usage of up to 15% of all signs used during a conversational discourse [Fenlon et al., 2014, McKee and Kennedy, 2006, Nicodemus et al., 2017]. We use fingerspelling as our target use-case, not only because of its high usage level, but also because it represents a challenge for any spatial recognition model [Wheatland et al., 2016].

There are, however, several challenges that impede the relative progress of sign language recognition and translation systems when compared with those for spoken languages. One of the principal difficulties is the sparsity of datasets and corpora content of sufficient scale to be used as training input for deep learning models,

[1] `http:/www.irishdeafsociety.ie/irish-sign-language/`

 ISBN 978-0-9934207-7-1

in particular for low resource languages [Bragg et al., 2019]. Additional challenges to sign language recognition are subject variability, occlusion problems, different lighting and camera perspectives, as well as variations in image resolution and background [Rastgoo et al., 2021, Ming Jin Cheok and Jaward, 2019]. In an attempt to mitigate these issues, we propose a novel set of data augmentation techniques for synthetic data generation, whereby the individual skeletal bones are coloured differently, based on their depth in 3D space, in order to increase the classification performance of models. Additionally, we augment our training datasets with specific ranges of variations in camera perspective and hand size to improve the generalisation of the models. We also apply small random changes to the individual angular rotations of the bones in the hand pose to cater for variations in ISL dialects and fluency. In order to mitigate against the domain-specific influences of skin texture and image background, we feed images derived from the abstracted hand pose landmarks to the models during training. We augment MediaPipe keypoints as in [Fowley and Ventresque, 2021] to obtain the same input representation during testing as that used in training. In our experiments, we break down the effects of different features on classification performance and show that this method of guided data augmentation can help models obtain state-of-the-art performance using solely synthetic data.

The paper is structured as follows: Section 2 details some related work and our research methods are outlined in Section 3. The experimental results are reported in Section 5 and we conclude with a discussion of the results and future directions in Section 6.

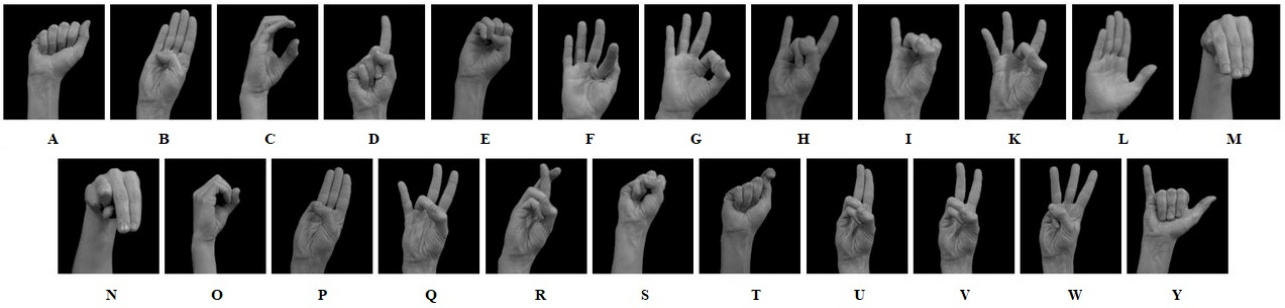


Figure 1: ISL FingerSpelling Alphabet - Static signs. (Source: Irish Deaf Society)

2 Related work

One of the primary challenges of sign language recognition is the lack of labelled data, especially for lower resourced languages such as ISL. In order to effectively capture the variability amongst letters being signed, and individuals signing, a large number of training examples are needed to train deep learning models in such a way that they generalise effectively to new signers. Moreover, signer-independent models, those where the signers present in the training set are different to those used for evaluation, generally perform significantly worse than those where the signers are common throughout the training and test set [Kim et al., 2016]. This problem is compounded once we consider that the number of individuals in sign language datasets is typically low. Bragg et al. [2019] highlights the lack of adequately annotated datasets along with the need to cater for variance in fluency and dialect, camera quality and occlusion issues.

To address these issues, the use of transfer learning for sign language recognition has been proposed [Halvardsson et al., 2021, Mocialov et al., 2020]. Other works have proposed retraining the higher layers of common pre-trained convolutional neural networks (CNN) using synthetic imagery for an object recognition task [Rajpura et al., 2018, Carneiro et al., 2021]. The use of transfer learning is sensible in these scenarios as it enables these models to use complex features even without a significant number of training examples. Notably, however, the test datasets used in these studies were derived from the same environments as the training datasets, which means that it is unknown how these models would perform in a completely novel environment.

Data augmentation is also a potential remedy to the data sparsity challenge as it allows us to create permutations of existing examples to supplement the original dataset. Data augmentation and pre-processing techniques

have been applied to SLR tasks in several works [Xie and Ma, 2019, Carneiro et al., 2021, Cao et al., 2022, Takayama and Takahashi, 2020]. Nunnari et al. [2021] extract 3D pose landmarks from sign language videos and train an end-to-end keypoint-to-text model, augmenting their training set by applying coordinate manipulations to simulate different camera perspectives, anatomical variations and other features. Tao et al. [2018] 2018, propose a multi-view data augmentation method whereby they apply transformations to a 3D point cloud, obtained using Microsoft Kinect devices, to mimic different camera positions and occlusion settings. Park and Sohn [2020] have implemented a similar approach to ours, using the skeletal images as training data for their SLR model. Their data augmentation techniques include the random removal of key-points, the variation of finger lengths and camera perspective transformations. They achieve significant success but only report on training accuracies.

A related topic is synthetic data generation where synthetic data is used in place of real training data. Fowley and Ventresque [2021] created a synthetic data generator for ISL that can generate ISL handshapes with several variations to help create generalisable models. Notably, the models in this work are trained exclusively on synthetic data.

3 Extended Synthetic Data Generation

We extend the synthetic data generation framework proposed by [Fowley and Ventresque, 2021] which produces training datasets for fingerspelling classification models. This system is built with Blender[2], an open source graphical rendering and animation package and a skeletally-rigged hand avatar mesh. The framework outputs both greyscale and colour images along with sequences. It also produces hand joint coordinates, hand bone rotations and palm alignments – along with their ground-truth labels. It can output allophonic variations –which represent different articulations of a given sign – programmatically. The system generates images in the formats of the leading pose estimation models to enable these systems to be used for inference. These "wireframe" samples are skeletal image representations of the poses. In this section, we give a description of extensions we have added to this framework along with their motivations. We then go on to explain the training and testing pipelines that our method uses.

3.1 Pre-Processing Techniques

Our training data generation can be configured to introduce specific variations into datasets. This controlled data augmentation includes allophonic variations, different camera perspectives, hand-to-palm size ratio, hand-width to hand-height ratio, minor fluctuations in individual bone rotations as well as random rotations of the hand pose about the wrist. These are introduced to the datasets to improve the accuracy and generalisation of models. We can also control the signing speed, frames-per-second of videos and image dimension. The range of variations is constrained to be within anatomical statistical limits, specifically we used those reported by [Park and Bae, 2020]. Models can be trained with the resulting skeletal images ("wireframes") constructed from the pose coordinates following this augmentation stage. The image pixels are manipulated purposely to augment their class discrimination. For instance, there are inherent similarities between some ISL fingerspelling letter signs, such as the 'F' and 'G' shapes, especially when viewed as a 2-dimensional image. To increase the discriminatory power of the models, we experimented with different manipulations of the wireframe graphical elements. These deliberate geometric and graphical modifications represent a "*feature injection*" which aims to add potentially discriminable artificial features into the training dataset samples.

3.2 Training and Test Pipelines

Figure 2 outlines the model training and evaluation processes for our proposed framework. The framework is built to use MediaPipe in the test pipeline to extract hand pose landmarks which are then converted into skeletal wireframe images. These images are then used to train models. The conversion from RGB pixel images to

[2] http://www.blender.org

 ISBN 978-0-9934207-7-1

wireframes alleviates some domain adaptation issues arising from real-world variations in image background, skin texture and colour. Furthermore pose estimation models, such as MediaPipe, are typically trained on much larger datasets than would be available for sign language recognition. MediaPipe is a good choice here due to its wide use and availability of Model Card [3] information. The pre-processing of the wireframe images, including depth-coded colouring of finger bones and the elimination or inclusion of particular joint vectors, aims to resolve the inter-class similarity between some signs.

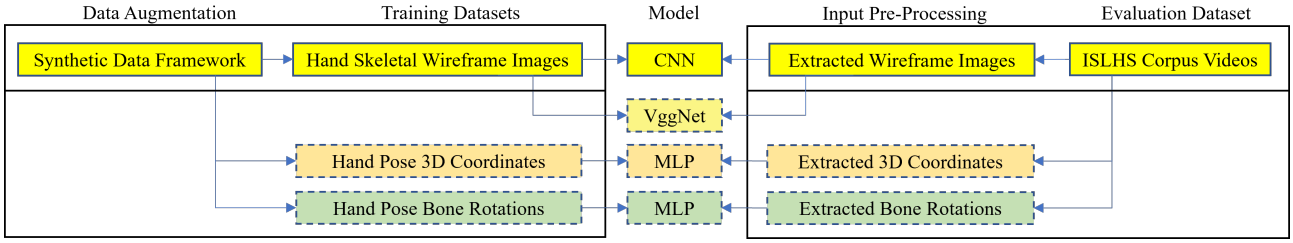


Figure 2: Training and Test Pipelines

4 Ablation Study

In order to understand the efficacy of each data augmentation method for classification of fingerspelling, we performed an ablation study. The following is a description of the augmentations applied along with the models and dataset that we used to evaluate these augmentations methods.

4.1 Augmentations

The data augmentations performed to produce the training datasets are outlined in table 1. The pose variations include slight changes in the finger bone rotations to represent variability in signer fluency and dexterity in the training data. Our proposed extended data augmentations were then applied to this data as illustrated in Table 3.

Allophonic Variants	Camera Perspectives	Pose/Anatomical Variations	Training Dataset Size
69	180	288	2,384,640

Table 1: Training Data Augmentation

4.2 Models

The models were trained with a learning rate of 0.001 and used an Adam [Kingma and Ba, 2015] optimiser. Batch-sizes of 16, 32 and 64 were compared in the hyper-parameter adjustments with 32 proving to be optimal. Our models were trained solely on synthetic training datasets. Similarly to [Cai et al., 2021], who use synthetic pose data to train 3D hand pose estimation models, we split our real-world ISL data between validation and test sets and use the validation dataset for model early-stopping and hyperparameter comparison.

We compared the performance of a custom CNN model with a VGGNet16 network pre-trained on ImageNet. We also compare this CNN to a multi-layer perceptron (MLP) which we trained to classify the fingerspelling signs based on non-image geometric pose data, namely, 3D landmark coordinates and angular rotations.

[3] `https://mediapipe.page.link/handmc`

4.3 Dataset

Our models were tested on the ISL-HS corpus of ISL fingerspelling signs [Oliveira et al., 2017] recorded by six native signers. The dataset is available as 468 short videos of static ISL alphabet letters. RGB colour frame samples from the test dataset videos are shown in Figure 3.

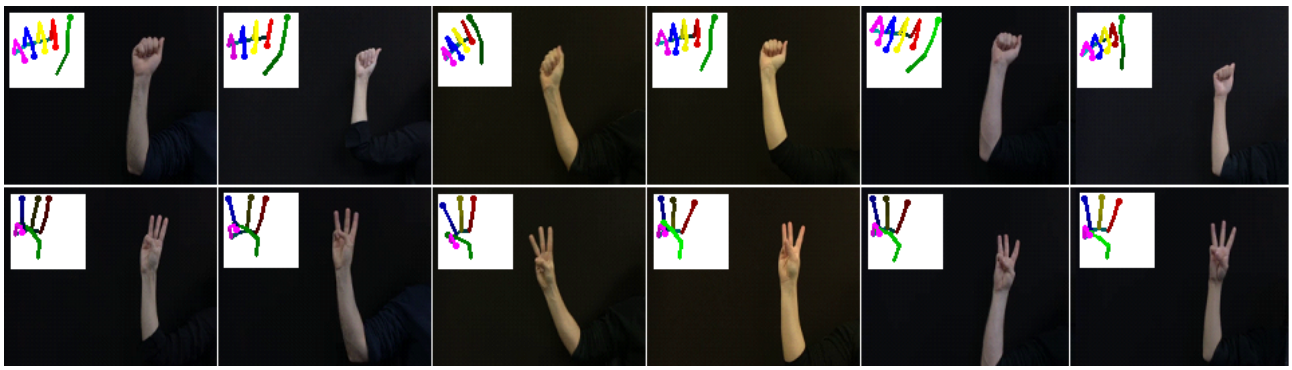


Figure 3: Test Dataset: Samples of letters 'A' (top) and 'W' (below) from the six signers in the test dataset. The corner insets contain the pre-processed image for each sign showing the effect of depth-colouring.

5 Results Evaluation

Our *feature injection* began with simple changes to finger bone colour and line widths. The reduction in the image dimensions produced a significant step increase in performance when predicting against the test dataset. As well as the removal of the palm bones, the introduction of depth-coded bone colours provided a large increase in the discrimination between some letters. The increased granularity of this eventually showed an improvement in the overlap between the letters "F" and "G" which has proved the most difficult to mitigate. The introduction of artificial bone lines between the finger metacarpophalangeal joints (MP) or "palm top bones" resulted in mitigating the confusion between "M" and "N" and other letters. The addition of the fingertip graphical element resolved the inter-class confusion between "A" and "T", and "E" and "S".

The model performance depends on the accuracy of MediaPipe, which had difficulty recognising samples with severely occluded key-points. The low recognition accuracy of the "R" letter is partially due to the miscalculation of the "index" and "middle" finger key-points by MediaPipe.

	Person1	Person2	Person3	Person4	Person5	Person6	**Overall**
Top-1 Accuracy %	86.4	87.8	86.1	80.2	89.7	89.8	**86.7**
Top-2 Accuracy %	97.1	94.8	91.5	94.3	98.1	97	**95.5**
Top-3 Accuracy %	99	96.7	94.8	98.2	99.6	98.2	**97.8**

Table 2: The table shows the model performance when tested on data from each of the 6 signers individually. The Overall figure is where the model is tested on a dataset that includes all 6 signers. The Top-1 Accuracy denotes that the highest model prediction was correct. The Top-2 Accuracy is where the correct test sample label was in the top 2 highest predictions by the model.

5.1 Feature Injection Comparison

Table 3 shows the impact of adding specific geometric and graphical effects or "features" into the training dataset samples. It also includes examples of the pre-processing that was applied. These changes were deliber-

ately introduced to try to resolve some of the inter-letter confusion seen in the model performance on the test set. Obviously, there is a need for the model to discriminate between letters that appear similar from a front elevation and only discriminable by adding depth information or a side elevation view. The main example of this is the similarity between the signs for “F” and “G”. However, as features were added, their introduction gave rise to new inter-letter confusion. The addition of fingertips improved the recognition of letters where the fingertip is naturally occluded by the pose but reduced the accuracy of the letter “T”. This was improved by making the fingertip the same colour as its corresponding finger, avoiding the tip from over-writing the pixels of the top finger bone. The letter recognition performance of the best performing model, used to recognise all 23 static ISL fingerspelling letters, is shown in Table 2.

Training Sample	Pre-processing Applied	Model	Top-1 Accuracy	Worst Letter Confusion (<60% recognition)
	“Wireframe" images: Different colours per finger.	VGGNet	71.3%	‘A’,‘F’,‘R’,‘S’,‘T’
	56 x 56 Resized Wireframe images	CNN	79.8%	‘F’,‘R’,‘S’,‘T’
	Bones with depth-coded colours; palm bones removed; top palm vectors added.	CNN	82.9%	‘F’,‘I’,‘R’,‘S’
	Fingertips added.	CNN	83.3%	‘F’,‘H’,‘I’,‘R’,
	More fine-tuned depth-coded colours.	CNN	86.7%	‘N’,‘O’,‘R’
Non-Image	21 x 3D Pose Coordinates.	MLP	85.1%	‘F’,‘N’,‘R’,‘S’
Non-Image	25 x Pose Angles.	MLP	66.5%	‘C’,‘F’,‘M’,‘N’,‘R’,‘S’

Table 3: "Feature Injection" - Performance Comparison. The table highlights the experimentation steps that yielded the most significant improvements in overall recognition accuracy and reduced inter-class confusion.

6 Conclusion

The above results demonstrate that neural networks, trained solely on synthetic pose images or pose geometric data, can successfully classify non-synthetic fingerspelling signs when deployed alongside a pose estimation model in the recognition pipeline. This suggests that the inherent domain adaptation issues, normally seen with synthetic data [Cai et al., 2021] can be mitigated given sufficient levels of variation in the generated data. In terms of a static computer vision based model for manual sign recognition, the maximum performance is bound by a small number of pairs of overlapping classes. Such inter-class similarity exists in other sign languages [Tao et al., 2018]. However, a practical sign language application must be capable of recognising all sign shapes. Therefore, any successful SLR solution for these may need to have a different architecture depending on how prevalent this issue is within a given sign language.

Our future research will focus on remedying this class confusion problem, through the use of ensemble models or the introduction of more fine-tuned feature-injection in our pre-processing. Furthermore, given that the augmentations performed were guided by the ISL testing dataset, some bias towards this data was likely introduced. We intend to evaluate whether the best performing augmentations also enhance the class

 ISBN 978-0-9934207-7-1

discriminability in other datasets. We plan to extend our datasets and models to recognise the full set of ISL manual sign shapes as well as creating temporal models to cater for signing sequences.

Acknowledgement

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (D-REAL) under Grant No. 18/CRT/6224, and supported, in part, by Science Foundation Ireland grant 13/RC/2094. This work has also been conducted within the SignON project, a European Union's Horizon 2020 research and innovation programme under grant agreement No 101017255.

References

[Bragg et al., 2019] Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., and Ringel Morris, M. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.

[Cai et al., 2021] Cai, Y., Ge, L., Cai, J., Thalmann, N. M., and Yuan, J. (2021). 3D hand pose estimation using synthetic data and weakly labeled RGB images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3739–3753.

[Cao et al., 2022] Cao, Y., Li, W., Li, X., Chen, M., Chen, G., Hu, L., Li, Z., and Kaii, H. (2022). Explore more guidance: A task-aware instruction network for sign language translation enhanced with data augmentation. *arXiv 2204.05953*.

[Carneiro et al., 2021] Carneiro, A. C., Silva, L. B., and Salvadeo, D. P. (2021). Efficient sign language recognition system and dataset creation method based on deep learning and image processing. In *Thirteenth International Conference on Digital Image Processing (ICDIP 2021)*, volume 11878, pages 11–19. SPIE.

[Cormier et al., 2008] Cormier, K., Schembri, A., and Tyrone, M. (2008). One hand or two? Nativisation of fingerspelling in ASL and BANZSL. *Sign Language and Linguistics*, 11:3–44.

[Dyzel et al., 2020] Dyzel, V., Oosterom-Calo, R., Worm, M., and Sterkenburg, P. S. (2020). Assistive technology to promote communication and social interaction for people with deafblindness: A systematic review. *Frontiers in Education*, 5:164.

[Fenlon et al., 2014] Fenlon, J., Schembri, A., Rentelis, R., Vinson, D., and Cormier, K. (2014). Using conversational data to determine lexical frequency in British Sign Language: The influence of text type. *Lingua*, 143:187–202.

[Fowley and Ventresque, 2021] Fowley, F. and Ventresque, A. (2021). Sign language fingerspelling recognition using synthetic data. *Proc. 29th Irish Conferenfe on Artificial Intelligence and Cognitive Science*, pages 1–6.

[Halvardsson et al., 2021] Halvardsson, G., Peterson, J., Soto-Valero, C., and Baudry, B. (2021). Interpretation of Swedish Sign Language using Convolutional Neural Networks and transfer learning. *SN Comput. Sci.*, 2:207.

[Kim et al., 2016] Kim, T., Wang, W., Tang, H., and Livescu, K. (2016). Signer-independent fingerspelling recognition with deep neural network adaptation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6160–6164. IEEE.

[Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*.

[Leeson et al., 2020] Leeson, L., Sheridan, S., Cannon, K., Murphy, T., Newman, H., and Veldheer, H. (2020). Hands in motion: Learning to fingerspell in irish sign language. *TEANGA the Journal of the Irish Association for Applied Linguistics*, 11:120–141.

[Matthews, 1996] Matthews, P. A. (1996). Extending the lexicon of Irish Sign Language (ISL). *TEANGA: The Irish Yearbook of Applied Linguistics*, 16:135–67.

[McKee and Kennedy, 2006] McKee, D. and Kennedy, G. (2006). The distribution of signs in New Zealand Sign Language. *Sign Language Studies*, 6(4):372–390.

[Ming Jin Cheok and Jaward, 2019] Ming Jin Cheok, Z. O. and Jaward, M. (2019). A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 10:131–153.

[Mocialov et al., 2020] Mocialov, B., Turner, G., and Hastie, H. (2020). Transfer learning for British Sign Language modelling. *arXiv preprint arXiv:2006.02144*.

[Nicodemus et al., 2017] Nicodemus, B., Swabey, L., Leeson, L., Napier, J., Petitta, G., and Taylor, M. M. (2017). A cross-linguistic analysis of fingerspelling production by sign language interpreters. *Sign Language Studies*, 17(2):143–171.

[Oliveira et al., 2017] Oliveira, M., Chatbri, H., Little, S., Ferstl, Y., Oconnor, N. E., and Sutherland, A. (2017). Irish Sign Language recognition using principal component analysis and Convolutional Neural Networks. *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*.

[Park and Bae, 2020] Park, Y. and Bae, J. (2020). A three-dimensional finger motion measurement system of a thumb and an index finger without a calibration process. *Sensors*, 20(3):756.

[Petitta et al., 2016] Petitta, G., Halley, M., and Nicodemus, B. (2016). Managing metalinguistic references in bimodal interpreted discourse: an analysis of an american sign language–english interpretation. *Rivista da Psicolinguistica Applicata*, 16:53–69.

[Rajpura et al., 2018] Rajpura, P., Aggarwal, A., Goyal, M., Gupta, S., Talukdar, J., Bojinov, H., and Hegde, R. (2018). Transfer learning by finetuning pretrained cnns entirely with synthetic images. *Communications in Computer and Information Science Computer Vision, Pattern Recognition, Image Processing, and Graphics*, page 517–528.

[Rastgoo et al., 2021] Rastgoo, R., Kiani, K., and Escalera, S. (2021). Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794.

[Takayama and Takahashi, 2020] Takayama, N. and Takahashi, H. (2020). Data augmentation using feature interpolation of individual words for compound word recognition of sign language. In *2020 International Conference on Cyberworlds (CW)*, pages 137–140. IEEE.

[Tao et al., 2018] Tao, W., Leu, M. C., and Yin, Z. (2018). American Sign Language alphabet recognition using Convolutional Neural Networks with multiview augmentation and inference fusion. *Engineering Applications of Artificial Intelligence*, 76:202–213.

[Wheatland et al., 2016] Wheatland, N., Abdullah, A., Neff, M., Jörg, S., and Zordan, V. (2016). Analysis in support of realistic timing in animated fingerspelling. In *2016 IEEE Virtual Reality (VR)*, pages 309–310.

[Xie and Ma, 2019] Xie, M. and Ma, X. (2019). End-to-end residual neural network with data augmentation for sign language recognition. In *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 1, pages 1629–1633. IEEE.

A machine vision system for avian song classification with CNN's

Gabriel R. Palma[1], Ana C. M. M. Aquino[2], Patricia F. Monticelli[2], Luciano M. Verdade[2], Charles Markham[1], and Rafael A. Moral[1]

[1]*Maynooth University, Maynooth, Ireland*
[2]*University of São Paulo, Brazil*

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/MHZN4111

Abstract

Soundscape ecologists aim to study the acoustic characteristics of an area that reflects natural processes [Schafer, 1977]. These sounds can be interpreted as biological (biophony), geophysical (geophony), and human-produced (anthrophony) [Pijanowski et al., 2011]. A common task is to use sounds to identify species based on the frequency content of a given signal. This signal can be further converted into spectrograms enabling other types of analysis to automate the identification of species. Based on the promising results of deep learning methods, such as Convolution Neural Networks (CNNs) in image classification, here we propose the use of a pre-trained VGG16 CNN architecture to identify two nocturnal avian species, namely *Antrostomus rufus* and *Megascops choliba*, commonly encountered in Brazilian forests. Monitoring the abundance of these species is important to ecologists to develop conservation programmes, detect environmental disturbances and assess the impact of human action. Specialists recorded sounds in 16 bit wave files at a sampling rate of 44Hz and classified the presence of these species. With the classified wave files, we created additional classes to visualise the performance of the VGG16 CNN architecture for detecting both species. We end up with six categories containing 60 seconds of audio of species vocalisation combinations and background only sounds. We produced spectrograms using the information from each RGB channel, only one channel (grey-scale), and applied the histogram equalisation technique to the grey-scale images. A comparison of the system performance using histogram equalised images and unmodified images was made. Histogram equalisation improves the contrast, and so the visibility to the human observer. Investigating the effect of histogram equalisation on the performance of the CNN was a feature of this study. Moreover, to show the practical application of our work, we created 51 minutes of audio, which contains more noise than the presence of both species (a scenario commonly encountered in field surveys). Our results showed that the trained VGG16 CNN produced, after 8000 epochs, a training accuracy of 100% for the three approaches. The test accuracy was 80.64%, 75.26%, and 67.74% for the RGB, grey-scaled, and histogram equalised approaches. The method's accuracy on the synthetic audio file of 51 minutes was 92.15%. This accuracy level reveals the potential of CNN architectures in automating species detection and identification by sound using passive monitoring. Our results suggest that using coloured images to represent the spectrogram better generalises the classification than grey-scale and histogram equalised images. This study might develop future avian monitoring programmes based on passive sound recording, which significantly enhances sampling size without increasing cost.

Keywords: Soundscape, Spectrogram, Deep Learning, Machine Vision

1 Introduction

Soundscape ecology is a multidisciplinary research area aiming to understand how organisms interact with their environments by relating acoustic characteristics of a site to biological, geophysical, and human sounds [Schafer, 1977, Pijanowski et al., 2011, Gasc et al., 2017]. This multidisciplinary discipline involves studies of bioacoustics, landscape ecology, community ecology and engineering to answer research questions. The combination of those disciplines benefits the zoological and ethological communities by providing a theoretical framework grounded in a broad ecological context, a wealth of long-term soundscape collections from around the

world, methods optimising acoustic monitoring, and the analysis of acoustic big-data [Gasc et al., 2017]. The collection of acoustic data involves a high volume, high acquisition rate and a variety of signals [Gasc et al., 2017, Pijanowski et al., 2011], showing the necessity of new methods to analyse acoustic big-data [Emmanuel and Stanier, 2016, Brunsdon and Comber, 2020]. Consequently, the zoological and ethological communities can also benefit from Deep Learning (DL) techniques.

Creating new methods that automate identifying avian species based on spectrograms will provide several gains for the zoological and ethological communities. Among the possible ways to analyse signals from natural sounds, the Fourier transformation coupled with the Gabor transformation offers many benefits for a visual interpretation of avian songs. However, understanding spectrograms is a challenge for biologists requiring effort to train researchers to detect species. This necessity of rigorous training to achieve high accuracy in classifying avians led to an opportunity for the application of deep learning methods [Zhang et al., 2019, Ruff et al., 2020, Ruff et al., 2021, Hidayat et al., 2021, Permana et al., 2021, Bravo Sanchez et al., 2021]. Brazilian biomes represent excellent opportunities to study the application of these methods given the vast diversity of avian species. Therefore, this paper has two main objectives: i) to implement an algorithm capable of detecting two Brazilian species of nocturnal avians, and ii) to evaluate the impact of image pre-processing on species classification.

Figure 1: *Antrostomus rufus* and *Megascops choliba* species. These pictures were provided respectively by Rafael Cerqueira and Rafael Martos Martins.

2 State of the Art

Deep Learning has been applied in soundscape ecology, zoology and ethology research projects were primarily interested in species identification [Selin et al., 2006, Chou et al., 2007, Sprengel et al., 2016, Lasseck, 2018a, Christin et al., 2018, Sankupellay and Konovalov, 2018, Lasseck, 2018b, Zhang et al., 2019, Koh et al., 2019, Ruff et al., 2020, LeBien et al., 2020, Ruff et al., 2021, Huang and Basanta, 2021, Campos Paula et al., 2022]. Widely used algorithms in this context are Deep Neural Networks and Convolutional Neural Networks (CNNs) [Ruff et al., 2020, Christin et al., 2018, Zhang et al., 2019, Ruff et al., 2021, Hidayat et al., 2021, Kahl et al., 2021, Permana et al., 2021, Disabato et al., 2021]. Over the last year, studies showed good Deep Learning performances on avian species identification based on their sounds using mainly the Deep Neural Networks and CNN architectures [Kahl et al., 2021, Ruff et al., 2021, Hidayat et al., 2021, Permana et al., 2021]. These researchers focussed their effort on classifying species and applying different image pre-processing techniques. An example of this approach was evaluating the effect of grey-scale and jet colour map Spectrogram on the accuracy of avian species classification [Incze et al., 2018].

3 Methods

We experimented with 20 autonomous acoustic recording units in Angatuba (São Paulo - Brazil) to monitor two nocturnal species. Specialists recorded sounds in 16-bit wave files at a sampling rate of 44Hz. These recordings followed a discontinuous protocol (1-minute recordings in 3-minute intervals) 24 hours a day, for 15 days during 6 months. Specialists classified the presence of *Antrostomus rufus* and *Megascops choliba* using the Raven Pro 1.4 software (Bioacoustic Research Program, 2011). We selected nocturnal avians because, at night, other sounds such as biophony, and anthrophony are not so evident [Gasc et al., 2017]. Also, these species are common in Brazil and produce sounds with almost no variation and constant amplitude. With the classified wave files, we created additional classes to visualise the performance of the VGG16 CNN method (see Figure 3

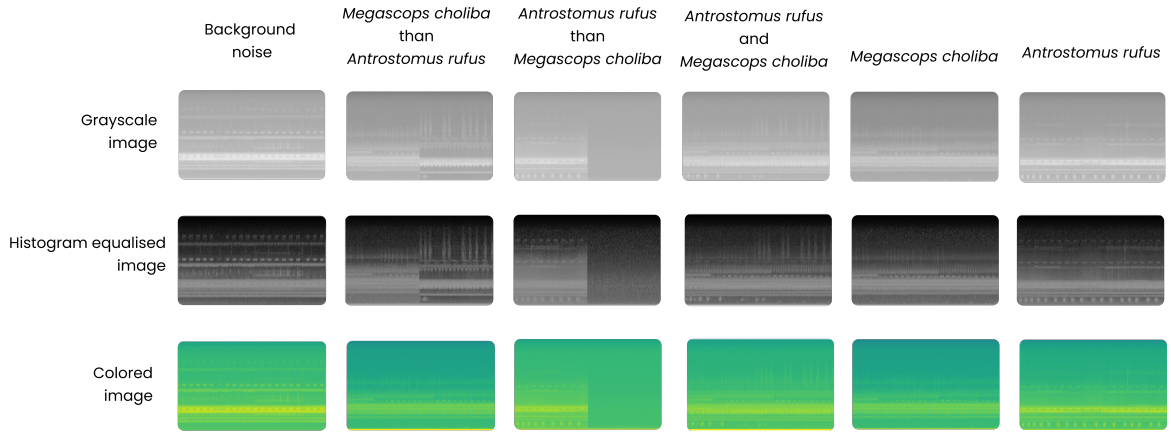


Figure 2: spectrograms of *Antrostomus rufus*'s vocalisation, *Megascops choliba*'s vocalisation, both species vocalising together, *Antrostomus rufus* than *Megascops choliba* vacalization, *Megascops choliba* than *Antrostomus rufus* vocalisation. The diagram also shows the differences between grey-scale, histogram equalised and coloured images.

for details) for detecting both species. We created 6 classes containing 60-second audio clips of (1) *Antrostomus rufus*'s vocalisation, (2) *Megascops choliba*'s vocalisation, (3) both species vocalising together, (4) *Antrostomus rufus* then *Megascops choliba* vocalisation, (5) *Megascops choliba* then *Antrostomus rufus* vocalisation, and (6) background only sounds.

Then, based on the Gabor transformation of the audio frequency data, we produced spectrograms using the information from each RGB channel, only one channel (grey-scale), and applied the histogram equalisation technique on the grey-scale images to better visualise the contrasts of the spectrograms. Histogram equalisation was used to maintain the spatial and dynamic range properties of the image. Other enhancements, such as Gaussian filtering and thresholding were rejected in this study as they reduce the information of the image. Thus, we compiled a data set containing 505 images of size 150×150 (Figure 2). The feature extraction was based on convolution and pooling operations, and the parameters were optimised using the ImageNet data set ([Chollet, 2018]; [Wani, 2020]; [Simonyan and Zisserman, 2014]). Then, using the feature maps with size 4×4 provided by the pre-trained model, we trained two additional densely connected layers with a dropout rate of 0.5 to identify the nocturnal avian species (see Figure 3 for an illustration of these operations). We used 80% of the data (412 images) for training and 20% (93 images) to test the proposed CNN architecture's performance. We used 8000 epochs to compute the method's accuracy, precision, and recall.

Finally, to show the practical application of our work, we created two audio files of 51 minutes, which contains more noise than the presence of both species. This is a common scenario encountered in field surveys, and it consumes a representative amount of time from researchers that analyse such data. The first audio file contains both species singing together more frequently, and in the second one, both species sing alone more regularly. To construct them, we selected spectrograms from the test data and reordered them to create the 51-minute audio file.

4 Results

Figure 4 shows the accuracy, precision and recall metrics obtained from the CNN VGG16 architecture using the selected pre-processing spectrograms. The trained VGG16 CNN produced, after 8000 epochs, a training accuracy of 100% for the three approaches. The test accuracy was 80.64%, 75.26%, and 67.74% for the RGB, grey-scaled, and histogram equalised approaches. It indicates that the CNN trained with coloured images provided a better generalisation than the one based on histogram equalised or grey-scale images. Considering

 ISBN 978-0-9934207-7-1

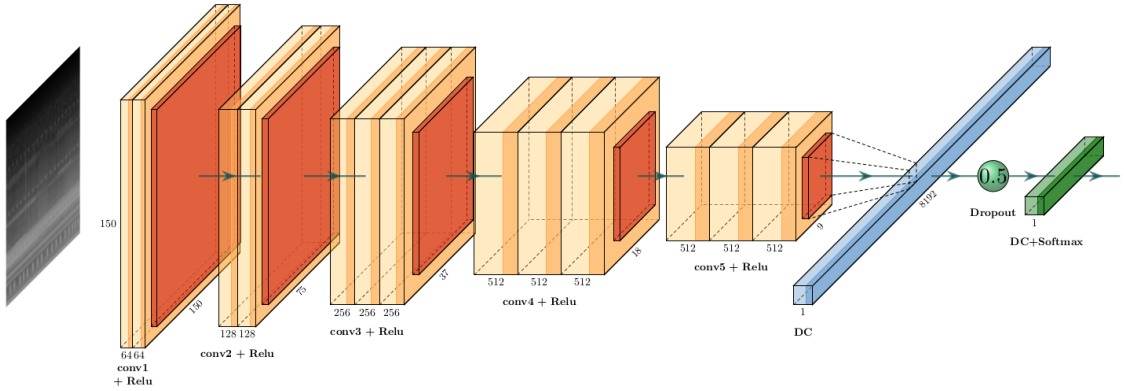


Figure 3: Scheme representing the pretrained VGG16 CNN architecture and the additional layers used to train the model. The convolution, max-pooling and dropout operations are represented in orange, red and green, respectively. Finally, densely connected (DC) layers are added with two activation functions: ReLU before the dropout and softmax afterwards.

the experimental optimised parameters of the CNN introduced with coloured pictures, the accuracy on the synthetic audio file of 51 minutes presented in Figure 5 was 92.15% and in Figure 6 was 76.47%. The sequence created for Figure 5 investigated the system's performance when the species are vocalising at the same time. Figure 6 focusses on the performance when the species sing at different times. It indicates that the system with this approach would help the researcher identify the presence of these two species on a large dataset containing more noise than the individuals themselves. In Table 1, we present the confusion matrix provided by the VGG16 architecture trained utilising the information from each RGB channel. Finally, this pilot study shows promising results indicating an optimistic scenario for improved DL studies of this kind to automate the detection of nocturnal avian species in Brazil.

Predicted	Observed					
	Antrostomus rufus before *Megascops choliba*	*Antrostomus rufus*	Both species	*Megascops choliba* before *Antrostomus rufus*	*Megascops choliba*	Noise
Antrostomus rufus before *Antrostomus rufus*	19	0	0	1	0	0
Antrostomus rufus	0	11	5	0	1	0
Both species	0	3	13	0	3	0
Megascops choliba before *Antrostomus rufus*	1	0	0	17	0	0
Megascops choliba	0	0	3	0	4	0
Noise	0	0	1	0	0	11

Table 1: A confusion matrix produced based on the predictions of the CNN VGG16 architecture trained utilising the information from each RGB channel.

 ISBN 978-0-9934207-7-1

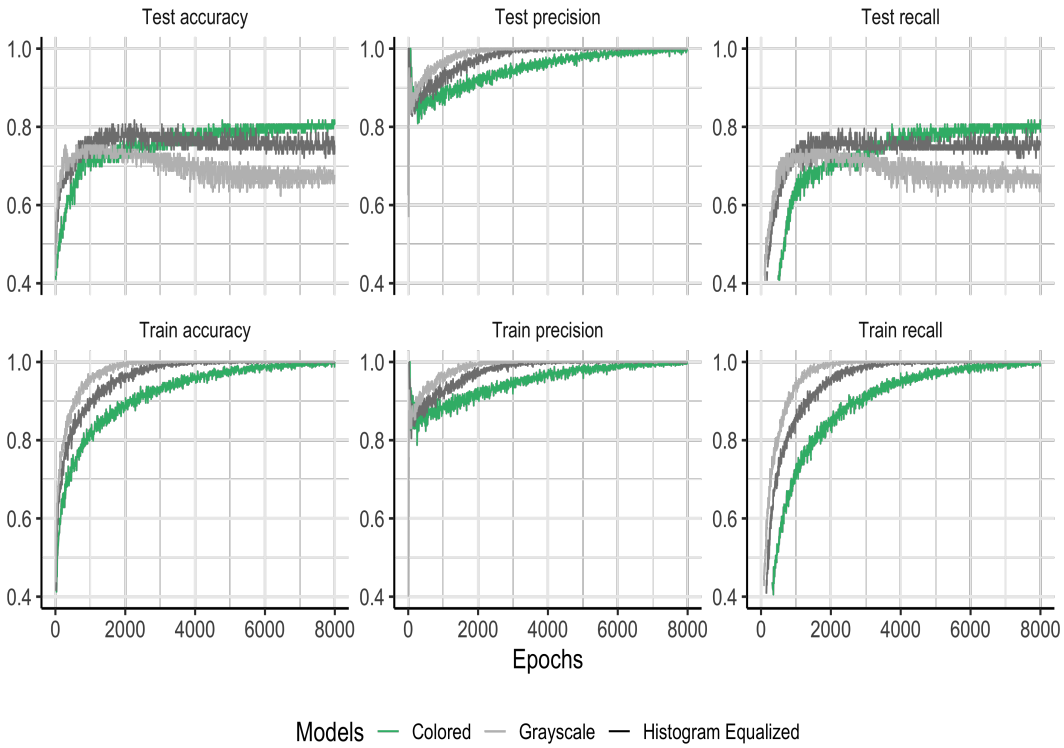


Figure 4: Accuracy, precision and recall metrics of the CNN VGG16 architecture for the train and test data using 8000 epochs. Green, grey and black lines result from architectures trained utilising the information from each RGB channel, only one channel (grey-scale), and applied the histogram equalisation technique on the grey-scale images.

5 Conclusion

The obtained validation accuracy shows the feasibility of the pre-trained VGG16 architecture in detecting the studied avian species. Also, given the number of classes presented in this paper, including the presence of both species and noise, our results show a good perspective for further investigation of soundscape studies, including other species. Our results suggest that using coloured images to represent the spectrogram generalises the classification better than grey-scale and histogram equalised images. This study will serve as a basis for developing a future animal monitoring program based on passive recording sound, which significantly enhances sampling efforts without increasing cost.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6049. We would like to thank three anonymous referees for their valuable comments which helped improve the manuscript. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Science Foundation Ireland.

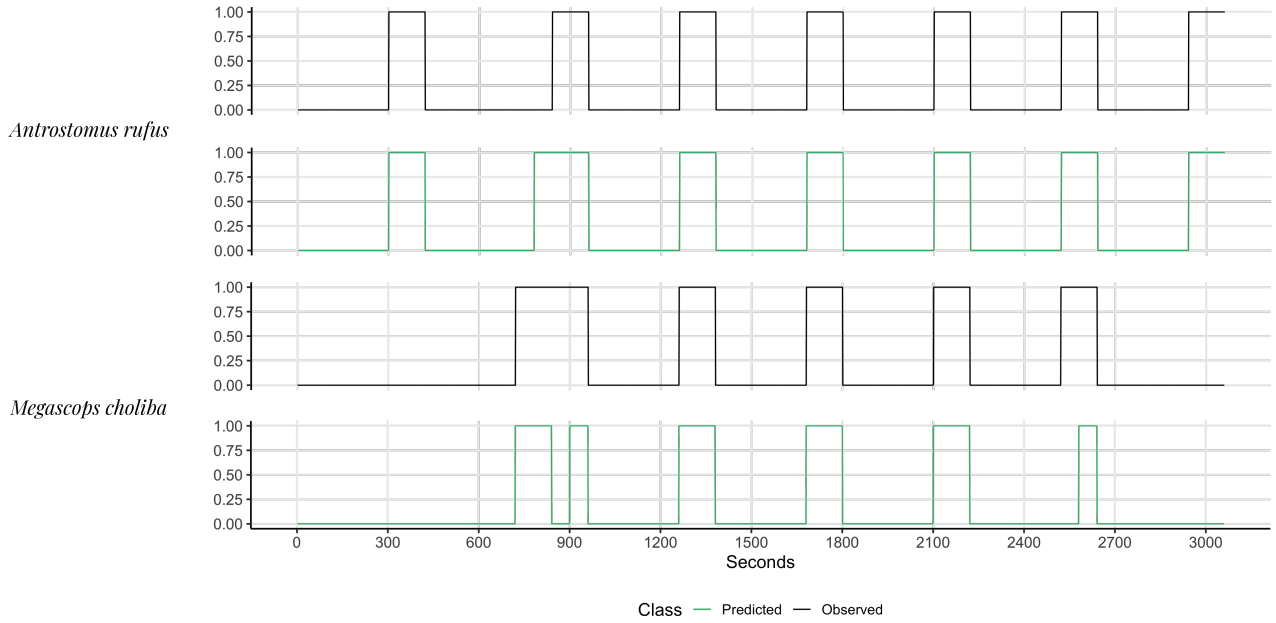


Figure 5: Detection of the studied species based on the CNN VGG16 architecture using 51 minutes of audio showing the practical application of our results. The green line represents the detection of the CNN and the black line is the real class detected by the specialists. For this dataset, we obtain an accuracy of 92.15%.

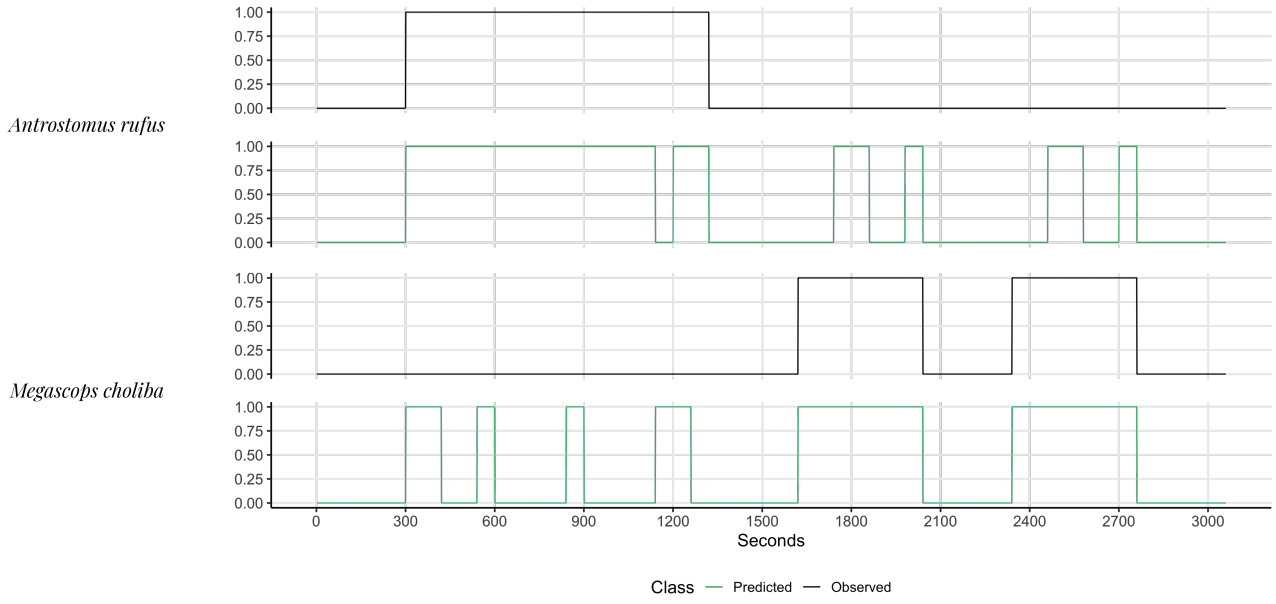


Figure 6: Detection of the studied species based on the CNN VGG16 architecture using 51 minutes of audio showing the practical application of our results. The green line represents the detection of the CNN and the black line is the real class detected by the specialists. For this dataset, we obtain an accuracy of 76.47%.

References

[Bravo Sanchez et al., 2021] Bravo Sanchez, F. J., Hossain, M. R., English, N. B., and Moore, S. T. (2021). Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture. *Scientific Reports*, 11(1):1–12.

[Brunsdon and Comber, 2020] Brunsdon, C. and Comber, A. (2020). Big issues for big data: challenges for critical spatial data analytics. *arXiv preprint arXiv:2007.11281*.

[Campos Paula et al., 2022] Campos Paula, B., Luchesi, L., and Monticelli, P. (2022). Railway noise and long-distance calls of free-living maned wolves in ecological station of itirapina, são paulo, brazil. *The Journal of the Acoustical Society of America*, 151(4):A147–A147.

[Chollet, 2018] Chollet, F. (2018). *Deep Learning with Python*.

[Chou et al., 2007] Chou, C.-H., Lee, C.-H., and Ni, H.-W. (2007). Bird species recognition by comparing the hmms of the syllables. pages 143–143.

[Christin et al., 2018] Christin, S., Hervet, E., and Lecomte, N. (2018). Applications for deep learning in ecology.

[Disabato et al., 2021] Disabato, S., Canonaco, G., Flikkema, P. G., Roveri, M., and Alippi, C. (2021). Birdsong detection at the edge with deep learning. In *2021 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 9–16. IEEE.

[Emmanuel and Stanier, 2016] Emmanuel, I. and Stanier, C. (2016). Defining big data. In *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*, pages 1–6.

[Gasc et al., 2017] Gasc, A., Francomano, D., Dunning, J. B., and Pijanowski, B. C. (2017). Future directions for soundscape ecology: The importance of ornithological contributions. *The Auk: Ornithological Advances*, 134(1):215–228.

[Hidayat et al., 2021] Hidayat, A. A., Cenggoro, T. W., and Pardamean, B. (2021). Convolutional neural networks for scops owl sound classification. *Procedia Computer Science*, 179:81–87.

[Huang and Basanta, 2021] Huang, Y.-P. and Basanta, H. (2021). Recognition of endemic bird species using deep learning models. *IEEE Access*, 9:102975–102984.

[Incze et al., 2018] Incze, A., Jancso, H.-B., Szilagyi, Z., Farkas, A., and Sulyok, C. (2018). Bird sound recognition using a convolutional neural network. pages 000295–000300.

[Kahl et al., 2021] Kahl, S., Wood, C. M., Eibl, M., and Klinck, H. (2021). Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236.

[Koh et al., 2019] Koh, C.-Y., Chang, J.-Y., Tai, C.-L., Huang, D.-Y., Hsieh, H.-H., and Liu, Y.-W. (2019). Bird sound classification using convolutional neural networks. In *CLEF (Working Notes)*.

[Lasseck, 2018a] Lasseck, M. (2018a). Acoustic bird detection with deep convolutional neural networks. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, pages 143–147.

[Lasseck, 2018b] Lasseck, M. (2018b). Audio-based bird species identification with deep convolutional neural networks. *CLEF (working notes)*, 2125.

[LeBien et al., 2020] LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J. P., Dodhia, R., Ferres, J. L., and Aide, T. M. (2020). A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics*, 59:101113.

[Permana et al., 2021] Permana, S. D. H., Saputra, G., Arifitama, B., Caesarendra, W., Rahim, R., et al. (2021). Classification of bird sounds as an early warning method of forest fires using convolutional neural network (cnn) algorithm. *Journal of King Saud University-Computer and Information Sciences*.

[Pijanowski et al., 2011] Pijanowski, B., Villanueva-Rivera, L., Dumyahn, S., Farina, A., Krause, B., Napoletano, B., Gage, S., and Pieretti, N. (2011). Soundscape ecology: The science of sound in the landscape. *BioScience*, 61.

[Ruff et al., 2020] Ruff, Z. J., Lesmeister, D. B., Appel, C. L., and Sullivan, C. M. (2020). A convolutional neural network and r-shiny app for automated identification and classification of animal sounds. *bioRxiv*.

[Ruff et al., 2021] Ruff, Z. J., Lesmeister, D. B., Appel, C. L., and Sullivan, C. M. (2021). Workflow and convolutional neural network for automated identification of animal sounds. *Ecological Indicators*, 124:107419.

[Sankupellay and Konovalov, 2018] Sankupellay, M. and Konovalov, D. (2018). Bird call recognition using deep convolutional neural network, resnet-50. In *Proceedings of ACOUSTICS*, volume 7.

[Schafer, 1977] Schafer, R. M. (1977). *The Tuning of the World*.

[Selin et al., 2006] Selin, A., Turunen, J., and Tanttu, J. T. (2006). Wavelets in recognition of bird sounds. *EURASIP Journal on Advances in Signal Processing*, 2007:1–9.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*.

[Sprengel et al., 2016] Sprengel, E., Jaggi, M., Kilcher, Y., and Hofmann, T. (2016). Audio based bird species identification using deep learning techniques. Technical report.

[Wani, 2020] Wani (2020). *Advances in Deep learning*.

[Zhang et al., 2019] Zhang, P., Chen, H., Bai, H., and Yuan, Q. (2019). Deep scattering spectra with deep neural networks for acoustic scene classification tasks. *Chinese Journal of Electronics*, 28(6):1177–1183.

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/KUPA8487

High-Fidelity Face Swapping with Style Blending

Xinyu Yang, Zhijin Guo, Chengxi Zeng, Mowen Xue, and Zijian Shi

Department of Computer Science, Univeristy of Bristol

Figure 1: **Face Swap Visualisation.** Replace face in target image with the source face. Result of ours appears in the right.

Abstract

Face swapping is gaining significant traction, boosted by the plethora of human face synthesis with the deep learning methods. Recent works based on Generative Adversarial Nets (GAN) for face swapping often suffer from blending inconsistency, distortions and artefacts, as well as instability in training. In this work, we propose a novel end-to-end framework for high-fidelity face swapping, leveraging the high photorealistic face generation techniques from StyleGAN. Firstly, we invert the facial images into the style latent space by purposing a novel facial attributes encoder that is capable of extracting face essentials from the face image and projecting them to the style code in the latent space. We show that such inverted style code encapsulates facial attributes that are indispensable for face swapping task. Secondly, a carefully designed style blending module (SBM) is introduced for transferring the identity from a source image to the target by the multi-head attention (MHA) mechanism. We propose relevant constraints for guiding the learning of the SBM, leading to the effective blending of the Face ID from the source face to the target image. Finally, the blended style code can be translated back to the image space via the style decoder, benefiting from the training stability and the high quality of the generative capability of the style-based decoder. Extensive experiments demonstrate the superior quality of the face synthesis results (illustrated in Figure 1) of our face-swapping system compared with other state-of-the-art methods.

Keywords: Face Synthesis, Face Swap, Multi-head Attention, GAN

1 Introduction

Face swapping, an emerging field as well as one of the most controversial topic of face synthesis, has broad positive application prospects in entertainment, film and television making, human-computer interaction, privacy protection *etc.*, despite its notorious application like generating synthetic fake media and news through deepfake [Bitouk et al., 2008]. Nevertheless, in addition to developing accurate and robust deepfakes detection models, it is worthwhile to explore and understand the scheme of advanced methods for face swapping.

The task of face swapping is to transfer the identity of a source face image to the target image while the illumination, head posture, facial expression, background, and other attribute information of the target image hold intact. This task is challenging due to the unavailability of the ground truth image as well as the lack of evaluation method for the synthesised images.

The early works of the face-swapping method train a model that requires many images of source identity A and images of target identity B [Bitouk et al., 2008, Korshunova et al., 2017]. The learnt model can only be

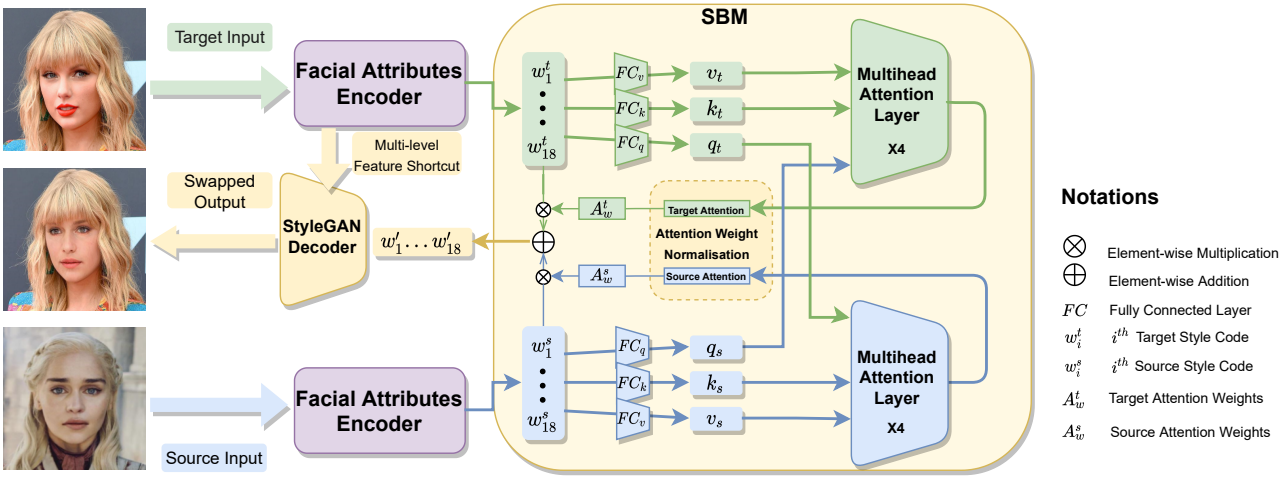


Figure 2: **Face Swapping Generator and Style Blending Module Overview.** We utilise an encoding-blending-decoding paradigm for learning where the encoder produces the style code $w \in \mathbb{R}^{18\times512}$ in the latent space $\mathcal{W}$ for both the target input and the source input. The Style Blending Module (SBM) fuses the target style code w^t and source style code w^s by attention weights A^t_w and A^s_w which are generated by the multi-head attention (MHA) layers. We apply five constraints on the training of the model so that the attention weight can be adaptively adjusted to the desirable swapped face. To retain the details from the target input, we reuse the low-level features in the decoder.

used for swapping the face of identity A and identity B. However, such a model demonstrated low transferability to unseen identities. Benefiting from the great generative ability of GAN, the recent face-swapping methods achieve high transferability on unseen identities, generating images with a good mixture of the source and target identities and facial attribute information [Zhu et al., 2021, Li et al., 2019, Xu et al., 2021]. Unfortunately, these methods do not perform well in balancing the identity similarity and the retaining of the low-level details, such as illumination and background, thus leading to the frequent artefacts and the low visual perception.

The development of image generation methods [Karras et al., 2019, Karras et al., 2017, Choi et al., 2018, Zhu et al., 2017] has greatly advanced the face synthesis task. Most notably, StyleGAN [Karras et al., 2019] can generate photorealistic faces with high resolution. They purpose the concept of style code which makes the process of face synthesis more controllable. The style code in the potential latent space $\mathcal{W}$ is projected by a random vector in the initial latent space $\mathcal{Z}$. They demonstrate that the projected style code is rich in facial attributes, which is crucial for guiding the generation of the face image.

Inspired by the face manipulation with the style code, we introduce a novel face-swapping pipeline, taking the advantage of high-quality face generation controlled by the style code, addressing the issue of balancing the identity similarity and the retention of the low-level details of the target image in the face-swapping process. We draw lessons from previous face-swapping systems [Xu et al., 2021, Zhu et al., 2021], presenting three key designs for high-fidelity face swapping: facial attributes encoder, style blending module and style decoder. Firstly, the facial attributes encoder encodes facial images to the embeddings in the style latent space $\mathcal{W}$ at multiple levels via the feature pyramid. The embeddings encapsulate the facial attributes and identity information. We generate both source embeddings and target embeddings and reuse the multi-level feature representations from the target input later in the decoding phase. The feature pyramid that preserves low-level details of the target image is fed into the decoder for face generation; this design is crucial for the retention of details of the target image in the face-swapping process. Secondly, the source embeddings and the target embeddings are blended with a style blending module (SBM) with multi-head attention (MHA). By satisfying the appropriate constraints in the training, MHA would be guided to learn the desirable attention for facial attributes elements from the target embeddings and the source embeddings, then fuse them as the embeddings for the output. Finally, the pre-trained style decoder translates the blended embeddings in the latent space back to the image space while leveraging the multi-level feature representations from the target image to retain as many low-level details as possible. The whole model is trained end-to-end with six carefully designed loss functions.

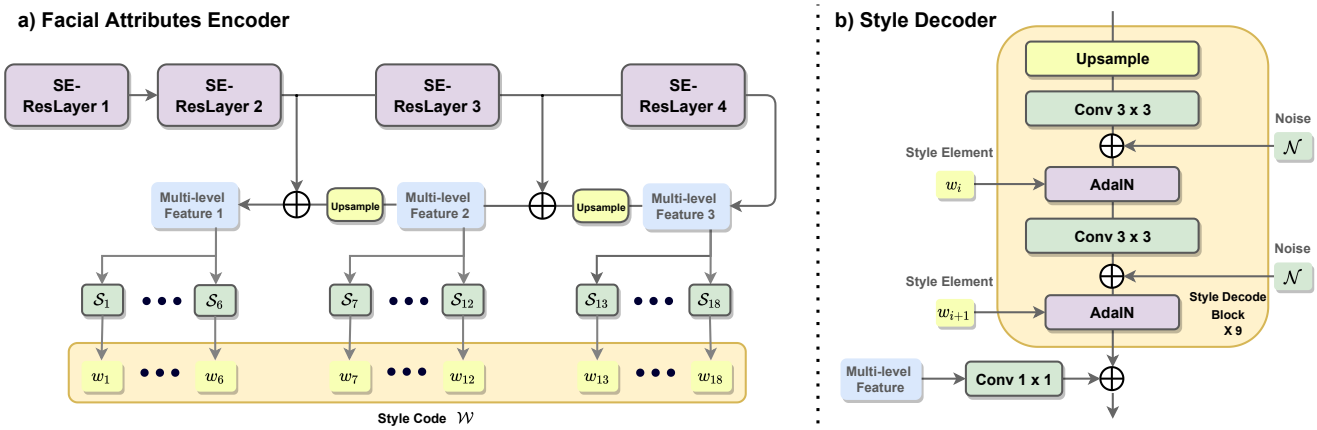


Figure 3: **Architectures of Facial Attributes Encoder and Style Decoder**. a) shows the detail of the facial attributes encoder E. It contains ResNet backbone with Squeeze-and-Excitation (SE) Module [Hu et al., 2018] for each layer, feature pyramids network (FPN) for feature encoding at multiple levels and 18 different style heads $\mathcal{S}_i$ for transforming image features to latent space $\mathcal{W}$. b) demonstrates the style decoder consists of 9 style decode blocks; each operates at a different resolution, ranging from 4^2 to 1024^2. For each style decode block, two corresponding style elements are injected into the block for controlling the face synthesis via the Adaptive Instance Normalisation layer (AdaIN) [Huang and Belongie, 2017]. Following StyleGAN, we use scaled Gaussian noise $\mathcal{N}$ injected after the convolution layer to separate the high-level semantics from stochastic variation. We also reuse the target image feature as the residual input for the style decoder for better retention of the low-level detail of the target image.

2 Methods

We introduce an end-to-end framework for high-fidelity face swapping in wild faces. Our framework follows the traditional GAN methods with both Generator G and discriminator V. The generator G is designed with the encoding-blending-decoding pipeline. Let us consider the input target image $\mathcal{I}^t$ and input source image $\mathcal{I}^s$ are randomly sampled from a face dataset $\mathfrak{D}$. The facial attributes encoder E encodes both the target input image $\mathcal{I}^t$ and the source input image $\mathcal{I}^s$ into the style latent space $\mathcal{W}$, denote as $\mathcal{W}^t$ and $\mathcal{W}^s$ respectively. Following the style code design in StyleGAN [Karras et al., 2019], we use 18 style elements w_i with 512 dimension for each style element, that is $\mathcal{W} \in \mathbb{R}^{18\times512}$. Then, a style blending module B is applied to fuse the target style code $\mathcal{W}^t$ and source style code $\mathcal{W}^s$ into a swapped style code $\mathcal{W}'$. Lastly, the swapped style code $\mathcal{W}'$ is translated back to the image $\hat{\mathcal{I}}^{s\to t}$ with the pretrained style decoder D. The face swapping generator G and the details of SBM are illustrated in Figure 2. We train the generator G adversarially along with the simultaneous training of discriminator V.

2.1 Facial Attributes Encoder

Face images $\mathcal{I}^t$ and $\mathcal{I}^s$ are projected into the latent space $\mathcal{W}$, denoted as $\mathcal{W}^s$ and $\mathcal{W}^t$ respectively, by the facial attributes encoder. The structure of the encoder is illustrated in Figure 3a.

Specifically, facial attributes encoder consists of a SE-ResNet50 backbone [Hu et al., 2018], a feature pyramid structure [Lin et al., 2017], and several style heads. The SE-ResNet50 backbone extracts feature representation for the input image with the Squeeze-and-Excitation (SE) module that adaptively recalibrates the channel-wise feature response [Hu et al., 2018]. This explicit model of inter-channel relations allows a more stable training and the separation of style elements. The feature pyramid structure enables the exploration of the facial attributes at different levels. Different levels of feature representations are responsible for the extraction of different style elements. Later in the decoder phase, these target multi-level features are reused in the decoder for low-level details retention.

Then, the feature representations are projected into style code that contains 18 style elements w. We use three-levels of feature representation, each of which corresponding to the extraction of 6 style elements. We introduce 18 different style heads $\mathcal{S}_i$ ($i \in (1 \cdots 18)$) operated on the multi-level feature representation for the generation of the style elements. Each style head $\mathcal{S}_i$ consists of several `Conv` with `LeakyReLU` layers

 ISBN 978-0-9934207-7-1

(depending on the input resolution) and a final FC layer for projecting to a unified $\mathbb{R}^{512}$ latent space. The latent code is the concatenation of the style elements $\mathcal{W}^t = \|_{i=1}^{18} w_i^t$, $\mathcal{W}^s = \|_{i=1}^{18} w_i^s$ for the target style code and the source style code respectively.

2.2 Style Blending Module

The style blending module (SBM) is introduced for the fusion of source facial attributes to the target context. Operated in the style latent space $\mathcal{W}$, SBM takes both source style code $\mathcal{W}^s$ and the target style code $\mathcal{W}^t$ as input, leveraging the cross-attention, fusing the target style elements to the source style elements adaptively.

As shown in Figure 2, the style code $\mathcal{W}^s$ and $\mathcal{W}^t$ are mapped to the values v_s, v_t, the keys k_s, k_t and the queries q_s, q_t with three different linear transformation functions respectively, $v_s = f_v^s(\mathcal{W}^s)$,$k_s = f_k^s(\mathcal{W}^s)$ and $q_s = f_q^s(\mathcal{W}^s)$ for source, $v_t = f_v^t(\mathcal{W}^t)$,$k_t = f_k^t(\mathcal{W}^t)$ and $q_t = f_q^t(\mathcal{W}^t)$, for target where the f represents the FC layers for encoding the style code and style code to the values, keys or queries. Then multi-head cross attention operation are applied on target value and key with source query, source value and key with target query,

$$\mathrm{Atten}(q_s, k_t, v_t) = \mathrm{softmax}\Big(\frac{q_s k_t}{\sqrt{d_{k_t}}}\Big) v_t \,, \qquad \mathrm{Atten}(q_t, k_s, v_s) = \mathrm{softmax}\Big(\frac{q_t k_s}{\sqrt{d_{k_s}}}\Big) v_s \,. \tag{1}$$

The multi-head attention (MHA) is repeated four times, arriving at the source attention A^s and the target attention A^t. The attention weights for each style element are normalised and concatenated (denoted as $\|$) respectively by

$$\mathrm{A}_w^t = \|_{i=1}^{18} \frac{e^{A_i^t}}{e^{A_i^s} + e^{A_i^t}} \,, \quad \mathrm{A}_w^s = \|_{i=1}^{18} \frac{e^{A_i^s}}{e^{A_i^s} + e^{A_i^t}} \,, \tag{2}$$

Finally, A_w^t and A_w^s are merged with the target and source style codes as the final swapped source style code with

$$\mathcal{W}' = A_w^t \otimes \mathcal{W}^t \oplus A_w^s \otimes \mathcal{W}^s \,, \tag{3}$$

where $\otimes$ denotes the element-wise broadcast multiplication and $\oplus$ denotes the element-wise broadcast addition, $\mathcal{W}'$ is the swapped style code.

2.3 High-Fidelity Face Generation

The style decoder projects style code in latent space $\mathcal{W}$ back to the image space. As illustrated in Figure 3b, it contains 9 layers of style decode blocks. Each layer is responsible for the synthesis at different resolution, ranging from 4^2 to 1024^2. Thus, it progressively adds the details and improves qualities along with the increase of the resolution.

The style element is injected via the adaptive instance normalisation AdaIN[Huang and Belongie, 2017] layers after each Conv layer. Let suppose the input feature to the AdaIN layer is x and the injection style element is w_i, the AdaIN operation is defined as

$$\mathrm{AdaIN}(x, w_i) = \frac{x - \mathbb{E}[x]}{\mathbb{V}[x] + \epsilon} * \gamma_{w_i} + \beta_{w_i} \,, \tag{4}$$

where $\mathbb{E}[x]$ is the instance mean for each channel, $\mathbb{V}[x]$ is the instance standard deviation for each channel. ϵ is a value added to the denominator for numerical stability. γ_{w_i} and β_{w_i} are the decomposition of the affine transformation of style element w_i: $[\gamma_{w_i}, \beta_{w_i}] = \mathrm{T} w_i$, where T is the learnable affine transformation matrix.

Following the StyleGAN [Karras et al., 2019], a scaled Gaussian noise $\mathcal{N}$ is injected in the pipeline for a better separation of the high-level semantics from stochastic variation. In order to recover as much as the low-level details of the target image as possible, we build a short pathway for target image feature adding after the style decode block, only if the output of the style decode block has the same resolution with target feature.

 ISBN 978-0-9934207-7-1

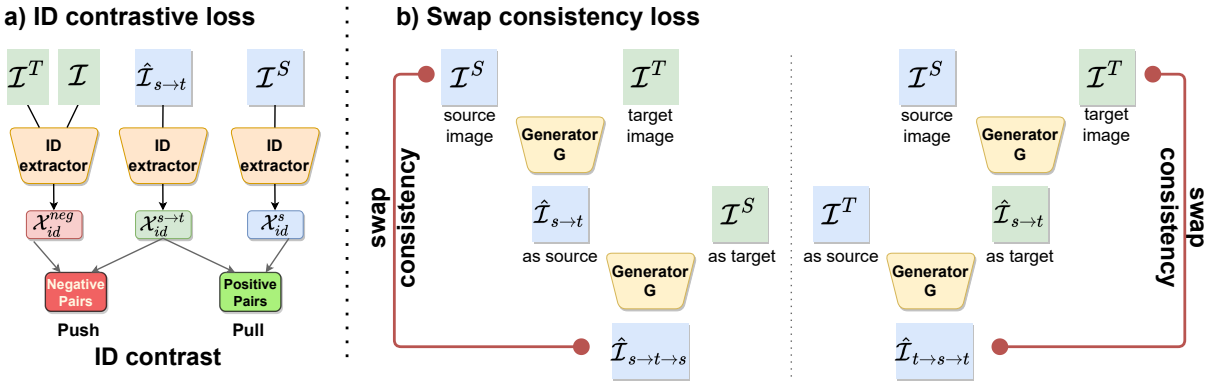


Figure 4: **ID contrastive loss and swap consistent loss**. a) shows the ID contrastive loss. This loss push the negative pairs and pull the positive pairs. b) illustrates the swap consistency loss. Generator G is applied twice to swap back to the original input. Two circumstances are inclued in this loss.

2.4 Training Objectives

The aim of the face-swapping task is to transfer the face ID of a source image to the target image while keeping the other low-level details intact. The output should satisfy i) photorealistic, no artefacts, ii) similarity face ID with the source input, iii) similarity in facial expression, head pose, background and illumination with the target input. We introduce 6 losses to fulfil these objectives.

Adversarial loss. We use adversarial training for our face swapping generator G. Let $\mathcal{L}_{adv}$ be the adversarial loss for making prediction image $\hat{\mathcal{I}}_{s\to t}$ realistic. Discriminator V is implemented as a multi-scale discriminator [Park et al., 2019] on the downsampled output images. We apply hinge loss [Lim and Ye, 2017] on multi-scale outputs of discriminator, the loss for G and V are defined as

$$\mathcal{L}_{adv}^{\mathrm{G}} = -\mathbb{E}_{\mathcal{I}^s,\,\mathcal{I}^t\sim\mathfrak{D}}\mathrm{D}(\mathrm{G}(\mathcal{I}^s,\mathcal{I}^t))\ , \tag{5}$$

$$\mathcal{L}_{adv}^{\mathrm{V}} = -\mathbb{E}_{\mathcal{I}\sim\mathfrak{D}}\big[\min(0,-1+\mathrm{D}(\mathcal{I}))\big] - \mathbb{E}_{\mathcal{I}^s,\,\mathcal{I}^t\sim\mathfrak{D}}\big[\min(0,-1-\mathrm{D}(\mathrm{G}(\mathcal{I}^s,\mathcal{I}^t)))\big]\ , \tag{6}$$

where target image $\mathcal{I}^t$, source image $\mathcal{I}^s$ and a random image $\mathcal{I}$ are sampled from dataset $\mathfrak{D}$.

Identity loss. The identity loss is used to preserve the identity of the source image. It is formulated as

$$\mathcal{L}_{id} = 1 - \cos\left(\mathcal{F}_{id}\left(\hat{\mathcal{I}}_{s\to t}\right), \mathcal{F}_{id}\left(\mathcal{I}^s\right)\right)\ , \tag{7}$$

where $\mathcal{F}_{id}$ is the ID extractor implemented with the pretrained ArcFace [Deng et al., 2019] model; cos represents the cosine similarity of two ID vectors.

ID contrastive loss. The ID contrastive loss is an additional restriction term in training aiming at addressing the issue of similarity of the target face ID with the output after the swapping. Following the InfoNCE loss [Van den Oord et al., 2018], we formulate our ID contrast as

$$\mathcal{L}_{contrast} = -\log\frac{\exp(\mathcal{S}_{\mathrm{sim}}(\mathcal{X}_{id}^s, \mathcal{X}_{id}^{s\to t})/\tau)}{\exp(\mathcal{S}_{\mathrm{sim}}(\mathcal{X}_{id}^t, \mathcal{X}_{id}^{s\to t})/\tau) + \sum_{n\sim p_{neg}}\exp(\mathcal{S}_{\mathrm{sim}}(\mathcal{X}_{id}^n, \mathcal{X}_{id}^{s\to t})/\tau)}\ , \tag{8}$$

where $\mathcal{X}_{id}^t$, $\mathcal{X}_{id}^s$ and $\mathcal{X}_{id}^{s\to t}$ denote the target, source, prediction ID vectors, generated by the pretrained ArcFace $\mathcal{F}_{id}$ ID extractor. $\mathcal{S}_{\mathrm{sim}}$ denotes the normalised dot product operation; τ is the InfoNCE temperature coefficient. Figure 4a demonstrates the details of ID contrastive loss.

Face reconstruction loss. When the source and target images are the same input, the model should be able to reconstruction the original image, this loss is widely applied in previous works [Li et al., 2019, Xu et al., 2021], It is defined as

$$\mathcal{L}_{rec} = \begin{cases} \frac{1}{2}\left\|\hat{\mathcal{I}}^{s\to t} - \mathcal{I}^t\right\|_2^2 & \text{if } \mathcal{I}^t = \mathcal{I}^s \\ 0 & \text{otherwise}\ . \end{cases} \tag{9}$$

We occasionally force the source input and the target input to be the same image $\mathcal{I}^t = \mathcal{I}^s$, so that the expected output should recover the input image in pixel level. This loss is disabled if the source input and the target input are different. The negative samples in ID contrastive loss (Equation 8) would not consider $\mathcal{I}^t$ as negatives if $\mathcal{I}^t = \mathcal{I}^s$. In practice, we use a dynamic controller P_π to control the probability of source input and the target input being the same image. We start with P_π=1 initially and gradually decrease this value to 0 in the end.

Facial landmark constraint loss. This loss considers the facial expression and poses in $\mathcal{I}^t$ stay the same after the face swapping. We leverage the landmark loss to ensure the consistency between $\mathcal{I}^t$ and $\hat{\mathcal{I}}^{s\to t}$:

$$\mathcal{L}_{lm} = \frac{1}{2}\left\|\mathcal{F}_{lm}(\mathcal{I}^t) - \mathcal{F}_{lm}\left(\mathcal{I}^{s\to t}\right)\right\|_2^2 , \tag{10}$$

where $\mathcal{F}_{lm}$ is the off-the-shelf facial landmark extractor [Sun et al., 2019] that generated 19 facial landmark keypoints for $\mathcal{I}^t$ and $\mathcal{I}^{s\to t}$. L2 loss is used for regularising the landmark keypoints.

Swap consistent loss. This loss aims to make the output consistent with the original image. The generator G is expected to reconstruct the input image if it is applied twice. As shown in Figure 4b, there are two circumstances. 1) considering the $\mathcal{I}^s$ as the target input, the first round swap output $\mathcal{I}^{s\to t}$ as the source input, the model needs to reconstruct $\mathcal{I}^s$ in the next round of swapping. 2) considering the $\mathcal{I}^t$ as the source input, $\mathcal{I}^{s\to t}$ as the target input, the model is expected to reconstruct $\mathcal{I}^t$, thus the loss can be formulate as

$$\mathcal{L}_{swap} = \frac{1}{2}\left\|\mathrm{G}\left(\mathrm{G}\left(\mathcal{I}^s, \mathcal{I}^t\right), \mathcal{I}^s\right) - \mathcal{I}^s\right\|_2^2 + \frac{1}{2}\left\|\mathrm{G}\left(\mathcal{I}^t, \mathrm{G}\left(\mathcal{I}^s, \mathcal{I}^t\right)\right) - \mathcal{I}^t\right\|_2^2 \tag{11}$$

Overall loss. The overall training loss of the discriminator is $\mathcal{L}_{\mathrm{V}} = \mathcal{L}_{adv}^{\mathrm{V}}$. The overall loss of the generator G is the weighted summation of adversarial loss $\mathcal{L}_{adv}^{\mathrm{G}}$, identity loss $\mathcal{L}_{id}$, ID contrastive loss $\mathcal{L}_{contrast}$, face reconstruction loss $\mathcal{L}_{rec}$, facial landmark constraint loss $\mathcal{L}_{lm}$, swap consistent loss $\mathcal{L}_{swap}$,

$$\mathcal{L}_{\mathrm{G}} = \mathcal{L}_{adv}^{\mathrm{G}} + \lambda_1\mathcal{L}_{id} + \lambda_2\mathcal{L}_{contrast} + \lambda_3\mathcal{L}_{rec} + \lambda_4\mathcal{L}_{lm} + \lambda_5\mathcal{L}_{swap} \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 are the loss weights.

3 Experiments

Datasets. CelebA-HQ [Karras et al., 2017] is a high-resolution facial dataset containing $30,000$ images at 1024^2 resolution. It is a subset of CelebA dataset, built for face detection, facial landmark localisation. FFHQ [Karras et al., 2019] is a dataset that contains 70,000 megapixel face images collected from Flickr. FFHQ has a large variations of gender, age, ethnicity.

Implementation Details. We use the mix of FFHQ and CelebA-HQ as the training set. For each facial image from FFHQ or CelebA-HQ, we first align the face with the facial landmark, and then resize the resolution to 256^2 as the input. In all experiments, the learning rate for G and V is set to $lr = 4e^{-5}$ with a batch size of 32, and linearly scaled with different batch sizes. We set $\tau = 0.07$ in Equation 8 and $\lambda_1 = 10$, $\lambda_2 = 5$, $\lambda_3 = 1$, $\lambda_4 = 100$, $\lambda_5 = 1$ in Equation 12. The StyleGAN pretrained weights [Karras et al., 2019] are used to initialise style decoder D. For training stability, we freeze D and only train E and SBM. After that, we unfreeze D and fine-tune with a small learning rate. The model is trained for 2.4M iterations in total on 8 RTX2080Ti GPUs.

Qualitative Comparison. The qualitative comparisons are shown in Figure 5a. FSGAN [Nirkin et al., 2019] and FaceShifter [Li et al., 2019] are early GAN-based face-swapping works, they achieved remarkable results at that time, but they fail on face ID transfer, producing significant artefacts *e.g.* R3, R4, R6 in Figure 5a. [Nitzan et al., 2020] and [Yang and Qiao, 2021] are style-code-based methods. They demonstrate photorealistic results with fewer artefacts; however, they fail to preserve the low-level details like hairstyle, background, head pose *etc.* from the target image, such as R1, R3, R4, R5. Instead, our method well balanced the issues of face ID similarity and details preservation, demonstrating the most realistic swapping effect.

Ablation Studies. We conducted ablation study on two important components in our designs. Figure 5b visualised the results. It can be observed that if the swap consistent loss is dropped in the training, the low-level details such as skin colour and glasses are not preserved. Furthermore, if the model is trained without the

 ISBN 978-0-9934207-7-1

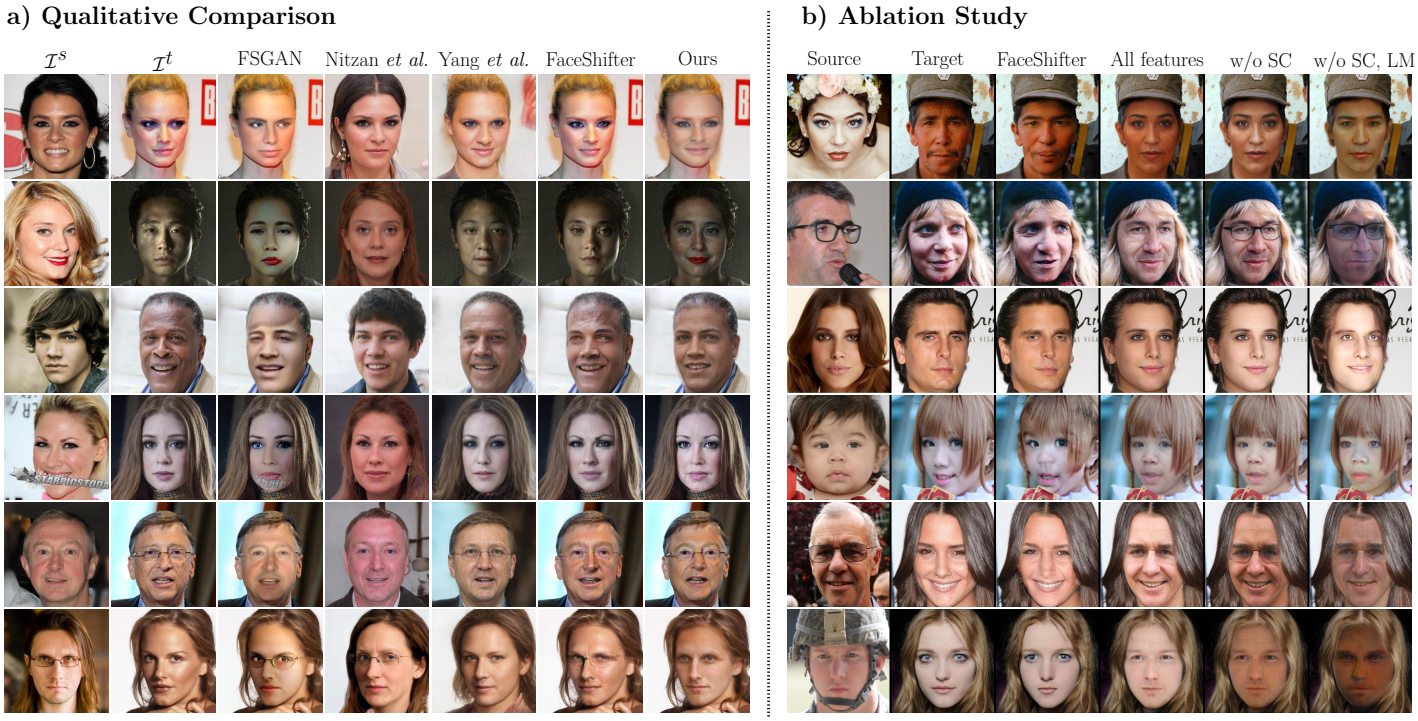


Figure 5: **Qualitative Comparison with the state-of-the-art Methods and Ablation Studies** a) demonstrates the Qualitative Comparison of our methods over the others on CelebA-HQ test set. From left to right, each column represents source input, target input, FSGAN [Nirkin et al., 2019], [Nitzan et al., 2020], [Yang and Qiao, 2021], FaceShifter [Li et al., 2019] and our results. b) shows the ablation study results. 'all features' represents our model with all aforementioned features, 'w/o SC' denotes our method without applying swap consistent loss, 'w/o SC,LM' denotes without the swap consistent loss and facial landmark constraint loss.

facial landmark constraint, we can see the head pose and facial expressions are not preserved and a significant decrease of visual perception.

User Studies. We conducted user studies on the quality of our synthesised faces compared with the state-of-the-art methods. The quality of the image is evaluated in three aspects: i) *score for similar identity with source face* ii) *score for the preservation of details (head pose, face expression, illuminations etc.) with the target image* iii) *score for the image realism*. The users are asked to score each aspect from 1 to 5. 20 generated images are randomly selected from 200 generated images on the test set of CelebA-HQ for each user study. 100 people took part in this study, out of which 87 valid results are used. Table 1 shows the results. We can see our method outperforms the other methods with a significant margin on all three aspects.

Further Analysis. We take a further study on how the face swap system works. Figure 6 illustrates the example of face swapping in our system. We visualised the intermediate output $\mathcal{W}^s$ and $\mathcal{W}'$ from the model. We observe that the face attributes encoder first learns the essential face attributes from the source image. Then, the SBM edits the face attributes adaptively by the guidance of the target style code. We can see the SBM effectively synthesises the facial expression and the head pose *etc.*. Lastly, it is rendered leveraging the multi-level features that contain low-level details from the target image.

4 Conclusion

In this paper, we introduced a novel face-swapping method that leverages the powerful style-code-based model for photorealistic face synthesis. We built a novel style blending module operated in the latent space; Our method demonstrated superior face synthesis quality and addressed the issue of balancing the identity similarity with the source, at the same time retaining the low-level details with the target. The work holds the promise that a style-code-based model can be effectively tuned for face-swapping task and thereby help unlock the high-resolution, high-photorealistic face swapping.

 ISBN 978-0-9934207-7-1

method	id.	attr.	real.
DeepFakes	2.3	3.1	2.6
FSGAN	2.7	2.4	3.0
Nitzn *et al.*	2.9	2.1	3.5
Yang *et al.*	3.1	2.7	3.6
FaceShifter	2.9	3.7	3.0
Ours	**3.4**	**4.2**	**3.9**

Table 1: **User Study Results.** We conduct user study on visual results of the state-of-the-art methods and ours on three aspects: ID similarity, attribute preservation and image realism

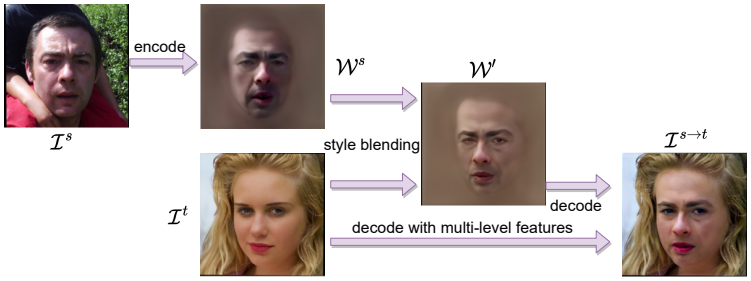


Figure 6: **Visualisation on How the Method Works.** The visualisation of $\mathcal{W}^s$ and $\mathcal{W}^t$ is projected to the image space via decoder without the multi-level features. The model learns the face attributes from the source image and then edits the face attributes adaptively by the target input. Later, it synthesises the facial expression and the head pose *etc.* from target input. Lastly it is translated back to the image leveraging the multi-level features to recover low-level details.

References

[Bitouk et al., 2008] Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., and Nayar, S. K. (2008). Face swapping: automatically replacing faces in photographs. In *ACM SIGGRAPH 2008 papers*, pages 1–8. 1

[Choi et al., 2018] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797. 2

[Deng et al., 2019] Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699. 5

[Hu et al., 2018] Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141. 3

[Huang and Belongie, 2017] Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510. 3, 4

[Karras et al., 2017] Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*. 2, 6

[Karras et al., 2019] Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410. 2, 3, 4, 6

[Korshunova et al., 2017] Korshunova, I., Shi, W., Dambre, J., and Theis, L. (2017). Fast face-swap using convolutional neural networks. In *ICCV*, pages 3677–3685. 1

[Li et al., 2019] Li, L., Bao, J., Yang, H., Chen, D., and Wen, F. (2019). Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*. 2, 5, 6, 7

[Lim and Ye, 2017] Lim, J. H. and Ye, J. C. (2017). Geometric gan. *arXiv preprint arXiv:1705.02894*. 5

[Lin et al., 2017] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125. 3

[Nirkin et al., 2019] Nirkin, Y., Keller, Y., and Hassner, T. (2019). Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, pages 7184–7193. 6, 7

[Nitzan et al., 2020] Nitzan, Y., Bermano, A., Li, Y., and Cohen-Or, D. (2020). Face identity disentanglement via latent space mapping. *ACM Transactions on Graphics (TOG)*, 39(6):1–14. 6, 7

[Park et al., 2019] Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346. 5

[Sun et al., 2019] Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., and Wang, J. (2019). High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*. 6

[Van den Oord et al., 2018] Van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807. 5

[Xu et al., 2021] Xu, Z., Yu, X., Hong, Z., Zhu, Z., Han, J., Liu, J., Ding, E., and Bai, X. (2021). Facecontroller: Controllable attribute editing for face in the wild. *arXiv preprint arXiv:2102.11464*. 2, 5

[Yang and Qiao, 2021] Yang, S. and Qiao, K. (2021). Shapeediter: a stylegan encoder for face swapping. *arXiv preprint arXiv:2106.13984*. 6, 7

[Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232. 2

[Zhu et al., 2021] Zhu, Y., Li, Q., Wang, J., Xu, C.-Z., and Sun, Z. (2021). One shot face swapping on megapixels. In *CVPR*, pages 4834–4844. 2

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/FEVR2516

On the Feasibility of Privacy-Secured Facial Authentication for low-power IoT Devices - Quantifying the Effects of Head Pose Variation on End-to-End Neural Face Recognition

Wang Yao[1], Viktor Varkarakis[2], Joseph Lemley[2], and Peter Corcoran[1]

[1]*School of Engineering, National University of Ireland, Galway.*
[2]*Xperi Corporation, Galway.*

Abstract

Recent low-power neural accelerator hardware provides a solution for end-to-end privacy and secure facial authentication, such as smart refueling machine locks in shared accommodation, smart speakers, or televisions that respond only to family members. This work explores the impact that head pose variation has on the performance of a state-of-the-art face recognition model. A synthetic technique is employed to introduce head pose variation into data samples. Experiments show that the synthetic pose variations have a similar effect on face recognition performance as the real samples with pose variations. The impact of large variations of head poses on the face recognizer was then explored by further amplifying the angle of the synthetic head pose. It is found that the accuracy of the face recognition model deteriorates as the pose increases. After fine-tuning the network, the face recognition model achieves close to the accuracy of frontal faces in all pose variations, indicating that the face recognition model can be tuned to compensate for the effect of large poses.

Keywords: Head Pose, Face Recognition, Face Rotation Method

1 Introduction

Face recognition (FR) is a long-standing research topic in the field of computer vision. In recent years there have been significant improvements in accuracy with the emergence of deep learning networks and associated neural FR algorithms.The latest research has focused on network architectures [Wu et al., 2018, Sandler et al., 2018, Yang et al., 2021] and loss functions [Deng et al., 2019, Wang et al., 2018, Liu et al., 2017, Deng et al., 2020], which have promoted increased accuracy and robustness. In earlier research near-frontal faces are commonly used, and many researchers have focused on transforming faces to a frontal orientation in order to compensate for the pose variation. This technique is known as face-frontalization and is frequently adopted in the literature [Zhang et al., 2019].

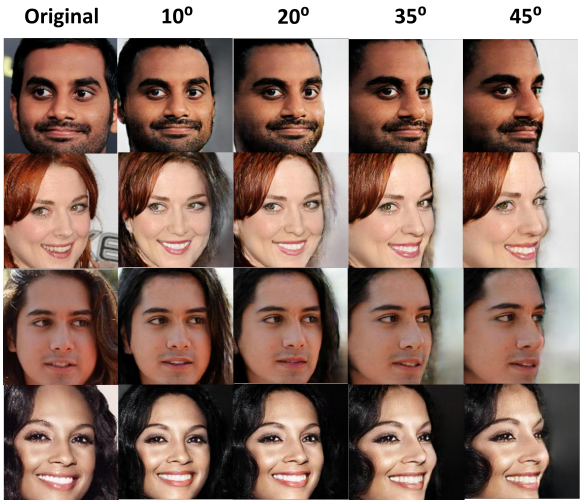


Figure 1: Samples of head pose variations injected in the original CelebA-HQ images using the rotation and rendering method.

In practical applications, such as user authentication, facial images taken in the wild may have different variations (poses, lighting, expressions and etc)

 ISBN 978-0-9934207-7-1

that are not constrained. It is shown that factors such as pose, lighting, expression, and age can have a significant impact on the performance of FR algorithms. Of these challenges the problem of extreme facial lighting has been recently addressed [Varkarakis et al., 2021]. PoseFace [Meng et al., 2021] was proposed to solve the data imbalance between frontal and profile images of faces, mainly for profile faces with yaw angle greater than 60 degrees. While this work [Meng et al., 2021] suggests that less extreme pose angles will not significantly affect FR algorithms it is important for our consumer use-case to accurately quantify any potential effects on authentication accuracy. In this way, whether they need to be compensated in an end-to-end neural FR will be determined.

In recent years, the development of end-to-end neural FR and the emergence of low-power neural accelerators [Corcoran et al., 2019] have made it possible to implement high-performance biometric user authentication for low-power devices [Corcoran et al., 2019], such as smartphones, IoT, etc. It has been shown that state-of-the-art neural FR architectures can be implemented in ultra-low power consumer appliances [Fleischer et al., 2018]. However state-of-art neural accelerators require an end-to-end neural algorithm to realize low power consumptions, thus it is not possible to pre-process input images with image enhancement pipelines, or face frontalization algorithms.

This research studies the effects of the facial pose, in particular larger pose extremes of facial data samples in an end-to-end fully neural FR system. The key research questions are as follows.

Figure 2: Head rotation with yaw, pitch and roll.[1].

1) How robust is the latest FR algorithm to pose variations? Is there a critical pose angle beyond which significant loss in accuracy/performance occurs?

2) Does a synthetic dataset created by a pose-adapting GAN behave in the same way as a real multi-pose dataset?

3) Can a state-of-the-art (SOTA) FR algorithm be trained/finetuned with synthetic data to improve its accuracy on a real multi-pose dataset?

This work quantifies the impact of multi-pose faces on the SOTA FR algorithm using receiver operating characteristic curve (ROC) techniques. Due to the lack of sufficient pose variations in the existing real public datasets, a method for synthesizing multi-pose faces with the addition of yaw variations is used to complement the available public datasets. The capability of handling pose variations by fine-tuning the neural FR network illustrates the potential of this work to provide end-to-end full neural face authentication in the field.

2 Related Works

Face recognition has been one of the most important research topics in the field of computer vision. The ability to recognize faces in various poses presents a unique challenge to face recognition systems[Zhang and Gao, 2009, Asthana et al., 2011]. Generally, two approaches have been used to tackle the problem. One is to normalize the face image to a frontal face and the other is to perform pose invariant face recognition (PIFR).

Face frontalization is a method to synthesize a frontal face directly from an input non-frontalized face, i.e., to calibrate an unconstrained face to a standard view in order to improve the accuracy of face recognition. Existing face frontalization methods are 3D fitting-based methods [Feng et al., 2018], deep learning methods

 ISBN 978-0-9934207-7-1

with encoder-decoder architectures [Zhang et al., 2013], and generative adversarial networks (GAN) methods [Zhang et al., 2019]. While these techniques demonstrate encouraging results, the quality of the synthesized faces is dependent on the training set and recent research [Varkarakis et al., 2020] shows that the FR identities of synthesized facial identities can be unreliable. In addition, two independent neural networks for facial frontalization and for face recognition are needed to implement the full neural frontalization FR technique.

PIFR refers to the ability to recognize face images in an arbitrary pose. Typically faces captured in the wild exhibit a wide variation in poses and lighting [Varkarakis et al., 2021]. Accurate FR across such variations is important for low-power consumer authentication systems, such as portable consumer and IoT devices. For example, electronic devices such as doorbell cameras and electronic locks that do not have keyboards for input but need to verify the user's identity, where biometric authentication such as facial verification will play an important role. Traditional methods generally focus on extracting face invariant features, such as face landmark points [Ding et al., 2015], statistical features [Kim and Kittler, 2006], template matching [De Marsico et al., 2012], machine learning methods [Tao et al., 2008] etc. With the large-scale application of deep learning, it has become an effective method to train deep learning models to solve PIFR using large pose datasets of faces. However, such data-driven approaches rely heavily on well-annotated data, a process that is quite expensive.

3 Methodology

In this section, techniques of this research are presented, including the used datasets, a method for synthetic multi-pose generation, the FR method and the evaluation metric.

3.1 Datasets

To discuss the questions proposed in this study, two types of datasets are used: synthetic dataset and 'real-world' dataset. The following is a detailed explanation of the datasets used in this research.

CelebA-HQ: CelebA-HQ [Karras et al., 2017] contains 30k images of 1024x1024 faces with roughly 6000 identities. It is a high-quality face dataset derived from the CelebA [Liu et al., 2015] dataset. CelebA-HQ is a dataset that we used to generate the synthetic dataset with multiple poses to answer the research question of this paper. Note that since the CelebA-HQ dataset does not contain information on yaw, pitch and roll, FSA-Net [Yang et al., 2019] was employed to evaluate the pose in the CelebA-HQ dataset.

BIWI: The Biwi Kinect Head Pose Dataset [Fanelli et al., 2013] contains around 15.8k images of 20 people (6 females and 14 males, 4 people were recorded twice). The resolution of each sample is 640x480. It is a 'real-world' dataset, which was captured by a Kinect at about one meter distance. The BIWI has many real head pose variations, and it is used in this work to generate synthetic head poses and evaluate the second research question "How well does a GAN-based head pose transformation match real data?".

3.2 Synthetic Multi-Pose Generation Method

In this work, synthetic multiple poses are applied to CelebA-HQ and BIWI datasets using the rotate-and-render technique [Zhou et al., 2020]. This multi-pose generation technique represents SOTA GAN to generate various multi-view faces. Unlike other similar methods that need paired multi-view training data, this method only uses the single-view image as input. Figure 1 shows samples generated from CelebA-HQ.

Head pose rotation is divided into yaw, pitch and roll, as shown in Figure 2[1]. Roll and pitch angles are not considered in this research because the effect of roll will be eliminated by face alignment, while face images

[1]Image from datasets: https://www.unavarra.es/gi4e/databases/hpdb

with large pitch angles are rare [Meng et al., 2021]. We restrict our experiments to generating a series of fixed-angle faces to obtain a better understanding of synthetic head pose data. In this paper, the CelebA-HQ and BIWI datasets are used to generate multi-view faces and validate the relationship between the synthetic and real datasets. To better understand the generated various head pose images, we select 'frontal' faces from the datasets as input to synthesize profile faces at arbitrary-angle.

3.3 Face Recognition Method

ArcFace [Deng et al., 2019] is selected as FR model in this work, which has a public reference implementation[2]. ArcFace is one of the SOTA FR models, achieving 99.83% accuracy on the LFW benchmark. The proposed method uses FR to verify the validity of synthetic head pose images. Firstly, MTCNN [Zhang et al., 2016] is used for detection and cropped, and 112x112 face images are acquired. Then the preprocessed faces are fed into the ArcFace network and 512-embedding will be computed corresponding to the faces. Finally, the identity similarity of two faces is obtained by calculating the cosine similarity, which is using two 512-embeddings.

3.4 ROC Metric

The ROC curve illustrates the ability of the classifier under different thresholds, which can visually demonstrate the uniqueness of the identity and is widely used to evaluate the performance of FR models [Varkarakis et al., 2020]. In this work, first, different pose angles are classified based on the original ground-truth seed dataset to generate front-side positive pairs and negative pairs, respectively. Then, the similarity score is obtained through FR, and the ROC curve of the original pose is drawn for analysis. Due to the limited identities and poses in the original real datasets, two methods are used for discussion, one is to directly synthesize multi-pose face data, and the other is to use synthetic CelebA-HQ to synthesize multi-pose samples.

4 Quantifying the Effects of Head Pose on the FR model

Experiment Setting: There are two parts in this experiment. Section 4.1 is the first part, the corresponding positive pairs (PPs) with various yaw angles and negative pairs (NPs) with various yaw angles respectively are created by different pose angles of real dataset. The PPs are one-to-one image pairs of all faces with the same identity. The faces with different identities are selected to form image pairs which are NPs. Then the related face images of the PPs and NPs are fed into the face recognizer. Finally, the ROC curves are plotted by the embeddings from FR. Section 4.2 and 4.3 are second parts, firstly, frontal faces are selected to synthesize multi-pose face images, and the corresponding PPs and NPs are created respectively, and then the relevant synthetic faces are fed into the face recognizer and the ROC curves are plotted. Note that each pair of the PPs and NPs for synthetic poses are one image from the real face and the other image from the synthetic head pose.

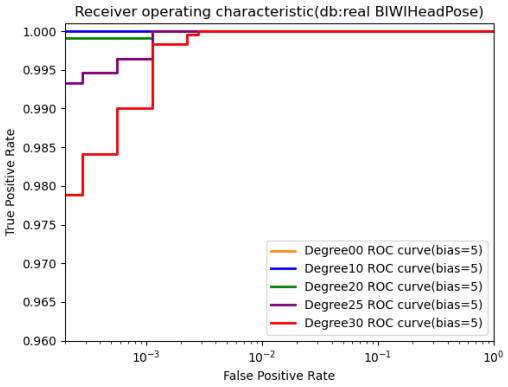


Figure 3: The ROC of real BIWI head pose. Note the classification of various degree in this experiment is only the variation of yaw are selected from BIWI, while pitch and roll are kept at (-5,5).

4.1 The effects of real-world multi-pose BIWI on FR

This experiment introduces ROCs to measure and quantify the impact of real multi-pose head data on FR. Due to the limitation of the number of images and identities, we only acquire angles within 30 degrees for

[2]https://github.com/deepinsight/insightface

discussion here. Taking the ground-truth degrees from the BIWI dataset as the baseline and 5 degrees as the bias, the relevant ROC curves are plotted in Figure 3. Bias is applied to extend the angle range, e.g. a 30 degree angle with a bias of 5 has a group of images with yaw angles in (25, 35), pitch and roll angles in (-5,5). The ROCs from Figure 3 show that the change in the real-world pose will lead to a small insignificance drop in FR. Figure 3 indicates that the FR model has high robustness for real faces rotated within 30 degrees along the yaw direction.

4.2 The effects of synthetic multi-pose BIWI on FR

This experiment quantifies the effect of multi-pose synthetic faces on FR. First, we validate the effect of yaw variation from 0 to 35 degrees on FR to explore the relationship between the real dataset and the synthetic dataset on FR. Second, we extend the angular range of the BIWI dataset to study the effect of large yaw variation on FR.

From Figure 4 the variation in the head pose will lead to a difference in FR performance. The Original-BIWI, Degree00-BIWI, Degree10-BIWI, Degree20-BIWI, Degree25-BIWI and Degree30-BIWI ROC curves in Figure 4 show that the FR model performs close to the performance of the FR model on the original frontal dataset. The comparison shows that the synthetic faces in Figure 4 are consistent with the performance of the real faces in Figure 3 in terms of FR.

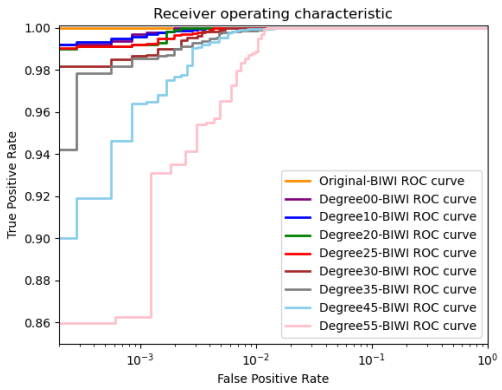


Figure 4: The ROC of synthetic head pose on BIWI. Here the data are generated by original BIWI, which yaw, pitch & roll in (-5,5) degrees.

Observing Figure 4, the ROC curve at 35 degrees, we can see that the accuracy starts to drop obvious. As the yaw angle increases, the accuracy decreases more and more obviously. It shows that the pose of 55 degrees is the most challenging one for the FR task in this experiment. This is consistent with our expectation that the FR performance decreases significantly with increasing input head pose, especially when the angle is greater than 35 degrees.

4.3 The effects of synthetic multi-pose CelebA-HQ on FR

In this experiment, we will explore the effects of multi-pose face data generated by another synthetic dataset, CelebA-HQ, on the face recognizer. This dataset has many identities with less variation in the head pose. The frontal face images from CelebA-HQ (yaw, pitch and roll in (-5,5) degrees) are employed to generate head poses in our experiments. The ROC curves for the synthetic poses from CelebA-HQ were plotted in Figure 5.

Through observation, Figure 5 is consistent with the degradation trend in Figure 4, verifying that changes in head pose lead to changes in face recognition performance. The original-CelebAHQ ROC curve in Figure 5 has the best performance. We observe that the performance of the FR model slightly decreases in Degree20-CelebAHQ ROC and Degree25-CelebAHQ ROC from Figure 5, which is similar to Figure 4. In Figure 5, there is a noticeable decreasing trend in the Degree35-CelebAHQ ROC curve, indicating that the 35-degree is the angle that causes the performance of the FR model to start decreasing more significantly. This behavior is consistent with the behavior

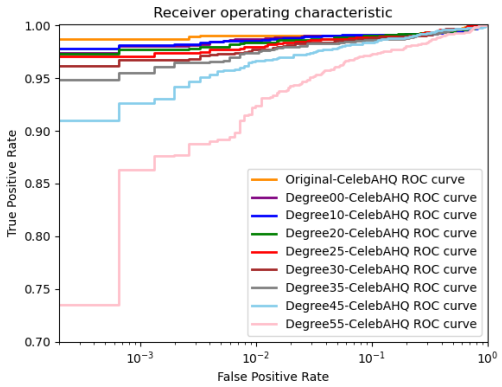


Figure 5: The ROC of synthetic head pose on CelebA-HQ. Here the data are generated by original CelebA-HQ, which yaw, pitch & roll in (-5,5) degrees. The degrees are detected by FSA[Yang et al., 2019].

 ISBN 978-0-9934207-7-1

of the Degree35-BIWI curve in Figure 4, indicating that the 35-degree synthetic head pose begins to become a challenging angle for FR tasks.

Observing the accuracy of the ROC curve for the 35-degree face in Figure 5 is close to its 0.95 TPR (true positive rate), which is slightly higher than the ROC curve for the 35 degree face in Figure 4. When the synthetic head pose is 45 degree Figure 5 is slightly higher than 0.9 TPR, in Figure 4 is roughly 0.9 TPR. When the synthetic head pose angle is 55 degree, Figure 5 is below 0.75 TPR, while Figure 4 is slightly above 0.85 TPR. This experiment illustrates a certain consistency in the behavior of the synthetic head pose. As the degree increases further, the FR model starts to degrade rapidly, which is consistent with the real-world dataset.

5 Fine-tuning the FR model with Head Pose

Experiment Setting: The pretrain ResNet50 network from arcface[3] was selected in this section. The training set utilizes the original CelebA-HQ dataset as well as synthetic samples of various head poses (0, 10, 20, 25, 30, 35,45 and 55) with a total of 184193 training images. Original arcface loss is used, the learning rate is set to 0.00125, and the batch size is 32. Due to the relatively large number of images, all network layers were unfrozen for fine-tuning. The fine-tuned network is used to calculate the embeddings of the BIWI samples.

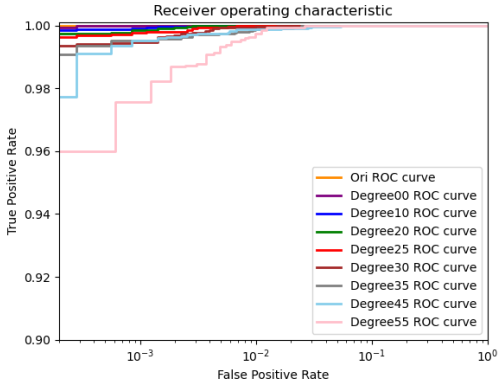


Figure 6: Fine-tune results on BIWI.

Figure 6 shows the ROC of the original BIWI image and 8 head poses under the fine-tuned model, and we selected the model with the epoch of 40. As shown in Figure 6, the ROCs corresponding to the synthesized BIWI pose face images after fine-tuning, compared to the ROCs (Figure 4) corresponding to these head poses on the initial network, are significantly improved. Since no BIWI-related data were used in the fine-tuning process, and the results are all at a high level, it shows the ability of the network to be able to generalize to other pose datasets. Overall, the FR model trained with head pose variations was able to adapt and handle face samples with the different head poses, achieving high accuracy results. This experiment illustrates that pose can be compensated by fine-tuning methods without using pre-processing methods such as face frontalization, which is beneficial for use in a neuro-accelerator.

6 Conclusion

This work uses the ROC metric to measure the impact of head pose variations on face recognition, designs experiments and discusses several research questions presented in this paper. Experiments have shown that small poses do not affect the FR model, and there is a significant effect on the FR model when the head pose is larger than 35 degrees. Large angle poses will lead to severe degradation of the FR model. The synthetic pose resembles the effect of the real pose on the performance of the FR model, indicating that the synthetic pose dataset can replace or enhance the real dataset to some extent. The fine-tuning experiment shows that the fine-tuned model can learn changes in head pose and recover to a performance level close to the original baseline with some generalization ability in the multi-pose head data condition, thus illustrating the effectiveness of the fine-tuning process. The experiments improved the performance level of the model for pose data, which means that fine-tuning could be used to eliminate the need for pre-processing techniques that correct the pose which is good for implementing the FR model in neuro accelerators.

The modest scale identification in BIWI suggests that we need to find or generate a larger dataset to better understand how much the effect of the pose has on FR, especially for more extreme head pose angles (e.g. more

[3]https://github.com/deepinsight/insightface

 ISBN 978-0-9934207-7-1

than 55 degrees). Furthermore, this study illustrates that it is feasible to build large-scale synthetic datasets by synthesizing multi-pose head data for large-scale training datasets in FR or other application scenarios in future work. In addition, multi-pose data is used to fine-tune the face recognition model that improves the sensitivity of the original face recognizer to pose data, and increases the robustness of face recognition in the wild and other challenges. For most consumer applications, users will look at the camera when using the system, so it can be concluded that FR is performing adequately for most consumer use cases.

Acknowledgments

This research is supported by (i) Irish Research Council Enterprise Partnership Ph.D. Scheme (Project ID: EPSPG/2020/40) and, (ii) Xperi Corporation, Ireland.

References

[Asthana et al., 2011] Asthana, A., Marks, T. K., Jones, M. J., Tieu, K. H., and Rohith, M. (2011). Fully automatic pose-invariant face recognition via 3d pose normalization. In *2011 International Conference on Computer Vision*, pages 937–944. IEEE.

[Corcoran et al., 2019] Corcoran, P., Lemley, J., Costache, C., and Varkarakis, V. (2019). Deep learning for consumer devices and services 2—ai gets embedded at the edge. *IEEE Consumer Electronics Magazine*, 8(5):10–19.

[De Marsico et al., 2012] De Marsico, M., Nappi, M., Riccio, D., and Wechsler, H. (2012). Robust face recognition for uncontrolled pose and illumination changes. *IEEE transactions on systems, man, and cybernetics: systems*, 43(1):149–163.

[Deng et al., 2020] Deng, J., Guo, J., Liu, T., Gong, M., and Zafeiriou, S. (2020). Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision*, pages 741–757. Springer.

[Deng et al., 2019] Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

[Ding et al., 2015] Ding, C., Choi, J., Tao, D., and Davis, L. S. (2015). Multi-directional multi-level dual-cross patterns for robust face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):518–531.

[Fanelli et al., 2013] Fanelli, G., Dantone, M., Gall, J., Fossati, A., and Van Gool, L. (2013). Random forests for real time 3d face analysis. *Int. J. Comput. Vision*, 101(3):437–458.

[Feng et al., 2018] Feng, Y., Wu, F., Shao, X., Wang, Y., and Zhou, X. (2018). Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 534–551.

[Karras et al., 2017] Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

[Kim and Kittler, 2006] Kim, T.-K. and Kittler, J. (2006). Design and fusion of pose-invariant face-identification experts. *IEEE transactions on circuits and systems for video technology*, 16(9):1096–1106.

[Liu et al., 2017] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220.

[Liu et al., 2015] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738.

[Meng et al., 2021] Meng, Q., Xu, X., Wang, X., Qian, Y., Qin, Y., Wang, Z., Zhao, C., Zhou, F., and Lei, Z. (2021). Poseface: Pose-invariant features and pose-adaptive loss for face recognition. *arXiv preprint arXiv:2107.11721*.

[Sandler et al., 2018] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

[Tao et al., 2008] Tao, D., Li, X., Wu, X., and Maybank, S. J. (2008). Geometric mean for subspace selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):260–274.

[Varkarakis et al., 2020] Varkarakis, V., Bazrafkan, S., Costache, G., and Corcoran, P. (2020). Validating seed data samples for synthetic identities–methodology and uniqueness metrics. *Ieee Access*, 8:152532–152550.

[Varkarakis et al., 2021] Varkarakis, V., Yao, W., and Corcoran, P. (2021). Towards end-to-end neural face authentication in the wild–quantifying and compensating for directional lighting effects. *arXiv preprint arXiv:2104.03854*.

[Wang et al., 2018] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274.

[Wu et al., 2018] Wu, X., He, R., Sun, Z., and Tan, T. (2018). A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896.

[Yang et al., 2019] Yang, T.-Y., Chen, Y.-T., Lin, Y.-Y., and Chuang, Y.-Y. (2019). Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1087–1096.

[Yang et al., 2021] Yang, X., Jia, X., Gong, D., Yan, D.-M., Li, Z., and Liu, W. (2021). Larnet: Lie algebra residual network for face recognition. In *International Conference on Machine Learning*, pages 11738–11750. PMLR.

[Zhang et al., 2016] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.

[Zhang et al., 2019] Zhang, S., Miao, Q., Zhu, X., Chen, Y., Lei, Z., Wang, J., et al. (2019). Pose-weighted gan for photorealistic face frontalization. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2384–2388. IEEE.

[Zhang and Gao, 2009] Zhang, X. and Gao, Y. (2009). Face recognition across pose: A review. *Pattern recognition*, 42(11):2876–2896.

[Zhang et al., 2013] Zhang, Y., Shao, M., Wong, E. K., and Fu, Y. (2013). Random faces guided sparse many-to-one encoder for pose-invariant face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2416–2423.

[Zhou et al., 2020] Zhou, H., Liu, J., Liu, Z., Liu, Y., and Wang, X. (2020). Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5911–5920.

Texture improvement for human shape estimation from a single image

Jorge González Escribano, Susana Ruano, Archana Swaminathan, David Smyth, and Aljosa Smolic

V-SENSE, Trinity College Dublin, Dublin, Ireland

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/SOWW6683

Abstract

Current human digitization techniques from a single image are showing promising results when it comes to the quality of the estimated geometry, but they often fall short when it comes to the texture of the generated 3D model, especially on the occluded side of the person, while some others do not even output a texture for the model. Our goal in this paper is to improve the predicted texture of these models without requiring any other additional input more than the original image used to generate the 3D model in the first place. For that, we propose a novel way to predict the back view of the person by including semantic and positional information that outperforms the state-of-the-art techniques. Our method is based on a general purpose image-to-image translation algorithm with conditional adversarial networks adapted to predict the back view of a human. Furthermore, we use the predicted image to improve the texture of the 3D estimated model and we provide a 3D dataset, V-Human, to train our method and also any 3D human shape estimation algorithms which use meshes such as PIFu.

Keywords: human shape estimation, neural networks, dataset

1 Introduction

Human shape estimation has been traditionally tackled with classic computer vision techniques which require multiple viewpoints as input. Methods to reduce the number of different viewpoints needed have been explored and nowadays, machine learning techniques can use a single viewpoint to predict the shape and color of a person. One of the most successful approaches makes use of implicit functions to represent a surface, which is much more efficient in terms of memory needed to store the 3D asset than others such as voxel-based ones. One of the most relevant technique which has been considered the baseline for many others is PIFu [Saito et al., 2019]. Nevertheless, although this method has been improved in many works [Saito et al., 2020, Huang et al., 2020, He et al., 2020, Hong et al., 2021], the majority of them focus on improving only the shape whereas the color and appearance of the reconstructed model is not taken into account. Deep learning techniques for tasks involving 3D data require copious amounts of- 3D data for the training of these learning-based approaches, especially if they are supervised- [Zolanvari et al., 2019, Ruano and Smolic, 2021]. The most popular datasets typically used to train learning-based 3D human reconstruction methods requiring images as input are RenderPeople and Twindom. These datasets contain 3D scans of people from different ethnicities, wearing different fashion styles with a substantial level of detail. However, these are expensive commercial datasets and therefore their accessibility is limited to companies that have the financial resources to purchase them [Zhang et al., 2021]. Consequently, researchers have created other databases of 3D human models to train deep learning methods [Zheng et al., 2019, Zhang et al., 2017,

Figure 1: Examples from V-Human dataset

 ISBN 978-0-9934207-7-1

Yu et al., 2021, Pumarola et al., 2019, Gabeur et al., 2019, Caliskan et al., 2020]. But these datasets are limited in the quality of the models and the number poses because of the effort and equipment needed for the capture and preparation of the data.

Our contribution in this paper consists of a method that improves the color prediction of 3D reconstructions methods which only needs a single image as input and does not rely on parametric models. The novelty is the use of the semantic information and UV positional to predict the back view of the person. We show how it outperforms state-of-the-art methods and we also show how it also improves commercial solutions. Furthermore, we contribute with the creation and release of a freely available synthetically generated dataset of 3D human models (samples shown in Fig. 1) called V-Human, which is used to train our method but it can also be used for training other deep learning reconstruction methods that use meshes for training.

2 Related work

Human shape and color estimation. Human shape estimation has been widely studied in the literature and classic techniques to solve 3D reconstruction problems require a huge number of viewpoints but nowadays, there are methods that use a single image as input. Initial approaches for estimating the 3D shape of a human from a single image used a parametric model of the body [Loper et al., 2015]. The main drawback of these techniques is that the model represents a naked human. Although these parametric approaches have been proven to be effective strategies for capturing motion with accurate body proportions, they cannot handle clothes and props. Other methods such as BodyNet [Varol et al., 2018] use voxel-based approaches but a well-known disadvantage of this data format is the high storage requirement necessary to capture fine details. In contrast to volumetric approaches, implicit functions are a memory efficient way to represent a surface since there is no need to store the space in which the surface is enclosed. PIFu [Saito et al., 2019] is one of the first methods that successfully reconstructed humans from a single image with a pixel-aligned implicit function strategy. They not only provide a solution for the shape but also a method to predict the colors on the surface of the geometry. However, the color estimation comes with a price not only the surface is considered but also a relatively small 3D space around it. Consequently, it has advantages for estimating color in the occluded parts but it does not allow for having sharp definition. PIFuHD [Saito et al., 2020] and ARCH [Huang et al., 2020] are extensions of this work. The former provides a more detailed reconstruction due to a multi-level architecture but no color estimation is provided. The latter produces animatable reconstructions by incorporating body semantic knowledge and takes into account the color estimation but the approach is similar to PIFu. Many other techniques build upon the PIFu baseline such as Geo-PIFu [He et al., 2020] and StereoPIFu [Hong et al., 2021] but they do not consider the color estimation. DIMNet [Zhang et al., 2021] improves PIFu's sampling strategy but it uses several views to perform feature fusion and improve the color estimation.

Training datasets. Learning-based techniques are heavily dependent on the data used for the training, and estimating 3D human shapes is especially data demanding. On the one hand, there are several commercial datasets that are used in human shape estimation papers: RenderPeople, Twindom and AXYZ. Those datasets contain a great variety of human scans with people from different ethnicities wearing a broad variety of clothes and hair styles, and acting with natural poses. Around 500 models from RenderPeople were used in PIFu [Saito et al., 2019] and PIFuHD [Saito et al., 2020]. More than double the number of scans (in particular, 1016) are used in [Zins et al., 2021]. Scans from Twindom are also used to train many algorithms [Chibane et al., 2020, Zheng et al., 2021] (1600 and 1700 models, respectively).

A notable drawback of these datasets is that they are not available for many researchers because they are commercial and financially expensive, as noted in [Zhang et al., 2021]. Consequently, the creation of freely available datasets has been of interest to the wider research community. THuman [Zheng et al., 2019] has been used in many published works [Zhang et al., 2021, He et al., 2020]. It has approximately 7000 3D scans of people with 100 different subjects. The meshes lack detail, so despite the large number of models, the accuracy of the reconstruction will be limited if this dataset is used for training. A new version, the THuman2.0 Dataset [Yu et al., 2021] was recently released and contains 500 high-resolution scans of people. In comparison with the first version, the quality is improved but the variety of the poses is drastically reduced.

 ISBN 978-0-9934207-7-1

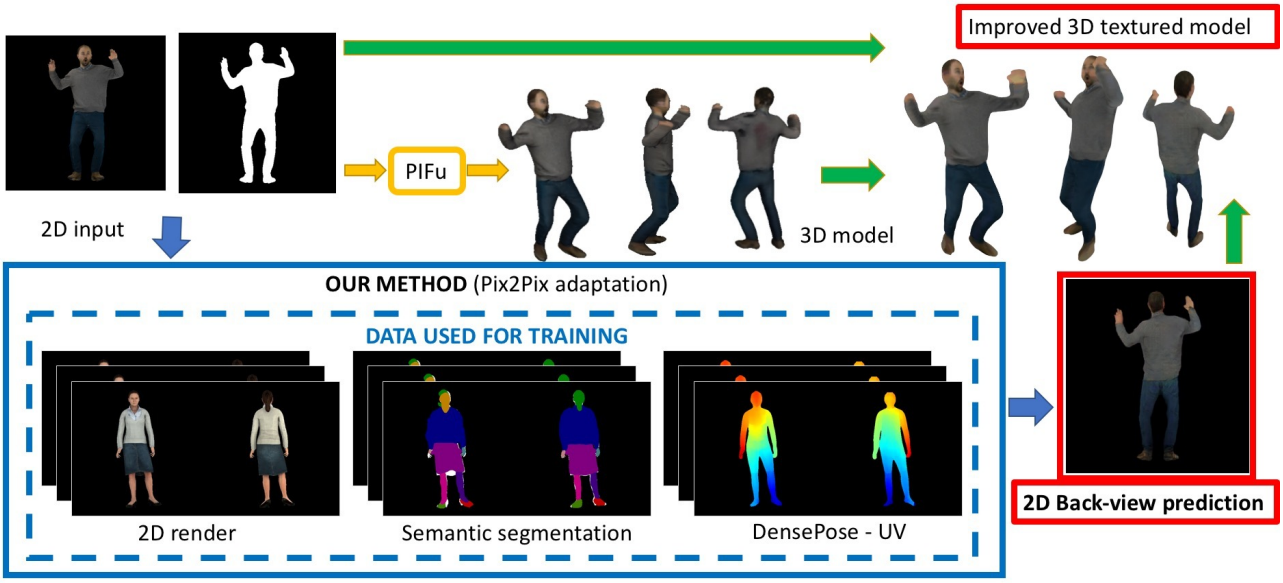


Figure 2: Pipeline overview

The MGN dataset [Bhatnagar et al., 2019] contains 96 models with segmented clothes and is also used in [Zhang et al., 2021], but they complement the training set with other models. 3D-HUMANS [Gabeur et al., 2019] also have scans of people performing different activities but it is limited to 21 subjects. 3Dpeople [Pumarola et al., 2019] has 80 people performing 70 actions but they do not share the 3D meshes due to copyright reasons. Finally, 3DVH [Caliskan et al., 2020] is a synthetic dataset but it is focused on providing a set of realistic renderings of the models with different backgrounds. The number of 3D models is not specified and is still to be released.

3 Proposed method

Our goal in this paper is to improve the color estimation of PIFu without requiring any additional information and we show an overview of our system in Figure 2. In particular, we develop a method which uses the same input as in PIFu to predict the 2D back-view of the person (in blue), and then we use this prediction along with the 3D model generated with PIFu (in yellow) to improve its texture (in green). We first describe in Section 3.1 our novel strategy to predict the 2D back-view; then, in Section 3.2 we present how to improve the 3D model texture with the predicted occluded side and finally, in Section 3.3 we describe the dataset used for training.

3.1 2D back-view prediction

Our idea is inspired by the work in [Natsume et al., 2019], where the back view of the person is predicted in the 2D space allowing for a more detail information about the clothed human. We observed that the silhouette of the person is the same in the front view and the back view, so we can apply the same idea of using an image-to-image translation method. As a difference, we increment the information used for the training. We examined the results from PIFu and observed that the parts corresponding to the back of the person were the more difficult to predict and it seemed to be related to where the extremities were located or the type of clothes they were wearing. Consequently, we thought that having some type of semantic information during the training could help to better predict the back side of the person. Also, to enforce a stronger specialization of a generic image-to-image translation algorithm into predicting the non-view side of a human, we added as input some positional information. Therefore, we can train the algorithm with the mapping between the pixels of the RGB images and the 3D surface of the body.

As the base neural model for our experiments, we chose pix2pix [Isola et al., 2017], an image-to-image translation cGAN that has been proven to perform well in a wide variety of image translation problems. The reason behind this choice of architecture is that generating the back side of a 3D model from its front side image is best framed as a conditional GAN problem, as the predicted texture is directly conditioned by the texture of the model on the opposite side, and the versatility of this network made it the best candidate. As semantic information, we use clothing segmentation data inferred by the neural network presented in [Li et al., 2020] encoded as an RGB image. More specifically, an implementation of this model trained with the ATR dataset, which includes 18 labels of different garments that the person in the input image may be wearing. To the output of the segmentation neural network we have added a mask of the silhouette of the person in the picture as the background, so in the case that the segmentation process is not able to recognise a portion of the input image it is not shown as empty. An example of the semantic input can be seen in the blue box in Figure 2 (middle). As positional information, we use the output of the DensePose neural network [Güler et al., 2018], which consists of the estimated UV coordinates (2D mapped texture coordinates) of a person shown in an image, encoded using the default 'rainbow' encoding, which makes use of the full RGB spectrum to represent this coordinates as a color image. An example of positional information can be seen in Figure 2 (right part inside the blue box).

3.2 3D texture improvement

We are able to achieve a higher quality texture on 3D models by combining the original input view, our generated back-side view and the 3D textured model output from a neural model such as PIFu. First, we perform the orthogonal projection of the input image onto the 3D model from the front, aligning the 3D model exactly with the texture. Then, we perform this same step using our generated back view texture from the exact opposite angle. After performing this step, we end up with a high resolution texture on the parts of the 3D model that are on the line of sight of the projection, but with the occluded parts showing either no texture or the texture belonging to the occluding geometry. To solve this issue, we perform occlusion detection to find those vertices which are not in the line of sight in either the front or the back view. We locate these vertices on the UV projection and find the triangles that they form on it, in order to create an occlusion mask. We use the occlusion mask to show the color-per-vertex from PIFu on the masked pixels, which correspond to those occluded by the model itself, while showing either the front or the generated back texture on the unmasked pixels. This way we can show the higher quality texture on those pixels for which it is available, while using the lower quality but available color-per-vertex on those occluded from the camera.

3.3 V-Human dataset

In order to alleviate the cost and quality issues with existing commercial and freely available datasets respectively, we have created a dataset from synthetic models suitable for training deep learning algorithms for 3D human reconstruction from images. To prepare the dataset we used the fully rigged avatars from the Microsoft Rocketbox library [Gonzalez-Franco et al., 2020], re-targeted Mixamo animations to them and refined them to make them suitable to be used as direct input for learning techniques with implicit functions. Following the aforementioned procedure we created our dataset, V-Human, which consists of a collection of 1620 models. These models were created with the 90 Microsoft Rocketbox avatars and to increase the variety of poses included in the dataset we do not always use all the frames of a single animation. Instead, we select a varying number, which is adapted depending on the pace of the action represented in the animation. Thus, we avoid having very similar poses if a particular action is almost static. Each avatar adopts 18 different poses, which makes a total of 1620 unique poses in the dataset.

4 Experiments

We design two different kind of experiments to test the performance of our texture method. The former has ground-truth associated data and the latter shows the performance in a realistic environment.

 ISBN 978-0-9934207-7-1

	V-Human		RenderPeople		Volograms	
	mean	median	mean	median	mean	median
LSGAN	76.1	**82.3**	65.0	69.5	73.7	80.8
WGAN-GP	71.8	80.6	62.0	69.1	72.8	79.4
128 filters	75.3	81.5	60.2	68.0	70.4	79.1
PatchGAN 9	**76.2**	81.5	**67.4**	**74.4**	**76.8**	**81.6**
PIFu	52.8	56.0	54.1	57.2	58.2	63.1
PIFu retrained	50.5	51.7	42.0	39.3	48.4	53.2

Table 1: Results of segmentation experiment in % of correctly classified pixels (4000 epochs)

Experiments with ground-truth. First, we evaluate the quality of the estimated back-side as an image and we compare ourselves with PIFu and PIFu trained with V-Human (18 epochs the shape training and 6 epochs color training). Furthermore, we perform an ablation study to fine-tune our pix2pix model by using four different variants: the first one uses the default parameters in pix2pix (LSGAN), the second one makes use of the WGAN-GP [Gulrajani et al., 2017] loss instead of LSGAN (WGAN-GP), the third one has 128 filters in the last layer of both the generative and the discriminator models instead of 64 (128 filters), and the fourth one changes the number of layers in the PatchGAN discriminator from 3 to 9 (PatchGAN). We train the models for 4000 epochs.

Evaluation metric. As quality metric, inspired by [Isola et al., 2017], we use a clothing segmentation model to classify each pixel of the ground-truth and the predicted image. Then, we calculate the percentage of pixels correctly classified, which are the ones that belong to the same feature type (eg., shirt, trousers, hat, skin - no clothing...) in the predicted and ground-truth images. We discard the background pixels.

Training and test sets. On the one hand, as training set we use 1458 models from V-Human (90%) and we leave a 10% aside for testing purposes as it is done in [Saito et al., 2019, Saito et al., 2020]. Those ones correspond to 81 different identities with 18 different poses. On the other hand, for testing we use data from three different sources: our V-Human dataset, RenderPeople and Volograms. The nine subjects left from the complete V-Human dataset were used to create the test set, which has 162 models with 18 different poses per subject. We used nine rigged RenderPeople avatars and created a dataset of 162 posed models following a similar pipeline with Mixamo animations. Finally, we used 162 models from nine different volumetric video sequences captured by Volograms using their studio technology [Pagés et al., 2021].

Experiments in the wild. We also explore how our method performs in a realistic environment. For that, we apply our 2D back-view prediction to images captured with a smartphone. This situation is different from the previous experiments because those ones are weak-perspective renderings of 3D models and not images from a standard camera. Furthermore, we qualitatively compare the result with the back view of the 3D texture model created with Volograms' mobile technology[1].

5 Results

The results from the image segmentation experiments are shown in Table 1. We can see that all of our models used in the ablation study outperform PIFu consistently, with the best one achieving 24% better mean and 26% median accuracy than PIFu. From our four models tested, we can see that the one using 9 layers on a PatchGAN discriminator (PatchGAN 9) has the best results in all of the datasets.

In the left part of Figure 3 we show a closeup of a texture ground truth and its corresponding predictions made by PIFu and our method. In it we can clearly see how PIFu produces no wrinkles while our method does, increasing the perceived quality, although they do not match with those from the ground truth. Clothing wrinkles can be seen as a type of "noise" because it is very unlikely that estimated and ground truth wrinkles match, therefore PSNR and SSIM are not adequate as evaluation metrics. Nevertheless, the wrinkles on the

[1] www.volograms.com

Figure 3: Texture results comparison. On the left, wrinkle texture closeup: groundtruth, PIFu, ours. On the right, example results comparison: PIFu (top row), ours (bottom row)

Figure 4: 3D Result Example. In order from left to right: our texture, PIFu texture, combined texture, PIFu back view, ours back view

clothes give it a more textile and realistic appearance, meanwhile the texture of the clothes generated by PIFu looks a little more like modeling clay, as it does not preserve high frequency details. On the right part of this image, some more examples of comparison between PIFu and our method can be seen.

Regarding the results of applying our full pipeline to improve 3D model textures, Figure 4 shows a 3D model who is holding their hand in front of the chest, occluding some of the clothing, from a perspective that shows this occluded part. The first image on the left shows the high resolution texture generated using our approach, but with the hand texture filtering onto the occluded part of the clothing. In the next image, the 3D model is textured using the output from PIFu, which has some noticeable artifacts, and right next to it is a comparison with our method without them.

Results in the wild. We show in Figure 5 an example of the experiments done with images captured with a smartphone. In particular, Figure 5a shows the input of our system which is a picture taken with a smartphone with the background removed and Figure 5b the output, our predicted back view. We do not have the information of how the person looks like from the back but, as we can observe, the predicted back view is quite plausible. In particular, we can see that the prediction is very good for the jeans, because it creates some wrinkles which are pretty credible. Furthermore, we can compare the prediction with the results obtained with vologramsʼ mobile technology. Although their 3D predicted model is very sharp in the front-view, the back part, shown in Figure 5c, is not. We can see that our 2D prediction is much more detailed, so potentially, it can be use to improve the view.

 ISBN 978-0-9934207-7-1

6 Conclusion

In this paper we have presented a new method for improving the texture of 3D models predicted from a single image of a person. We presented an strategy to predict the 2D back-view of a person which includes semantic information about the person and its clothes along with positional information. Moreover, we show how this result can be incorporated in the 3D model output of available systems, demonstrating how we can improve the state-of-the-art solutions by generating a sharper prediction. Furthermore, we provide an open-source dataset to train the model which is also suitable for training deep learning algorithms that uses 3D meshes as input. Finally, we show how our method helps to improve existing commercial solutions in natural environments.

(a) (b) (c)

Figure 5: **Qualitative results in the wild.**(a) input image, (b) 2D prediction with our method, (c) back of the 3D model with vologram's mobile technology

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776. We thank Vologrames for providing their data.

References

[Bhatnagar et al., 2019] Bhatnagar, B. L., Tiwari, G., Theobalt, C., and Pons-Moll, G. (2019). Multi-garment net: Learning to dress 3d people from images. In *ICCV*, pages 5420–5430.

[Caliskan et al., 2020] Caliskan, A., Mustafa, A., Imre, E., and Hilton, A. (2020). Multi-view consistency loss for improved single-image 3d reconstruction of clothed people. In *ACCV*.

[Chibane et al., 2020] Chibane, J., Alldieck, T., and Pons-Moll, G. (2020). Implicit functions in feature space for 3d shape reconstruction and completion. In *CVPR*, pages 6970–6981.

[Gabeur et al., 2019] Gabeur, V., Franco, J.-S., Martin, X., Schmid, C., and Rogez, G. (2019). Moulding humans: Non-parametric 3d human shape estimation from single images. In *ICCV*, pages 2232–2241.

[Gonzalez-Franco et al., 2020] Gonzalez-Franco, M., Ofek, E., Pan, Y., Antley, A., Steed, A., Spanlang, B., Maselli, A., Banakou, D., Pelechano Gómez, N., Orts-Escolano, S., et al. (2020). The rocketbox library and the utility of freely available rigged avatars. *Frontiers in virtual reality*, 1(article 561558):1–23.

[Güler et al., 2018] Güler, R. A., Neverova, N., and Kokkinos, I. (2018). Densepose: Dense human pose estimation in the wild. In *CVPR*, pages 7297–7306.

[Gulrajani et al., 2017] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. *CoRR*, abs/1704.00028.

[He et al., 2020] He, T., Collomosse, J., Jin, H., and Soatto, S. (2020). Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *arXiv preprint arXiv:2006.08072*.

[Hong et al., 2021] Hong, Y., Zhang, J., Jiang, B., Guo, Y., Liu, L., and Bao, H. (2021). Stereopifu: Depth aware clothed human digitization via stereo vision. In *CVPR*, pages 535–545.

[Huang et al., 2020] Huang, Z., Xu, Y., Lassner, C., Li, H., and Tung, T. (2020). Arch: Animatable reconstruction of clothed humans. In *CVPR*, pages 3093–3102.

[Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134.

[Li et al., 2020] Li, P., Xu, Y., Wei, Y., and Yang, Y. (2020). Self-correction for human parsing. *TPAMI*.

[Loper et al., 2015] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16.

[Natsume et al., 2019] Natsume, R., Saito, S., Huang, Z., Chen, W., Ma, C., Li, H., and Morishima, S. (2019). Siclope: Silhouette-based clothed people. In *CVPR*, pages 4480–4490.

[Pagés et al., 2021] Pagés, R., Zerman, E., Amplianitis, K., Ondřej, J., and Smolic, A. (2021). Volograms & V-SENSE Volumetric Video Dataset. *ISO/IEC JTC1/SC29/WG07 MPEG2021/m56767*.

[Pumarola et al., 2019] Pumarola, A., Sanchez-Riera, J., Choi, G., Sanfeliu, A., and Moreno-Noguer, F. (2019). 3dpeople: Modeling the geometry of dressed humans. In *ICCV*, pages 2242–2251.

[Ruano and Smolic, 2021] Ruano, S. and Smolic, A. (2021). A benchmark for 3D reconstruction from aerial imagery in an urban environment. In *VISIGRAPP (5: VISAPP)*, pages 732–741.

[Saito et al., 2019] Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., and Li, H. (2019). Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, pages 2304–2314.

[Saito et al., 2020] Saito, S., Simon, T., Saragih, J., and Joo, H. (2020). Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, pages 84–93.

[Varol et al., 2018] Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., and Schmid, C. (2018). Bodynet: Volumetric inference of 3d human body shapes. In *ECCV*, pages 20–36.

[Yu et al., 2021] Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., and Liu, Y. (2021). Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *ICCV*.

[Zhang et al., 2017] Zhang, C., Pujades, S., Black, M. J., and Pons-Moll, G. (2017). Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *CVPR*.

[Zhang et al., 2021] Zhang, S., Liu, J., Liu, Y., and Ling, N. (2021). Dimnet: Dense implicit function network for 3D human body reconstruction. *Computers & Graphics*.

[Zheng et al., 2021] Zheng, Y., Shao, R., Zhang, Y., Yu, T., Zheng, Z., Dai, Q., and Liu, Y. (2021). Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. *arXiv preprint arXiv:2105.00261*.

[Zheng et al., 2019] Zheng, Z., Yu, T., Wei, Y., Dai, Q., and Liu, Y. (2019). Deephuman: 3d human reconstruction from a single image. In *ICCV*, pages 7739–7749.

[Zins et al., 2021] Zins, P., Xu, Y., Boyer, E., Wuhrer, S., and Tung, T. (2021). Learning implicit 3d representations of dressed humans from sparse views. *arXiv preprint arXiv:2104.08013*.

[Zolanvari et al., 2019] Zolanvari, S. I., Ruano, S., Rana, A., Cummins, A., da Silva, R. E., Rahbar, M., and Smolic, A. (2019). Dublincity: Annotated lidar point cloud and its applications. In *BMVC*.

 ISBN 978-0-9934207-7-1

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/AZWK8552

Box Supervised Video Segmentation Proposal Network

Tanveer Hannan*[1], Rajat Koner*[1], Jonathan Kobold[2], and Matthias Schubert[1]

[1] *Ludwig Maximilian University of Munich, Germany*
[2] *Hensoldt Analytics GmbH, Vienna, Austria*

Abstract

Bounding box supervision provides a balanced compromise between labeling effort and result quality for image segmentation. However, there exists no such work explicitly tailored for videos. Applying the image segmentation methods directly to videos produces sub-optimal solutions because they do not exploit the temporal information. In this work, we propose a box-supervised video segmentation proposal network. We take advantage of intrinsic video properties by introducing a novel box-guided motion calculation pipeline and a motion-aware affinity loss. As the motion is utilized only during training, the run-time remains fixed during inference time. We evaluate our model on Video Object Segmentation (VOS) challenge. The method outperforms the state-of-the-art self-supervised methods by 16.4% and 6.9% $\mathcal{J}\&\mathcal{F}$ score, and the majority of fully supervised ones on the DAVIS and Youtube-VOS dataset. Code is available at https://github.com/Tanveer81/BoxVOS.git.

Keywords: Video Object Segmentation, Motion Calculation, Pseudo Labelling

1 Introduction

Video Object Segmentation (VOS) primarily segments objects and tracks them throughout the video sequence. The two prior approaches for solving VOS are fully-supervise [Xie et al., 2021] using expensive mask annotation and self-supervise [Vondrick et al., 2018], which utilizes intrinsic video properties. However, the latter suffers from a significant performance drop compared to fully supervised once. This paper tries to bridge the gap by only using box annotation. To the best of our knowledge, we are the first to propose a box-supervised video segmentation proposal network. The central idea is to distinguish similar regions inside a box based on motion and color similarity. Foreground objects in a video constantly change their appearance or are camouflaged by the background. Thus, separating them using only color information can lead to a sub-optimal solution. In contrast to color, object motion in a video is an independent yet complementary feature that can help to distinguish the objects from the background. Figure 1 shows a panda sitting in a bamboo field, where its facial fur is camouflaged, i.e., it has a similar color compared to the wooden background stem. It poses a considerable challenge to distinguish the object pixel from the background solely based on color similarity. We introduce two new components to segregate the foreground from the background without an annotated mask. At first, we employed a train time bounding box guided motion calculation pipeline. Second, we propose a novel motion-aware affinity loss for generated pseudo masks, i.e., an approximation of the ground truth masks derived from the combined color and motion as shown in figure 1. Motion computation is inherently noisy due to global camera and other minute movements. To mitigate this, our motion pipeline uses an affine transformation on the current frame for alignment of the backgrounds of subsequent frames. Moreover, the pixels outside the bounding box are exploited for better background characteristics and transformation matrix computation. In addition, suppressing pixels outside a box ensures foreground and moving objects are the primary sources of motion response. Afterward, the generated motion is fused with color to create a pseudo mask for supervision. Each pixel pair located inside

*Equal contribution. Correspondence to: Tanveer Hannan <hannan@dbs.ifi.lmu.de>

 ISBN 978-0-9934207-7-1

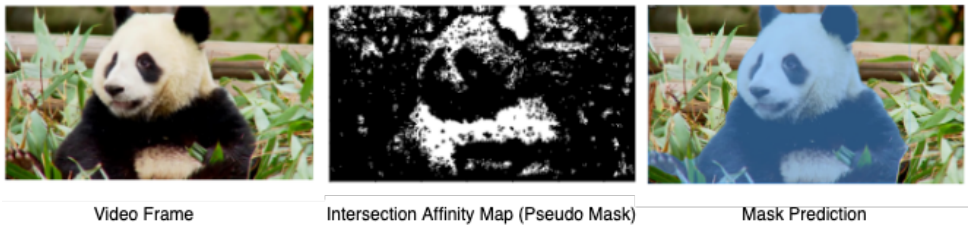


Figure 1: The pseudo mask is created where pixel pairs share similar motion and color characteristics. The intersection of color and motion improves performance by reducing background noise and generating a sparse pseudo mask.

a box are compared to its surroundings. Finally, a pixel pair with similar motion and color characteristics are considered to be positively affined. It helps to compute a tight and precise mask for a perspective foreground object. The contribution of our work can be summarized as follows:

- We are the first to propose a video segmentation proposal network that employs bounding box supervision and is compatible with most of the existing VOS frameworks.
- An improved motion compensation technique is proposed that effectively utilizes the available bounding box coordinates for reducing global camera movement.
- A novel motion-aware affinity loss is introduced for frame-level segmentation with a pseudo mask generated from pair-wise color and motion similarity.
- Our method achieved absolute 16.4% and 6.9% J&F improvements on the DAVIS and Youtube-VOS datasets compared to self-supervised methods.

2 Related Work

Box Supervised Image Instance Segmentation The state-of-the-art box supervised instance segmentation methods are BoxInst [Tian et al., 2021] and DiscoBox [Lan et al., 2021]. BoxInst clusters nearby pixels based on the color similarity and forces the neural network to predict the similar class for the similar color pixels. The pairwise loss and projection loss aid the network in learning segmentation in a weakly supervised manner. DiscoBox, on the other hand, generates pseudo labels through multi-instance learning. However, DiscoBox is not directly applicable to VOS as it requires class-level labels to train the correspondence matching. This paper explores box supervision on video data, allowing our model to exploit inter-frame dependency for VOS.

Motion Generation The motion map in a video is calculated by subtracting the background from two consecutive frames [Ellenfeld et al., 2021]. In the case of global camera movement, motion compensation [Hartley and Zisserman, 2003] is used. A forward-backward sparse optical flow [Wan et al., 2014] or three-frame difference [Sommer et al., 2016] is used for further refinement of motion. Our method employs multiple filters and incorporates bounding box coordinates to improve the motion.

Segmentation using motion Existing works utilize optical flow to approximate the motion of objects. For example, FlowIRN [Liu et al., 2021] uses class activation maps and dense optical flow to generate pseudo supervision. In Motion Grouping [Yang et al., 2021], dense optical flow is used to cluster pixels into foreground and background. In Deep Fusion[Ellenfeld et al., 2021], a motion map calculated from the sparse optical flow is used as network input. Dense optical flow is used to detect camouflaged animals in [Lamdouar et al., 2020]. The optical flow-generated motion map is used as input in the mentioned works. In contrast, our work uses motion only during training time as supervision.

Semi-Supervised VOS is a recent challenge introduced by DAVIS [Pont-Tuset et al., 2017] and adopted by Youtube-VOS [Xu et al., 2018a]. The term “semi-supervised" does not refer to the level of supervision used during training. Instead, it only refers to the availability of annotation masks for the first frames of test videos.

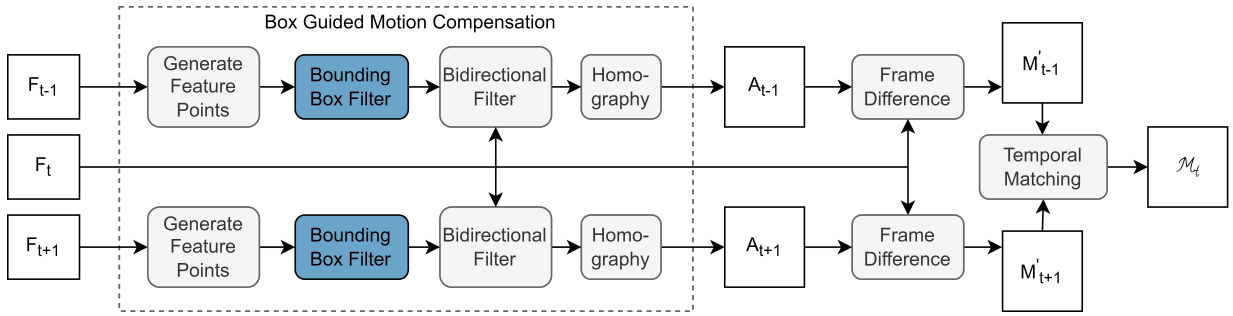


Figure 2: Proposed box supervised motion calculation pipeline with affine transformation.

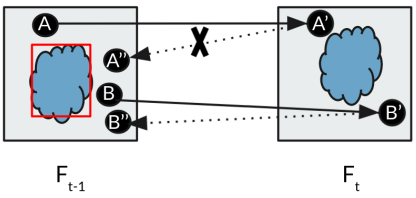


Figure 3: Box and Bidirectional Filters. Bounding box is colored red.

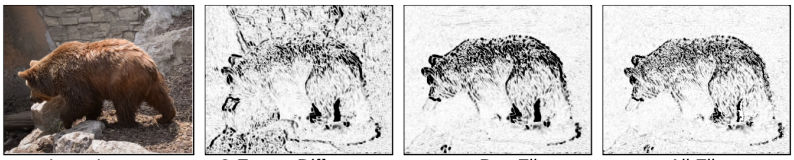


Figure 4: Improvement of motion map with additional filters.

The challenge is to generate segmentation in the subsequent frames from those given in the first frame. Some methods [Xie et al., 2021, Oh et al., 2019, Luiten et al., 2018] solve the task in a fully-supervised setup, where segmentation masks are used for supervision. On the contrary, [Vondrick et al., 2018, Lai and Xie, 2019, Lu et al., 2020] trained their models in a self-supervised way to generate segmentation.

3 Method

3.1 Proposed Motion Calculation Pipeline

Box Guided Motion Compensation: Videos are often captured or recorded with a moving camera, which introduces an inherent noisy background motion in addition to foreground objects. This erroneous background motion can be reduced if the previous and current frames are co-registered. At time t the previous frame F_{t-1} is aligned to the current frame F_t with a homography matrix H. The homography matrix H is the affine transformation between the point pairs from F_{t-1} and F_t. We calculate H by sampling feature points from F_{t-1} and F_t, which are matched with the RANSAC algorithm. Our feature point extraction is similar to [Shi and Tomasi, 1994]. Notably, this co-registration or background alignment works well if these points come only from the background, not the foreground. Hence, we utilize the foreground box information using a Box Filter(BF) to sample feature points from outside the box. However, annotations usually only comprise primary foreground objects. Thus remaining objects and background clutter remain a big source of motion. The projection error for a pixel propagating through forward and backward projection is used for clutter detection. A Bi-Directional Filter (BDF) uses this forward-backward projection to refine feature points. Figure 3, depicts our BDF, where A and B are two points from frame F_{t-1} along with their forward $\{A^{'}, B^{'}\}$ and backward $\{A^{''}, B^{''}\}$ projection obtained from sparse optical flow[Lucas and Kanade, 1981]. Afterwards, we remove points that have an L2 distance D_A between the original point (i.e., A) and backward projected point (i.e.,$A^{''}$) greater than τ where, $\tau = median(\{D_P | P \in \text{feature points}\})$. Finally, the remaining points are more likely from the background, which are suitable for calculating the homography matrix H, as shown in Figure 2. The final motion between two consecutive frames is obtained as $M'_{t-1} = |F_t - A_{t-1}|$ corresponds to a background subtraction. Figure 4 shows the improvement of the motion map for moving objects with the proposed filters.

 ISBN 978-0-9934207-7-1

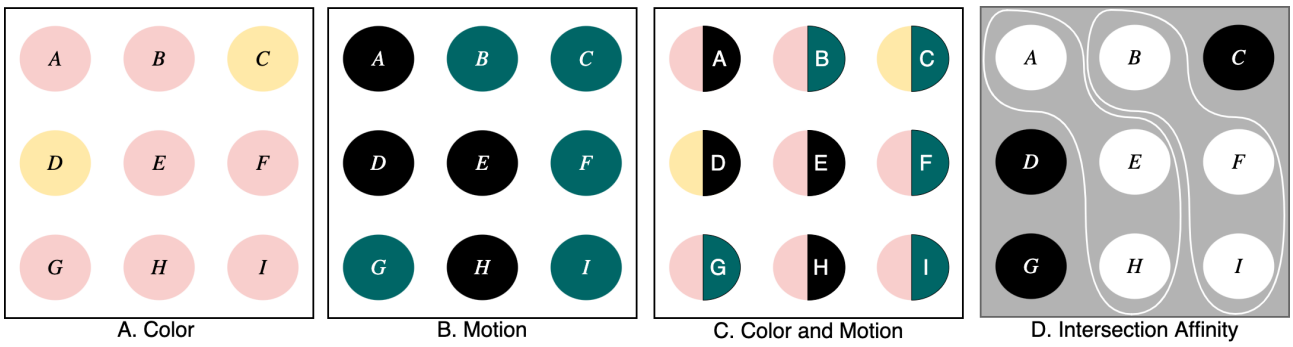


Figure 5: An illustration of the affinity loss in a 3x3 pixel grid. In Figure A and B, pixels with a similar color/motion are indicated by the same color. Figure C combines the color(left) and motion(right). Figure D shows the final affinity map. Pixels with the same color/motion are marked by white boundaries.

Temporal Matching: Motion calculated only from two consecutive frames often suffers from indiscriminate movement and erroneous optical flow, thus resulting in a low-quality motion map. We incorporate motion from both forward and backward temporal directions to alleviate these problems and merge them. Here, the next frame's motion map ($M^{'}_{t+1}$) from F_t to F_{t+1} is computed similarly to the earlier two consecutive frames difference. Finally, the motion map M_t with temporal bidirectional matching is obtained with the equation. 1, where τ_2 is a predefined threshold.

$$M_t = \begin{cases} M^{'}_{t-1} & \text{if } M^{'}_{t-1} - M^{'}_{t+1} < \tau_2 \\ 0, & \text{else} \end{cases} \tag{1}$$

3.2 Motion Aware Affinity loss

The core idea of this work relies on the generation of a pseudo mask for the foreground objects and associated affinity loss combining the color and motion information. Figure 5 shows our proposed loss in a local neighborhood of 3×3. Consider pixels A and B and their corresponding motion values are M_A and M_B. C_A and C_B are their color values. The motion and color similarities between these pixels are $\psi_{A,B} \in [0,1]$, and $\phi_{A,B} \in [0,1]$ and η is a hyper-parameter. For a pair of pixels, $\Psi_{A,B}$ and $\Phi_{A,B}$ are the motion and color affinity. τ_m and τ_c are the motion and color similarity thresholds. The proposed loss is calculated only for pixel pairs with high similarity, i.e., positive affinity. It essentially means the network is trained to predict segmentation with the correct pixel pairs. The affinity between two pixel (A,B) is defined with $I_{A,B}$. Let the confidence score of the segmentation network for pixels A and B being foreground be P_A and P_B. Then the confidence of these two pixels being predicted as the same class (either foreground or background) is $P_{A,B}$, where the confidence of the pixel pair being foreground is $P_A * P_B$, or being background is $(1-P_A) * (1-P_B)$. Let S be the set of pixel pairs where at least one pixel falls inside the bounding box. For a pixel pair A and B, our motion aware affinity loss, $L_{A,B}$ and for all pairs of pixels S, the loss is $L_{affinity}$.

$$\psi_{A,B} = exp(-\|M_A - M_B\| * \eta); \quad \phi_{A,B} = exp(-\|C_A - C_B\| * \eta) \tag{2}$$

$$\Psi_{A,B} = \begin{cases} 1 & \psi_{A,B} > \tau_m \\ 0 & \text{else} \end{cases}; \quad \Phi_{A,B} = \begin{cases} 1 & \phi_{A,B} > \tau_c \\ 0 & \text{else} \end{cases}; \quad I_{A,B} = \begin{cases} 1 & \text{if } \Psi_{A,B} = \Phi_{A,B} \\ 0, & \text{else} \end{cases} \tag{3}$$

$$P_{A,B} = P_A * P_B + (1-P_A) * (1-P_B); \quad L_{A,B} = -I_{A,B} * log(P_{A,B}); \quad L_{affinity} = \sum_{A,B \in S} L_{A,B} \tag{4}$$

3.3 Network Architecture

CondInst[Tian et al., 2020] is our default segmentation network. The fully supervised mask loss of CondInst is replaced with a projection loss[Tian et al., 2021] and the proposed motion-aware affinity loss. For VOS

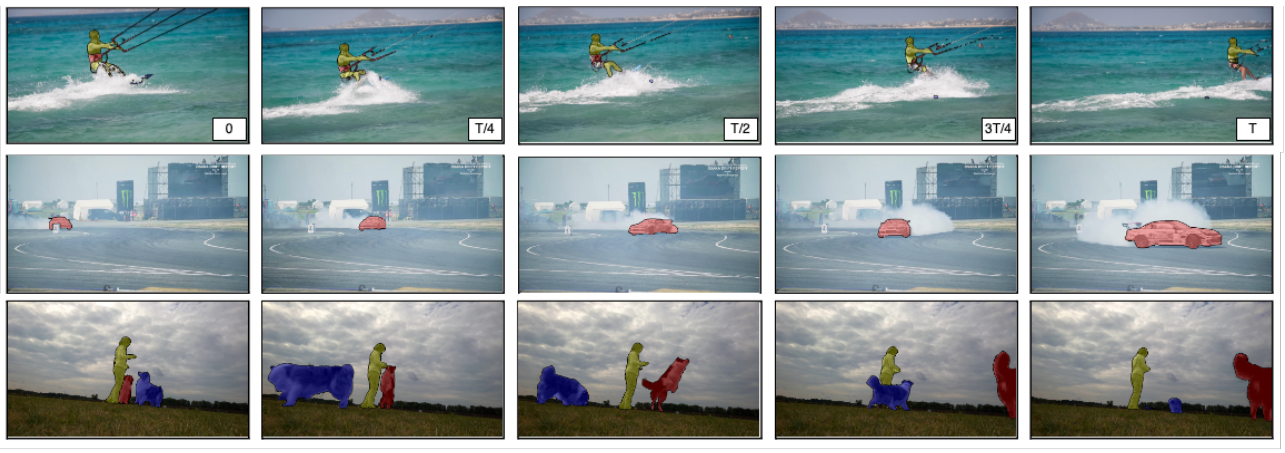


Figure 6: Predictions of our method on the DAVIS dataset. Due to semi-supervised evaluation, the first frame (0th) represents the ground truth annotation mask.

evaluation the segmented instances need to have a continuous ID over time. The segmentations are tracked with optical flow [Ilg et al., 2017] and object re-identification [Li et al., 2017] network. The ground-truth proposals from the first frame are matched with the next frame proposals using a bipartite matching mechanism. This matching utilizes a score from IoU overlap of predicted segmentation masks wrapped with the optical flow and re-identification score. To learn more about the tracker's implementation, we recommend reading PReMVOS[Luiten et al., 2018]. We use segmentation masks instead of boxes to calculate IoU for the tracker.

4 Experiments

4.1 Dataset and Evaluation Metrics

DAVIS-2017 is one of the most widely used VOS datasets where every frame is annotated with a ground truth mask. A predefined training and validation split consist of 60 and 30 videos.
Youtube-VOS is a recently proposed one of the largest video object segmentation datasets. It contains 3,471 training and 507 validation videos with 94 different object categories.
Dataset Preparation is done through generating frame-wise COCO [Lin et al., 2014] style annotations for both YoutubeVOS [Xu et al., 2018a] and DAVIS [Pont-Tuset et al., 2017] datasets, and trained for individuals frames. The bounding box is computed from the ground truth segmentation mask as no separate box information is available. The ground truth mask is not used for any other training purposes.
$\mathcal{J}$ **&** $\mathcal{F}$ **Score** The Jaccard index $\mathcal{J}$ is the intersection of prediction and ground truth mask over their union. It indicates the region accuracy of the prediction. The contour accuracy $\mathcal{F}$ calculates the f1-score of the boundary pixels with a bipartite matching algorithm. The mean of these $\mathcal{J}$ & $\mathcal{F}$ is reported as in the DAVIS 2017 [Pont-Tuset et al., 2017], for both DAVIS and Youtube-VOS.
mAP is the mean average precision is the area under the precision-recall curve where the IoU threshold determines the positive and negative examples. We use the implementation from COCO [Lin et al., 2014] for the frame-level evaluation in the ablation study.

4.2 Training Procedure

We have implemented our method on the detectron2 [Wu et al., 2019] or more specifically its adapted version, AdelaiDet [Tian et al., 2019] framework. Our segmentation network is trained on the YoutubeVOS and then finetuned on the DAVIS. Our model shares the similar configuration to CondInst [Tian et al., 2020] with some minute changes. A more detailed configuration can be found in our code. All the training parameters are determined by manual hyper-parameter tuning.
YoutubeVOS training: Our segmentation network train on the YoutubeVOS using the SGD with a gradual

Method	Sup.	PT	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Vid. Color.[Vondrick et al., 2018]	Self	✓	34.0	34.6	32.7
CorrFlow[Lai and Xie, 2019]	Self	✓	50.3	48.4	52.2
Mug[Lu et al., 2020]	Self		56.1	54.0	58.2
UVC[Li et al., 2019]	Self	✓	59.5	57.7	61.3
BoxInst[Tian et al., 2021] †	Box	✓	68.9	68.2	69.6
Ours	**Box**	✓	**72.5**	**71.5**	**73.5**
OSMN[Yang et al., 2018]	Full	✓	54.8	52.5	57.1
OSVOS[Caelles et al., 2017]	Full	✓	60.3	56.6	63.9
SiamMask[Wang et al., 2018]	Full		56.4	54.3	58.5
OSVOS-S[Maninis et al., 2018]	Full		68.0	64.7	71.3
GC[Li et al., 2020]	Full		71.4	69.3	73.5
FEELVOS[Voigtlaender et al., 2019]	Full	✓	71.5	69.1	74.0
PReMVOS[Luiten et al., 2018]	Full	✓	77.8	73.9	81.7
STM[Oh et al., 2019]	Full	✓	81.8	79.2	84.3
RMNet[Xie et al., 2021]	**Full**	✓	**83.5**	**81.0**	**86.0**

Table 1: State-of-the-art comparison on DAVIS [Pont-Tuset et al., 2017] validation set. †: baseline method which uses only color. PT: pretrain on other datasets.

warmup of 10000 iterations and a batch size of 12. The base learning rate is set to 0.01, and further, it reduces by 0.1 factor after 60000 and 80000 iterations. Our model is trained till 100K iterations with three Quadro RTX 8000 and the ResNet-101 as a backbone feature.
DAVIS Fine-tuning. For finetuning on the DAVIS, we set the learning rate very low at 0.001. We evaluate our model after 350 iterations which constitute one epoch, and report our best model at epoch number 18.

4.3 Results

Comparison with State-of-the-Art VOS methods, Table 1, and 2 show that our method outperforms all the self-supervised methods on both DAVIS and YouTube-VOS. Furthermore, we demonstrate superior performance compared to the most fully supervised approaches and achieved a competitive performance with the top-performing models. Figure 6 shows some qualitative examples.
Comparison with Box Supervised Color only Baseline To demonstrate the importance of motion utilization, we compare our method with the state-of-the-art box supervised image segmentation network BoxInst [Tian et al., 2021]. For a fair comparison, our network architecture and other parameters remain the same as BoxInst and are trained on the same dataset.

4.4 Ablation Studies

Unless otherwise specified, the ablation studies use the ResNet-50 backbone and DAVIS dataset. Table 3, shows the improvement and influence due to different motion filters and their combination as discussed in Section 3.1. Table 4 shows how the combination of color and motion during supervision lead to superior results as stated in Section 3.2. Here, the third row is the intersection affinity which is replaced with a union operation in the first row's experiment. Table 5, analyzes the type of input backbone and data-sets used during training. Table 6, shows contribution of filters in extracting high-precision foreground pixels. Ground truth segmentation masks are used here for evaluation only. However, ground truth masks have never been used in training or performing filter optimization. Pixels greater than the threshold value of 0.5 are labeled as foreground, with the motion being normalized between 0 and 1. Finally, to understand image segmentation quality, we report frame-wise mAP in table 7. For Youtube-VOS, we hold out 20% training data as validation masks are not publicly available. The split contains roughly the same category distribution as the training set.

 ISBN 978-0-9934207-7-1

Method	Sup.	Overall ↑	Seen		Unseen	
			$\mathcal{J}$ ↑	$\mathcal{F}$ ↑	$\mathcal{J}$ ↑	$\mathcal{F}$ ↑
Vid. Color.[Vondrick et al., 2018]	Self	38.9	43.1	38.6	36.6	37.4
CorrFlow[Lai and Xie, 2019]	Self	46.6	50.6	46.6	43.8	45.6
BoxInst[Tian et al., 2021] †	Box	48.2	51.6	52.3	44.7.8	44.9
Ours	**Box**	**53.5**	**58.7**	**59.2**	**46.3**	**48.4**
OSMN[Yang et al., 2018]	Full	51.2	60.0	60.1	40.6	44.0
MSK[Khoreva et al., 2016]	Full	53.1	59.9	59.5	45.0	47.9
RGMP[Oh et al., 2018]	Full	53.8	59.5	–	45.2	–
OnAVOS[Voigtlaender and Leibe, 2017]	Full	55.2	60.1	62.7	46.6	51.4
RVOS[Ventura et al., 2019]	Full	56.8	63.6	67.2	45.5	51.0
OSVOS[Caelles et al., 2017]	Full	58.8	59.8	60.5	54.2	60.7
S2S[Xu et al., 2018b]	Full	64.4	71.0	70.0	55.5	61.2
PReMVOS[Luiten et al., 2018]	Full	66.9	71.4	75.9	56.5	63.7
STM[Oh et al., 2019]	Full	79.4	79.7	84.2	72.8	80.9
RMNet[Xie et al., 2021]	**Full**	**81.5**	**82.1**	**85.7**	**75.7**	**82.4**

Table 2: State-of-the-art comparison on Youtube-VOS[Xu et al., 2018a]. † denotes our baseline.

Filters	$\mathcal{J}\&\mathcal{F}$
None	68.5
BF*	69
BF* + TM	69.5
BF* + BDF + TM	70.5

Table 3: Impact of the filters. BF*: Box Filter, BDF: Bi-Directional Filter, TM: Temporal matching

Supervision	$\mathcal{J}\&\mathcal{F}$
Motion Only	69.4
Color ∪ Motion	68.8
Color ∩ Motion	70.5

Table 4: Effect of combined motion and color utilization on pseudo mask generation.

Backbone	T	$\mathcal{J}\&\mathcal{F}$
ResNet-50	D	70.5
ResNet-50	DY	72.1
ResNet-101	**DY**	**72.5**

Table 5: Effect of various backbones and training data. T: Train Data, D: DAVIS, Y: YVOS

Filters	F1	Precision	Recall	Avg. FPS
None	0.404	0.318	0.553	**10**
BF*	**0.482**	0.397	**0.612**	4
BF* + TM + BDF	0.460	**0.421**	0.507	2

BF*: Box Filter, BDF: Bi-Directional Filter, TM: Temporal matching

Table 6: Proposed filters' capability to extract foreground pixels for DAVIS trainval data-set.

Sup	Method	mAP	
		YVOS-train-val	DAVIS-val
Box	BoxInst [Tian et al., 2021] †	31.3	24.2
Box	Ours	34	28.4
Full	BoxInst with Mask Annotation †	41.8	35.2

Table 7: Performance evaluation on frame level segmentation proposal generation. † denotes that network is only supervised with only color information.

5 Conclusion

We are the first to explore the potential of incorporating motion with box supervision for video segmentation proposal generation. Our network has demonstrated that exploring intrinsic video cues like motion with weak box supervision significantly boosts pseudo mask generation and subsequent improvement on segmentation proposal. It helps our network to outperform all self-supervised methods. Furthermore, it also significantly reduces the performance gap with top-performing, fully supervised methods. We sincerely hope our weak box supervised VOS framework will pave the way for new video segmentation research.

6 Acknowledgments

This work was funded by Hensoldt Analytics GmbH.

References

[Caelles et al., 2017] Caelles, S., Maninis, K.-K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., and Van Gool, L. (2017). One-shot video object segmentation. In *Computer Vision and Pattern Recognition (CVPR)*.

[Ellenfeld et al., 2021] Ellenfeld, M., Moosbauer, S., Cardenes, R., Klauck, U., and Teutsch, M. (2021). Deep fusion of appearance and frame differencing for motion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4339–4349.

[Hartley and Zisserman, 2003] Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge university press.

[Ilg et al., 2017] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470.

[Khoreva et al., 2016] Khoreva, A., Perazzi, F., Benenson, R., Schiele, B., and Sorkine-Hornung, A. (2016). Learning video object segmentation from static images. *CoRR*, abs/1612.02646.

[Lai and Xie, 2019] Lai, Z. and Xie, W. (2019). Self-supervised learning for video correspondence flow. *CoRR*, abs/1905.00875.

[Lamdouar et al., 2020] Lamdouar, H., Yang, C., Xie, W., and Zisserman, A. (2020). Betrayed by motion: Camouflaged object discovery via motion segmentation. In *Proceedings of the Asian Conference on Computer Vision*.

[Lan et al., 2021] Lan, S., Yu, Z., Choy, C., Radhakrishnan, S., Liu, G., Zhu, Y., Davis, L. S., and Anandkumar, A. (2021). Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3406–3416.

[Li et al., 2019] Li, X., Liu, S., De Mello, S., Wang, X., Kautz, J., and Yang, M.-H. (2019). Joint-task self-supervised learning for temporal correspondence. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Li et al., 2017] Li, X., Qi, Y., Wang, Z., Chen, K., Liu, Z., Shi, J., Luo, P., Tang, X., and Loy, C. C. (2017). Video object segmentation with re-identification. *arXiv preprint arXiv:1708.00197*.

[Li et al., 2020] Li, Y., Shen, Z., and Shan, Y. (2020). Fast video object segmentation using the global context module. *CoRR*, abs/2001.11243.

[Lin et al., 2014] Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

[Liu et al., 2021] Liu, Q., Ramanathan, V., Mahajan, D., Yuille, A., and Yang, Z. (2021). Weakly supervised instance segmentation for videos with temporal mask consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13968–13978.

[Lu et al., 2020] Lu, X., Wang, W., Shen, J., Tai, Y.-W., Crandall, D. J., and Hoi, S. C. (2020). Learning video object segmentation from unlabeled videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8960–8970.

[Lucas and Kanade, 1981] Lucas, B. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision (ijcai). volume 81.

[Luiten et al., 2018] Luiten, J., Voigtlaender, P., and Leibe, B. (2018). Premvos: Proposal-generation, refinement and merging for video object segmentation. *CoRR*, abs/1807.09190.

[Maninis et al., 2018] Maninis, K.-K., Caelles, S., Chen, Y., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., and Van Gool, L. (2018). Video object segmentation without temporal information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

[Oh et al., 2019] Oh, S. W., Lee, J., Xu, N., and Kim, S. J. (2019). Video object segmentation using space-time memory networks. *CoRR*, abs/1904.00607.

[Oh et al., 2018] Oh, S. W., Lee, J.-Y., Sunkavalli, K., and Kim, S. J. (2018). Fast video object segmentation by reference-guided mask propagation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7376–7385.

[Pont-Tuset et al., 2017] Pont-Tuset, J., Perazzi, F., Caelles, S., Arbelaez, P., Sorkine-Hornung, A., and Gool, L. V. (2017). The 2017 DAVIS challenge on video object segmentation. *CoRR*, abs/1704.00675.

[Shi and Tomasi, 1994] Shi, J. and Tomasi (1994). Good features to track. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600.

[Sommer et al., 2016] Sommer, L. W., Teutsch, M., Schuchert, T., and Beyerer, J. (2016). A survey on moving object detection for wide area motion imagery. *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*.

[Tian et al., 2019] Tian, Z., Chen, H., Wang, X., Liu, Y., and Shen, C. (2019). AdelaiDet: A toolbox for instance-level recognition tasks. https://git.io/adelaidet.

[Tian et al., 2020] Tian, Z., Shen, C., and Chen, H. (2020). Conditional convolutions for instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 282–298. Springer.

[Tian et al., 2021] Tian, Z., Shen, C., Wang, X., and Chen, H. (2021). Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5452.

[Ventura et al., 2019] Ventura, C., Bellver, M., Girbau, A., Salvador, A., Marqués, F., and Giró-i-Nieto, X. (2019). RVOS: end-to-end recurrent network for video object segmentation. *CoRR*, abs/1903.05612.

[Voigtlaender et al., 2019] Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., and Chen, L. (2019). FEELVOS: fast end-to-end embedding learning for video object segmentation. *CoRR*, abs/1902.09513.

[Voigtlaender and Leibe, 2017] Voigtlaender, P. and Leibe, B. (2017). Online adaptation of convolutional neural networks for video object segmentation. *CoRR*, abs/1706.09364.

[Vondrick et al., 2018] Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., and Murphy, K. (2018). Tracking emerges by colorizing videos. *CoRR*, abs/1806.09594.

[Wan et al., 2014] Wan, Y., Wang, X., and Hu, H. (2014). Automatic moving object segmentation for freely moving cameras. *Mathematical Problems in Engineering*, 2014.

[Wang et al., 2018] Wang, Q., Zhang, L., Bertinetto, L., Hu, W., and Torr, P. H. S. (2018). Fast online object tracking and segmentation: A unifying approach. *CoRR*, abs/1812.05050.

[Wu et al., 2019] Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2.

[Xie et al., 2021] Xie, H., Yao, H., Zhou, S., Zhang, S., and Sun, W. (2021). Efficient regional memory network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1286–1295.

[Xu et al., 2018a] Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., and Huang, T. (2018a). YouTube-VOS: A large-scale video object segmentation benchmark. *arXiv*, pages 1–10.

[Xu et al., 2018b] Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., and Huang, T. S. (2018b). Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327.

[Yang et al., 2021] Yang, C., Lamdouar, H., Lu, E., Zisserman, A., and Xie, W. (2021). Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7177–7188.

[Yang et al., 2018] Yang, L., Wang, Y., Xiong, X., Yang, J., and Katsaggelos, A. K. (2018). Efficient video object segmentation via network modulation. *CoRR*, abs/1802.01218.

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/QTUV2945

KinePose: A temporally optimized inverse kinematics technique for 6DOF human pose estimation with biomechanical constraints

Kevin Gildea[1], Clara Mercadal-Baudart[1], Richard Blythman[1,2], Aljosa Smolic[2], and Ciaran Simms[1]

[1]*School of Engineering, Trinity College Dublin*
[2]*V-SENSE, School of Computer Science & Statistics, Trinity College Dublin*

Abstract

Computer vision/deep learning-based 3D human pose estimation methods aim to localize human joints from images and videos. Pose representation is normally limited to 3D joint positional/translational degrees of freedom (3DOFs), however, a further three rotational DOFs (6DOFs) are required for many potential biomechanical applications. Positional DOFs are insufficient to analytically solve for joint rotational DOFs in a 3D human skeletal model. Therefore, we propose a temporal inverse kinematics (IK) optimization technique to infer joint orientations throughout a biomechanically informed, and subject-specific kinematic chain. For this, we prescribe link directions from a position-based 3D pose estimate. Sequential least squares quadratic programming is used to solve a minimization problem that involves both frame-based pose terms, and a temporal term. The solution space is constrained using joint DOFs, and ranges of motion (ROMs). We generate 3D pose motion sequences to assess the IK approach both for general accuracy, and accuracy in boundary cases. Our temporal algorithm achieves 6DOF pose estimates with low Mean Per Joint Angular Separation (MPJAS) errors (3.7°/joint overall, & 1.6°/joint for lower limbs). With frame-by-frame IK we obtain low errors in the case of bent elbows and knees, however, motion sequences with phases of extended/straight limbs results in ambiguity in twist angle. With temporal IK, we reduce ambiguity for these poses, resulting in lower average errors. Code and supplementary material are available[1].

Keywords: Human Pose Estimation, Computer Vision, Inverse Kinematics, Motion Capture, Biomechanics

1 Introduction

Human motion capture technologies are widely used for animation, physiotherapy, sports biomechanics, ergonomics, robotics, and augmented/virtual reality. However, many of these systems currently involve the use of calibrated multi-sensor setups, making them prohibitively expensive and unsuitable for potential in-the-wild applications. In recent years, a wide variety of computer vision based 3D pose estimation methods have been developed, which present opportunities for a form of free/low-cost markerless motion-capture for use on in-the-wild video footage, e.g., [Dinesh Reddy et al., 2021, Pavllo et al., 2019].

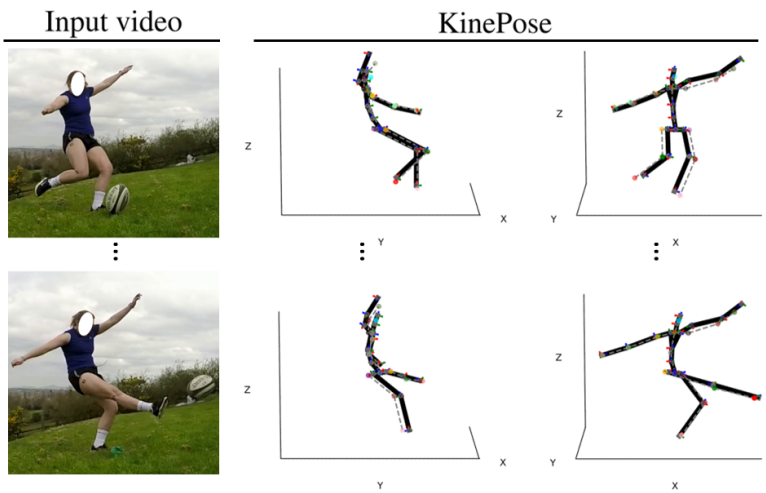


Figure 1: KinePose applied to an example case. 3D poses from [Liu et al., 2020] (multicolor keypoints and gray dashed lines).

[1]https://kevgildea.github.io/KinePose/

 ISBN 978-0-9934207-7-1

State-of-the-art position-based 3D pose estimators are capable of inferring root-relative joint positions with remarkable accuracy, however, joint/link orientations are often not included, limiting their practical uses. Approaches that indirectly regress for joint positions through applying joint orientations within a kinematic chain currently have sub-par performance, and do not allow for user definition of chain properties, i.e., limb lengths and joint ranges of motion (ROMs). Therefore, we propose a temporal optimization technique for mapping/retargeting the motion of an ordered set of 3D joint positions obtained from a pose estimator, to a user-defined open kinematic chain. The kinematic solution space (see 3.2) is constrained using joint DOFs and ROMs and limiting joint orientation changes between adjacent frames for temporal consistency (see 3.3). The proposed approach can be applied as a post-processing technique for position-based 3D pose estimates (particularly useful for monocular pose estimators), allowing for refinement of 3D poses for a subject-specific biomechanical model.

2 State of the Art

Both multi-camera and single-camera (monocular) 3D human pose estimation methods exist. Generally, they employ a 2D keypoint pose estimator as a backbone and use computer vision/machine learning techniques to 'lift' the 2D pose into 3D. A simple calibrated multi-camera technique involves classical algebraic triangulation of keypoint/joint positions. Many machine learning/regression-based multi-camera pose estimators have also been developed, which use neural networks trained and tested on widely used motion capture datasets (e.g. Human3.6M) to regress for both 2D and 3D keypoints directly from images/videos. Regression-based multi-camera pose estimators have achieved high accuracies of below 2cm Mean Per Joint Position Error (MPJPE) over a variety of tasks in Human3.6M [Dinesh Reddy et al., 2021, Iskakov et al., 2019]. Considering the inherent depth ambiguities, remarkably high accuracy has also been achieved from monocular pose estimators (in the order of 3-5cm MPJPE) [Dinesh Reddy et al., 2021, Pavllo et al., 2019]. However, position-based pose representation suffers from several limitations. In particular, joint/link orientations are not included, subject size estimates are inaccurate, and predicted limb lengths are often quite variable, greatly limiting its practical uses, e.g., in biomechanics applications. Other regression-based pose estimators address a portion of these limitations through predictions of joint rotations within a kinematic chain [Pavllo et al., 2018, Zhou et al., 2016], i.e., through forward kinematics (FK) (see 3.1). There are various ways to parameterize these rotations (e.g., sequential Euler/Cardan angles, Euler/Screw axis-angle), though unit quaternions are generally preferred as they overcome singularities associated with operations in $\mathbb{R}^3$. Orientation-based 3D pose representation allows for orientation losses to be included, i.e., MPJAS (see 3.3.4). Furthermore, model-based pose estimators allow for other biomechanical constraints such as joint DOFs. Though orientation-based pose representation is more physically meaningful than position-based representation, the results are mostly sub-par in terms of MPJPE. However, these are promising approaches meriting further development. A popular human biomechanics software OpenSim contains an IK tool for prescribing joint positions to a biomechanical model, which has recently been applied to obtain joint orientations from triangulated 2D predictions, i.e., Pose2Sim [Pagnon et al., 2021]. However, many practical applications require a similar tool for 3D pose estimates from a monocular method, which have characteristics that restrict its application. Namely, subject dimensions are inaccurately scaled (since there is no context for the size of the subject), and limb lengths are often variable. Therefore, our approach instead prescribes link directions (see 3.3.2), thus allowing for retargeting of incongruent skeletons.

3 Methods

In this section we give an overview of our IK approach for mapping kinematic chains, i.e., kinematic retargeting. In 3.1 we describe the process of FK, which is used in parameterization. In 3.2 we describe the indeterminate nature of the problem. In 3.3, we describe the proposed IK optimization procedure incorporating chain DOFs, ROMs, and kinematic loss terms.

3.1 Forward kinematics

Kinematic chains can be described as a hierarchical system of links and joints. They are represented by: 1) the locations and orientations of each of the joints, and 2) the hierarchy of the joints in the system. The Denavit-Hartenberg (DH) FK convention allows for compact representation and convenient manipulation of

kinematic chains as a series multiplication of 4x4 transformation matrices along the path to the joint. DH-FK uses a parent-child convention, whereby positions and orientations of child joints are expressed in the parent coordinate systems, i.e., the position of an arbitrary joint (j) is expressed as a vector in the coordinate system of its parent joint/body (j_{parent}), i.e., $\{\vec{r}\}_{j,j_{parent}}^{j_{parent}}$, and the orientation of the joint is expressed as a 3x3 rotation matrix which specifies the orientation with respect to the parent, i.e., $[R^{j,j_{parent}}]$. For each joint, the parent-child transformation matrix is then defined in Equation 1. The set of these for an N-joint chain C (with '0' denoting the global coordinate system) is expressed in Equation 2.

$$\{T_j\}^{j,j_{parent}} = \begin{bmatrix} [R^{j,j_{parent}}] & \{\vec{r}\}_{j,j_{parent}}^{j_{parent}} \\ 0\,0\,0 & 1 \end{bmatrix}. \quad (1) \qquad C = \{\{T_1\}^{1,0}, \ldots, \{T_N\}^{N,N_{parent}}\}. \quad (2)$$

To perform kinematic operations on an open chain with branching (e.g., the human body) we need also use a directed graph (G) specifying its hierarchical structure. We can then perform FK to express joint/body positions and orientations in the global coordinate system (Equation 3).

$$\begin{gathered} \{T_j\}^{j,0} = \prod_{i=j}^{1} \{T_i\}^{i,i_{parent}} \\ \text{for i on path to root} = \mathbf{FK}(C, G). \end{gathered} \quad (3)$$

The utility of defining a kinematic chain in this manner can be demonstrated if we would like to edit or reorient our kinematic chain. It makes physical sense to apply the rotation in the local coordinate system, e.g., a rotation of the knee is kinematically constrained (due to its DOFs) to occur about a fixed axis defined in the local coordinate system (usually chosen to be a cardinal axis). Furthermore, performing FK with the inclusion of a reorientating transformation matrix $\{T_{j_{\Delta R}}\}^{j,j_{parent}}$ will automatically reorient and reposition all down-chain joints (Equation 4). The kinematic chain is parameterized in this way.

$$\begin{gathered} \{T_j\}^{j,0} = \prod_{i=j}^{1} \{T_i\}^{i,i_{parent}} \{T_{j_{\Delta R}}\}^{j,j_{parent}} \\ \text{for i on path to root} = \mathbf{FK}(C, G, \{T_{j_{\Delta R}}\}^{j,j_{parent}}). \end{gathered} \quad (4)$$

3.2 Kinematic indeterminacy

In the Supp.Mat.[1], we formulate all rotation matrices representing all possible Euler axis-angle combinations for mapping/retargeting a vector $\vec{a}$ to point in the same direction as a vector $\vec{b}$ (Equation 5).

$$\left[R^{\hat{a}\to\hat{b}}\right]^{EAS} = \begin{bmatrix} n_{\alpha,x}^2 + (n_{\alpha,y}^2 + n_{\alpha,z}^2)cos(\Phi_\alpha) & n_{\alpha,x}n_{\alpha,y}(1-cos(\Phi_\alpha)) - n_{\alpha,z}sin(\Phi_\alpha) & n_{\alpha,x}n_{\alpha,z}(1-cos(\Phi_\alpha)) + n_{\alpha,y}sin(\Phi_\alpha) \\ n_{\alpha,x}n_{\alpha,y}(1-cos(\Phi_\alpha)) + n_{\alpha,z}sin(\Phi_\alpha) & n_{\alpha,y}^2 + (n_{\alpha,x}^2 + n_{\alpha,z}^2)cos(\Phi_\alpha) & n_{\alpha,y}n_{\alpha,z}(1-cos(\Phi_\alpha)) - n_{\alpha,x}sin(\Phi_\alpha) \\ n_{\alpha,x}n_{\alpha,z}(1-cos(\Phi_\alpha)) - n_{\alpha,y}sin(\Phi_\alpha) & n_{\alpha,y}n_{\alpha,z}(1-cos(\Phi_\alpha)) + n_{\alpha,x}sin(\Phi_\alpha) & n_{\alpha,z}^2 + (n_{\alpha,x}^2 + n_{\alpha,y}^2)cos(\Phi_\alpha) \end{bmatrix}. \quad (5)$$

Where $\vec{n}_\alpha$ and Φ_α represent the Euler axis-angle solution space (EAS), and α is a continuous angular value spanning $-\pi$ to π used to define unit vectors on the plane symmetrically bisecting the vectors, i.e., candidate Euler axes for which corresponding Euler angles can be computed.

We can use Equation 5 to analytically define the solution space for retargeting/mapping one kinematic chain (C) to an ordered set of 3D positions (P). The Euler axis-angle solution spaces for all joints in the chain are computed by combining Equations 4&5. This is visualized in Figure 2 for a 3-joint serial chain. The problem has an indeterminate solution (infinite solution space), therefore, optimization is needed. Specifically, the indeterminacy of this problem is due to the ambiguous 'twist' angle. The solution space can

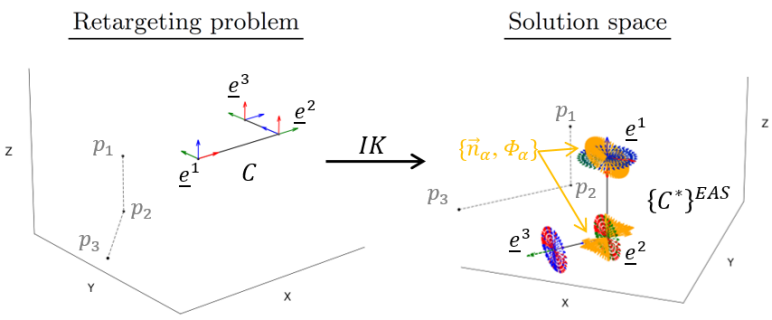


Figure 2: Solution space for mapping/retargeting a kinematic chain to an incongruent hierarchical set of 3D positions.

be somewhat constrained by specifying anatomical joint DOFs and ROMs (see 3.3.1). The problem is complicated in a chain such as the human body with branching, i.e., the mapping process used to generate the solution space is faced with a dilemma when reorienting branching joints. Therefore, our approach involves IK optimization, with parameterization of the reorientation transformation matrix in Equation 4 (i.e., $\{T_{j_{\Delta R}}\}^{j,j_{parent}}$). In general terms, since motion sequences involve kinematic indeterminacy our hypothesis is that two metrics can be used to converge to an optimal solution: 1) a frame-based pose loss term, and 2) a temporal consistency loss term (see 3.3.2).

3.3 Inverse kinematics optimization

3.3.1 Kinematic chain

Many position-based 3D pose estimators employ a similar chain configuration. We define a kinematic chain based on this, using common characteristics to optimize for joint orientations. Joints/bodies are defined to correspond to keypoints, i.e., the mid pelvis (defined as the root), hips, knees, ankles, mid spine, neck, head, shoulders, elbows, and ankles. Additional joints are also defined for greater biomechanical fidelity, i.e. the lower spine, and clavicles (see Figure 3). The chain has 14 joints for reorientation, and a total of 28 rotational DOFs. Initially, the kinematic chain is defined in a rest/quiet standing pose, i.e., we define a set of transformation matrices (see Equations 2&3) to describe a human in a relaxed standing pose with their arms by their sides. Limb/link lengths in the chain are constant and may be user defined. Changes in pose from rest are implemented using FK and an additional transformation matrix (see Equation 4). Joint ROMs may be specified in the local coordinate system as allowable angle ranges about each cardinal axis of the joint coordinate system, where θ_j, ϕ_j, and ψ_j correspond to sequential rotations about the local X, Y, and Z axes respectively (see Figure 3). For assessment, generalized ROM limits are used, however, they can be further customized. Since there is no 3D keypoint that can be used to directly target the lower spine joint, the lower and mid spine joints are assigned equally proportioned ROMs which add to the overall ROM of that area of the spine. Though not without limitations [Baerlocher and Boulic, 2001], Cardan/Euler angles present a convenient and intuitive approach for parameterizing the optimization weights associated with joint rotations within the kinematic chain, whilst also allowing for specification of joint ROMs using bounds to constrain the solution space. Gimbal lock singularities limit the of use of Cardan/Euler angles for parameterization of human pose. However, the application of anatomical joint DOFs and ROMs in the form of Cardan/Euler angle ranges naturally avoids singularities in our formulation by limiting $-\frac{\pi}{2} < \phi_j < \frac{\pi}{2}$ (see Supp.Mat.[1]). However, since the root joint retains a full ROM, we parameterize the root orientation with respect to the global coordinate system using the Euler/Screw axis-angle convention ($\vec{n} = \{n_x, n_y, n_z\}, \Phi$). Using ROMs throughout the chain and bounding the Screw angle in the root joint also overcomes rotation periodicity, i.e., $\sin x = \sin(x + 2\pi$, and $\cos x = \cos x + 2\pi$.

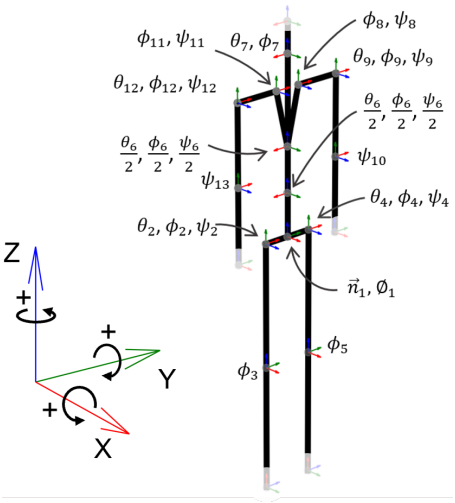


Figure 3: Kinematic chain parameters.

3.3.2 Losses

Two frame-based pose error terms are defined, which both minimize the angular difference in vectors between the 3D pose and the reoriented kinematic chain. We apply a local pose error term (Equation 6) which penalizes angular difference in link vector directions (shown in green in Figure 4), where $\{\hat{r}\}^0_{j,j_{parent}}$ is the normalized vector in the kinematic chain linking joint j to its parent in the global coordinate system (denoted by superscript '0') and $\{\hat{p}\}^0_{j,j_{parent}}$ is the corresponding normalized vector on the 3D pose skeleton. Since the kinematic chain includes clavicles which do not correspond to 3D pose keypoints, and there is no appropri-

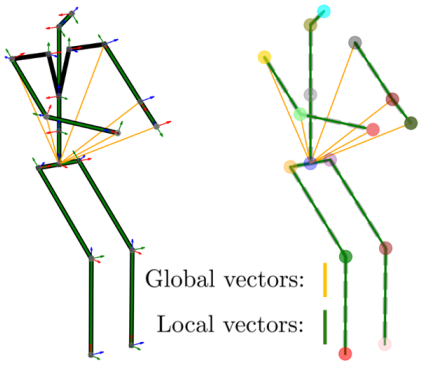


Figure 4: Pose loss vectors.

 ISBN 978-0-9934207-7-1

ate local pose vector to use to constrain the orientations of these joints, we also include a global pose error term (Equation 7) which considers angular difference in global vectors in the arms of the kinematic chain $\{\hat{\mathrm{r}}\}^0_{j,1}$, i.e., stemming from the root joint (shown in orange in Figure 4).

$$\mathrm{E}_{\angle(\vec{r},\,\vec{p})_{local}} = \frac{1}{N}\sum_{j=1}^{N}\left\|\cos^{-1}(\{\hat{\mathrm{r}}\}^0_{j,j_{parent}}\cdot\{\hat{\mathrm{p}}\}^0_{j,j_{parent}})\right\|. \quad (6) \qquad \mathrm{E}_{\angle(\vec{r},\,\vec{p})_{global}} = \frac{1}{A}\sum_{j=1}^{A}\left\|\cos^{-1}(\{\hat{\mathrm{r}}\}^0_{j,1}\cdot\{\hat{\mathrm{p}}\}^0_{j,1})\right\|. \quad (7)$$

In cases where there are co-linear sequential links, i.e., when there are fully extended limbs (knees or elbows), there will be kinematic indeterminacy. Our hypothesis is that appropriately weighted minimization of joint orientation changes between frames ($\Delta\Theta_j$) can constrain the solution space toward an optimal solution with temporal consistency, i.e., optimizing poses across frames with extended and bent limbs (Equation 8) (for a sequence of poses M frames long).

$$\mathrm{E}_{\Delta\Theta} = \frac{1}{M}\sum_{m=1}^{M}\left\{\sum_{j=1}^{N}\left\|\Delta\Theta_j\right\|\right\}_m, \quad (8)$$

where $\Delta\Theta_j = \Theta_j^{m+1} - \Theta_j^m$ for $m = 1$, $\Delta\Theta_j = (\Theta_j^{m+1} - \Theta_j^{m-1})/2$ for $1 < m < M$, and $\Delta\Theta_j = \Theta_j^m - \Theta_j^{m-1}$ for $m = M$.

We define a loss function L_{frame} for frame-by-frame IK (Equation 9), and combining Equations 6,7&8 with scalar weighting for the temporal term which competes with the pose error terms, we define $\mathrm{L}_{temporal}$ for temporal IK (Equation 10).

$$\mathrm{L}_{frame} = \mathrm{E}_{\angle(\vec{r},\,\vec{p})_{local}} + \mathrm{E}_{\angle(\vec{r},\,\vec{p})_{global}}, \quad (9) \qquad \mathrm{L}_{temporal} = \frac{1}{M}\sum_{m=1}^{M}\mathrm{L}_{frame=m} + \lambda\mathrm{E}_{\Delta\Theta}. \quad (10)$$

3.3.3 Algorithms

A sequential least squares programming procedure is used to solve the minimization problem [Kraft, 1988]. This procedure is commonly used for robust nonlinear programming solutions to kinematic retargeting and IK. All algorithms take the kinematic chain C in its rest/quiet pose, and a sequence of 3D poses for F frames as inputs (P). The algorithms output optimized model parameters or joint DOFs (Θ^*) using IK (Equation 11). These can then be used to specify a sequence of retargeted chains $\{\mathrm{C}_1^*,\ldots,\mathrm{C}_F^*\}$ using Equation 4.

$$\Theta^* = \mathbf{IK}(C, G, P, \Theta_0) \text{ subject to } \arg\min_{\Theta}(\mathrm{L}). \quad (11)$$

Since convergence of local optimization methods depends on the initial guess (Θ_0), we perform a relatively computationally inexpensive frame-by-frame pre-optimization as input to our temporal optimization, using L_{frame} (Equation 9) as the objective function (Algorithm 1). This addition also speeds up the optimization time significantly.

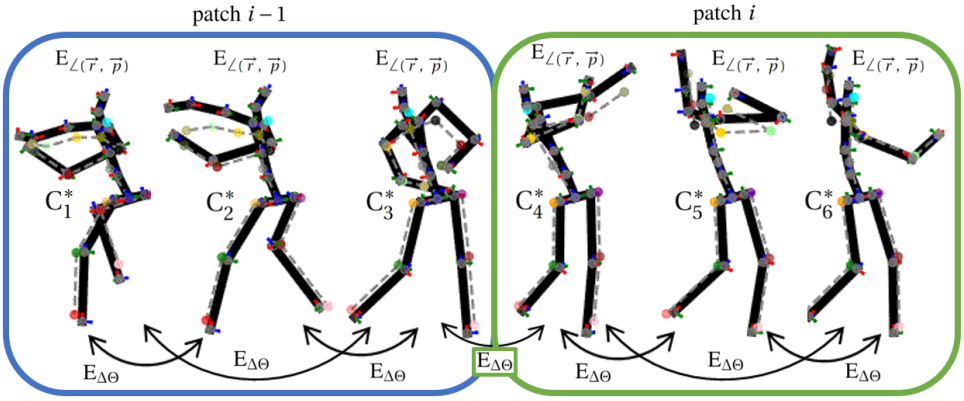


Figure 5: Overview of the proposed IK approach (patch length M=3).

The temporal algorithm (Algorithm 2) involves blocks, or patches of M frames (out of a total of F frames) in which sequences of poses are retargeted in tandem with the inclusion of both pose and temporal error terms (Equation 10). Patches are temporally linked or 'stitched' with an additional temporal error term, i.e., the difference in model parameters ($\mathrm{E}_{\Delta\Theta}$) between the final frame of patch $i-1$ and the first frame of patch i are used in the loss function for patch i (Figure 5).

 ISBN 978-0-9934207-7-1

Algorithm 1: Frame-based retargeting

Data: C, G, P
Result: $\{\Theta_1^*, \ldots, \Theta_F^*\}$
for $f = 1, \ldots, F$ **do**
  if $f = 1$ **then**
    $\Theta_f^* = \mathbf{IK}(C, G, P, \Theta_0) \rightarrow$ Subject to $\text{argmin}_\Theta(\text{L}_{frame})$
    Where Θ_0 corresponds to the chain in its rest pose.
  else
    $\Theta_f^* = \mathbf{IK}(C, G, P, \Theta_{f-1}^*) \rightarrow$ Subject to $\text{argmin}_\Theta(\text{L}_{frame})$
  end
end

Algorithm 2: Temporal retargeting

Data: $C, G, P, \{\Theta_1, \ldots, \Theta_F\}$
$\rightarrow \Theta_F$ obtained from Algorithm 1
Result: $\{C_1^*, \ldots, C_F^*\}$
Reshape $\{\Theta_1, \ldots, \Theta_F\}$ into $\left\{\Theta_1, \ldots, \Theta_{\frac{F}{M}}\right\}$
for $i = 1, \ldots, \frac{F}{M}$ **do**
  $\Theta_i^* = \mathbf{IK}(C, G, P, \Theta_i) \rightarrow$ Subject to $\text{argmin}_\Theta(\text{L}_{temporal})$
end
Reshape $\left\{\Theta_1^*, \ldots, \Theta_{\frac{F}{M}}^*\right\}$ into $\{\Theta_1^*, \ldots, \Theta_F^*\}$
for $f = 1, \ldots, F$ **do**
  for $j = 1, \ldots, N$ **do**
    Express $\Theta_{f,j}^*$ as $\{\text{T}_{\text{j}\Delta\text{R}}\}_{f,j}^{j,j_{parent}}$
  end
  $C_f^* = \left\{\{\text{T}_1\}_1^{1,0}\{\text{T}_{1\Delta\text{R}}\}_{\text{f},1}^{1,0}, \ldots, \{\text{T}_\text{N}\}_N^{N,N_{parent}}\{\text{T}_{\text{N}\Delta\text{R}}\}_{f,N}^{N,N_{parent}}\right\}$
end

3.3.4 Assessment

For algorithmic assessment of the IK approaches, we generate twenty-four 30-frame long motion sequences within joint ROMs for the kinematic chain, with three levels of motion speed, with maximum cardan angle changes between frames of a) $\frac{\pi}{500}$, b) $\frac{\pi}{200}$, and c) $\frac{\pi}{70}$. The latter corresponding approximately to full movement within ROMs. Half of the sequences consist of only bent limbs, and the other half have a combination of prescribed phases of either bent or extended limbs. The motion sequences have been defined such that each sequence with bent limbs has a corresponding sequence with phases of bent limbs and extended limbs, where motion in the rest of the skeleton is identical. These phases are: 1) 3 frames extended, 2) 9 frames bent, 3) 6 frames extended, 4) 9 frames bent, 5) 3 frames extended. Generation of motion in this manner allows for analysis of the effect on accuracy of extended limbs (which results in ambiguity in twist angle) for various IK implementations. We apply a series of IK algorithms to the resulting joint pose sequences, i.e., with ground truth joint orientations hidden. We compare Algorithm 1 with and without feeding of optimized model parameters from previous frames (1_a and 1_b respectively), and Algorithm 2 with different patch sizes, i.e., 2_3 (M=3) and 2_5 (M=5). Computation time on an Intel® Core™ i7-9700 processor is recorded for each implementation. Overall agreement was assessed using Mean Per Joint Angular Separation (MPJAS_N) for all joints that are reoriented (N=14) across frames, i.e., the average angular difference about the Screw axis between inferred and ground truth (GT) joint orientations $\{\Phi_\text{j}\}_{pred,GT}$ (Equation 12). For comparison, we also calculate MPJAS for each individual joint.

$$\text{MPJAS}_N = \frac{1}{F}\sum_{f=1}^{F}\left\{\frac{1}{N}\sum_{j=i}^{N}\left\|\{\Phi_\text{j}\}_{pred,GT}\right\|\right\}_f. \tag{12}$$

An ablation study was performed on a sample of 12 additional motion sequences to determine the optimal scalar weight for the temporal model (λ) for different motion speeds ($\lambda = 0.3$ for speed c, $\lambda = 0.5$ for speed b, and $\lambda = 0.7$ for speed a).

4 Results

Table 1 shows a comparison of average angular errors and optimization times for Algorithm 1 with feeding of weights between frames for initialization (1_b), or no feeding (1_a), and Algorithm 2 with patches of length 3 (2_3) or 5 (2_5) for three levels of motion speed.

Table 1: Accuracy of IK Algorithms 1 and 2. MPJAS_{14} per frame for three levels of motion speed (fps: Optimization frames per second, E: 14-joint Mean Per Joint Angular Separation (rad/joint)).

	Speed a		**Speed b**		**Speed c**		**Average**	
Algorithm	fps	E	fps	E	fps	E	fps	E
1_a	$7.44\text{x}10^{-2}$	$9.45\ \text{x}10^{-2}$	$6.96\text{x}10^{-2}$	$9.13\ \text{x}10^{-2}$	$7.48\text{x}10^{-2}$	$8.12\text{x}10^{-2}$	$7.29\text{x}10^{-2}$	$8.90\ \text{x}10^{-2}$
1_b	$9.63\text{x}10^{-2}$	$7.65\text{x}10^{-2}$	$9.28\text{x}10^{-2}$	$7.45\text{x}10^{-2}$	$8.61\text{x}10^{-2}$	$5.68\text{x}10^{-2}$	$9.15\text{x}10^{-2}$	$6.93\text{x}10^{-2}$
2_3	$2.03\text{x}10^{-2}$	$8.28\text{x}10^{-2}$	$2.33\text{x}10^{-2}$	$7.69\text{x}10^{-2}$	$5.74\text{x}10^{-2}$	$5.07\text{x}10^{-2}$	$2.74\text{x}10^{-2}$	$7.01\text{x}10^{-2}$
2_5	$8.94\text{x}10^{-3}$	$7.07\text{x}10^{-2}$	$1.32\text{x}10^{-3}$	$7.40\text{x}10^{-2}$	$3.90\text{x}10^{-2}$	$5.13\text{x}10^{-2}$	$1.41\text{x}10^{-2}$	$6.53\text{x}10^{-2}$

 ISBN 978-0-9934207-7-1

Table 2 shows a comparison of average angular errors for motions with 1) bent limbs only, or 2) a combination of bent and extended limbs. Motion sequences with only bent limbs have low errors for all algorithms, and for sequences with phases of extended limbs algorithm 1_b and both temporal algorithms have lower errors.

Table 2: $MPJAS_{14}$ per frame for motion sequences with 1) only bent limbs, or 2) phases of extended limbs.

Algorithm	Bent limbs only	Extended & bent limbs
1_a	$5.84x10^{-2}$	$1.20x10^{-1}$
1_b	$6.05x10^{-2}$	$7.80x10^{-2}$
2_3	$5.39x10^{-2}$	$8.64x10^{-2}$
2_5	$5.64x10^{-2}$	$7.43x10^{-2}$

Figure 6 shows a comparison of average average angular errors for parameterized joints in the chain for motion sequences with phases of bent and extended limbs. Algorithm 1_a produces higher errors in the clavicles (due to a lack of corresponding keypoints), and shoulders and hips (associated with ambiguity in twist angles). Results for motion sequences with only bent limbs are provided in the Supp.Mat.[1] Furthermore, the relationship between phases of extended/bent limbs are investigated further for boundary cases.

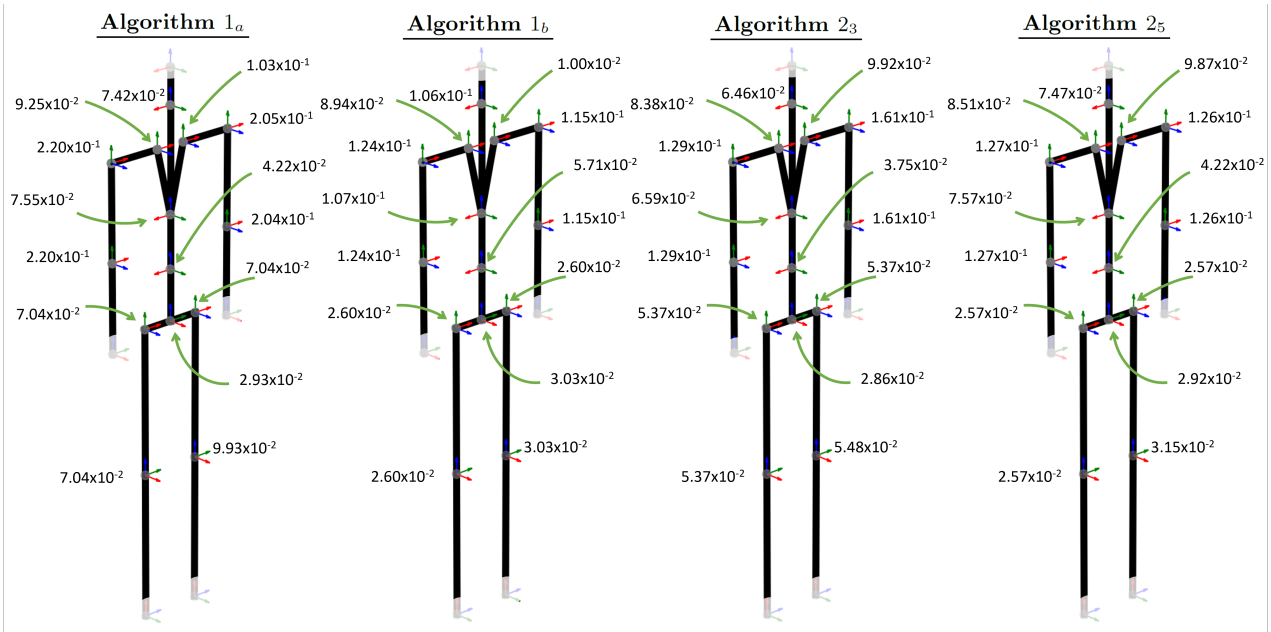


Figure 6: $MPJAS_1$ by joint and algorithm, for motion sequences with phases of bent and extended limbs.

5 Discussion

$MPJAS_{14}$ errors are low for all algorithms and motion speeds, ranging from average errors of $8.90x10^{-2}$ rad/joint for algorithm 1_a (5.1°), to $6.53x10^{-2}$ rad/joint for algorithm 2_5 (3.7°) (Table 1). Differences observed between algorithm 1_a and 1_b highlight the importance of the initialization guess for weights, i.e., feeding weights from previous frames for initialization in a frame-by-frame optimization reduces time to convergence (Table 1), and adds a degree of temporal consistency. The results provide a clear justification for the temporal model. While algorithm 1_b outperforms 2_3 on average, temporal algorithm 2_5 is the best performer across all indicators. The benefits of the temporal model relate to twist angle ambiguity (Table 2 & Figure 6), and are of particular benefit for a specific boundary case involving an initial phase of extended limbs, followed by a phase of bent limbs (see Supp.Mat.[1]). Large fluctuations in twist angles are observed for algorithm 1_a, whereas, algorithm 1_b and both variations of the temporal algorithm result in lower errors both overall, and in boundary cases. Algorithm 2_5 achieves the highest accuracy for all motion speeds, and obtains particularly low errors for the lower limbs (hips & knees), i.e., $2.71x10^{-2}$ rad/joint (1.6°) for motion sequences involving phases of extended limbs (Figure 6), and $1.63x10^{-3}$ rad/joint (0.1°) for motion with only bent limbs (see Supp.Mat.[1]). Temporal implementations with sufficiently large patch sizes are effective for reducing ambiguity in twist angle, which is a key benefit. Nevertheless, algorithm 1_b is suitable for applications requiring faster processing. The temporal scalar weight should be tuned for both motion speed and video framerate. For fast motion or a low framerate, relative joint orientation changes between frames will naturally be higher, therefore requiring a lower weighting. Similarly, for slow motion or high framerate a higher weighting is required.

 ISBN 978-0-9934207-7-1

Due to a lack of relevant keypoints, twist angles of the forearms and neck, and joint reorientations of the ankles and wrists are currently not included. However, an extension could include these for 3D pose inputs with more keypoints in the extremities. A large proportion of the errors are associated with those joints in the chain without corresponding keypoints, i.e., clavicles and lower spine. Parameterization of the lower/mid spine with equally proportioned rotations in two joints may often not be biofidelic, indeed, this relationship is often the subject of analyses in biomechanics. These limitations are imposed by the fact that current 3D pose estimators do not include sufficient keypoints. Synthetic generation of motion, as per [Beeson and Ames, 2015], allows for a detailed assessment of the mathematical procedure including boundary cases of interest. However, for potential practical applications further testing should include GT activity-specific motion capture footage. Furthermore, practical applications rely on the accuracy of the 3D pose estimator used; calibrated multi-camera pose estimators have higher accuracy, whereas monocular pose estimators are more practically useful. KinePose is applicable to either multi-camera, or monocular 3D pose estimators with incorrectly scaled skeletons and variable link lengths.

6 Conclusions

This paper proposes KinePose, a convenient post-processing technique for position-based 3D human pose estimation methods, to infer joint orientations in a subject-specific biomechanical model. We demonstrate that this technique allows for accurate prediction of joint orientations. This technique may be particularly useful for sports biomechanics, and telehealth applications. Though currently unsuited to real-time applications, we envisage many potential applications as a post-processing technique for 3D pose estimators.

Acknowledgements

This research was funded under the RSA-Helena Winters Scholarship for Studies in Road Safety.

References

[Baerlocher and Boulic, 2001] Baerlocher, P. and Boulic, R. (2001). Parametrization and Range of Motion of the Ball-and-Socket Joint. *Deformable Avatars*, pages 180–190.

[Beeson and Ames, 2015] Beeson, P. and Ames, B. (2015). TRAC-IK: An open-source library for improved solving of generic inverse kinematics. *International Conference on Humanoid Robots*, pages 928–935.

[Dinesh Reddy et al., 2021] Dinesh Reddy, N., Guigues, L., Pishchulin, L., Eledath, J., and Narasimhan, S. G. (2021). TesseTrack: End-to-End Learnable Multi-Person Articulated 3D Pose Tracking. In *Conference on Computer Vision and Pattern Recognition*, pages 15190–15200.

[Iskakov et al., 2019] Iskakov, K., Burkov, E., Lempitsky, V., and Malkov, Y. (2019). Learnable Triangulation of Human Pose. *International Conference on Computer Vision*, pages 7717–7726.

[Kraft, 1988] Kraft, D. (1988). *A software package for sequential quadratic programming*. Wiss. Berichtswesen d. DFVLR.

[Liu et al., 2020] Liu, J., Rojas, J., Li, Y., Liang, Z., Guan, Y., Xi, N., and Zhu, H. (2020). A Graph Attention Spatio-temporal Convolutional Network for 3D Human Pose Estimation in Video. In *International Conference on Robotics and Automation*, pages 3374–3380.

[Pagnon et al., 2021] Pagnon, D., Domalain, M., and Reveret, L. (2021). Pose2Sim: An End-to-End Workflow for 3D Markerless Sports Kinematics—Part 1: Robustness. *Sensors 2021, Vol. 21, Page 6530*, 21(19):6530.

[Pavllo et al., 2019] Pavllo, D., Feichtenhofer, C., Grangier, D., and Auli, M. (2019). 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition*, pages 7745–7754.

[Pavllo et al., 2018] Pavllo, D., Grangier, D., and Auli, M. (2018). QuaterNet: A Quaternion-based Recurrent Model for Human Motion. In *British Machine Vision Conference*.

[Zhou et al., 2016] Zhou, X., Sun, X., Zhang, W., Liang, S., and Wei, Y. (2016). Deep Kinematic Pose Regression. In *European Conference on Computer Vision*, pages 186–201.

Grad-CAM++ is Equivalent to Grad-CAM With Positive Gradients

Miguel Lerma*[1] and Mirtha Lucas†[2]

[1]*Northwestern University, Evanston, IL 60208, USA*
[2]*DePaul University, Chicago, IL 60604, USA*

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/AWJV6348

Abstract

The Grad-CAM algorithm provides a way to identify what parts of an image contribute most to the output of a classifier deep network. The algorithm is simple and widely used for localization of objects in an image, although some researchers have point out its limitations, and proposed various alternatives. One of them is Grad-CAM++, that according to its authors can provide better visual explanations for network predictions, and does a better job at locating objects even for occurrences of multiple object instances in a single image. Here we show that Grad-CAM++ is practically equivalent to a very simple variation of Grad-CAM in which gradients are replaced with positive gradients.

Keywords: Explainable Artificial Intelligence, Attribution Methods, Grad-CAM++

1 Introduction

Artificial Intelligence (AI) has progressed in the last few years at an accelerated rate, but many AI models behave as black boxes providing a prediction or solution to a problem without giving any information about how or why the model arrived to a given conclusion. This has a negative effect in the trust humans are willing to place in the output of AI systems. Explainable Artificial Intelligence (XAI) aims to remediate this problem by providing tools intended to explain the output of AI models. Here we look at two of them, Grad-CAM and Grad-CAM++, how they work, and how they are related.

We will start in section 2 by outlining how Grad-CAM and Grad-CAM++ work. In section 3 we discuss the methodology behind Grad-CAM++. In section 4 we show how Grad-CAM++ compares to a small variation of Grad-CAM that we call Grad-CAM$^+$.

2 Overview of Grad-CAM and Grad-CAM++

Grad-CAM and Grad-CAM++ can be used on deep convolutional networks whose outputs are differentiable functions. Their goal is to identify what parts of the network input contribute most to the output. In this section we explain how they work.

2.1 Grad-CAM algorithm

Here we present two versions of Grad-CAM. The first version is described in [Selvaraju et al., 2017]. The second version, that we name Grad-CAM$^+$, is a small variation we have found in some implementations of the Grad-CAM algorithm.

*mlerma@math.northwestern.edu
†mlucas3@depaul.edu

 ISBN 978-0-9934207-7-1

2.1.1 Grad-CAM

The Grad-CAM algorithm works as follows. First, we must pick a convolutional layer A, which is composed of a number of feature maps or "channels." Let A^k be the k-th feature map of the picked layer, and let A^k_{ij} be the activation of the unit in position (i, j) of the k-th feature map. Then, the localization map, also called heatmap, saliency map, or attribution map, is obtained by combining the feature maps of the layer using weights w^c_k that capture the contribution of the k-th feature map to the output y^c of the network that corresponds to class c.[1]

In order to compute the weights we pick a class c, and determine how much the output y^c of the network depends of each unit of the k-th feature map, as measured by the gradient $\frac{\partial y^c}{\partial A^k_{ij}}$, which can be obtained by using the backpropagation algorithm. The gradients are then averaged thorough the feature map to yield a weight w^c_k, as indicated in equation (1). Here Z is the size (number of units) of the feature map.

$$w^c_k = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A^k_{ij}}}_{\text{gradients via backprop}} . \tag{1}$$

The next step consists of combining the feature maps A^k using the weights computed above, as shown in equation (2). Note that the combination is also followed by a Rectified Linear function $\mathrm{ReLU}(x) = \max(x, 0)$, because we are interested only in the features that have a positive influence on the class of interest. The result $L^c_{\text{Grad-CAM}}$ is called *class-discriminative localization map* by the authors. It can be interpreted as a coarse heatmap of the same size as the chosen convolutional feature map.

$$L^c = \mathrm{ReLU}\ \underbrace{\left(\sum_k w^c_k A^k\right)}_{\text{linear combination}} . \tag{2}$$

After the heatmap has been produced, it can be normalized and upsampled via bilinear interpolation to the size of the original image, and overlapped with it to highlight the areas of the input image that contribute to the network output corresponding to the chosen class. Figure 1 shows the resulting heatmap (after applying a colormap) obtained in the same figure for classes 'cat' and 'dog' respectively. The red color indicates the areas in which the heatmaps have a higher intensity, which are expected to coincide with the location of the objects corresponding to the classes picked.

Figure 1: Grad-CAM locating a cat and a dog.

The method is very general, and can be applied to any (differentiable) network outputs.

[1] We note that the Grad-CAM algorithm as described in [Selvaraju et al., 2017] uses gradients of pre-softmax scores, however we have found implementations in which gradients of post-softmax scores are used instead—which in general is easier than using pre-softmax scores e.g. when the model is given by a third party and pre-softmax scores may not be easy to access. So in general we will understand that y^c may represent either the pre-softmax or the post-softmax output of the network, at the choice of the user.

2.1.2 Grad-CAM$^+$

In some implementations of Grad-CAM we have found that in the computation of the weights only terms with *positive* gradients are used, i.e., the weights are computed using the following formula:

$$w_k^c = \frac{1}{Z}\sum_i\sum_j \text{ReLU}\Big(\frac{\partial y^c}{\partial A_{ij}^k}\Big). \tag{3}$$

This can be justified by the intuition that negative gradients correspond to units where features from a class different from the chosen class are present (say an area showing a 'cat' when trying to locate a 'dog' in an image). Although in the implementations this version of the Grad-CAM algorithm with positive gradients is still named "Grad-CAM," for clarity here we denote it "Grad-CAM$^+$."

2.2 Grad-CAM++ algorithm

Grad-CAM++ [Chattopadhyay et al., 2018] has been introduced with the purpose of providing a better localization than Grad-CAM. When instances of an object appear in various places, or produce footprints in different feature maps, a plain average of the gradients at each feature map may not provide a good localization of the regions of interest because the units of each feature map may have different "importance" not fully captured by the gradient alone. The solution proposed in [Chattopadhyay et al., 2018] is to replace the plain average of gradients at each feature map with a weighted average:

$$w_k^c = \sum_i\sum_j \alpha_{ij}^{kc}\cdot\text{ReLU}\Big(\frac{\partial y^c}{\partial A_{ij}^k}\Big), \tag{4}$$

where the α_{ij}^{kc} represents the importance of each individual unit in a feature map. Note also that the gradients are fed to a ReLU function, so only positive gradients are considered at this step. The rest of the algorithm is the same as the original Grad-CAM, in particular the final heatmap is still computed using equation (2).

In order to determine the values of the alphas the following equation is posed:

$$Y^c = \sum_k w_k^c \cdot \sum_{i,j} A_{ij}^k, \tag{5}$$

i.e., the global pooling of the feature maps of the layer, combined using the weights w_k^c, should produce the class scores Y^c. Next, plugging (4) in (5) we get:

$$Y^c = \sum_k\Big(\Big\{\sum_{a,b}\alpha_{ab}^{kc}\cdot\text{ReLU}\Big(\frac{\partial Y^c}{\partial A_{ab}^k}\Big)\Big\}\Big[\sum_{i,j}A_{ij}^k\Big]\Big). \tag{6}$$

The summation indices (i, j) in the first sum have been renamed (a, b) to avoid confusion. Next, in order to isolate the alphas, partial derivatives w.r.t. A_{ij}^k are computed on both sides of (6) (without the ReLU function), obtaining (according to the original paper):

$$\frac{\partial Y^c}{\partial A_{ij}^k} = \sum_{a,b}\alpha_{ab}^{kc}\cdot\frac{\partial Y^c}{\partial A_{ab}^k} + \sum_{a,b}A_{ab}^k\Big\{\alpha_{ij}^{kc}\cdot\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2}\Big\}. \tag{7}$$

Taking partial derivative w.r.t. A_{ij}^k again and isolating α_{ij}^{kc} the following equation is obtained:

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2}}{2\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2} + \sum_{ab}A_{ab}^k\{\frac{\partial^3 Y^c}{(\partial A_{ij}^k)^3}\}}. \tag{8}$$

 ISBN 978-0-9934207-7-1

The final expression for the alphas involve second and third order partial derivatives. Assuming that the score Y^c is an exponential of the pre-softmax output of the network S^c, i.e., $Y^c = \exp(S^c)$, the expression yielding the alphas becomes:

$$\alpha_{ij}^{kc} = \frac{\left(\frac{\partial S^c}{\partial A_{ij}^k}\right)^2}{2\left(\frac{\partial S^c}{\partial A_{ij}^k}\right)^2 + \sum_{ab} A_{ab}^k \left(\frac{\partial S^c}{\partial A_{ij}^k}\right)^3}, \tag{9}$$

which does not involve high order partial derivatives. This final equation (9) is the one used in actual implementations. Note that if $\frac{\partial S^c}{\partial A_{ij}^k} = 0$ then the right hand side of (9) becomes $0/0$, which is undefined. In this case α_{ij}^{kc} is assigned value zero. After the alphas are computed the weights are obtained using equation (4), and then the saliency map using equation (2).[2]

3 Discussion

In this section we discuss several issues we found in the algorithm for Grad-CAM++ and its derivation.

3.1 The math is unclear

Here we focus on mathematical issues.

1. Equation (6) is a system of linear equations with more unknowns (the α_{ij}^{kc}) than equations (just one per class), which makes it underdetermined. More specifically, equation (6) can be written:

$$\sum_{k,a,b} C_{ab}^k \cdot \alpha_{ab}^{kc} = Y^c, \tag{10}$$

   where $C_{ab}^k = \mathrm{ReLU}\left(\frac{\partial Y^c}{\partial A_{ab}^k}\right)\left[\sum_{i,j} A_{ij}^k\right]$, or $C_{ab}^k = \frac{\partial Y^c}{\partial A_{ab}^k}\left[\sum_{i,j} A_{ij}^k\right]$ if we remove the ReLU. Note that (10) has infinitely many solutions. A general set of solutions can be written like this:

$$\alpha_{ab}^{kc} = \beta_{ab}^{kc} \cdot Y^c \cdot \left(\sum_{k,a,b} C_{ab}^k \cdot \beta_{ab}^{kc}\right)^{-1}, \tag{11}$$

   where β_{ab}^{kc} are arbitrary numbers constrained only by the condition $\sum_{k,a,b} C_{ab}^k \cdot \beta_{ab}^{kc} \neq 0$. Consequently, isolating the α_{ij}^{kc} to get a single solution like in equation (8) is impossible without adding additional constrains.

2. In the derivation of equation (7) the α_{ij}^{kc} are treated as constants, but that cannot be correct because they depend on the A_{ij}^k. In fact the method used to isolate the alphas is equivalent to "solving" the equation $\alpha x = x^2$ for α by differentiating on both sides w.r.t. x and concluding that $\alpha = 2x$, which is obviously incorrect. The actual result of differentiating on both sides of this equation should read $\frac{d\alpha}{dx}x + \alpha = 2x$, which is correct, but does not help isolate α, it only complicates unnecessarily the original equation.

3. Even if we assume that the α_{ij}^{kc} are constant, taking partial derivative of (6) w.r.t. A_{ij}^k does not yield (7). The actual result is the following:[3]

$$\frac{\partial Y^c}{\partial A_{ij}^k} = \sum_{a,b} \alpha_{ab}^{kc} \cdot \frac{\partial Y^c}{\partial A_{ab}^k} + \sum_l \left(\left[\sum_{u,v} A_{uv}^l\right]\left\{\sum_{a,b} \alpha_{ab}^{lc} \cdot \frac{\partial^2 Y^c}{\partial A_{ij}^k \partial A_{ab}^l}\right\}\right) \tag{12}$$

[2]Note that it does not matter if we use $y^c = \exp(S^c)$, hence $\frac{\partial y^c}{\partial A_{ij}^k} = \exp(S^c)\frac{\partial S^c}{\partial A_{ij}^k}$, or $y^c = S^c$, $\frac{\partial y^c}{\partial A_{ij}^k} = \frac{\partial S^c}{\partial A_{ij}^k}$. The heatmaps obtained one way or the other differ by the constant factor $\exp(S^c)$, which will have no effect after min-max normalization of the heatmaps. It fact, because of the rapid grow of the exponential function, which may lead to numerical instability, the choice $y^c = S^c$ is arguably better than $y^c = \exp(S^c)$.

[3]The computations are pretty straightforward, they are just long and tedious. Space limitations prevent us from including them here, but we would be glad to provide them by request.

 ISBN 978-0-9934207-7-1

4. Another aspect that interferes with the process of solving equation (6) is that second degree derivatives kill linearities, so if after computing the alphas we replace Y^c with Y^c plus a linear function of the A^k_{ij}, i.e., $Y^c \to Y^c + \sum_{i,j,k} \lambda_{ijk} A^k_{ij} + C$, where λ_{ijk} and C are constants, the process would lead to the same values for the α^{kc}_{ij}, even though they cannot be solutions to the original and new equations simultaneously.

Leaving aside the question of the derivation of the formula for the alphas, we next discuss the formula actually used in the implementations to compute the alphas.

3.2 Formula used in actual implementations

The formula for α^{kc}_{ij} actually used in most code implementations of Grad-CAM++ we have found, including the official code distributed by the authors, is (9). Note that if we divide numerator and denominator by $(\partial S^c / \partial A^k_{ij})^2$ we get

$$\alpha^{kc}_{ij} = \frac{1}{2 + \sum_{ab} A^k_{ab} \left(\frac{\partial S^c}{\partial A^k_{ij}} \right)}, \tag{13}$$

which is mathematically equivalent to (9), except when the gradients are zero, in which case $\alpha^{kc}_{ij} = 0$.

Although the formula does not contain second and third powers anymore (which reduces the risk of under or overflow) the expression is still numerically unstable, because nothing prevents the denominator from getting close to zero. In our experiments we have observed that this in fact happens occasionally. Also, if the second term in the denominator of (13) remains small compared to 2, the alphas will be approximately constant. Next, we discuss this two issues.

3.3 In practice the alphas are nearly constant

We have observed that the absolute value of the second term in the denominator of (13) is usually small compared to 2, and consequently the alphas tend to take values around 1/2. This happened consistently for all individual images we tried. The distribution of the alphas obtained after feeding a VGG16 network pretrained on ImageNet [Simonyan and Zisserman, 2015] with 10,000 images from ImageNetV2 MatchedFrequency [Recht et al., 2019] is shown in Figure 2. For the boxplot we have considered non-zero values of α^{kc}_{ij} only, and removed outliers. The reason to consider only non-zero values for the alphas is that α^{kc}_{ij} is set to zero precisely when the gradients vanish, and in that case the values of α^{kc}_{ij} do not play any role because if $\frac{\partial S^c}{\partial A^k_{ij}} = 0$ then $\alpha^{kc}_{ij} \cdot \mathrm{ReLU}\left(\frac{\partial S^c}{\partial A^k_{ij}} \right) = 0$ regardless of the value of α^{kc}_{ij}. For the histogram on the right we have included again all non-zero values of alpha, but without removing outliers, centered at 0.5 and squeezed with an hyperbolic tangent, i.e., $\tanh(\alpha^{kc}_{ij} - 0.5)$. The tanh-squeezing allows to display all the frequency bars within the interval $[-1, 1]$ even though the outliers may take values very far away from 0.5.

Consequently, we have $\alpha^{kc}_{ij} \approx \frac{1}{2}$, and $w^c_k \approx \frac{1}{2} \sum_i \sum_j \mathrm{ReLU}\left(\frac{\partial y^c}{\partial A^k_{ij}} \right)$, hence the weights computed are (except for a multiplicative constant) approximately the same as the weights computed for the simple variation of Grad-CAM that we call Grad-CAM$^+$.

3.4 The computation of the alphas is numerically unstable

Even though the alphas tend to take values around 0.5 we also noticed the presence of outliers α^{kc}_{ij} reaching very large values. On the whole ImageNetV2 the extreme values were -1398101.4 and 559240.56 respectively, which confirmed the fact that the computation of the alphas using equation (13) is numerically unstable. Furthermore, we cannot think of any theoretical justification to assign extremely large values to the "pixel-importance" (as measured by α^{kc}_{ij}) when the value of $\sum_{ab} A^k_{ab} \left(\frac{\partial S^c}{\partial A^k_{ij}} \right)$ happens to be close to -2.

 ISBN 978-0-9934207-7-1

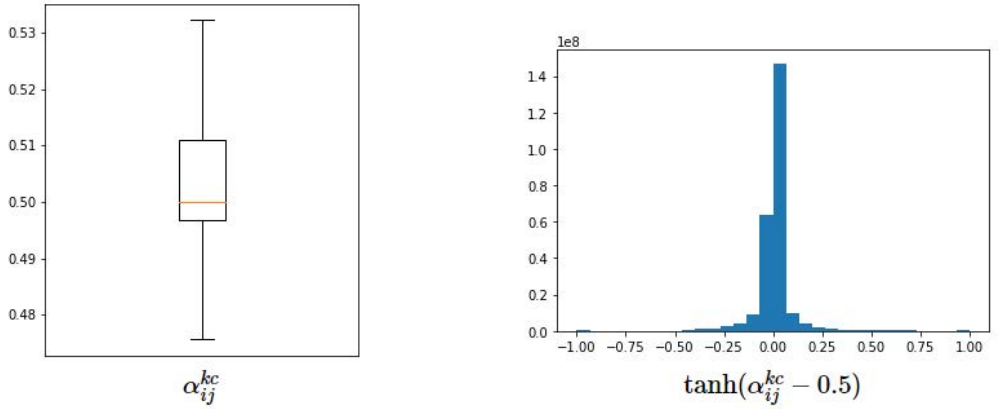


Figure 2: Distribution of (non-zero) α_{ij}^{kc} across 10,000 images from ImageNetV2. On the left a boxplot of the values of α_{ij}^{kc} (with outliers removed) is shown. On the right there is a histogram of the values of $\tanh(\alpha_{ij}^{kc} - 0.5)$.

4 Empirical tests

The previous section contains mainly theoretical considerations, but the fact remains that the literature shows Grad-CAM++ usually performing better than Grad-CAM. Here we examine a possible explanation. Our hypothesis is that the main factor that makes Grad-CAM++ better than the original version of Grad-CAM is not the Grad-CAM++ algorithm per se, but the fact that in the computation of the weights only positive gradients are used. A result that can be presented in support of this hypothesis is section 7 of [Chattopadhyay et al., 2018], which shows that the performance of Grad-CAM++ drops significantly when the requirement of using only positive gradients is dropped. We will see that using only positive gradients in the computation of the weights is not only necessary, it is also sufficient in the sense that (in most cases) Grad-CAM++ does not do significantly better than Grad-CAM with positive gradients (Grad-CAM$^+$), i.e., the version of Grad-CAM in which weights are computed using equation (3). Note that the different coefficients in front of the sum are inconsequential because the heatmaps are ultimately min-max normalized.

To compare the performances of the three methods we use a metric loosely inspired in the average drop, increase in confidence, and win-percentage metrics described in [Chattopadhyay et al., 2018], with a difference: rather than averaging percentages with an arithmetic mean we will average proportions using geometric mean. The reason for this choice is that, in general, adding, subtracting, and averaging percentages is not a good idea and can lead to wrong results.

The first step is to produce explanation maps, defined as the Hadamard (point-wise) product of images and their (min-max normalized) corresponding heatmaps generated by the attribution methods:

$$E^c = L^c \odot I\,. \tag{14}$$

The explanation map E^c can be interpreted as the original image I in which the areas with least contribution to the output of the network have been obscured. Hence, if we feed the explanation map E^c to the network, we expect that the output of the network will be larger when the explanation map correctly captures the relevant areas of the image compared to the output obtained if the relevant areas are poorly captured.

Then, the relative performance of two attribution methods can be assessed by comparing the network outputs after feeding the network with each of the corresponding explanation maps. More specifically, let I_i be the ith image from the dataset, let E_i^c, $E_i'^c$ be explanation maps produced for I_i using attribution methods M and M', and let O_i^c, $O_i'^c$ be the network outputs (after softmax) obtained when feeding the network with E_i^c, $E_i'^c$. We call the quotient $O_i'^c / O_i^c$ *relative performance* of M' vs M for image I_i. Note that the relative performance will be larger than 1 if the heatmap produced by method M' captures the relevant areas of I_i better than the heatmap produced by method M does, otherwise $O_i'^c / O_i^c$ will be less than 1.

 ISBN 978-0-9934207-7-1

The relative performance of method M' versus M across a dataset can be measured as the geometric mean of the product of $O_i'^c / O_i^c$ across the given dataset:

$$\text{relative performance of } M' \text{ vs } M \text{ (across a dataset)} = \sqrt[n]{\prod_{i=1}^{n} \frac{O_i'^c}{O_i^c}}. \tag{15}$$

Note that we average the ratio $O_i'^c / O_i^c$ using a geometric rather than arithmetic mean. This choice is consistent with the use of the geometric mean in fields in which amounts are compared using ratios rather than differences (e.g. growth rates, financial indices, etc.). For some statistics we also use the log relative performance per image defined as $\log(O_i'^c / O_i^c) = \log O_i'^c - \log O_i^c$. Note that the relative performance across a dataset is the exponential of the arithmetic mean of the log relative performance across that dataset:

$$\text{log relative performance of } M' \text{ vs } M = \exp\Big\{\frac{1}{n}\sum_{i=1}^{n} \log\Big(\frac{O_i'^c}{O_i^c}\Big)\Big\}. \tag{16}$$

The testing dataset used is ImageNetV2 MatchedFrequency, which contains 10,000 images from 1,000 categories [Recht et al., 2019]. In order to separate network performance from attribution method performance, we restrict the statistics to a subset of 4219 images for which the network assigns a (post-softmax) score of more than 0.5 to the right class.

After applying the relative performance metric to each pair of algorithms of (original) Grad-CAM, Grad-CAM$^+$, and Grad-CAM++ across the dataset, we get the results shown in Table 1. The two column relative performance shows results obtained with explanation maps produced using gradients of pre-softmax and post-softmax scores respectively. As we can see, the relative performance of Grad-CAM++ and Grad-CAM$^+$ are very similar (relative performance ≈ 1), supporting our hypothesis that Grad-CAM++ is practically equivalent to Grad-CAM$^+$.

methods (M' vs M)	**relative performance**	
	pre-softmax expl. maps	post-softmax expl. maps
Grad-CAM++ vs Grad-CAM	1.24	1.27
Grad-CAM$^+$ vs Grad-CAM	1.16	1.25
Grad-CAM++ vs Grad-CAM$^+$	**1.06**	**1.01**

Table 1: Relative performances of Grad-CAM methods across the ImageNetV2 dataset for images for which the network assigns a (post-softmax) score of more than 0.5 to the right class.

We get also measures of dispersion from the distributions of log relative performances per image across the dataset. We show the means and standard deviations of the distributions in Table 2.

methods (M' vs M)	**log relative performance**			
	pre-softmax expl. maps		post-softmax expl. maps	
	mean	std	mean	std
Grad-CAM++ vs Grad-CAM	0.21	0.64	0.24	0.68
Grad-CAM$^+$ vs Grad-CAM	0.15	0.66	0.22	0.68
Grad-CAM++ vs Grad-CAM$^+$	**0.06**	**0.41**	**0.01**	**0.11**

Table 2: Log relative performances of Grad-CAM methods across the ImageNetV2 dataset for images for which the network assigns a (post-softmax) score of more than 0.5 to the right class.

Figure 3 shows boxplots of relative performance per image across the same dataset. A value of 1 means same performance, larger than 1 means the first method has better performance compared to the second, and less than 1 if the second does better than the first. We note that the distribution of the relative performance again shows that Grad-CAM++ and Grad-CAM$^+$ yield very similar results, particularly if we use explanation maps obtained using gradients of post-softmax scores.

 ISBN 978-0-9934207-7-1

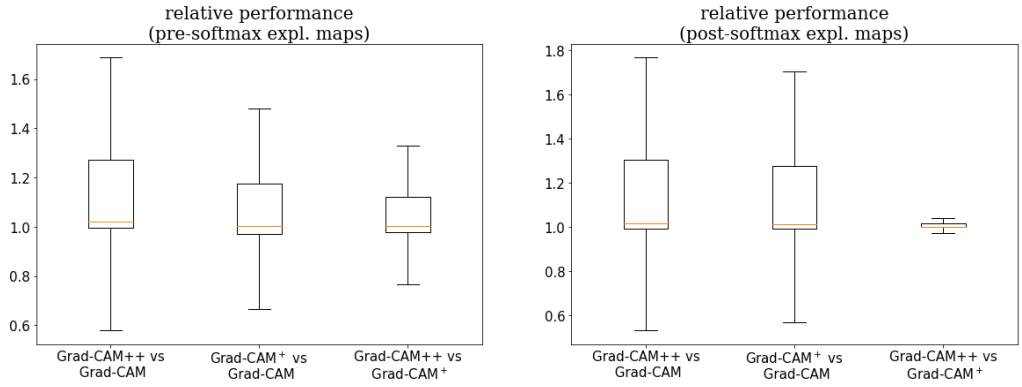


Figure 3: Boxplots of distributions of relative performance per image.

5 Conclusions and Future Work

We have examined the way Grad-CAM and Grad-CAM++ work, and compared them to a version of Grad-CAM, that we call Grad-CAM$^+$, in which gradients are replaced with positive gradients in the computation of the weights used to combine the feature maps that produce heatmaps. We have critically examined the methodology behind the design of the Grad-CAM++ algorithm, uncovering a number of weak points in it, and then showed that the algorithm is in fact very approximately equivalent to the much simpler Grad-CAM$^+$. The tests still show a minor performance improvement of Grad-CAM++ over Grad-CAM$^+$ (about 1% post-softmax and 6% pre-sofmax), but we do not consider it to be significant enough to modify the conclusion that both methods are *approximately* equivalent.

In future work additional effort can be dedicated to the comparison of Grad-CAM++ vs Grad-CAM$^+$ using different datasets and network models, and using other evaluation techniques. Also, although the methodology used to design the Grad-CAM++ is unclear, the idea of a point-wise weighting (the alphas) during the computation of the weights (w_k^c) may still be salvaged, but equation (13) would need to be replaced with a new one that does not suffer from numerical instability and is properly justified.

References

[Chattopadhyay et al., 2018] Chattopadhyay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*.

[Recht et al., 2019] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? In *ICML*.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-Cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[Simonyan and Zisserman, 2015] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *ICLR*.

Dynamic Channel Selection in Self-Supervised Learning

Tarun Krishna*, Ayush K. Rai*, Yasser A. D. Djilali, Alan F. Smeaton,
Kevin McGuinness and Noel E. O'Connor

Insight Centre for Data Analytics, Dublin City University, Dublin, Ireland

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/LKLI8696

Abstract

Whilst computer vision models built using self-supervised approaches are now commonplace, some important questions remain. Do self-supervised models learn highly redundant channel features? What if a self-supervised network could dynamically select the important channels and get rid of the unnecessary ones? Currently, convnets pre-trained with self-supervision have obtained comparable performance on downstream tasks in comparison to their supervised counterparts in computer vision. However, there are drawbacks to self-supervised models including their large numbers of parameters, computationally expensive training strategies and a clear need for faster inference on downstream tasks. In this work, our goal is to address the latter by studying how a standard channel selection method developed for supervised learning can be applied to networks trained with self-supervision. We validate our findings on a range of target budgets t_d for channel computation on image classification task across different datasets, specifically CIFAR-10, CIFAR-100, and ImageNet-100, obtaining comparable performance to that of the original network when selecting all channels but at a significant reduction in computation reported in terms of FLOPs.

Keywords: Dynamic Neural Networks, Self-Supervised Learning (SSL), Computer Vision

1 Introduction

Self-supervised pre-training of convolutional neural networks, as mentioned in [Chen et al., 2020], has almost matched the performance of supervised pre-training on the ImageNet [Deng et al., 2009] image classification task, but at a cost of a huge number of parameters and inefficient training and inference methods. In the supervised learning setting, as described in [Veit and Belongie, 2018], it is accepted that networks with dynamic data dependent (conditional) channel computation architectures during inference can lead to enhanced representation power, adaptivity, interpretability and can greatly reduce computation cost and memory resources without compromising on the accuracy by a significant margin. This motivates us to investigate the behaviour of neural networks with a channel selection mechanism trained under self-supervision. We hypothesise that self-supervised models are an ideal candidate for such dynamic network structures as they capture highly redundant channel features during pre-training. In addition, there is also a great need to explore more efficient inference methods on downstream tasks for SSL.

In order to establish the trade off between computation and performance, there are two well established research directions when it comes to introducing channel sparsity using dynamic structure in neural networks: channel pruning and channel conditional computation. Dynamic channel pruning, as reported in [Gao et al., 2018], estimates channel saliency measures and allows a network to learn and prioritise certain channels and ignore the irrelevant ones given a fixed target density. Models based on pruning usually learn sparsity through a three stage pipeline i.e, pretrain-prune-finetune while in other works like [Tiwari et al., 2021] the pruning stage itself consists of two steps, namely soft pruning and hard pruning. Conditional channel computation as proposed in [Herrmann et al., 2020] learns to compute only a subset of channels in every layer for the

*Equal contribution.

given input and hence benefits inference time efficiency and provides an insight into dataset specific network behaviour. Both channel pruning and conditional channel computation are categorical decisions that cannot be optimised by gradient descent methods; however, using the Gumbel-Softmax trick from [Jang et al., 2016] provides a way to overcome this challenge. Adafuse [Meng et al., 2020] proposed an adaptive temporal fusion network that learns a decision policy to dynamically fuse channels from current and history feature maps (i.e. dynamically deciding which channels to keep, reuse or skip per layer and per instance) for action recognition. Notwithstanding these works, the use of dynamic networks for channel selection has to date been mostly limited to supervised learning settings only.

To the best of our knowledge, there is no study on the impact of conditional channel selection on SSL. The work described in [Caron et al., 2020b] studies the effect of standard pruning techniques developed for supervised learning on a network trained with self-supervision. In particular they use an iterative magnitude based pruning technique described in [Han et al., 2015], which compresses the network by alternatively minimising a training objective and pruning the network parameters with smallest magnitude. The weights of the resulting sub-network are reset based on a weight initialisation scheme: the lottery winning ticket [Frankle and Carbin, 2018, Frankle et al., 2020]. We adopt a similar strategy but focus on exploring the application of standard conditional channel selection methods, as proposed in [Li et al., 2021], to self-supervised models during the pre-training stage and do not include any re-training. Our contributions can be summarised as follows:

1. *Do self-supervised models learn redundant channel features*? Through our exhaustive evaluation we demonstrate that the SSL model (SimSiam) does indeed learn redundant channel features.

2. We show in Table 1 that exploiting this redundancy leads to a drop in computational complexity (FLOPs), reducing inference time without excessively increasing training time, as we learn from scratch and on-the-fly, unlike competing approaches [Caron et al., 2020b] that involve re-training.

3. We demonstrate that this channel selection mechanism preserves the feature quality when evaluated on the task of image classification and gives comparable performance when compared with a vanilla (no channel selection) SSL approach.

2 Related Works

2.1 Self-Supervised Representation Learning

SSL has recently matched the performance of supervised learning on several computer vision benchmarks [Chen et al., 2020, Djilali et al., 2021, Krishna et al., 2021, Bachman et al., 2019, Grill et al., 2020].

Contrastive Learning. Contrastive learning refers to learning by comparison [Oord et al., 2018, Chen et al., 2020] where the final objective is based on some variation of Noise Contrastive Estimation (NCE). The main intuition is to bring similar instances closer in the embedding space while contrasting them with other negative samples to avoid feature collapse. These methods are usually trained in a Siamese setting with shared weights using a large batch size or memory bank [Chen et al., 2020, Oord et al., 2018, Wu et al., 2018, Misra and Maaten, 2020]. Here we present summaries of a series of important aspects that contribute to this.

Clustering Methods. One category of self-supervised methods for representation learning is based on clustering [Caron et al., 2018, Asano et al., 2019, Caron et al., 2020a], which alternates between clustering the representations and learning to predict the cluster assignment. These clustering method are also based on contrastive approaches but at cluster level, which also makes the training computationally expensive.

Distillation Methods. Recent approaches like BYOL [Grill et al., 2020] and SimSiam [Chen and He, 2021], need no negative samples, yet they learn useful representations and perform on-par with other SSL methods. They learn in a student-teacher setting and consequently avoid feature collapse. However, why and how they avoid collapse is still unclear and an open research area.

Information Maximization. A more principled way to avoid feature collapse is to capture information bottlenecks as in Barlow Twins [Zbontar et al., 2021] and VICReg [Bardes et al., 2021].

 ISBN 978-0-9934207-7-1

It is unclear how many channel redundant features are learned by these self-supervised approaches. In this work we aim to study this redundancy by exploiting a dynamic channel selection mechanism from the literature.

2.2 Dynamic Channel Computation

Channel Pruning. Channel pruning estimates channel saliency measures and eliminates all input and output connections from unimportant channels. The approach reported in [Wen et al., 2016] added group Lasso on channel weights to the model's training loss function resulting in a reduction of the magnitude of channel weights during training. The authors in [He et al., 2018] proposed pruning channels using thresholds by setting unimportant channels to zero. Network Slimming [Liu et al., 2017] used Lasso regularisation with global thresholds. However, deep models pruned with structured sparsity methods lose their capabilities and connections permanently. As a result, dynamic channel pruning methods were devised that learn sparsity through a three-stage pipeline pretrain-prune-finetune or use pretrained models. The authors of [Gao et al., 2018] propose feature boosting and suppression (FBS) to dynamically amplify and suppress output channels computed by convolutional layers. [Tiwari et al., 2021] presents a deterministic pruning strategy using the continuous heaviside function and *crispness loss* to identify a highly sparse subnetwork from an existing dense network.

Conditional Channel Computation. Regarding conditional computation at the channel level, the work proposed in [Lin et al., 2017] generates decisions to skip computation for a subset of output channels. The channel gating network [Hua et al., 2019] finds regions among the features that contribute less to the classification result and skips computation on a subset of the input channels for these ineffective regions. ConvAIG [Veit and Belongie, 2018] introduced a network with a hard attention mechanism that adaptively selects specific layers of importance for each input image to assemble an inference graph by specifying a target rate for each layer. The authors of [Herrmann et al., 2020] also study conditional computation at the channel level and extend ConvAIG by learning target rates for each gate by specifying the target rate for the whole network. DGNet [Li et al., 2021] proposed a dual gating mechanism by introducing sparsity along two separate dimensions, spatial and channel, in order to reduce model complexity at run time. For a more detailed background on sparsity, pruning and conditional computation, we recommend the review work presented in [Hoefler et al., 2021].

While [Caron et al., 2020b] studied the behaviour of self-supervised models under standard pruning techniques, we investigate the effect of standard channel selection methods described in DGNet on self-supervised models. We also analyse whether networks trained under self-supervision with channel selection can preserve performance on downstream tasks.

3 Method

3.1 Self-supervised Module

In this work we consider SimSiam [Chen and He, 2021] as our self-supervised objective. We use ResNet18 as a base encoder*, which takes two augmented views $\mathbf{x}_1$ and $\mathbf{x}_2$ from an anchor view $\mathbf{x}$ by applying stochastic augmentation from a set of augmentations $\mathcal{P}$. $\mathcal{P}$ comprises random resized crop, color jitter, random gray scale, Gaussian blur and random horizontal flip. These augmented views are processed through $f_{\boldsymbol{\theta}}$ to get a compact representation of $f_{\boldsymbol{\theta}}(\mathbf{x}_1), f_{\boldsymbol{\theta}}(\mathbf{x}_2) \in \mathbb{R}^{512}$. One view is further processed by a prediction MLP head (bottleneck architecture) $g_{\boldsymbol{\phi}}$ giving rise to an asymmetric architecture i.e. $\mathbf{p}_1 \triangleq g_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(\mathbf{x}_1)))$ and $\mathbf{z}_2 \triangleq f_{\boldsymbol{\theta}}(\mathbf{x}_2)$. As a standard practise, a base encoder is augmented with a projection head MLP i.e., $f_{\boldsymbol{\theta}} = h \circ m$, where m and h represents ResNet18 and projection layers respectively. The SimSiam learning objective simplifies to a symmetric cosine similarity:

$$\mathcal{L}_{\texttt{SSL}} = \frac{1}{2}\mathcal{D}(\mathbf{p_1}, \texttt{SG}(\mathbf{z_2})) + \frac{1}{2}\mathcal{D}(\mathbf{p_2}, \texttt{SG}(\mathbf{z_1})), \tag{1}$$

where $\mathcal{D}(\mathbf{a}, \mathbf{b}) = -\mathbf{a}^T\mathbf{b}$, with $\mathbf{a}$ and $\mathbf{b}$ being L_2 normalised vectors*. `SG` stands for `Stop-Grad()`.

*across all experiments

*i.e. $\mathcal{D}(\mathbf{a}, \mathbf{b})$ is negative cosine similarity.

 ISBN 978-0-9934207-7-1

3.2 Channel Selection via Gating

Preliminaries. Channel selection or conditional computation (data dependent gates) is often realised through a gating mechanism. A typical output for an input $\mathbf{x}$ from a convolutional (conv) layer l is given by $f_l(\mathbf{x}_{l-1}) \in \mathbb{R}^{C\times H\times W}$ where $f_l(\mathbf{x}_{l-1})$ consists of a convolution operation with kernel size k followed with a batch normalization layer (BN) and relu $((\cdot)_+)$ non-linearity with $\mathbf{x}_{l-1}$ being the output from the previous layer. The output from a gated convolutional network can be realised as: $\hat{f}_l(\mathbf{x}_{l-1}) = \pi_l(\mathbf{x}_{l-1}) \cdot \text{BN}(\texttt{conv}_l(\mathbf{x}_{l-1}))_+$, where $\pi_l(\mathbf{x}_{l-1})^* \in \{0,1\}^C$ is a gate dependent on input $\mathbf{x}_{l-1}$, which decides whether to keep ("on") or discard ("off") a particular channel. This can be seen as a form of *hard attention* (mask). This masking imposes a discrete structure over the network, making a computational graph for training and inference different. During training this structure is realised through stochastic gradient descent (SGD), while during inference it works as *hard attention*. One of the main reasons for channel selection is to induce sparsity i.e. operate on a lower computational budget (less FLOPs) during inference. In this work we closely follow DGNet using ResNet18 as our base encoder.

Channel Selection (Gating). In order for gates[*] to be effective, they need to estimate the importance of input features. This *importance* is often referred to as relevance/saliencies (vectors) of the input feature map (along the channels) in the literature. This relevance is crucial in order for the network to avoid trivial solutions. A simpler way is to use SE block [Hu et al., 2018], as was used in DGNet, to create a relevance vector. This usually requires getting a context vector $\mathbf{z} \in \mathbb{R}^C$ via global average pooling to accumulate spatial information. Finally, this context vector $\mathbf{z}$ is passed through a lightweight network to get channel attention $g_l(\mathbf{x}_{l-1})$, which can be summarized as:

$$g(\mathbf{x}_{l-1}) = \mathbf{W}_1\Big(\text{BN}\big(\mathbf{W}_0\mathbf{z}\big)\Big)_+, \quad \mathbf{W}_1 \in \mathbb{R}^{C_l\times\frac{C_{l-1}}{r}}, \mathbf{W}_0 \in \mathbb{R}^{\frac{C_l}{r}\times C_l}, \tag{2}$$

where r is a reduction ratio. For more details please refer to [Hu et al., 2018]. Finally, to achieve binary mask $\pi_l(\mathbf{x}_{l-1})$ we can use the channel attention $g_l(\mathbf{x}_{l-1})$ and set $\pi_l^i(\mathbf{x}_{l-1}) = 1$ if $g_l^i(\mathbf{x}_{l-1}) \geq 0$ and $\pi_l^i(\mathbf{x}_{l-1}) = 0$ otherwise. This discrete selection works during inference but it breaks the computational graph during training. To make the training possible, the Gumbel-SoftMax Trick [Jang et al., 2016] is adopted. The Gumbel-Trick has been widely used as reparameterisation technique for the task of dynamic channel selection [Li et al., 2021, Herrmann et al., 2020, Veit and Belongie, 2018, Meng et al., 2020]. A gating block is introduced after the first convolution in `Basic Block` of Resnet18 following DGNet. Intuitively, the channel selection network could be interpreted as learning a policy whether to keep (compute) or discard (skip) a particular channel.

3.3 Optimisation

To remove unimportant channels and induce sparsity in the gating mask $\pi_l(\mathbf{x}_{l-1})$ we need to add an objective based on some budget t_d. To this end we use regularisation, a term used in DGNet as sparsity objective, which is a combination of sparsity and a bound regularisation term:

$$\mathcal{L}_G = \lambda \underbrace{\left(\frac{\sum_{l=1}^{L} F_l^R}{\sum_{l=1}^{L} F_l^O} - t_d\right)^2}_{\text{Sparsity}} + \gamma\mathcal{L}_{Bound},$$

where F_l^R is the average FLOPs over the batch along with FLOPs computation of the gating block, while F_l^O is the original FLOPs without a gating module. Only the blocks with gating modules take part in FLOP computation as they are responsible for any sort of sparsity introduced in the network. The purpose of L_{Bound} is to control early optimisation as detailed in DGNet.

Final Objective. Overall training objective is defined as: $\mathcal{L} = \mathcal{L}_{\texttt{SSL}} + \mathcal{L}_G$, with $\lambda = 5$ and $\gamma = 1$ across all the datasets and training regimes.

[*] a vector of dimension equivalent to number of channels with ones and zeroes

[*] channel selection

 ISBN 978-0-9934207-7-1

4 Experimental Setup

Table 1: Performance comparison of SimSiam with dynamic channel selection during inference. Evaluated with k-nearest neighbours ($k = 1$) on the validation set of CIFAR-10, CIFAR-100 and ImageNet-100 across various target budgets t_d. * denotes baseline.

Dataset	Budget (t_d)	Acc%	FLOPs
CIFAR-10	*	85.46%	7.03E8
	10%	76.72%	6.64E7 (90.55%↓)
	20%	78.78%	1.25E8 (82.11%↓)
	30%	80.82%	2.01E8 (71.41%↓)
	40%	81.35%	2.66E8 (62.18%↓)
	50%	81.93%	3.29E8 (53.15%↓)
	60%	82.96%	3.94E8 (43.89%↓)
	70%	**83.08**%	4.58E8 (34.76%↓)
CIFAR-100	*	52.96%	7.03E8
	10%	46.84%	6.88E7 (90.21%↓)
	20%	49.50%	1.48E8 (78.88%↓)
	30%	50.70%	2.03E8 (71.01%↓)
	40%	52.05%	2.65E8 (62.32%↓)
	50%	52.54%	3.25E8 (53.67%↓)
	60%	53.18%	3.95E8 (43.73%↓)
	70%	**53.50**%	4.66E8 (33.69%↓)
ImageNet-100	*	64.34%	1.81E9
	30%	56.08%	5.30E8(70.43%↓)
	40%	57.86%	7.13E8 (60.66%↓)
	50%	**60.38**%	8.78E8 (51.55%↓)

Implementation Details. We closely follow the approach in DGNet for channel selection. For training, we use SimSiam as a self-supervised model with ResNet18 as a base encoder whose objective is modified as explained in section 3.3. We train the model with varying target densities t_d. The implementation of SimSiam is based on the solo-learn library [da Costa et al., 2022]. The base encoder is *randomly initialised*[*] and is trained with SGD for 500 epochs (for a given target budget t_d) with a batch size of 256 on 2 Nvidia 2080Ti GPUs, with a warm-start of 10 epochs following a cosine decay with base learning rate of 0.01. Since we are using a very lightweight model as our gating network, there is no significant computational overhead during training. We report the inference speedup in terms of the hardware-independent theoretical metric of FLOPs and not wall-clock time as we are not using any hardware accelerators to utilise sparsity during training. Code is made available here.

Evaluation. Training and evaluation is carried on train and validation data of CIFAR-10, CIFAR-100 and ImagNet-100 respectively. For Cifar-10/100 we train for $t_d = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ while for ImageNet-100 we restrict t_d to only $\{0.3, 0.4, 0.5\}$ due to computational constraints. We use k-nearest neighbours as our evaluation metric evaluated with $k = 1$. For *baseline* we train Simsiam with standard objective without any channel selection for each of the datasets under consideration (* in Table 1).

Results. Our main findings based on the evaluation criteria validate our initial hypothesis that self-supervised models can learn highly redundant channel features.

Table 1 shows that in the case of CIFAR-10, by keeping only 70% of the channels across the whole network, SimSiam achieves 83.08% accuracy on the KNN task, which is a minor drop from the baseline performance of 85.46% but at an ample reduction of 34.76% in FLOPs. Furthermore, we also find that an enormous 90.55% of FLOPs can be reduced by using only 10% of the channels across the whole network causing a drop of only 8.74% in KNN accuracy. For CIFAR-100, we found that by restricting the channel usage to only 60% over the whole network, SimSiam surpasses the baseline KNN accuracy of 52.95% by 0.22% reaching 53.18%. Additionally, FLOP computation can be reduced by 90.21% by keeping only 10% of the channels, leading to a drop of only 6.12% in KNN accuracy. On ImageNet-100, 50% of the channel usage in the entire network results in 60.38% KNN accuracy, which is 3.96% less than the baseline. However, this decrease in accuracy is compensated by ∼51.55% percent drop in FLOPs. Aside from this, we get a substantial 70.43% drop in FLOPs by fixing channel utilization to only 30% in the whole model. Therefore, channel selection can be thought of as a way to take advantage of the trade-off between performance and computation depending the downstream task and individual use case. These results also show that SSL models trained with channel selection preserve the performance in downstream tasks.

Figure 1 shows the channel activation distribution for CIFAR-10, CIFAR-100 and ImageNet-100 datasets, revealing a deeper insight into the dataset specific behaviour of the channel selection network by visualising how many channels in each ResNet18 blocks are always off (skipped), always on (computed), or input dependent.

[*] default initialisation in Pytorch

 ISBN 978-0-9934207-7-1

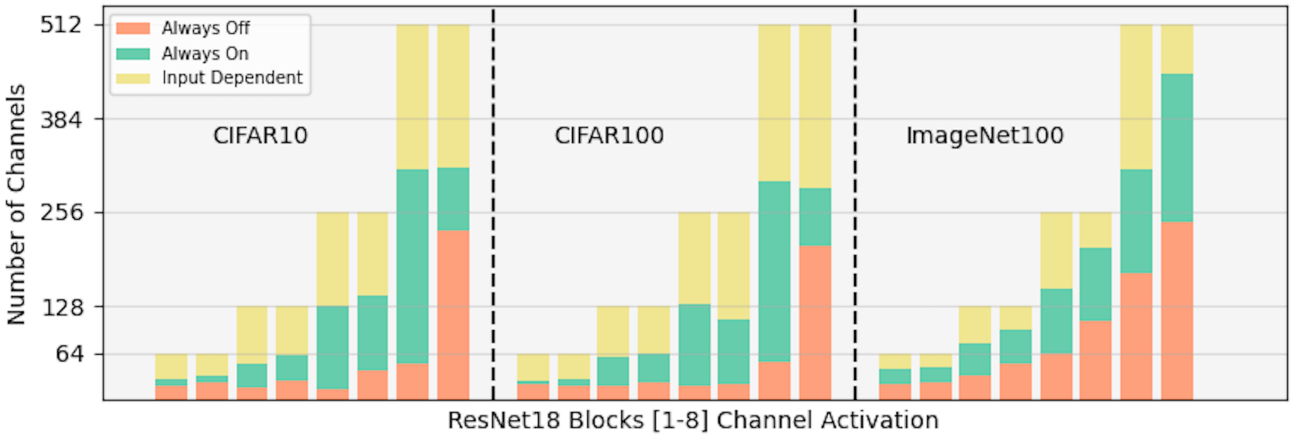


Figure 1: Channel distribution over validation set for $t_d = 0.5$ on CIFAR-10, CIFAR-100, ImageNet-100

We notice that significant number of channels are switched off and switched on all the time in all the three datasets and while others are input dependent. The channel distribution for CIFAR-10 and CIFAR-100 are very similar, which might be due to the fact that image statistics in both of these datasets are similar.

5 Conclusion

In this paper, we studied the behaviour of self-supervised learning when integrated with channel selection networks given a global target budget for computational cost. Our empirical results provided interesting insights about self-supervised learning when trained with channel selection. First, self-supervised models learn highly redundant channel features that can be discarded to reduce computational overhead (Figure 1, Table 1). Second, we showed that channel selection modules can significantly reduce FLOP computation and make inference more efficient (Table 1). Third, our results also provide intuition that representations learnt by self-supervised networks with channel selection can also be transferred to downstream tasks.

There are, however, some limitations with our work. First, we still need to evaluate the tranferability of learned representations beyond classification to other downstream tasks such as object segmentation, detection and instance retrieval to name a few. Second, the SSL training objective involves maximizing the agreement between augmented views of the same object or scenes (instance discrimination) and this forces them to have similar representations in the embedding space. In this work, we do not account for this by enforcing some consistency aware constraints for channel selection in the training objective. These limitations will be addressed in future work.

Acknowledgments

This work has emanated from research supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund and Xperi FotoNation.

References

[Asano et al., 2019] Asano, Y. M., Rupprecht, C., and Vedaldi, A. (2019). Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*. 2

[Bachman et al., 2019] Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32. 2

[Bardes et al., 2021] Bardes, A., Ponce, J., and LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*. 2

[Caron et al., 2018] Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149. 2

[Caron et al., 2020a] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020a). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924. 2

[Caron et al., 2020b] Caron, M., Morcos, A., Bojanowski, P., Mairal, J., and Joulin, A. (2020b). Pruning convolutional neural networks with self-supervision. *arXiv preprint arXiv:2001.03554*. 2, 3

[Chen et al., 2020] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR. 1, 2

[Chen and He, 2021] Chen, X. and He, K. (2021). Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758. 2, 3

[da Costa et al., 2022] da Costa, V. G. T., Fini, E., Nabi, M., Sebe, N., and Ricci, E. (2022). solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56):1–6. 5

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee. 1

[Djilali et al., 2021] Djilali, Y. A. D., Krishna, T., McGuinness, K., and O'Connor, N. E. (2021). Rethinking 360deg image visual attention modelling with unsupervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15414–15424. 2

[Frankle and Carbin, 2018] Frankle, J. and Carbin, M. (2018). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*. 2

[Frankle et al., 2020] Frankle, J., Dziugaite, G. K., Roy, D., and Carbin, M. (2020). Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR. 2

[Gao et al., 2018] Gao, X., Zhao, Y., Dudziak, Ł., Mullins, R., and Xu, C.-z. (2018). Dynamic channel pruning: Feature boosting and suppression. *arXiv preprint arXiv:1810.05331*. 1, 3

[Grill et al., 2020] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284. 2

[Han et al., 2015] Han, S., Pool, J., Tran, J., and Dally, W. (2015). Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28. 2

[He et al., 2018] He, Y., Kang, G., Dong, X., Fu, Y., and Yang, Y. (2018). Soft filter pruning for accelerating deep convolutional neural networks. *arXiv preprint arXiv:1808.06866*. 3

[Herrmann et al., 2020] Herrmann, C., Bowen, R. S., and Zabih, R. (2020). Channel selection using gumbel softmax. In *European Conference on Computer Vision*, pages 241–257. Springer. 1, 3, 4

[Hoefler et al., 2021] Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., and Peste, A. (2021). Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124. 3

[Hu et al., 2018] Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141. 4

[Hua et al., 2019] Hua, W., Zhou, Y., De Sa, C. M., Zhang, Z., and Suh, G. E. (2019). Channel gating neural networks. *Advances in Neural Information Processing Systems*, 32. 3

[Jang et al., 2016] Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*. 2, 4

[Krishna et al., 2021] Krishna, T., McGuinness, K., and O'Connor, N. (2021). Evaluating contrastive models for instance-based image retrieval. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 471–475. 2

[Li et al., 2021] Li, F., Li, G., He, X., and Cheng, J. (2021). Dynamic dual gating neural networks. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5310–5319. 2, 3, 4

[Lin et al., 2017] Lin, J., Rao, Y., Lu, J., and Zhou, J. (2017). Runtime neural pruning. *Advances in neural information processing systems*, 30. 3

[Liu et al., 2017] Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., and Zhang, C. (2017). Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE international conference on computer vision*, pages 2736–2744. 3

[Meng et al., 2020] Meng, Y., Panda, R., Lin, C.-C., Sattigeri, P., Karlinsky, L., Saenko, K., Oliva, A., and Feris, R. (2020). Adafuse: Adaptive temporal fusion network for efficient action recognition. In *International Conference on Learning Representations*. 2, 4

[Misra and Maaten, 2020] Misra, I. and Maaten, L. v. d. (2020). Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717. 2

[Oord et al., 2018] Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*. 2

[Tiwari et al., 2021] Tiwari, R., Bamba, U., Chavan, A., and Gupta, D. K. (2021). Chipnet: Budget-aware pruning with heaviside continuous approximations. *arXiv preprint arXiv:2102.07156*. 1, 3

[Veit and Belongie, 2018] Veit, A. and Belongie, S. (2018). Convolutional networks with adaptive inference graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18. 1, 3, 4

[Wen et al., 2016] Wen, W., Wu, C., Wang, Y., Chen, Y., and Li, H. (2016). Learning structured sparsity in deep neural networks. *Advances in neural information processing systems*, 29. 3

[Wu et al., 2018] Wu, Z., Xiong, Y., Yu, S., and Lin, D. (2018). Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*. 2

[Zbontar et al., 2021] Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR. 2

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/VHMQ4826

Unsupervised Scale-Invariant Multispectral Shape Matching

Idan Pazi, Dvir Ginzburg, Dan Raviv

Tel Aviv University

Abstract

Alignment between non-rigid stretchable structures is one of the most challenging tasks in computer vision, as the invariant properties are hard to define, and there is no labeled data for real datasets. We present unsupervised neural network architecture based upon the spectral domain of scale-invariant geometry. We build on top of the functional maps architecture, but show that learning local features, as done until now, is not enough once the isometry assumption breaks. We demonstrate the use of multiple scale-invariant geometries for solving this problem. Our method is agnostic to local-scale deformations and shows superior performance for matching shapes from different domains when compared to existing spectral state-of-the-art solutions.

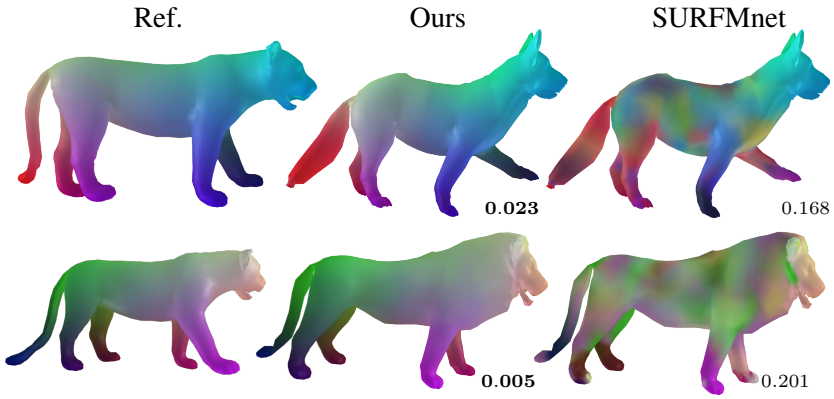


Figure 1: Unsupervised functional maps correspondence between highly non-isometric shapes from the SMAL dataset. Visualization of texture transfer. Mean geodesic error is marked next to each correspondence.

Keywords: Computer Vision, 3D Shape Matching, Deep Learning

1 Introduction

Shape correspondence is the problem of aligning features between two shapes and providing a pointwise map that reveals the underlying common anatomy of the shapes. The correspondence problem is especially challenging in the non-rigid case where stretching and bending are allowed, as shape features may differ substantially. The solution to this problem and its derivatives is vital for many applications, such as texture transfer, pose transfer, and space-time reconstruction. Trying to solve this problem directly, matching every point on the spatial domain yields a vast solution space, non-convex, and highly non-linear [Mémoli, 2007]. The complexity of non-rigid shape correspondence problems led the efforts in this field to try and simplify the problem, primarily by exploiting invariants between shapes, focusing on metric-based methods where isometry or near-isometry is assumed between shapes [Bronstein et al., 2006].

Spectral domain methods for shape correspondence had increased popularity in recent years, commonly using the Laplace-Beltrami operator (LBO) eigendecomposition for spanning the spectral domain. The functional maps framework [Ovsjanikov et al., 2012], showed it is possible to find a linear transformation between the spectral domains of different shapes and laid the foundations of recent advancements. Functional

 ISBN 978-0-9934207-7-1

maps efficiently optimize the descriptors correspondence problem, shifting the problem from estimating similarity between descriptors on the complex spatial domain to solving a least-squares problem on a compact spectral domain. In practice, the optimized linear functional map between LBO decompositions is sufficiently descriptive only between near-isometric shapes. Recent spectral methods for solving the correspondence problem are concentrated on matching near-isometric shapes with similar local and global properties [Litany et al., 2017, Roufosse and Ovsjanikov, 2018, Donati et al., 2020]. Only limited attempts were made to adjust axiomatically the spectral correspondence mechanism for dealing with non-isometric shapes. Cyclic distortion [Ginzburg and Raviv, 2020] neglects the need for a consistent metric between shapes and successfully correspond stretchable domains but is still unable to correspond shapes from different domains.

[Aflalo et al., 2013] presented a scale-invariant metric that is mathematically consistent under global scaling and showed that the LBO on this geometry is invariant to local scaling. One can argue that non-isometric shapes differ by a set of affine transformations and can be approximated by such a finite set; after inducing a scale-invariant metric, shapes can be treated as near-isometric, and a distance preserving map between the shapes can be obtained when measured by the alternative pseudo-metric.

Contributions

- We introduce a new architecture for non-rigid model alignment based on multispectral scale-invariant strategy, relaxing the near-isometry assumption.
- Provide superior alignment results, by a large margin, on scenarios where extensive stretching exists and open new possibilities to match non-isometric shapes using the functional maps framework.
- We report improved alignment results also on the near-isometry non-rigid benchmarks.

2 Background

2.1 Riemannian Manifold

A Riemannian manifold $(\mathcal{X}, g)$ in the 3-dimensional space is a parameterized surface $\mathcal{X} : \Omega_{\mathcal{X}} \subset \mathbb{R}^2 \to \mathbb{R}^3$ with g being a metric that varies differentiably with $p \in \Omega_{\mathcal{X}}$. The metric g can be viewed as a family of inner products $g_p(p_1, p_2) = \langle p_1, p_2 \rangle_p$ between vectors p_1, p_2 in the neighborhood of p. A *map* $T : \Omega_{\mathcal{X}} \to \Omega_{\mathcal{Y}}$ from one Riemannian manifold to another is a function that maps every point from one shape to the other.

2.2 Laplace-Beltrami operator spectral domain

The *Laplace-Beltrami operator* (LBO) of $f : \Omega_{\mathcal{X}} \to \mathbb{R}$, a differentiable function over a manifold, is defined as:

$$\Delta_g f = -\frac{1}{\sqrt{\det g}} \sum_{ij} \frac{\partial}{\partial x_i} (g^{-1} \sqrt{\det g} \, \frac{\partial}{\partial x_j} f) \tag{1}$$

with x_1, x_2 are coordinates in $\Omega_{\mathcal{X}}$ and g is the metric tensor $g = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}$, $g_{ij} = \langle p_i, p_j \rangle_p$. Hence, the operator is entirely defined by the metric tensor g. It is a positive semi-definite operator and admits an eigendecomposition $\Delta_g \phi_i = \lambda_i \phi_i$ of an orthonormal basis $\langle \phi_i, \phi_j \rangle = \delta_{ij}$ and a function on the manifold can be presented as a Fourier series $f = \sum \langle \phi_i, f \rangle_{\mathcal{X}} \phi_i$. It has been shown that the first k eigenfunctions of the LBO decomposition are the optimal k-dimensional basis for representing smooth functions on $\mathcal{X}$ [Aflalo et al., 2014].

2.3 Scale-invariant geometry

[Aflalo et al., 2013] constructed a metric on a parameterized differentiable surface that is invariant to scaling. Such metric $\tilde{g}$ could be achieved by adjusting the Euclidean metric g by the magnitude of the Gaussian curvature K:

$$\tilde{g} = |K| \, g \tag{2}$$

With this pseudo metric, distances on the surface $\mathcal{X}$ are preserved when locally scaled by some factor. This pseudo-metric can be used to define a scale-invariant LBO $\Delta_{\tilde{g}}$ by substituting Eq. (2) into Eq. (1).

 ISBN 978-0-9934207-7-1

2.4 Functional maps

Functional maps [Ovsjanikov et al., 2012], are spectral maps between pairs of shapes that allow efficient inference and manipulation. Given two parameterized manifolds $\mathcal{X} : \Omega_{\mathcal{X}} \subset \mathbb{R}^2 \to \mathbb{R}^3$ and $\mathcal{Y} : \Omega_{\mathcal{Y}} \subset \mathbb{R}^2 \to \mathbb{R}^3$, each with a space of real valued functions on the domain, $\mathcal{F}(\Omega_{\mathcal{X}}, \mathbb{R})$ and $\mathcal{F}(\Omega_{\mathcal{Y}}, \mathbb{R})$, our goal is to represent a bijective mapping between the two manifolds, namely $T : \Omega_{\mathcal{X}} \to \Omega_{\mathcal{Y}}$. Given some real-valued function $f : \Omega_{\mathcal{X}} \to \mathbb{R}$ on the first manifold, we may apply the mapping to obtain a corresponding function on the second manifold by $f \circ T^{-1} : \Omega_{\mathcal{Y}} \to \mathbb{R}$. This function composition can be viewed as a function transformation, represented by $T_F : \mathcal{F}(\Omega_{\mathcal{X}}, \mathbb{R}) \to \mathcal{F}(\Omega_{\mathcal{Y}}, \mathbb{R})$. The goal mapping T can be recovered from the function transformation T_F. Assuming the function spaces $\mathcal{F}(\Omega_{\mathcal{X}}, \mathbb{R})$ and $\mathcal{F}(\Omega_{\mathcal{Y}}, \mathbb{R})$ have orthonormal basis Φ and Ψ respectively, any function $f : \Omega_{\mathcal{X}} \to \mathbb{R}$ can be represented as a linear combination $f = \sum_i a_i \phi_i$, and T_F is a linear transformation:

$$T_F(f) = T_F\left(\sum_i a_i \phi_i\right) = \sum_i a_i T_F(\phi_i) = \sum_i \sum_j a_i c_{ij} \psi_j \tag{3}$$

Where c_{ij} is the j'th coefficient in the basis span by Ψ of the transformation of the i'th basis function, hence $T_F(\phi_i) = \sum_j c_{ij}\psi_j$, meaning that $T_F(\phi_i)$ determined only by the basis and the map T. On orthonormal basis functions we have $c_{ij} = \langle T_F(\phi_i), \psi_j \rangle$. By representing c_{ij} as a matrix $C = (c_{ij})$ and f as a vector of coefficients $\vec{a} = (a_0, a_1, ..., a_i, ...)$ the transformation takes the form of:

$$T_F(\vec{a}) = C\vec{a} \tag{4}$$

The problem of matching descriptors of two shapes can be transformed from non-convex highly non-linear constraints about every point in each shape to a linear least-squares problem over the elements of the matrix C. Given a set of descriptors (real valued functions) on each shape F and G and their projection to the spectral domains - F_Φ and G_Ψ, finding $C_{\Phi\Psi}$, a linear transformation from the spectral domain Φ to Ψ, is a least-squares problem:

$$C_{\Phi\Psi} = \underset{C}{\operatorname{argmin}} ||CF_\Phi - G_\Psi||_F^2 \tag{5}$$

C of significantly smaller size compared to the number of vertices can successfully describe the full correspondence with a low error when the LBO is used to span the spectral domain on each shape [Ovsjanikov et al., 2012]. So far, the research in the field of functional maps is mainly focused on the standard LBO decomposition with the Euclidean metric as the basis of choice. The choice of basis is a fundamental component for the solution, as each shape descriptor is projected into it, and the capacity of C to efficiently describe complex structures is derived from the strength of the space spanned by the basis. In this work, the limitations of this choice are researched and relaxed by using an alternative scale-invariant LBO.

2.5 Deep functional maps

[Litany et al., 2017] made a paradigm shift in FMNet, by training a supervised neural network for solving the shape correspondence problem without solving a labeling problem for each point but by refining baseline SHOT descriptors [Salti et al., 2014] and matching them using the functional maps framework. A soft error loss term was introduced for training the network, estimating the geodesic distance error at each point for a given functional map C using the ground truth correspondence and the geodesic distances between points on each shape.

In SURFMNet [Roufosse and Ovsjanikov, 2018], an unsupervised and more axiomatic approach for learning descriptors is suggested, where a set of numerical properties of the mapping matrix C has been shown to have a strong correlation to superb alignment results in terms of geodesic error. Not only this architecture is fully unsupervised and reduces the need for expensive point labeling, but it is also completely independent of the geodesic distances on the shapes, which makes this architecture compelling for corresponding non-isometric shapes where the geodesic distance between corresponding points may differ substantially between shapes.

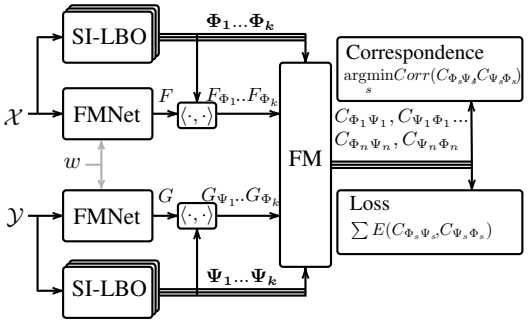


Figure 2: **Multispectral deep functional maps architecture.** Given two manifolds $\mathcal{X}, \mathcal{Y}$, each with a set of spectral domains $\{\Phi_s\}, \{\Psi_s\}$ and inferred descriptors F, G from the FMNet architecture. Correspondence between the descriptors is computed on each spectral domain yielding a set of transformations between the spectral domains of each shape (on both directions) $\{C_{\Phi_s\Psi_s}\}, \{C_{\Psi_s\Phi_s}\}$. The loss accumulated over all transformations.

3 Method

This work facilitates multispectral scale-invariant geometry for non-isometric shape correspondence, inspired by recent deep learning spectral-domain architectures. Section 3.3 and 3.2 describe our implementation of the discrete scale-invariant LBO and how the multispectral basis is constructed. Section 3.5 describes the multispectral fusion stage of the pipeline, where we gather correspondences from multiple spectral domains into a single spatial correspondence. The complete architecture of our solution is presented in Figure 2, our method is the first to propose a multispectral learnable approach for the dense non-rigid correspondence problem.[1]

3.1 Descriptors

We have used the FMNet [Litany et al., 2017] deep learning paradigm for generating local descriptors on the shapes. In the preprocessing stage, the hand-crafted SHOT algorithm was used for generating the baseline descriptors, where 352 initial descriptors are generated per point for each shape. During training and inference, each set of descriptors is refined through a series of 7 fully connected residual layers [He et al., 2015], with shared weights between the two shapes. To ease the comparison to previous methods, the descriptor refinement stage was left unchanged, and our architecture bears the same amount of learned weights as in [Halimi et al., 2018, Roufosse and Ovsjanikov, 2018]. Nonetheless, our inferred descriptors are more informative as a result of evaluating their performance on richer spectral domains when training.

3.2 Scale-invariant spectral domain

While scale-invariant LBO induces robustness to local deformations, experiments show that inducing scale-invariant geometry impairs the global properties of the decomposition. To overcome this problem, [Aflalo et al., 2014] introduced exponent parameter α to interpolate between the Euclidean metric ($\alpha = 0$) to the scale-invariant pseudo-metric ($\alpha = 1$), defining an intermediate scale-invariant geometry:

$$\tilde{g_\alpha} = |K|^\alpha \langle p_i, p_j \rangle \tag{6}$$

A visualization of the impact of α on the spectral domain is shown in Figure 3. We are the first to introduce a framework that infers correspondence by observing a set of different spectral domains.

The estimation of the Gaussian curvature was done by first applying Laplacian smoothing [Vollmer et al., 1999], then estimating the Gaussian curvature on the modified mesh according to the second fundamental form, following the algorithm presented in [Rusinkiewicz, 2004]. Curvature was clipped to a minimal value, to avoid vanishing distances on flat areas, and trimmed to a maximal value, to reduce the quantization noise of small triangles.

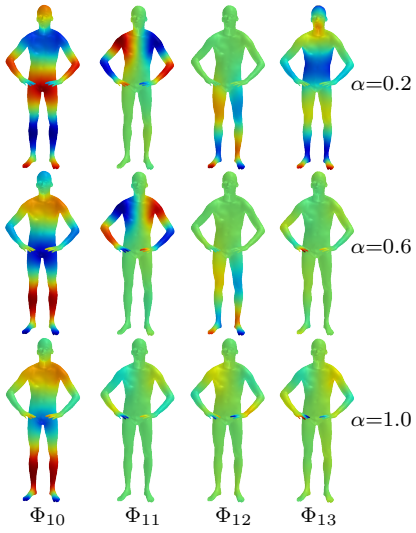


Figure 3: The 10th to 13th eigenfunctions of scale-invariant LBO with different α values. On more scale-invariant metrics (higher α), the eigenfunctions concentrates on local features.

[1] Implementation is available at github.com/idanpa/SURFMNet-Multispectral

3.3 Discrete scale-invariant LBO

The general approach for the discretization of the LBO on a triangulated surface with vertices $\{\vec{v_i}\}_{i=0}^{n-1}$ each with immediate neighbors $\{N(i)\}_{i=0}^{n-1}$ is to represent the operator as:

$$\Delta f(\vec{v_i}) := \frac{1}{d_i} \sum_{j \in N(i)} w_{ij}(f(\vec{v_i}) - f(\vec{v_j})) \tag{7}$$

with some *mass* d_i associated to each vertex and a *weight* w_{ij} associated to each edge. We use the cotangent weights [Pinkall and Polthier, 1993] and mass $d_i = a(i)/3$ where $a(i)$ is the scale-invariant area of all triangles at vertex i. Our implementation is based on the finite-element approach from [Reuter et al., 2009], achieving numerical stability by formulating the problem as a generalized eigenvalues problem:

$$A\vec{\phi} = -\lambda B\vec{\phi} \tag{8}$$

$$A(i,j) := \begin{cases} \frac{\cot(\alpha_{ij})+\cot(\beta_{ij})}{2} & (i,j) \text{ edge} \\ -\sum_{k \in N(i)} A(i,k) & i = j \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$B(i,j) := \begin{cases} \frac{|K_1|^\alpha |t_1| + |K_2|^\alpha |t_2|}{12} & (i,j) \text{ edge} \\ \frac{\sum_{l \in N(i)} |K_l|^\alpha |t_l|}{6} & i = j \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

α_{ij} and β_{ij} denote the two angles opposite to the edge (i,j), $|t|$ is the Euclidean area of the triangle t, K_i is the Gaussian curvature associated to the triangle t_i. t_1, t_2 are the triangles sharing the edge (i,j) and $\{t_l\}_{l \in N(i)}$ are the triangles in the immediate neighborhood of vertex i. α is the interpolation parameter from Eq. (6).

3.4 Loss function

Our loss function follows the unsupervised loss term presented in [Roufosse and Ovsjanikov, 2018], where a set of regularization terms applied on the functional maps $C^i_{\Phi\Psi}$ and $C^i_{\Psi\Phi}$ from Eq. (5).

Bijectivity achieved by enforcing the transformation from one spectral domain to the other to identity:

$$E_1 = ||C_{\Phi\Psi}C_{\Psi\Phi} - I||_2^2 + ||C_{\Psi\Phi}C_{\Phi\Psi} - I||_2^2 \tag{11}$$

Area preservation enforced by orthogonality:

$$E_2 = ||C_{\Phi\Psi}^\top C_{\Phi\Psi} - I||_2^2 + ||C_{\Psi\Phi}^\top C_{\Psi\Phi} - I||_2^2 \tag{12}$$

A pointwise map expresses an intrinsic isometry if and only if the functional map commutes with the LBO [Ovsjanikov et al., 2012], let Λ_Φ and Λ_Ψ the diagonal matrices of the eigenvalues of Φ and Ψ respectfully:

$$E_3 = ||C_{\Phi\Psi}\Lambda_\Phi - \Lambda_\Psi C_{\Phi\Psi}||^2 + ||C_{\Psi\Phi}\Lambda_\Psi - \Lambda_\Phi C_{\Psi\Phi}||^2 \tag{13}$$

Functional map T can represent a pointwise map if and only if it preserves the pointwise product $\bigodot$ between functions, namely $T(f \bigodot g) = T(f) \bigodot T(g)$ [Singh and Manhas, 1993]. [Nogneng and Ovsjanikov, 2017] followed this observation to present a penalty for descriptor preservation via commutativity:

$$\begin{gathered} E_4 = \sum_{f_i \in F_\Phi; g_i \in G_\Phi} ||C_{\Phi\Psi}M_{f_i} - M_{g_i}C_{\Phi\Psi}||^2 + ||C_{\Psi\Phi}M_{g_i} - M_{f_i}C_{\Psi\Phi}||^2 \\ M_{f_i} = \Phi^+ Diag(f_i)\Phi,\ M_{g_i} = \Psi^+ Diag(g_i)\Psi \end{gathered} \tag{14}$$

Where F_Φ and G_Ψ are the shape descriptors with the spectral domains Φ and Ψ respectfully, and Φ^+ is the Moore-Penrose pseudoinverse of Φ.

Given a set of functional maps $Sp = \{(C_{\Phi_s\Psi_s}, C_{\Psi_s\Phi_s})\}_{s=1}^k$ for each of the spectral domain pairs $\Phi_s\Psi_s$. The loss function accumulates loss over all spectral domains:

$$E = \sum_{C_{\Phi_s\Psi_s}, C_{\Psi_s\Phi_s} \in Sp} \sum_{i \in \{1,2,3,4\}} \omega_i E_i(C_{\Phi_s\Psi_s}, C_{\Psi_s\Phi_s}) \tag{15}$$

 ISBN 978-0-9934207-7-1

with the heuristic weights ω_i from [Roufosse and Ovsjanikov, 2018], where this loss term has been shown to strongly correlate to the desired low geodesic error of the pointwise correspondence.

The loss function is differentiable with respect to the weights and depends only on the functional maps themselves [Roufosse and Ovsjanikov, 2018], allowing unsupervised network training without a need for the ground truth correspondence, measuring expensive geodesic distances on shapes, or even inferring the pointwise correspondence while training. Instead of solving the non-convex assignment problem over thousands of distances between each point, we solve the least-squares problem over small spectral matrices. Unlike other works that presented similar penalties on the functional maps [Ovsjanikov et al., 2012, Rustamov et al., 2013, Eynard et al., 2016], the structural penalty on the functional maps is decoupled from the descriptor similarity optimization of Eq. (5).

3.5 Pointwise correspondence

Obtaining the pointwise correspondence between shapes from the functional map C is not trivial [Rodolà et al., 2015]. In [Ovsjanikov et al., 2012], an efficient method to find the pointwise correspondence is presented, based upon proximity search in the spectral domain. Given a spectral domain for each shape represented by matrices Φ on $\mathcal{X}$ and Ψ on $\mathcal{Y}$, where each column corresponds to a point and each row to an eigenfunction, for vertex i on $\mathcal{Y}$, the corresponding point on $\mathcal{X}$ is the nearest neighbor:

$$Corr(i) = \underset{j}{\operatorname{argmin}} ||col_j(C\Phi) - col_i(\Psi)||_2 \tag{16}$$

Where $Corr(i)$ is the index of the corresponding vertex to the i'th vertex and $col_i(M)$ is the i'th column of the matrix M. Given multiple spectral maps, we wish to blend the individual maps and infer a single pointwise correspondence. Direct proximity search within all spectral domains is not feasible as distances between domains are not comparable. To overcome this problem, the multispectral scheme finds the optimal correspondence by normalizing the mean and range of all distances in the different spectral domains and selects the best fit in terms of normalized distance on multiple pairs of spectral domains Φ_s, Ψ_s with functional map C_s between them:

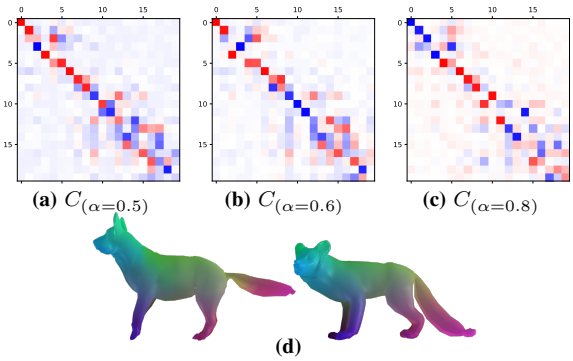
$$Corr(i) = \underset{j}{\operatorname{argmin}} \min_s ||col_j(C_s\Phi_s) - col_i(\Psi_s)||_2^* \tag{17}$$

Where $||\cdot||_2^*$ is the L_2 distance with the following normalization, on each spectral-domain we linearly normalize the distances to $[0, 1]$ and then subtract the mean of all the distances in the spectral domain, establishing a comparable distance between spectral domains. By fusing the correspondences, we can compensate for the degraded performance of scale-invariant metrics while withstanding non-isometric deformations. Figure 4 illustrates this stage.

4 Experiments

We present evaluation of our method on different datasets, showing an improvement over recent state-of-the-art solutions and highlighting the strength of our framework to overcome challenging cross-domain problems yet to be addressed using spectral methods.

For each shape, three different spectral domains were generated on the preprocessing stage, with heuristic intermediate metric factor α. For the near-isometric datasets we have used $\alpha = 0, 0.6, 0.8$ and for the non-isometric datasets we have used $\alpha = 0.5, 0.6, 0.8$. The Gaussian curvature clipping was made to the 0.4% - 75% range on each shape. Training was done for 10k iterations on FAUST dataset, on SMAL dataset training was done for 10k iteration on top of the trained model, we have used a learning rate of 0.001 with an ADAM optimizer.



(a) $C_{(\alpha=0.5)}$ **(b)** $C_{(\alpha=0.6)}$ **(c)** $C_{(\alpha=0.8)}$ **(d)**

Figure 4: Multispectral functional maps fusion. (a)(b)(c) are the functional maps for the different intermediate scale-invariant geometry (showing the first 30 eigenvectors). (d) is the combined correspondence from a dog (right) to a wolf (left) on the spatial domain.

 ISBN 978-0-9934207-7-1

4.1 Synthetic FAUST

For evaluating our method on the near-isometry case, we compared our method to [Roufosse and Ovsjanikov, 2018] and [Halimi et al., 2018] on the synthetic FAUST dataset [Bogo et al., 2014]. This dataset contains synthetic scans of ten different human subjects, each on ten different poses. As suggested in [Litany et al., 2017], we trained our model on the first eight subjects and tested it on the other two subjects. In Figure 5 we present the geodesic error comparison. Even in the near-isometry case, which our method was not optimized for, we manage to outperform state-of-the-art solutions with the same amount of weights. This shows how recent spectral domain methods bears an isometry assumption that not only limiting the capabilities to correspond shapes from different domains but also has an impact on the near-isometry case, where these methods are found to be sensitive to the stretching between different poses and different subjects.

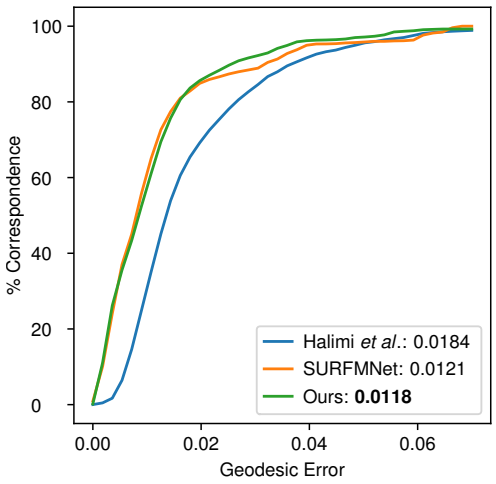


Figure 5: Error evaluation on FAUST dataset compared to [Roufosse and Ovsjanikov, 2018] and [Halimi et al., 2018] on inter-subject and intra-subject pairs. Our multispectral architecture shows improvement even on the near-isometric benchmark.

4.2 SMAL

The most significant strength of our architecture is the ability to correspond shapes from different domains, an ability we are the first to implement using an unsupervised functional maps framework. We have evaluated our method on pairs of shapes from the SMAL dataset [Zuffi et al., 2017, Zuffi et al., 2018]. In Figure 1 we present a visualization of the inter-class results. The results show how the relaxation of the near-isometry assumption gave the model the degree of freedom to generalize itself for corresponding non-isometric shapes. On the spectral domain of the scale-invariant geometry, our method managed to model inter-class relationships between shapes. We have observed how standard methods fail to converge in terms of the loss function and were giving completely broken matches for inter-class matching.

5 Conclusions

We have realized an end-to-end architecture for non-rigid and non-isometric shape correspondence, based upon scale-invariant geometry. Our architecture surpasses state-of-the-art methods on standard shape correspondence benchmarks, and it is the first to solve end-to-end inter-domain correspondence problems using the functional maps paradigm, where all other existing unsupervised spectral methods fail. This work demonstrates the advantages of using multiple scale-invariant spectral domains for the task of shape correspondence. Without questioning the optimality of the LBO for spanning spectral domains, we show how the Euclidean metric is suboptimal for the purpose of non-rigid shape correspondence. While presenting superb results on deformable pairs, challenging cases such as hippo and a cat can still be improved. We believe this work sets another step toward a fully general non-rigid shape correspondence method. The investigation of alternative metrics and bases should be further examined, such as the affine-invariant metric presented in [Raviv and Kimmel, 2015] and the task-driven decomposition suggested in [Azencot and Lai, 2019].

6 Acknowledgments

This work is partially funded by the Zimin Institute for Engineering Solutions Advancing Better Lives, the Israeli consortium for soft robotics and autonomous driving, the Nicholas and Elizabeth Slezak Super Center for Cardiac Research and Biomedical Engineering at Tel Aviv University and TAU Science Data and AI Center.

References

[Aflalo et al., 2014] Aflalo, Y., Brezis, H., and Kimmel, R. (2014). On the optimality of shape and data representation in the spectral domain. *CoRR*, abs/1409.4349.

[Aflalo et al., 2013] Aflalo, Y., Kimmel, R., and Raviv, D. (2013). Scale invariant geometry for nonrigid shapes. *SIAM Journal on Imaging Sciences*, 6(3):1579–1597.

[Azencot and Lai, 2019] Azencot, O. and Lai, R. (2019). Shape analysis via functional map construction and bases pursuit.

[Bogo et al., 2014] Bogo, F., Romero, J., Loper, M., and Black, M. (2014). Faust: Dataset and evaluation for 3d mesh registration. In *CVPR*.

[Bronstein et al., 2006] Bronstein, A. M., Bronstein, M. M., and Kimmel, R. (2006). Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5):1168–1172.

[Donati et al., 2020] Donati, N., Sharma, A., and Ovsjanikov, M. (2020). Deep geometric functional maps: Robust feature learning for shape correspondence. In *CVPR*.

[Eynard et al., 2016] Eynard, D., Rodola, E., Glashoff, K., and Bronstein, M. M. (2016). Coupled functional maps. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 399–407. IEEE.

[Ginzburg and Raviv, 2020] Ginzburg, D. and Raviv, D. (2020). Cyclic functional mapping: Self-supervised correspondence between non-isometric deformable shapes. In *Computer Vision – ECCV 2020*, pages 36–52, Cham. Springer International Publishing.

[Halimi et al., 2018] Halimi, O., Litany, O., Rodolà, E., Bronstein, A. M., and Kimmel, R. (2018). Self-supervised learning of dense shape correspondence. *CoRR*, abs/1812.02415.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[Litany et al., 2017] Litany, O., Remez, T., Rodolà, E., Bronstein, A. M., and Bronstein, M. M. (2017). Deep functional maps: Structured prediction for dense shape correspondence. *CoRR*, abs/1704.08686.

[Mémoli, 2007] Mémoli, F. (2007). On the use of gromov-hausdorff distances for shape comparison. In *Proceedings Point Based Graphics*, pages 81–90.

[Nogneng and Ovsjanikov, 2017] Nogneng, D. and Ovsjanikov, M. (2017). Informative descriptor preservation via commutativity for shape matching. *Computer Graphics Forum*, 36:259–267.

[Ovsjanikov et al., 2012] Ovsjanikov, M., Ben-Chen, M., Solomon, J., Butscher, A., and Guibas, L. (2012). Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics*, 31(4):1–11.

[Pinkall and Polthier, 1993] Pinkall, U. and Polthier, K. (1993). Computing discrete minimal surfaces and their conjugates. *Experimental Mathematics*, 2(1):15–36.

[Raviv and Kimmel, 2015] Raviv, D. and Kimmel, R. (2015). Affine invariant geometry for non-rigid shapes. *International Journal of Computer Vision*, 111(1):1–11.

[Reuter et al., 2009] Reuter, M., Biasotti, S., Giorgi, D., Patanè, G., and Spagnuolo, M. (2009). Discrete laplace–beltrami operators for shape analysis and segmentation. *Computers & Graphics*, 33(3):381 – 390. IEEE International Conference on Shape Modelling and Applications 2009.

[Rodolà et al., 2015] Rodolà, E., Moeller, M., and Cremers, D. (2015). Point-wise map recovery and refinement from functional correspondence. *arXiv preprint arXiv:1506.05603*.

[Roufosse and Ovsjanikov, 2018] Roufosse, J. and Ovsjanikov, M. (2018). Unsupervised deep learning for structured shape matching. *CoRR*, abs/1812.03794.

[Rusinkiewicz, 2004] Rusinkiewicz, S. (2004). Estimating curvatures and their derivatives on triangle meshes. In *2nd International Symposium on 3D Data Processing, Visualization and Transmission*, pages 486–493. IEEE Computer Society.

[Rustamov et al., 2013] Rustamov, R. M., Ovsjanikov, M., Azencot, O., Ben-Chen, M., Chazal, F., and Guibas, L. (2013). Map-based exploration of intrinsic shape differences and variability. *ACM Transactions on Graphics*, 32(4):1–12.

[Salti et al., 2014] Salti, S., Tombari, F., and Di Stefano, L. (2014). Shot: Unique signatures of histograms for surface and texture description. *Computer Vision and Image Understanding*, 125:251 – 264.

[Singh and Manhas, 1993] Singh, R. and Manhas, J. (1993). *Composition Operators on Function Spaces*. ISSN. Elsevier Science.

[Vollmer et al., 1999] Vollmer, J., Mencl, R., and Mueller, H. (1999). Improved laplacian smoothing of noisy surface meshes. In *Computer graphics forum*, volume 18, pages 131–138. Wiley Online Library.

[Zuffi et al., 2018] Zuffi, S., Kanazawa, A., and Black, M. J. (2018). Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *CVPR*, pages 3955–3963.

[Zuffi et al., 2017] Zuffi, S., Kanazawa, A., Jacobs, D. W., and Black, M. J. (2017). 3d menagerie: Modeling the 3d shape and pose of animals. *CVPR*.

An NLP approach to Image Analysis

Guillermo Martínez

Universidad de Valencia, Spain

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/KFBI5107

Abstract

In Natural Language Processing, measuring word frequency combined with word distribution can yield a precise indicator of lexical relevance, a measure of great value in the context of Information Retrieval. Such detection of keywords exploits the structural properties of text as revealed notably by Zipf's Law which describes frequency distribution as a 'long tailed' phenomenon. Can such properties be found in images? If so, can they serve to distinguish high content items (particular colours coded as RGBs) from low information items? To explore this possibility, we have applied NLP algorithms to a corpus of satellite images in order to extract a number of linguistic-type features in bitmaps so as to augment the original corpus with distributional information regarding its RGBs and observe if this addition improves accuracy throughout a Machine Learning pipeline tested with several Transfer Learning models.

Keywords: Natural Language Processing, Computer Vision, Zipfian Distribution, Average Reduced Frequency

1 Introduction

Natural Language Processing (NLP) and Computer Vision (CV) have been among the most actively developing research areas in Machine Learning. Yet, until recently they have been treated as separate areas without many ways to interact and benefit from each other. With the expansion of multimedia, researchers have started exploring the possibilities of applying both approaches to achieve better results. A combinational approach in the form of multimodal methodologies is favoured to analyse mixed types of data. For example, in healthcare where medical x-ray images are accompanied with their radiology reports, natural language information brings great potential for image analysis ([Shin, Lu, & Summers, 2017] or [Donnelly, Grzeszczuk & Guimaraes, 2022]).

Such translation between low-level pixels and high-level verbal description is known as "bridging the semantic gap" ([Zhao & Grosky, 2001]) where the text-image relation is considered as the transposition from textuality to visuality. Conversely the text can be seen as an illustration of an image, clarifying its content, adding information or developing an idea.

In this paper we present a different form of integration of NLP and CV techniques whereby images are read - literally - as text and RGB values are tallied as words. We will show how the application of algorithms designed for - and usually associated with - textual data can improve the processing of visual data in the context of Machine Learning. To illustrate the value of this approach we will test our 'textified' dataset of images (The UC-Merced Land Use Archive[1]) with 7 renowned models from Transfer Learning.

[1] The UC-Merced archive contains images that were manually extracted from large images from the United States Geological Survey National Map Urban Area Imagery collection for various urban areas around the country. The image data set consists of 2100 scene images pertaining to 21 categories such as *agricultural*, *airplane*, *baseball diamond*, *beach*, *buildings*, etc. ([Yang & Newsam, 2010]).

 ISBN 978-0-9934207-7-1

2 State of the Art

Essentially, the motivation behind any combination of NLP and CV techniques has been so far to establish relations between pictorial representation and verbal language with a view to improving our (automated) understanding of a dataset of images associated with a textual commentary. Whilst the objective of multimodal approaches has been the analysis of mixed data in the form of picture-text relations, the methods used have not been, strictly speaking, combinatorial. Rather, the approach typically consists in analysing, in parallel, the textual data via NLP and the pictorial data via CV. Of course, the distinct algorithms underlying these two domains define them as separate branches of Machine Learning. For example, a supervised model for image recognition would typically use pooling layers to reduce spatial size whilst a neural network for language modelling and grammar induction would seek to exhaustively identify all recurrent properties in text.

It is therefore fair to say that both the goal(s) and the method(s) have not favoured a genuine combination of NLP and CV algorithms for image analysis. The few NLP methods that have been used so far to analyse images are not specifically linguistic methods nor have they been exploited to the full extent with a view to image interpretation. For instance, the study of the Zipfian distribution of RGB colours in images has served as a reliable basis for image compression and to a certain degree for image segmentation ([Caron, 2004]). Also, co-occurrence analysis in the form of gray-level co-occurrence matrices (GLCMs) encodes the distribution of neighbouring pixel values at a given offset and serves typically to measure image texture ([Haralick et al., 1975]). Similarly, a specifically linguistic approach such as ngram analysis has been successfully used for image classification ([Kulkami et al., 2016]).

Hereafter we will demonstrate the adaptability of NLP methods to Machine Learning for deeper and more profitable results in the context of image analysis.

3 The Textuality of Images

In linguistics and literary studies, textuality is the property by which successive words and successive sentences form a coherent text as opposed to a random sequence. The term 'texture' covers the various devices used in establishing continuity of sense and thus making a sequence of sentences operational, i.e. both cohesive and coherent ([Halliday M. A. K. & Ruqaiya H., 1980]). This description could easily apply to images provided we overcome the inexistence of natural lexical and contextual units in images. Letters, syllables, words, phrases, sentences, paragraphs are not defined therefore we will approach pixel context as an artificial construct by considering horizontal and vertical lines in images as natural contexts for the RGB under study[2] (as opposed to the usual 4-connected and 8-connected contexts).

As the scientific method of language analysis, linguistics requires empirical evidence in the form of data drawn from language corpora in support of any statement made about language. The basis of this evidence are word frequencies which supply reliable measures for all content analysis procedures. Indeed, frequency and more importantly frequency effect pervade all language representation. As was established by H. P. Luhn (3.1), the Zipfian distribution of word frequency (3.2) enables a simple categorization of

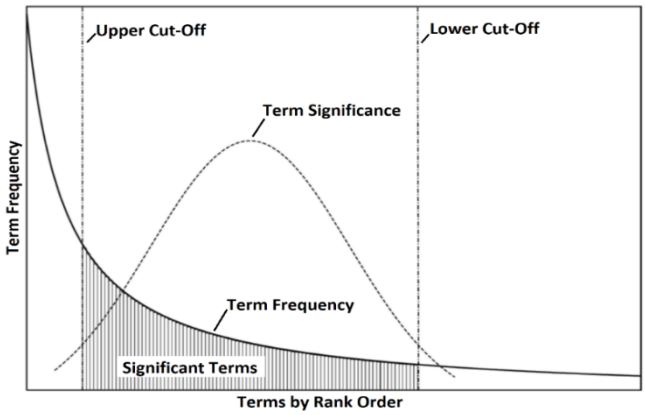


Figure 1: Relevance Ranking by H. P. Luhn

[2] The etymology of the word *text* (past participle stem of the Latin verb *texere*, to weave) is apparent in expressions that refer to the 'weaving' of a story, the 'thread' of an argument, or the 'texture' of a piece of writing.

 ISBN 978-0-9934207-7-1

significant items versus non-informative items. Can such a relevance ranking function be adapted to images?

3.1 Luhn's Thesis

In the 1950s, IBM researcher and Information Science pioneer, Hans Peter Luhn suggested that for the purpose of document automatic abstracting both extremely common and extremely uncommon words were of little if any use ([Luhn, 1958]). His diagnosis is based simply on term frequency as a significance factor and is applied in term-weighting reliant schemes such as the vector space model where weights assigned to the document terms are of crucial importance to the accuracy of the retrieval system when searching for "significant words". These global weighting schemes are shown to adhere to Luhn's resolving power as middle frequency terms are assigned the highest weight (Figure 1).

3.2 Zipfian Distributions

The best-known statistical regularity of language is Zipf's law, originally proposed by American linguist George Kingsley Zipf (1902–50) for the frequency of usage of different words in the English language. This is a mathematical rule that states that a word's frequency is inversely proportional to its rank in frequency. This implies that the word with the highest frequency (rank 1) appears twice as often as the next most common word (rank 2), thrice as often as the word that is ranked in position 3, etc.

Defining rank *r* as the numerical position of a word in a list sorted by decreasing frequency *f*, Zipf's formula is:

$$f \cdot r = k\ (constant)$$

Indeed, Zipf's Law reveals what is known as a power law describing frequency distribution as a 'long tailed' distribution because most of the probability mass is in the tail compared to an exponential distribution. From a statistical perspective the power scaling of Zipf's law ensures that the majority of words occur very infrequently and generate a severe sparse data problem. As a logical consequence, words with the lowest frequency - 1 occurrence - are far more common than we may realize[3]. Therefore, words appearing only once, *hapax legomena* (from Ancient Greek "something said only once") are quite frequent and large events are rare, thus capturing the scaling properties of a universal linguistic phenomenon. Figure 2 shows the Zipf plot for vocabulary in the Brown corpus[4], the standard corpus for linguistic analysis of American English ([Francis et al., 1967]).

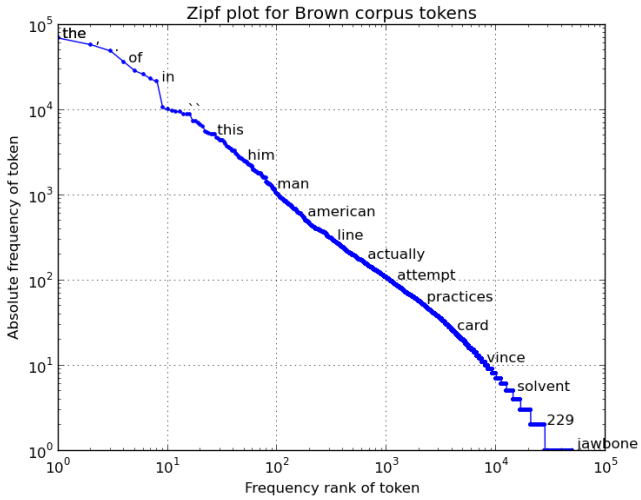


Figure 2: Zipf-like distribution for the Brown corpus

Although the results of the rank by frequency product differ from the idealized Zipfian distribution, this product shows some constancy. After the 10 most frequent words (*the, be, of, and, a, to, in, he, have* and *it*) which constitute over one quarter of all tokens in the corpus, frequency drops radically. At its other end, the curve flatlines

[3] The 'Brown University Standard Corpus of Present-Day American English' is an electronic collection of text samples of American English of varied genres (1006770 word-tokens, a vocabulary of 45215 word-types, in 500 samples of English compiled from works covering 15 text categories and published in the United States in 1961). While the first 135 most frequent words in the Brown corpus cover almost half of the amount of the total text, its 19130 hapax legomena represent 42% of the corpus' vocabulary.

[4] Plot taken from Finn Årup Nielsen's blog: Zipf plot for word counts in Brown corpus, posted on October 22, 2013 https://finnaarupnielsen.wordpress.com/2013/10/22/zipf-plot-for-word-counts-in-brown-corpus/

signaling the unique occurrences of many hapax legomena. However, Hapax, while only rarely encountered individually, are very common in the aggregate and on average represent about half of the words in a corpus. Therefore, it is common to disregard hapax legomena as they have little value for computational techniques.

While qualitative studies in literature may focus on hapax because of their rarity. At the other end of the frequency spectrum, the commonest words tend to be connectives, pronouns and articles, words whose meaning and use are much more prescribed and conventional than the case of rare words. These words are very frequent but relatively contentless. This is the statistical basis of Luhn's approach.

Zipfian distributions have been found in countless, seemingly unrelated areas among which the statistics of firm sizes, city sizes, the genome, family names, income, financial markets, Internet file sizes, or human behaviour ([Mitzenmacher, 2003]). Zipf's law has been found to apply to collections of written documents in virtually all languages, natural, artificial (Esperanto) or randomly generated ([Li, 2002]). In image analysis ([Crosier & Griffin, 2007]), Zipfian distributions have not been exploited with a view to image interpretation.

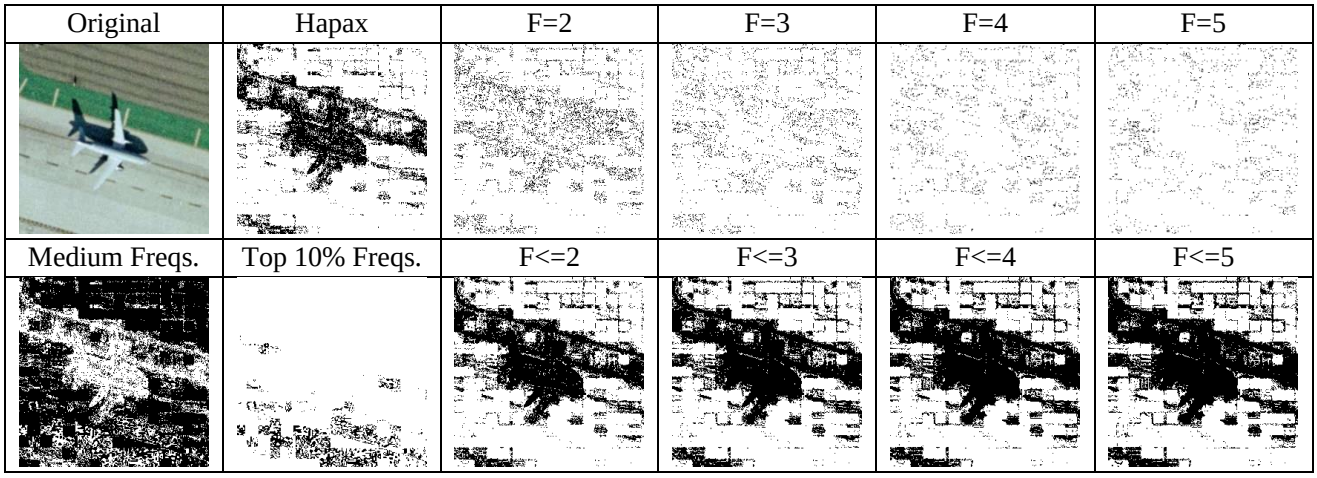


Figure 3: Filtering colours by frequency

Let us now consider the bitmap in Figure 3 and understand how its frequency distribution does not match Luhn's Relevance Ranking[5]. The original image in Figure 3 is a 256 x 256 bitmap (65536 pixels) with 21997 distinct RGB colors. There are a total of 16122 colors occurring only once (24.6% of hapax in 65536 RGBs). At the other end of the frequency spectrum we find the 10 most frequent RGBs totaling 5316 pixels (8.11% of image surface). In between these extremes we have a total of 5865 distinct colours covering 44098 pixels equaling 67.8% of the bitmap. The frequency-filtered interpretations in Figure 3 show that:

- high frequency colours do not carry any useful information regarding the original image
- medium frequencies do not convey significant information about the image
- hapax legomena draw a mass albeit blurry around the region of interest

The cumulative interpretations (F<= includes frequencies lower or equal to F) show that although frequency distribution follows Zipf's Law (the more frequent a colour, the less colours of this same frequency there are). Therefore as individual frequency increases, global frequency for this rank decreases: F2=5730, F3=3090, F4=2004, F5=1550, F6=1254, F7=1071, F8=760, F9=711, F10=570.

In phase with Zipf's Law but in contradiction with Luhn's ranking, the numerous Hapax legomena in bitmaps emerge as very informative items in the aggregate. Therefore, we shall extract hapax mappings for each of the 2100 bitmaps in the Merced corpus and add this information as a fourth channel after RGB in the images.

[5] The 16x16 paving observed in this bitmap corresponds to an artefact as this bitmap is a composite made of more detailed satellite photographs.

 ISBN 978-0-9934207-7-1

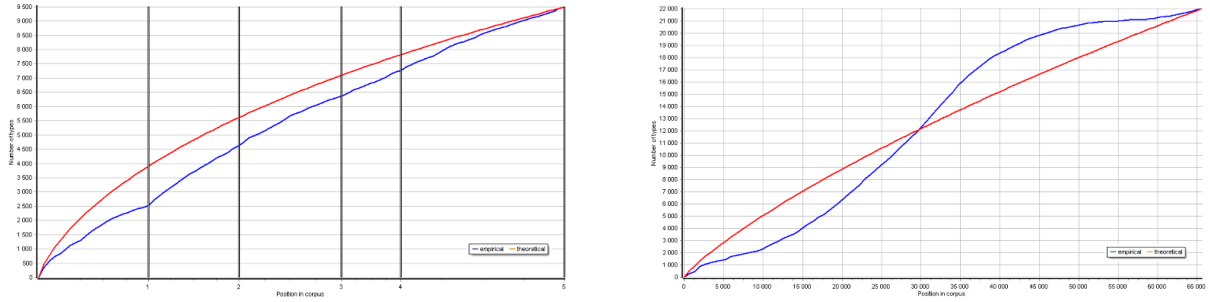


Figure 4: Vocabulary Growth in Text (left) and Image (right)

A possible explanation for the 'burstiness' of low frequency RGBs and the information they carry can be found in measuring vocabulary growth. Consider Figure 4. The first graph (left hand) represents the growth of vocabulary in James Joyce's novel, 'A Portrait of the Artist as a Young Man' (80000 tokens comparable to a bitmap's 65536). The red line represents theoretical growth of vocabulary as word tokens (axis x) are read in context. In the beginning, practically every word read in text order is a new and unknown type but quickly enough novelty (blue line) decreases despite more tokens being encountered. The slow drops are followed by spurting renewals where each vertical line corresponds to a new chapter in the novel and its corresponding semantic field. This pattern can be observed for any type of text, literature, newspaper articles, science textbooks, etc. However, the graph on the right describes a completely distinct phenomenon impossible in text because of the semantic and syntactic constraints of linguistic economy. In an image, vocabulary growth may occur explosively and exceed theoretical growth.

4 Average Reduced Frequency

Typically, a frequency list tallies items in a corpus following their occurrence ignoring all aspects of item distribution throughout the text. With Average Reduced Frequency (ARF), item dispersion is measured and integrated to item frequency in order to properly reflect the circulation of items in context. By attenuating incidental unevenness of word distribution, ARF prevents the result to be excessively influenced by one part of the corpus which contains a high concentration of the token ([Hlaváčová & Savický, 2002]).

In the context of NLP consider a text corpus split into *k* non-overlapping parts of the same length, where *k* is the corpus frequency of the word being analyzed. A reduced word frequency is the number of parts containing the word. There are as many reduced frequencies as the length of a part, since the start of the first part can be moved over the corpus until it reaches a split previously seen. The ARF is the average of these reduced frequencies.

To make the measure objective, ARF is calculated as the arithmetic mean over all possible beginnings of the first segment. To achieve this, the corpus is not read as a line segment but as a circle: after the last corpus position comes the first one. This way, the beginning of the first segment is moved all along the whole circle and reduced frequencies are estimated for each new position.

ARF is calculated according to the following formula:

$$ARF = \frac{1}{v}\sum_{i=1}^{f} min\{d_i, v\}$$

where $v=N/f$ (total number of tokens in corpus / frequency of current type) and di designate the distance between consecutive occurrences of the word in the corpus. Particularly, if $n1, n2, \ldots, nf$ are numbers of positions, where the word occurs, then $di=ni-ni-1$ for every $i= 2, \ldots, f$ and $d1=n1+(N-nf)$, which is the distance between the last and the first occurrence of the word in the cyclic order of the corpus.

To illustrate ARF, let us have a text corpus made of 60 tokens where a given word occurs 5 times (at positions 0,

 ISBN 978-0-9934207-7-1

11, 13, 16, 56). In Figure 5[6], the '+' symbol represents this specific word type which clearly has an unregular distribution in text with most of its occurrences at the beginning of the sequence. Now, let us split the corpus into 5 equal parts of length 12 (i.e., 60 divided by 5). On counting the distinct number of parts containing the word we observe the sequence *present|present|absent|absent|present* and obtain the reduced frequency of the word - in this case 3 (vertical bar indicates context splits).

In theory, there are 12 possible ways to split the corpus into 5 parts of equal length. Also, each partition will yield its own *Reduced Frequency*.

```
|+----------+|-+--+-------|------------|------------|--------+--- RF = 3
+|----------+-|+--+--------|------------|------------|--------+--- RF = 3
+-|---------+-+|--+---------|------------|------------|-------+--- RF = 3
+--|--------+-+-|-+----------|------------|------------|------+--- RF = 3
+---|-------+-+--|+-----------|------------|------------|-----+--- RF = 3
+----|------+-+--+|-----------|------------|------------|----+--- RF = 2
+-----|-----+-+--+-|-----------|------------|------------|---+--- RF = 2
+------|----+-+--+--|-----------|------------|------------|--+--- RF = 2
+-------|---+-+--+---|-----------|------------|------------|-+--- RF = 2
+--------|--+-+--+----|-----------|------------|------------+|--- RF = 3
+---------|-+-+--+-----|-----------|------------|-----------+-|-- RF = 3
+----------|+-+--+------|-----------|------------|----------+--|- RF = 3
```

Figure 5: Calculating Average Reduced Frequency

From these 12 results, we can calculate the *Average Reduced Frequency* as (3 + 3 + 3 + 3 + 3 + 2 + 2 + 2 + 2 + 3 + 3 + 3) / 12 = 2.67. This value discounts close occurrences of the word and estimates the word frequency would be only 2.67 in a homogeneous corpus. To adapt ARF to the non-linear nature of images, we averaged the results of horizontal and vertical calculations.

Again, the results obtained in image analysis differ greatly from those observed in textual data. A frequency dictionary of word-types may only be slightly altered by an ARF ranking as a moderate proportion of words will see their frequency moderated by their lack of dispersion in context. Therefore, their corrected rank is systematically lower than their original rank. With images, because of the bursty nature of colors, far less constrained than words, many RGBs may be demoted in rank because of their bad distribution just as many will be promoted. Indeed, many low frequency and therefore low-ranking RGBs may have such a balanced distribution throughout an image that their frequency is greatly increased by ARF. These frequency swapping phenomena can be interpreted as features in new channels where RGB promotion and demotion are coded in red and green respectively (Figure 6).

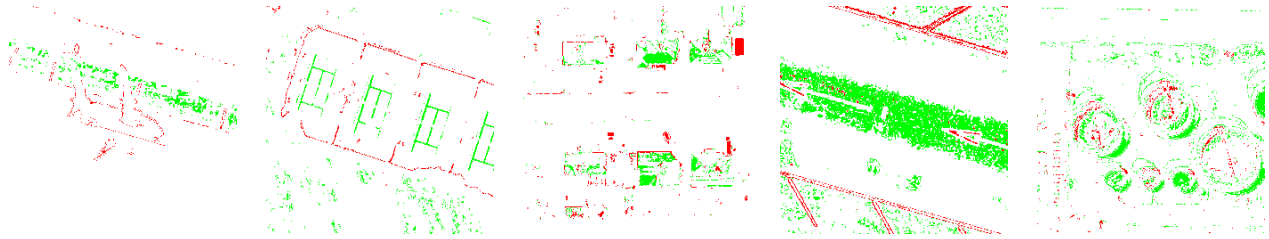


Figure 6: Features for Promoted and Demoted RGBs according to Average Reduced Frequency

[6] This example was taken from the documentation of Sketch Engine, an NLP software designed for text mining. The document's URL is: https://www.sketchengine.eu/documentation/average-reduced-frequency/

 ISBN 978-0-9934207-7-1

5 Deep Learning with NLP Features

Hereafter we present the results of Deep Learning with 7 distinct models available in the Keras framework as Transfer Learning models. The initial results[7] are presented in table 1 whilst the results after integration of NLP features as extra channels in our dataset appear in table 2. In between both processes we tested the NLP features on their own without the original data. The results were insufficient with major phenomena of overfitting in the order of 100% in training versus 30 to 40% in validation. This can be explained by the absence of optimization for these initial tests (no data augmentation, no dropout, no regularization) but we believe it also has to do with the over-simplified results that we obtained with these features. Indeed, these are blurry shapes or outlines pointing to the object of interest in each image but in terms of information or entropy the features by themselves are impoverished versions of the original bitmaps. This is why they serve much better as complementary channels for the original bitmaps as shown in table 2.

Model	Params.	Training acc.	Validation acc.	Training loss	Valid. loss
DenseNet201	20M	100.00%	92.38%	0.0002	0.27
ResNet101	45M	100.00%	89.76%	0.00003	0.38
Xception	23M	100.00%	87.37%	0.0181	0.42
MobileNetV3Large	4M	100.00%	85.48%	0.0002	0.61
InceptionV3	23M	100.00%	85.48%	0.0077	0.48
EfficientNetV2L	119M	99.94%	78.81%	0.0169	0.66
VGG16	15M	100.00%	67.14%	0.0004	2.15

Table 1: Performance measures for original dataset (ranked by validation accuracy)

Model	Training acc.	Validation acc.	Training loss	Valid. loss
DenseNet201	100.00%	93.01%	0.0002	0.46
ResNet101	100.00%	90.18%	0.0003	0.15
Xception	100.00%	88.11%	0.0091	0.33
MobileNetV3Large	100.00%	83.31%	0.0001	0.41
InceptionV3	100.00%	82.01%	0.0034	0.31
EfficientNetV2L	100.00%	71.54%	0.0085	0.43
VGG16	100.00%	93.01%	0.0001	1.13

Table 2: Performance measures for feature-augmented dataset (ranked by validation accuracy)

6 Conclusion and Future Work

NLP methods are considerably more diverse as compared to Computer Vision because of the structured complexity of the object of analysis. With syntax, morphology, compositionality, and semantics as areas of investigation, and letters, syllables, words, phrases, sentences and discourses, as items to be studied, Computational Linguistics has developed an array of algorithms for text analysis at various levels of exploration and understanding. In addition to highlighting similarities in their objects of analysis, this paper shows that a bridge exists between Natural Language Processing and Computer Vision. We have demonstrated that the integration of NLP methods to image analysis leads to significant improvement in many tasks. Hapax detection yields a feature that guarantees good discrimination from more common colours which occur frequently with all sorts of words and aren't

[7] All experiments were carried out with the same hyperparameters: batch size=16, epochs=100, no data shuffle, no data augmentation, loss=categorical cross-entropy. The optimizer is: SGD, learning rate=0.0001, momentum=0, decay=0.0, nesterov=false.

 ISBN 978-0-9934207-7-1

informative. Average Reduced Frequency underlines both important areas in images and superficial areas, both being characteristic of a given class or label. Future work aims to go beyond the basic tools of text quantification and probabilistic linguistics to explore the core of *textus* mechanics that is distributional and its transposition to image analysis in the form of cooccurrence analysis of pixels.

References

[Caron, 2004] Caron Y. (2004), *Contribution de la loi de Zipf à l'analyse d'images*, PhD Computer Science, Université de Tours, Dir. Nicole Vincent.

[Donnelly et al., 2022] Donnelly L.F., Grzeszczuk R., Guimaraes C.V. (2022). *Use of Natural Language Processing (NLP) in Evaluation of Radiology Reports: An Update on Applications and Technology Advances*. Semin Ultrasound CT MR. 2022 Apr;43(2):176-181.

[Francis et al., 1967] Francis, W. Nelson & Henry Kucera. 1967. *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

[Crosier & Griffin, 2007] Crosier M. & Griffin L. (2007). *Zipf's Law in Image Coding Schemes*. BMVC 2007 - Proceedings of the British Machine Vision Conference 2007.

[Halliday M. A. K. & Ruqaiya H., 1980] Halliday M. A. K. & Ruqaiya H., 1980, *Cohesion in English*, Style, Vol. 14, No. 1 (Winter 1980), pp. 47-50, Penn State University Press

[Haralick et al., 1975] Haralick R. M., Shanmugam K., Dinstein I., (1973), *Textural Features for Image Classification*, IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC 3, No. 6, pp 610-621.

[Hlaváčová & Savický, 2002] Hlavácová J. & Savický P. (2002) *Measures of Word Commonness*, Journal of Quantitative Linguistics, 9:3, 215-231, DOI: 10.1076/jqul.9.3.215.14124

[Kulkami et al., 2016] Kulkami P., Stranieri A. & Ugon J. (2016). *Texture Image Classification using Pixel N-grams*, IEEE International Conference in Signal and Image Processing, vol. 1, IEEE Press, 2016.

[Li, 2002] Li W. (2002). *Zipf's Law Everywhere*. Glottometrics. No. 5. International Quantitative Linguistics Association. Ram Verlag.

[Luhn, 1958] Luhn, H.P. (1958) *A Business Intelligence System*. IBM Journal of Research and Development, 2, 314-319. http://dx.doi.org/10.1147/rd.24.0314

[Mitzenmacher, 2003] Mitzenmacher M., (2003), *A Brief History of Generative Models for Power Law and Lognormal Distributions*, Internet Mathematics Vol. 1, No. 2: 226-251

[Salem, 2003] Salem A., (2003), *Lexico-User Manual*, https://lexi-co.com/ressources/L3-usermanual.pdf

[Shin et al., 2017] Shin, Hoo-Chang & Lu, Le & Summers, Ronald. (2017). *Natural Language Processing for Large-Scale Medical Image Analysis Using Deep Learning*.

[Shuiyuan et al. 2018] Yu, Shuiyuan, Chunshan Xu and Haitao Liu. (2018). *Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation*.

[Yang & Newsam, 2010] Yang Y. & Newsam S. (2010). *Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification*. ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS).

[Zhao & Grosky, 2001] Zhao, Rong & Grosky, William. (2001). *Bridging the Semantic Gap in Image Retrieval*.

[Zipf 1965] Zipf G. K. (1965). *The Psycho-Biology of Language. An Introduction to Dynamic Philology*. The MIT Press. Cambridge, MA, USA. 1965.

Classification of electromagnetic interference induced image noise in an analog video link

Anthony Purcell [1,2], Ciarán Eising [2]

[1] *Analog Devices, Inc*
[2] *Dept. Of Electronic and Computer Engineering, University of Limerick*

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/ZANB2627

Abstract

With the ever-increasing electrification of the vehicle showing no sign of retreating, electronic systems deployed in automotive applications are subject to more stringent Electromagnetic Immunity compliance constraints than ever before, to ensure the proximity of nearby electronic systems will not affect their operation. The EMI compliance testing of an analog camera link requires video quality to be monitored and assessed to validate such compliance, which up to now, has been a manual task. Due to the nature of human interpretation, this is open to inconsistency. Here, we propose a solution using deep learning models that analyse, and grade video content derived from an EMI compliance test. These models are trained using a dataset built entirely from real test image data to ensure the accuracy of the resultant model(s) is maximised. Starting with the standard AlexNet, we propose four models to classify the EMI noise level.

Keywords: Machine Learning, Image Noise, EMI, SNR, Regularisation

1 Introduction

The problem that this research aims to address is one in the domain of automotive electronics, more specifically, the domain of automotive video electronics. The addition of robust video interfaces to vehicles is something that has grown considerably over the past decade, but the idea has been around for much longer than that. Ever since the first reversing camera was fitted to the Buick Centurion back in the 1950's, the addition of video and camera interfaces to vehicles has progressed considerably and become much more commonplace since then. Most vehicles being sold today are being presented with reversing camera options, with many manufacturers including them as standard. In 2018, all new cars sold in the US were mandated to have a reversing camera fitted as standard [US, Department of Transportation 2014], and in the EU, reversing cameras along with a host of other safety applications for cameras will become mandatory in 2022 [European Commission 2019]. The number of cameras that have been deployed in new vehicles has grown year on year over the past five years and this growth is predicted to continue [Analog Devices 2018]. These interfaces will enable both safety critical and non-safety critical applications. For example, cameras are set to replace external and internal mirrors, perform critical Advanced Driver-Assistance Systems (ADAS) functionality, such as collision avoidance and driver status monitoring, as well as enabling full surround-view monitoring of the vehicle, blind spot detection and night vision, to name but a few novel applications. This increase in both the number of cameras in the vehicle, coupled with the more increasing safety critical nature of the applications that they service, means that the task of system level design, validation, and Electromagnetic Compatibility (EMC) compliance verification that needs to be addressed as part of the vehicles design is more important than ever. This research aims to address the topic of EMC compliance, specifically, the challenges often encountered during the Electromagnetic Immunity (EMI) validation process of a video link destined for an automotive application, and how the application of machine learning can enable more consistent testing methodologies, yielding more meaningful and comparable test results between different systems and system configurations.

2 State of the Art

This section will frame the problem being overcome by this research. Image noise is something that can exist in many forms, so the type of noise being considered will be formally defined. As well as this, a widely used method for determining an images noise level will be described, and how ML techniques can overcome its shortcomings.

In a camera system, image noise can come from many sources. For example, image sensors are known to introduce Gaussian noise to a system, and image processing operations can introduce quantisation noise [Kleinmann and Wueller 2007]. The noise being considered here is noise that is delivered via an external source or interferer. That interferer is introduced in the form of a Bulk Current Injection (BCI) EMI test [Mahesh and Subbarao 2008]. The BCI test method is performed by using a current probe acting as a transformer to inject a current of a specified magnitude and frequency onto the cable harness at defined positions relative to the Device Under Test (DUT). It tests immunity performance in the frequency range of 1 MHz to 400 MHz [ISO 2011]. This noise is periodic in nature, is well defined and introduced in a controlled manner during an immunity test. It is noise of specifically this nature that will be the focus of discussion here. Figure 1 gives an example of noise injection on a test pattern image.

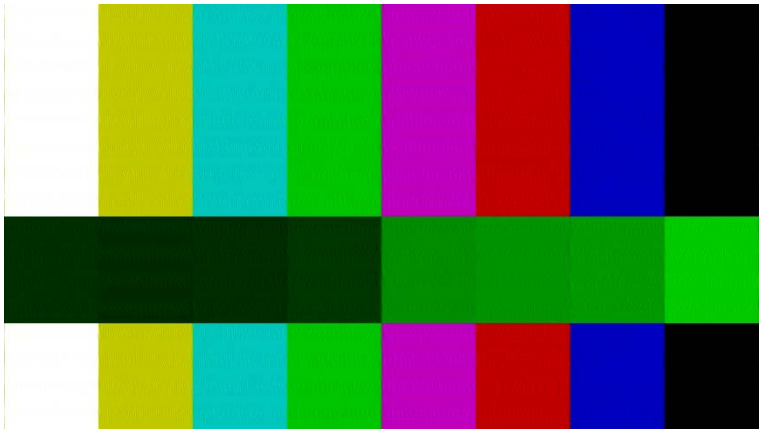


Figure 1: BCI Noise Injection on a Colour Bar Test Pattern

The classical approach to classifying noise in an image, given a reference image, is by using the peak signal-to-noise ratio (PSNR). For two Images I_1 and I_2 that have a 2-dimentional size of i and j and are composed of c channels, the PSNR in dB can be defined as:

$$PSNR = 10 \cdot log_{10}\left(\frac{{MAX_I}^2}{\frac{1}{c * i * j}\sum(I_1 - I_2)^2}\right) dB$$

Where MAX_I is largest valid pixel value in the Images I_1 and I_2, usually 255 for RGB images [OpenCV 2019]. If the images are identical, the divisor will be 0, and any difference in the images will yield a finite value for the image noise present. There are some short comings of the SNR method outlined above for the application that is the subject of this paper.

One advantage of using ML techniques to address the need here is that it will allow specific, visual noise to be targeted as the criteria for classifying the image quality. SNR quantification will highlight any variance between the pixel values of two images. Negligible changes in pixel values will lead to higher PSNR for, even though these differences may not be visible to the human eye. In effect, it has the potential to flag image variance that is not of interest to an inherently qualitative test, where visible noise is the main criteria to judge a pass/failure. Furthermore, the focus of this system will be on an analog video link. Analog video links are inherently not bit-accurate links, where from frame to frame, identical image content under perfect operating conditions is likely to show some least significant bit (LSB) variance. This variance would be considered as noise by SNR analysis, even if it is only ever likely to be a very low value. Thus, we can say, it is desirable to have a machine learning algorithm that models

subjective noise analysis. Others have investigated ways through which this gap between measurement and physical interpretation can be bridged. Kleinmann and Wueller have evaluated methods that focus on the difference between quantifying image noise based on the perception of the human eye as opposed to purely a measurement focussed approach. They investigated two algorithms that aim to emulate the process of the human visual system more closely than the SNR measurement approach:

1. A model for visual noise measurement where the process of human vision is simulated using opponent colour space and contrast sensitivity functions to come up with a visual noise value [Hung, Enomoto and Kozo 1996].
2. The S-CIELab model, which simulates human vision in approximately the same way as the model by Hung, et al, with the addition of an image comparison using the CIEDE2000 colour difference formula, which was designed for predicting the visual perceived difference between colour images [Fairchild and Johnson 2003].

Both models yielded good results, highlighting differences between the simulated visual noise value and the SNR measured value. In saying this, both models showed restrictions during their respective evaluations. The former model is effective at evaluating noise for uniform colour patches, so this places a restriction on the image content to be used in the testing. The performance of the latter model was deemed to be influenced by the kind of noise seen in the image, so is not expected to operate consistently across all image noise types. This would mean that it would need to be thoroughly evaluated to be used and trusted for a specific application [Kleinmann and Wueller 2007].

3 The Dataset

Fundamental to the effectiveness of any machine learning approach to solving a problem is an appropriate dataset. There was not an appropriate pre-existing dataset available to utilise for this work, so the dataset that will be used was created from scratch from video frames captured from a real test environment described in Figure 2.

3.1 Data capture

A common method for classifying image quality in an EMC environment involves capturing and transmitting the video that is sent over the camera link out of the test chamber via a HDMI optical extender. This system is a purpose-built system to ensure EMI does not affect the received video that is being sent out of the chamber for viewing. Outside the chamber, a HDMI monitor can be used to capture the video for live and post analysis. The setup is demonstrated in Figure 2. For a general introduction to BCI test methods, the reader is referred to [Mahesh and Subbarao 2008].

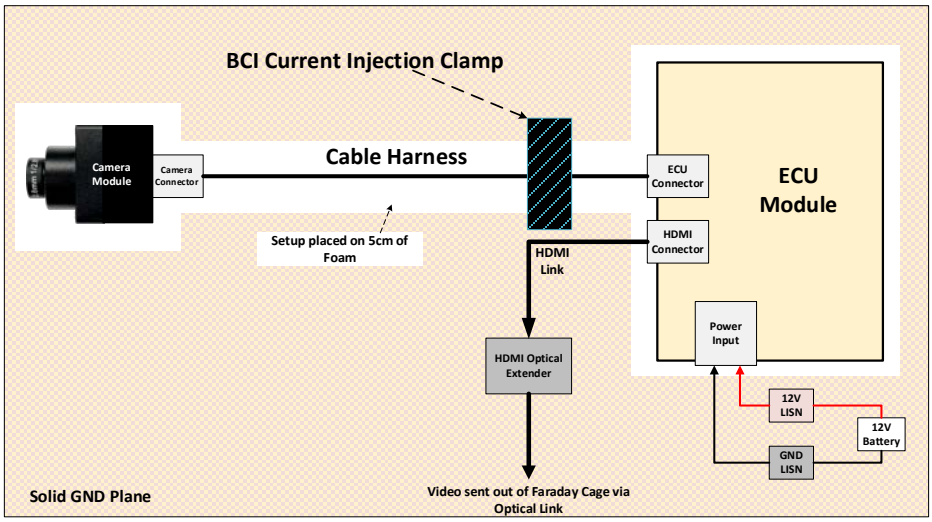


Figure 2: Sample BCI Test Setup for a Single Camera Link

 ISBN 978-0-9934207-7-1

3.2 The Video Content

The content of the video data being transmitted over the link is also an important factor to be considered in the creation of the dataset. Real-world image sensor data captured in an EMI chamber can prove difficult to execute consistent video quality analysis on, because the image is usually quite dark (or at least certainly darker than would be expected during normal operation), which can make the noise and the genuine image content difficult to distinguish. For this reason, video test patterns are used for image noise analysis. This approach comes with three significant benefits:

1. The image content of interest is consistent and does not change from a visual perspective from frame to frame. This makes identifying and analysing induced noise an easier task, and the test a more stringent one, because noise seen here may well be indistinguishable on a small in-vehicle display showing real world content.
2. The artificial nature of the test pattern content allows a wide spectrum of colour content to be examined, meaning a more thorough analysis can be done across a wider brightness and colour spectrum than would be possible with a natural image.
3. Finally, we are interested only in the noise introduced during the analog video transmission. The transmission of the test pattern isolates this source of noise from any other noise source.

A colour bar test pattern is selected as this contains low frequency video content which makes identifying external interference a less strenuous visual task.

3.3 The Dataset Structure

The dataset is formulated into training, validation, and test sets. We define the categories that the dataset will assume when communicating noise interference:

- Level 1: No visual noise detected
- Level 2: Low level of visual noise detected
- Level 3: Medium level of noise detected
- Level 4: High level of noise detected
- Level 5: Video link has ceased to operate, loss of lock event, flat blue field displayed.

These categories are, of course, arbitrary, but due to the subjective nature of the task, this is somewhat unavoidable. Care was taken to ensure consistency across all samples within a category, but this cannot be guaranteed from merely a visual segregation process. The dataset is organised into 800, 200 and 100 image samples within each category of the training, validation, and test sets respectively. Figure 3 shows examples of noise levels 2 to 4 (level 5 is not shown, as this is simply a flat blue field).

4 Video Noise Grading Model

This section will describe the implementation and verification of candidate CNN models to undertake the task of performing a video noise classification task. The creation of a custom dataset as described in section 3 will be utilised to train and validate these CNN’s. For reasons that will be discussed in detail, four candidate models are proposed and their difference in terms of complexity and performance is discussed comparatively. In addition to this, particular focus is given to techniques that result in improved model accuracy and generalisation.

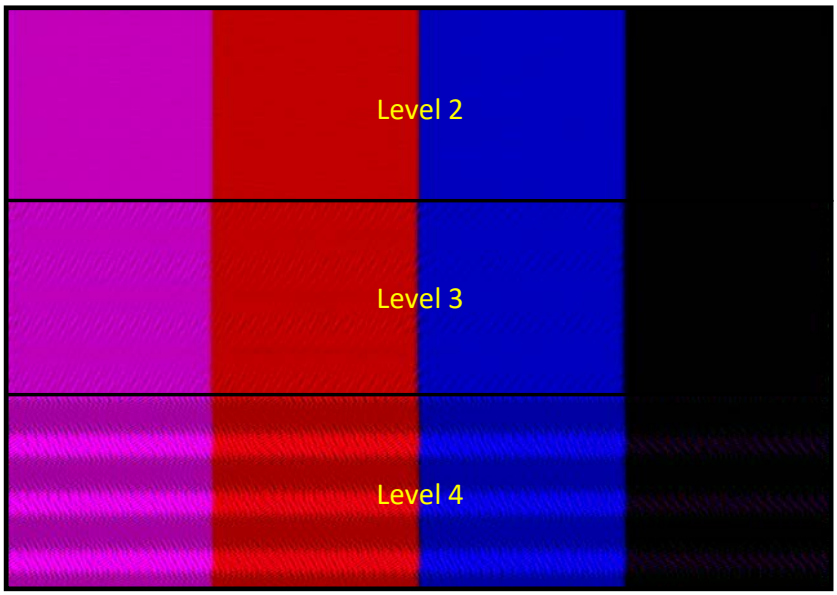


Figure 3: Example of the Interclass Image Quality Variance on a patch of the test pattern.

4.1 Data Pre-processing

The dataset is comprised of frames of 1280x720p video in the YCbCr colour space. These frames are reduced in size using nearest neighbour interpolation before being input to the network. Whilst a network that could take these images in their default state is certainly possible, it was advantageous from a complexity and training time perspective to reduce the size of the images appropriately before applying them to the network. A concern from performing such an operation is that once the images are resized, the features that are required to be detected (the image noise, in this case) should not be lost or degraded significantly in the process, which may prevent a model's ability to detect the feature. The AlexNet CNN architecture was used as a starting point for the model development. This architecture mandates a 227x227 size input, so the images were resized to meet this requirement. Across all models that are proposed, this input size remained unaltered for comparison. Rescaling of the image pixel values was also applied. By default, a pixel can occupy any value from 0 to 255 for 8-bit images. These pixel values are rescaled from [0…255] to [0… 1]. The final alteration that was made to the input dataset before being applied to a model was that the images were converted to greyscale, dropping the chroma channels of the image (Figure 4). This allows for the discarding of two feature maps in the input to the network, thus reducing the complexity of the required network. The encoding scheme on the video link used here ensures the luma and chroma portions of the signal are equally susceptible to EMI interference, thus using just the luma data is sufficient to characterise the noise level.

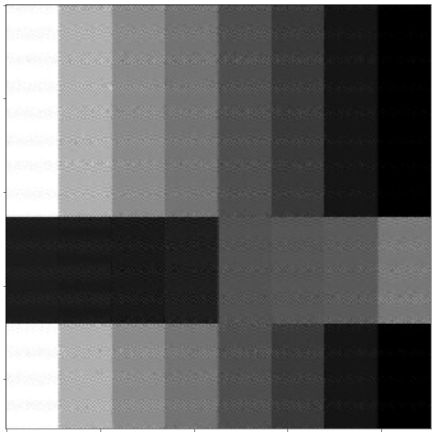


Figure 4: Greyscale Image Showing the Retention of Noise Features

 ISBN 978-0-9934207-7-1

4.2 Model Definition and Structure

Four CNN models will be defined for to carry out the task of classifying image quality. The models presented here start with the well-known AlexNet structure [Krizhevsky, Sutskever and Hinton 2012], and subsequent models with lower complexity are then defined. To start with, the AlexNet model was left as standard, apart from the following details:

1. The input size of the network. Originally, this model was applied to RGB images (227x227x3), but here the input size is reduced to 227x227x1 due to the conversion to greyscale applied to the images.
2. Data augmentation is applied to the input images before the convolutional layers. The images were randomly flipped in both the horizontal and vertical direction, which ensures the noise profile is maintained, but the active image content is differed, encouraging the models to better characterise the noise interference.

4.2.1 Lower Complexity CNN Models

For comparison with the AlexNet implementation, models that represent significant reductions in the number of trainable parameters are proposed to see if this task can be achieved more effectively using a less complex network. These networks are built from the same building blocks that constructed the AlexNet model, and use the same data augmentation step at the beginning. The main differences are in the convolution layers (both in the parameters specified and the number of layers) and in the number and size of the fully connected layers of the networks. A comparison of the three lower complexity models' structure is given in Figure 5. All activations were ReLU apart from the final fully connected layer in each model, which has a softmax activation.

CNN Model 2	
Input	227x227x1
Conv2D	55x55x32
Activation	55x55x32
Max Pool	27x27x32
Conv2D	27x27x96
Activation	27x27x96
Max Pool	13x13x96
Conv2D	13x13x128
Activation	13x13x128
Max Pool	6x6x128
Flatten	4608
FC	512
FC	5
Total Params:	2,504,741

CNN Model 3	
Input	227x227x1
Conv2D	55x55x32
Activation	55x55x32
Max Pool	27x27x32
Conv2D	27x27x64
Activation	27x27x64
Max Pool	13x13x64
Conv2D	13x13x128
Activation	13x13x128
Max Pool	6x6x128
Flatten	4608
FC	100
FC	5
Total Params:	557,661

CNN Model 4	
Input	227x227x1
Conv2D	55x55x16
Activation	55x55x16
Max Pool	27x27x16
Conv2D	27x27x16
Activation	27x27x16
Max Pool	13x13x16
Flatten	2704
FC	50
FC	5
Total Params:	137,777

Figure 5: Lower Complexity CNN Model Structure

5 Results

The training of each model was done with the Adam optimiser with a learning rate of 10^{-3}. All models converged within 30 epochs, and this number of training cycles was kept standard across all models for comparison. The performance of this training process is captured below in Table 1 and Figure 6.

	AlexNet Model			CNN Model 2			CNN Model 3			CNN Model 4		
Category	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Level 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Level 2	1.00	0.91	0.95	1.00	0.69	0.82	1.00	0.76	0.86	1.00	1.00	1.00
Level 3	0.64	1.00	0.78	0.97	1.00	0.99	0.96	1.00	0.98	0.94	1.00	0.97
Level 4	0.83	0.44	0.58	0.78	1.00	0.88	0.80	0.96	0.87	1.00	0.94	0.97
Level 5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Accuracy			**0.87**			**0.94**			**0.94**			**0.99**

Table 1: Performance Metrics on Test Set Data

 ISBN 978-0-9934207-7-1

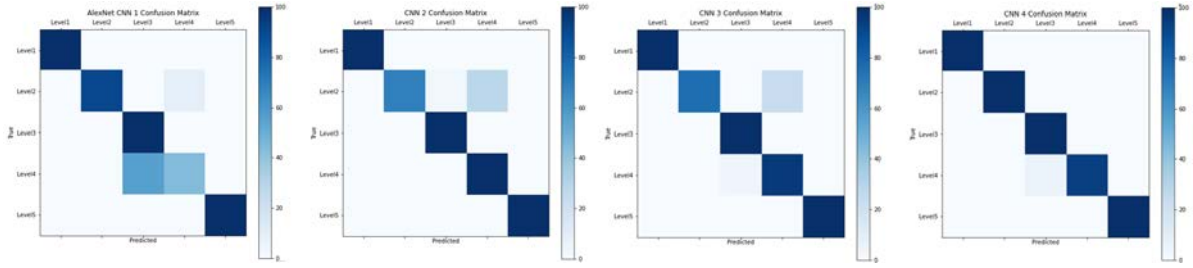


Figure 6: Confusion Matrix on Test Set Data

The metrics and confusion matrix comparisons show the lower complexity models generalise better, giving the best out of sample performance. We can see the issues experienced by model 1 (AlexNet) in classifying level 4 samples, clearly showing the reported recall value of < 50%. Models 2 and 3 are showing issues misclassifying some level 2 samples. These issues are resolved in the lowest complexity model, where near perfect performance is shown. The reduction in model complexity from the AlexNet model was necessary to achieve the best performance, and based on the complete analysis of model performance, model 4 would be the best model to use. This can be attributed to the regularising effect of reducing the number of parameters the model has to fit the data. This, however, does not mean that the AlexNet model cannot be made to work effectively for this task. To explore this further, the next section will discuss using explicit regularisation techniques to improve performance.

5.1 Regularisation

The more complex a model is, the more likely overfitting to the training data is to occur. To test if this is causing the more complex models' poorer out of sample performance, regularisation is applied. L2 regularisation is applied with a λ value of 0.01. Once this addition was made, all 4 models were re-trained. The performance metrics, for comparison, are shown below in Table 2 and Figure 7.

	AlexNet Model - L2 Reg			CNN Model 2 - L2 Reg			CNN Model 3 - L2 Reg			CNN Model 4 - L2 Reg		
Category	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Level 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Level 2	0.95	0.93	0.94	1.00	0.91	0.95	0.93	0.95	0.94	1.00	0.78	0.88
Level 3	1.00	0.94	0.97	0.99	0.99	0.99	0.62	1.00	0.77	0.76	1.00	0.87
Level 4	0.88	0.95	0.91	0.91	0.99	0.95	1.00	0.37	0.54	1.00	0.91	0.95
Level 5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Accuracy			**0.96**			**0.98**			**0.86**			**0.94**

Table 2: Performance Metrics on Test Set Data with L2 Regularisation

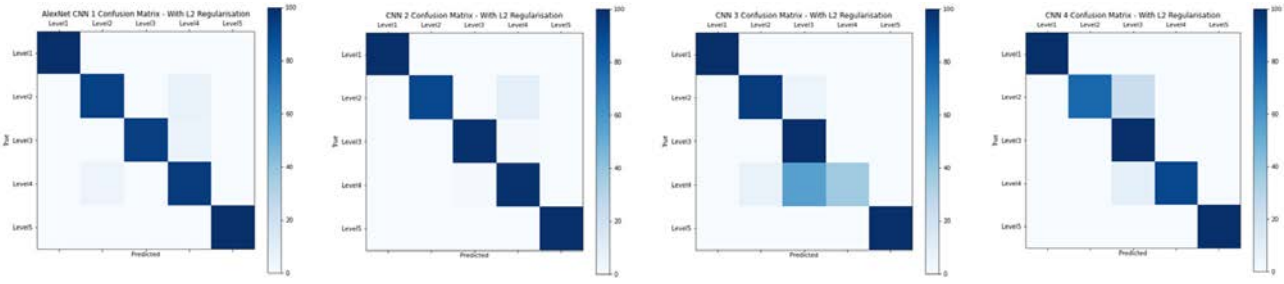


Figure 7: Confusion Matrix on Test Set Data with L2 Regularisation

From the above results, models 1 and 2 benefitted from the addition of L2 regularisation, with model 2 being the best performing of the set. Regularisation reduced overfitting such that out of sample performance is significantly improved, with an increase of 9% in accuracy in the case of the AlexNet model. Models 3 and 4 showed poorer out of sample performance because of the regularisation, the amount of regularisation applied ($\lambda = 0.01$) is clearly causing significant bias. Reducing this λ value would remedy this induced bias in these models.

 ISBN 978-0-9934207-7-1

6 Conclusions

This paper has shown that the task of assigning a class that represents image noise content is not only achievable but can be done quite effectively using ML techniques. Up until now, this has been a manual task for technicians. Furthermore, it is shown that all models that were initially presented as candidates can be made to perform better through some considered techniques to alter and constrain the training process to ensure an appropriate level of overfitting of the training data is realised. These models represent a wide range of complexity levels, the more complex models showing as good candidates for further development if the dataset to hand increases in size or variance. The exercise here has also proven that the task of classifying video quality in this fashion can be achieved and implemented on a wide range of hardware capabilities, from cost effective SoC's to large GPU devices.

References

[Analog Devices 2018] Analog Devices. (2018). "C2B: A New Car Camera Link for Automotive Applications." Accessed May 25, 2022. https://www.analog.com/en/technical-articles/c2b-a-new-car-camera-link-for-automotive-applications.html.

[European Commission 2019] European Commission. (2019). "New safety features in your car." 25 03. Accessed 05 25, 2022. https://ec.europa.eu/docsroom/documents/34588/attachments/1/translations/en/renditions/native.

[Fairchild and Johnson 2003] Fairchild, Mark D, and Garrett M Johnson. (2003). "A Top Down Description of S-CIELAB and CIEDE2000." *Color Research and Application* [Wiley] 28 [6]: 425-435.

[Hung, Enomoto and Kozo 1996] Hung, Po-Chieh, Hirochimi Enomoto, and Aoyama Kozo. (1996). "An evaluation of scanner noise based upon a human visual model." *IS&T's 49th Annual Conference.* Springfield, Ill.: Society for Imagine Science and Technology. 322-324.

[IS0 2011] ISO. (2011). "ISO 11452-4: Harness excitation methods." In *Road vehicles — Component test methods for electrical disturbances from narrowband radiated electromagnetic energy*.

[Kleinmann and Wueller 2007] Kleinmann, Johanna, and Dietmar Wueller. (2007). "Investigation of two Methods to quantify Noise in digital Images based on the Perception of the Human Eye." *Electronic Imagining: Image Quality and System Performance IV.* San Jose, Ca.: SPIE. 201-212.

[Krizhevsky, Sutskever and Hinton 2012] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. (2012). "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems.* Lake Tahoe, NV: Curran Associates, Inc.

[Mahesh and Subbarao 2008] Mahesh, G, and B Subbarao. (2008). "Comparison of Bulk Current Injection test methods of automotive, military and civilian EMC standards." *10th International Conference on Electromagnetic Interference & Compatibility.* Bangalore: IEEE. 547-551.

[OpenCV 2019] OpenCV. (2019). "Image Similarity - PSNR and SSIM." Accessed May 25, 2022. https://docs.opencv.org/2.4/doc/tutorials/highgui/video-input-psnr-ssim/video-input-psnr-ssim.html#image-similarity-psnr-and-ssim.

[US, Department of Transportation 2014] US, Department of Transportation. (2014). "NHTSA Announces Final Rule Requiring Rear Visibility Technology." 31 03. Accessed May 25, 2022. https://www.transportation.gov/briefing-room/nhtsa-announces-final-rule-requiring-rear-visibility-technology.

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/FUMF3414

Random Data Augmentation based Enhancement: A Generalized Enhancement Approach for Medical Datasets

Sidra Aleem[1], Teerath Kumar[2], Suzanne Little[2], Malika Bendechache[2], Rob Brennan[3], and Kevin McGuinness[1]

[1]*School of Electronic Engineering, Dublin City University, Ireland*
[2]*School of Computing, Dublin City University, Ireland*
[3]*ADAPT, School of Computer Science, University College Dublin, Ireland*

Abstract

Over the years, the paradigm of medical image analysis has shifted from manual expertise to automated systems, often using deep learning (DL) systems. The performance of deep learning algorithms is highly dependent on data quality. Particularly for the medical domain, it is an important aspect as medical data is very sensitive to quality and poor quality can lead to misdiagnosis. To improve the diagnostic performance, research has been done both in complex DL architectures and in improving data quality using dataset dependent static hyperparameters. However, the performance is still constrained due to data quality and overfitting of hyperparameters to a specific dataset. To overcome these issues, this paper proposes random data augmentation based enhancement. The main objective is to develop a generalized, data-independent and computationally efficient enhancement approach to improve medical data quality for DL. The quality is enhanced by improving the brightness and contrast of images. In contrast to the existing methods, our method generates enhancement hyperparameters randomly within a defined range, which makes it robust and prevents overfitting to a specific dataset. To evaluate the generalization of the proposed method, we use four medical datasets and compare its performance with state-of-the-art methods for both classification and segmentation tasks. For grayscale imagery, experiments have been performed with: COVID-19 chest X-ray, KiTS19, and for RGB imagery with: LC25000 datasets. Experimental results demonstrate that with the proposed enhancement methodology, DL architectures outperform other existing methods. Our code is publicly available at: `https://github.com/aleemsidra/Augmentation-Based-Generalized-Enhancement`.

Keywords: Classification, Data Augmentation, Generalized Enhancement, Segmentation

1 Introduction

DL algorithms are being used in many domains [Chandio et al., 2021, Turab et al., 2022, Kumar et al.,] and have revolutionized the field of medical image analysis. DL based applications are widely being used for computer assisted disease diagnosis to aid clinicians [Philipp et al., 2021]. Regardless of the DL model architecture, performance is strongly affected by the quality of the raw data [Zhou et al., 2019]. Particularly for medical images, data quality is a critical factor for reliable disease assessment and diagnosis [Chen et al., 2014]. Further issues arise due to the differences of acquisition protocols and the heterogeneity of data [Boyat and Joshi, 2015]. To work around these issues, research has been done both to introduce new DL architectures [Szegedy et al., 2015, Yadav and Jadhav, 2019, Cai et al., 2020] and to improve the data quality [Faes et al., 2019]. However, both of these solutions pose further problems. The use of new complex DL architectures has a number of drawbacks: (a) the issue of data quality remains persistent, which affects the final prediction; (b) it does not give an insight to the actual performance capability of traditional DL architectures; (c) it increases computational complexity without addressing the actual quality issue. To improve data quality, the optimal selection

 ISBN 978-0-9934207-7-1

of enhancement hyperparameters is extremely important. Due to the sensitive nature of medical data, this selection is even more important to have a robust performance independent of the data quality. The existing methods rely on fixed enhancement hyperparameters. These hyperparameters are chosen according to the dataset [Masud et al., 2021, Mittal et al., 2019, Hari et al., 2013]. Consequently, such enhancement methods are prone to overfitting. Moreover, these techniques are evaluated either on grayscale or RGB datasets [Gao et al., 2021, Wang et al., 2019, Zhou et al., 2019]. Thus, this the limited evaluation is not indicative of the strength of such methods. To prevent these issues, a computationally efficient and generalized contrast enhancement method can help.

To overcome above stated issues, this paper proposes random data augmentation based computationally efficient and generalized enhancement method, in which data quality is improved using random brightness and contrast hyperparameters. Unlike other existing methods, which use fixed hyperparameters for enhancement, we use a set of hyperparameters. The hyperparameters are randomly chosen from this set and hence are not reliant on the dataset. The random selection ensures that the hyperparameters do not overfit the data. The data-independent nature of these hyperparameters aids the proposed method to generalize well on different datasets. The range of brightness set is $[1.15, 1.35]$ and the range of contrast set is $[-0.1, 0.4]$. These specific ranges are chosen by performing experimental evaluation. First we started from -1.0 for both hyperparameters, where images have apparently no features. We evaluated the corresponding affect on enhancement by visualizing resultant data and kept on incrementing the value with interval of 0.15. It was observed that images started to show some feature at 1.15 and -0.1 brightness and contrast values respectively. Thus, these values were chosen as starting points for contrast and brightness sets. For the end point, we followed the same methodology and choose those values as end points before which images started to loose the features. The enhancement results achieved with starting and end points of selected range are shown in Figure 4. Beyond this particular range, image becomes darker or brighter and starts losing features. The performance is assessed by evaluating the resultant enhanced data with a variety of traditional DL architectures.

Contributions: The main contributions of our work are as follows:

- We propose a generalized and computationally efficient random data augmentation based enhancement approach for medical data.
- The enhancement hyperparameters are not manually selected according to the data; thus our enhancement method does not overfit a specific dataset.
- To check the effectiveness of our work, we perform extensive experiments on both gray scale and RGB datasets for classification and segmentation tasks.
- The proposed approach shows superior performance in terms of both accuracy and execution time over state-of-the-art techniques.

2 Related Work

Generally three types of contrast enhancement methods have been used: histogram methods, spectral methods and spatial methods [Pierre et al., 2017]. The histogram methods have remained very popular for contrast enhancement. Such methods transform gray scale images to an image with a specified histogram. However, such methods result in poor enhancement that can be attributed to both loss of information and over-enhancement of specific gray levels. Such methods are not adaptive and thus are inappropriate to provide contrast enhancement for the medical imaging domain [Reddy et al., 2018, Singh et al., 2016]. Spectral methods use wavelet transforms for quality enhancement. However, such methods fail to provide simultaneous enhancement to all the parts of of images. Moreover, it is difficult to automate enhancement using them [Wang et al., 2019]. The motive of the contrast enhancement in medical images is to aid clinicians with automated diagnosis, so such methods are also not best suited for medical domain.

A combination of adaptive histogram equalisation and discrete wavelet transform for contrast enhancement was used in [Lidong et al., 2015]. This process is comprised of three steps and gives a detailed output. However, the performance was affected due to contrast stretching and noise enhancement issues. Due to the sensitivity of

medical domain to errors, disease diagnosis can be affected by such noise enhancement. Furthermore, the filter based methods need to find the appropriate filter and their performance is highly reliant on the hyperparameters used. The hyperparameters are chosen according to the specific dataset. Thus these hyperparameters are not generalized to unseen datasets and consequently failed to perform well [Zhou et al., 2019]. So, such methods will fail to cope with the vast variety of medical data.

Gamma correction is one of the other enhancement methods used widely. Its performance is dependant on the γ coefficient. To deal with the varying illumination, an adaptive gamma correction technique to modify two non-linear functions has been proposed in [Shi et al., 2007]. However, these functions may be uniform for various regions and patterns of an image. An adaptive gamma correction method based on cumulative distribution to modify the histogram has been proposed in [Huang et al., 2012].

3 Proposed Method

In this section, we introduce our proposed approach for random data augmentation based enhancement. Let α and β be gain and bias. These parameters regulate contrast(α) and brightness (β). $f(i)$ is the input image and $g(i)$ is the resultant enhanced image [Steger et al., 2018] as shown in equation 1:

$$g(i) = \alpha f(i) + \beta. \tag{1}$$

Unlike other enhancement methods, which use fixed hyperparameters for enhancement and are prone to overfitting on particular dataset, our method randomly picks the value for α and β from the defined sets. To prevent overfitting, the set used consists of both positive and negative values for α and β. The values are randomly picked up from this set and are then used to augment data as shown in Algorithm 1. As α and β are data independent, this makes our proposed method a generalized enhancement approach that is suitable for different types of datasets.

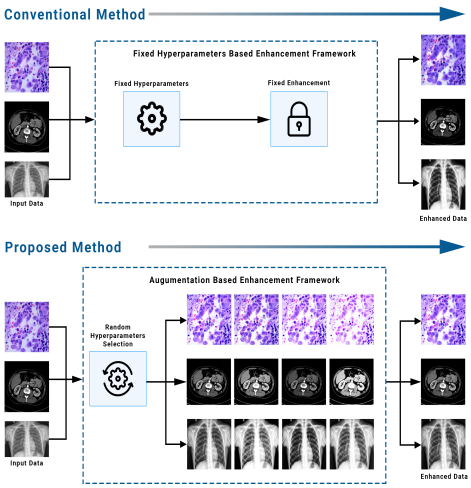


Figure 1: Overview of the proposed random augmentation based enhancement and the conventional enhancement

Algorithm 1: Random Data Augmentation based Enhancement

input : $f(i)$: batch of images
n: batch size
alpha (gain): randomly generated
beta (bias): randomly generated
output: $g(i)$: Enhanced images batch
for $f(i) \leftarrow 1$ **to** n **do**
$random\_index \leftarrow randrange(len(alpha))$
$r\_alpha \leftarrow alpha[random\_index]$
$random\_index \leftarrow randrange(len(beta))$
$r\_beta \leftarrow alpha[random\_index]$
$g(i) = clip(r\_alpha * f(i) + r\_beta, 0, 255)$
end

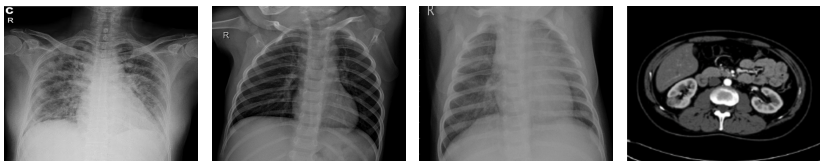
4 Experimental Results

4.1 Datasets

In this paper, four publicly available medical image datasets have been used. For evaluation on grayscale images: COVID-19 chest X-ray [P., 2020], KiTS19 [Heller et al., 2021] and for RGB images: LC25000 dataset [Borkowski et al., 2019] have been used. For a fair comparison, we used the same data splits as being used in [P., 2020, Heller et al., 2021, Borkowski et al., 2019]. COVID-19 chest X-ray, consists of 6432 images and have three classes: normal, pnuemonia and COVID-19 as shown in Figure 2. We divided it into stratified splits of 80% train, 10% validation and 10% test set. KiTS19 comprises of 300 gray scale abdominal scans of kidney patients with average of 216 slices (highest slice number is 1059) as shown in Figure 2(d). The ground truth for segmentation was created by experts with each pixel labeled as one of three classes: background, kidney or tumor. For validation three-fold cross validation has been used on 120 scans with their slices respectively. LC25000 dataset [Borkowski et al., 2019] comprises of 25000 histopathlogical images with 5 classes as shown in Figure. 3. There are 5000 images per class. We divided LC25000 with the ratio of 80:10:10 for train, validation and test set,respectively.

4.2 Implementation Details

SGD optimizer with a learning rate of $1e^{-4}$ was adopted and all models were trained for 100 epochs with batch size of 16. The cross entropy criterion is used. The codebase was setup in PyTorch framework and is available at github link [1].

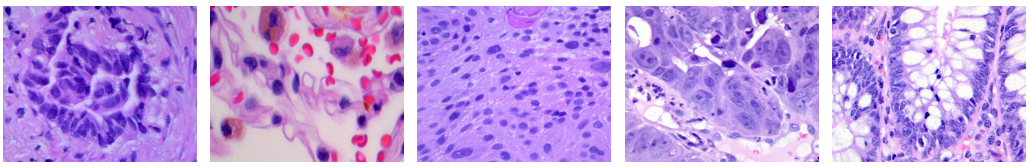


(a) (b) (c) (d)

Figure 2: (a) COVID (b) Normal (c) Pneumonia (d) Kidney CT scan.

[1] https://github.com/aleemsidra/Augmentation-Based-Generalized-Enhancement

 ISBN 978-0-9934207-7-1



(a) (b) (c) (d) (e)

Figure 3: (a) Lung Adenocarcinomas (b) Lung squamous cell carcinomas (c) Lung benign (d) Colon Adenocarcinomas (e) Colon Benign.

4.3 Statistical Evaluation

4.3.1 Evaluation on gray scale dataset

The proposed approach effectiveness is evaluated by conducting different experiments. Firstly, to evaluate the strength of the proposed enhancement technique, we used gray scale datasets: COVID-19 chest X-ray [P., 2020] and KiTS19 [Heller et al., 2021]. The performance is then compared with the other existing techniques. For COVID-19 chest X-ray [P., 2020], classification is performed and the accuracy score is used as evaluation criterion. For KiTS19 [Heller et al., 2021], segmentation is performed and performance is accessed using dice coefficient are used as an evaluation metric. In contrast to exisitng methods, our method achieved highest accuracy and dice coefficient score with COVID-19 chest X-ray and KiTS19, respectively.

Table 1: Accuracy comparison on the COVID-19 chest X-ray dataset

Model	ResNet18	ResNet50	ResNet101	VGG16	VGG19	Inception	DLH_COVID
Existing	0.9471	0.9346	0.9269	0.9486	0.9330	0.9144	0.9611
Proposed	**0.9642**	**0.9580**	**0.9486**	**0.9611**	**0.9517**	**0.9315**	**0.9626**

With our enhancement approach, state-of-the-art DL architectures achieved superior performance to that achieved with un-enhanced data as evident from Table 1 and Table 2. Thus, our enhancement approach aids the diagnostic performance of state-of-the-art DL architecture.

Table 2: Dice coefficent for segmentation on KiTS19 dataset

Model	U-Net	3D FCN	VB-Net	3D U-Net	MIScnn	**Proposed**
Kidney	0.9663	0.9805	0.9740	0.9743	0.9994	**0.9998**
Tumor	0.7778	0.8370	0.7890	0.8558	0.9319	**0.9411**
Background	-	-	-	-	0.6750	**0.6820**

4.3.2 Evaluation on RGB dataset

To further validate the effectiveness of the proposed method, we extended our work to RGB datasets. For this purpose, we used the LC25000 dataset [Borkowski et al., 2019]. To best of our knowledge, [Mangal et al., 2020] reported the highest accuracy with LC25000 dataset. So, we used its architecture to evaluate our proposed technique. With RGB datasets too, our method outperformed other existing techniques and achieved the highest accuracy as shown in Table 3. It justifies our claim of proposing a generalized contrast enhancement method for different types of datasets i.e. gray scale and RGB.

Table 3: Accuracy comparison of DL frameworks on the LC25000 dataset

Model	RF	Resnet50	CNN	CNN	CNN	**Proposed**
Lung	-	-	0.9720	0.9720	0.9789	**0.9844**
Colon	0.8530	0.9391	-	0.9720	0.9616	**0.9688**

4.4 Visual Evaluation

The visual analysis is equally important as clinicians will eventually use images for diagnosis. Firstly, to show the impact of the random hyperparameters, we selected the end points of brightness and contrast set range as discussed above in section 1. The impact is shown below in Figure. 4.

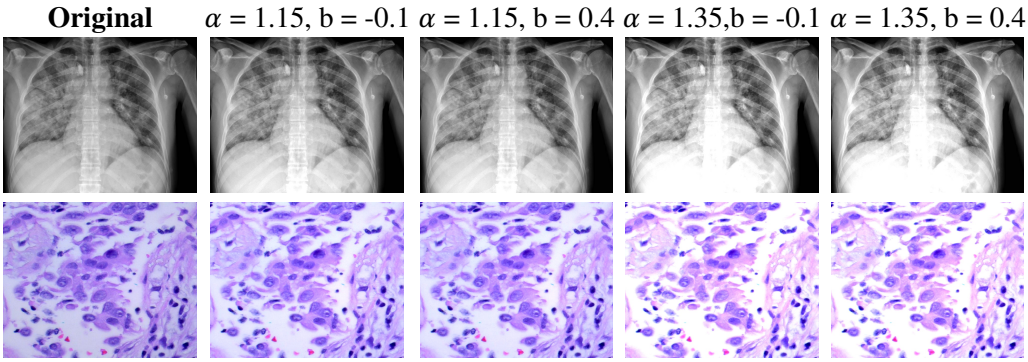


Figure 4: Ablation study to evaluate the impact of different hyper-parameters on data quality. The first column corresponds to original image and rest of the columns show imapct of different α and β values on image quality.

Further, to show the effectiveness of our enhancement method, we compare the original images with the images enhanced by our method as shown in Figure 5. It is evident that quality is significantly improved by our method.

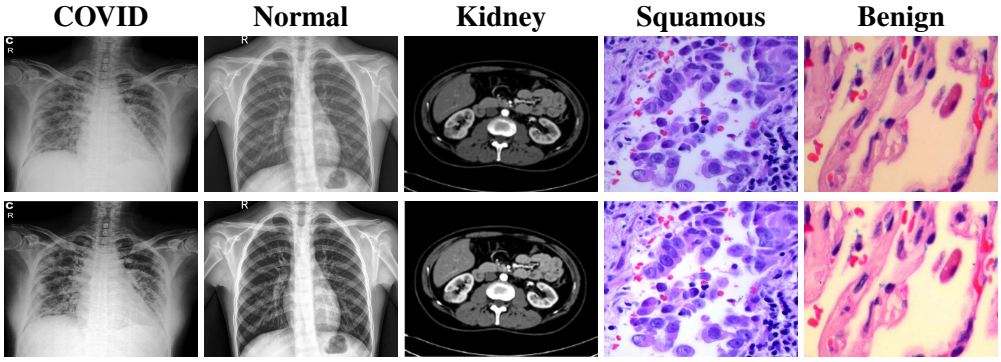


Figure 5: Visual results of enhancement: row 1 corresponds to input images, row 2 corresponds to enhanced images.

5 Conclusion

In this paper, we proposed a new enhancement technique based on data augmentation that improves medical image quality from brightness and contrast perspective. It uses a random hyperparameters selection

 ISBN 978-0-9934207-7-1

from the evaluated set of values. The proposed approach addressed the limitations of existing approaches including DL complex architecture and fixed hyperparameters based enhancement. Furthermore, to evaluate the generalization capability of the proposed approach, we used four datasets including grayscale and RGB with a wide variety of networks for both classification and segmentation tasks. The proposed approach, despite being a simple and easy to implement, outperforms existing approaches using state-of-the-art networks. To reproduce the results we made our code publicly available `https://github.com/aleemsidra/Augmentation-Based-Generalized-Enhancement`.

Acknowledgement

This research was supported by Science Foundation Ireland under grant numbers 18/CRT/6183 (ML-LABS Centre for Research Training),18/CRT/6223, SFI/12/RC/2289_P2 (Insight SFI Research Centre for Data Analytics), 13/RC/2094_*P*2 (Lero SFI Centre for Software) and 13/RC/2106_*P*2 (ADAPT SFI Research Centre for AI-Driven Digital Content Technology). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[Borkowski et al., 2019] Borkowski, A. A., Bui, M. M., Thomas, L. B., Wilson, C. P., DeLand, L. A., and Mastorides, S. M. (2019). Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*. [Online; accessed 28-Nov-2021].

[Boyat and Joshi, 2015] Boyat, A. K. and Joshi, B. K. (2015). A review paper: noise models in digital image processing. *arXiv preprint arXiv:1505.03489*.

[Cai et al., 2020] Cai, L., Gao, J., and Zhao, D. (2020). A review of the application of deep learning in medical image classification and segmentation. *Annals of translational medicine*, 8(11).

[Chandio et al., 2021] Chandio, A., Shen, Y., Bendechache, M., Inayat, I., and Kumar, T. (2021). Audd: audio urdu digits dataset for automatic audio urdu digit recognition. *Applied Sciences*, 11(19):8842.

[Chen et al., 2014] Chen, H., Hailey, D., Wang, N., and Yu, P. (2014). A review of data quality assessment methods for public health information systems. *International journal of environmental research and public health*, 11(5):5170–5207.

[Faes et al., 2019] Faes, L., Wagner, S. K., Fu, D. J., Liu, X., Korot, E., Ledsam, J. R., Back, T., Chopra, R., Pontikos, N., Kern, C., et al. (2019). Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *The Lancet Digital Health*, 1(5):e232–e242.

[Gao et al., 2021] Gao, G., Tong, S., Xia, Z., Wu, B., Xu, L., and Zhao, Z. (2021). Reversible data hiding with automatic contrast enhancement for medical images. *Signal Processing*, 178:107817.

[Hari et al., 2013] Hari, V., Raj, V. J., and Gopikakumari, R. (2013). Unsharp masking using quadratic filter for the enhancement of fingerprints in noisy background. *Pattern Recognition*, 46(12):3198–3207.

[Heller et al., 2021] Heller, N., Isensee, F., Maier-Hein, K. H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., and Han, M. (2021). The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821.

[Huang et al., 2012] Huang, S.-C., Cheng, F.-C., and Chiu, Y.-S. (2012). Efficient contrast enhancement using adaptive gamma correction with weighting distribution. *IEEE transactions on image processing*, 22(3):1032–1041.

[Kumar et al.,] Kumar, T., Brennan, R., and Bendechache, M. Stride random erasing augmentation.

[Lidong et al., 2015] Lidong, H., Wei, Z., Jun, W., and Zebin, S. (2015). Combination of contrast limited adaptive histogram equalisation and discrete wavelet transform for image enhancement. *IET Image Processing*, 9(10):908–915.

[Mangal et al., 2020] Mangal, S., Chaurasia, A., and Khajanchi, A. (2020). Convolution neural networks for diagnosing colon and lung cancer histopathological images. *arXiv preprint arXiv:2009.03878*.

[Masud et al., 2021] Masud, M., Sikder, N., Nahid, A.-A., Bairagi, A. K., and AlZain, M. A. (2021). A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. *Sensors*, 21(3):748.

[Mittal et al., 2019] Mittal, M., Goyal, L. M., Kaur, S., Kaur, I., Verma, A., and Hemanth, D. J. (2019). Deep learning based enhanced tumor segmentation approach for mr brain images. *Applied Soft Computing*, 78:346–354.

[P., 2020] P., P. (2020). Chest x-ray (covid-19 & pneumonia). `https://www.kaggle.com/prashant268/chest-xray-covid19-pneumonia`.

[Philipp et al., 2021] Philipp, M., Alperovich, A., Gutt-Will, M., Mathis, A., Saur, S., Raabe, A., and Mathis-Ullrich, F. (2021). Localizing neurosurgical instruments across domains and in the wild.

[Pierre et al., 2017] Pierre, F., Aujol, J.-F., Bugeau, A., Steidl, G., and Ta, V.-T. (2017). Variational contrast enhancement of gray-scale and rgb images. *Journal of Mathematical Imaging and Vision*, 57(1):99–116.

[Reddy et al., 2018] Reddy, P. S., Singh, H., Kumar, A., Balyan, L., and Lee, H.-N. (2018). Retinal fundus image enhancement using piecewise gamma corrected dominant orientation based histogram equalization. In *2018 International Conference on Communication and Signal Processing (ICCSP)*, pages 0124–0128. IEEE.

[Shi et al., 2007] Shi, Y., Yang, J., and Wu, R. (2007). Reducing illumination based on nonlinear gamma correction. In *2007 IEEE International Conference on Image Processing*, volume 1, pages I–529. IEEE.

[Singh et al., 2016] Singh, K., Vishwakarma, D. K., Walia, G. S., and Kapoor, R. (2016). Contrast enhancement via texture region based histogram equalization. *Journal of modern optics*, 63(15):1444–1450.

[Steger et al., 2018] Steger, C., Ulrich, M., and Wiedemann, C. (2018). *Machine vision algorithms and applications*. John Wiley & Sons.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

[Turab et al., 2022] Turab, M., Kumar, T., Bendechache, M., and Saber, T. (2022). Investigating multi-feature selection and ensembling for audio classification. *arXiv preprint arXiv:2206.07511*.

[Wang et al., 2019] Wang, W., Sun, N., and Ng, M. K. (2019). A variational gamma correction model for image contrast enhancement. *Inverse Problems & Imaging*, 13(3):461.

[Yadav and Jadhav, 2019] Yadav, S. S. and Jadhav, S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, 6(1):1–18.

[Zhou et al., 2019] Zhou, Y., Shi, C., Lai, B., and Jimenez, G. (2019). Contrast enhancement of medical images using a new version of the world cup optimization algorithm. *Quantitative imaging in medicine and surgery*, 9(9):1528.

Influence of Magnification in Deep Learning Aided Image Segmentation in Histological Digital Image Analysis

Kris D. McCombe[1], Stephanie G. Craig[1], Jacqueline A. James[1], Richard Gault[2]

1. Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Northern Ireland
2. School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Northern Ireland

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/RAKL2135

Abstract

The use of digital pathology has grown significantly for both healthcare and research purposes in recent years. With this comes opportunity to develop systems supported by computer vision (CV) and artificial intelligence (AI), with the potential to improve patient management and quality of care. The accessibility of CV and AI toolboxes have resulted in the rapid application of image analysis in this domain driven by accuracy related metrics. However, in this short paper we illustrate common pitfalls in the field through a semantic segmentation task, specifically how magnification can influence training data quality and demonstrate how this can ultimately affect model robustness.

Keywords: Digital Pathology, Segmentation, Medical Imaging

1 Introduction

Traditional pathology methods involve the examination of a patient tissue sample under a light microscope by a pathologist. Advances in computing hardware, image capture technology, and image compression algorithms have made it possible to scan, digitise, and store images of these patient samples at scale. The examination of computerised images has been termed Digital Pathology [Naizi et al., 2019]. Digital pathology holds many advantages over its traditional counterpart. For example, the digitisation of histopathology images allows for global remote access, which became particularly important over the past three years due to the COVID-19 pandemic [Browning et al., 2020]. Digital image analysis (DIA) can also be performed on these images for research and clinical biomarker quantification [Lara et al., 2021]. The rise of DIA has enabled the creation of more automated workflows through computer vision (CV) and Artificial Intelligence (AI) [Salto-Tellez et al, 2019]. Cancer Research UK estimates suggest that half of the UK population will get cancer in their lifetime [CRUK, 2015]. With increasing consultation requests and a decrease in qualified pathologists to assess them [George et al., 2020], the opportunity to implement an AI-assisted diagnostic workflow is prominent. Ultimately, these AI-driven tools have the potential to shorten time to diagnosis and improve patient outcomes [Steiner et al, 2020]. In this short paper we outline a workflow using semantic segmentation to automate annotation of histological structures in digitised cancer images. In addition, we examine how the choice of magnification can influence training data, and how this can impact on overall model performance and the potential for clinical translation.

2 Methods

This work aims to semantically segment complex structures from digital pathology images across multiple cancer types to examine the impact the magnification of the images can have on overall model performance and potential future clinical translation. Our hypothesis is that increased model performance may not always translate to the most applicable approach for clinical translation.

2.1 Data Description and Ground-truth annotation

Haematoxylin and Eosin-stained slides were obtained with ethical approval from the Northern Ireland Biobank (NIB21/0008) from cancer patients spanning four cancer types: Oropharyngeal, Lung, Breast and Oesophageal Cancer (151 patients in total). The slides from were digitised and saved using the Leica Aperio AT2 at x40 magnification (0.25 μm / pixel). For the purposes of this case study, Tertiary Lymphoid Structures (TLS), an ectopic immune structure that can occur during clinical scenarios of chronic inflammation, including cancer, were chosen as the target for segmentation. TLS have shown prognostic and predictive potential in cancer therapy, and are present across multiple cancer types [Schumacher et al 2022]. Identification of these complex structures is often time-consuming, providing motivation to develop an automated segmentation tool. TLS were manually annotated using QuPath (v0.3.2) [Bankhead et al. 2017] to generate ground-truth binary masks.

2.2 Image Tiling and Extraction

Whole Slide Images (WSIs) are very large often spanning 80k x 60k pixels making them prohibitive to perform deep-learning based image analysis at a whole image level. A common approach in the field to overcome this challenge is to apply tiling; breaking down these images into segments of manageable size that can be handled by commercial computers. This raises questions regarding wider image context and appropriate magnification for histological structures. To investigate this, we extracted tiles across 20x, 10x, 4x and 1x magnification, mimicking a physical microscope, to explore how this affects model training and performance. Tiles were extracted at 50% overlap with a resolution of 512px by 512px. This equated to a "real" tile size of 26.83mm^2, 1.68mm^2, 0.27mm^2 and 0.07mm^2 for 1x, 4x, 10x and 20x respectively. Only tiles that exhibited a TLS were used for training (2208 tiles at 1X, 7675 at 4X, 12841 at 10X and 51275 at 40X). Tile extraction was performed using a bespoke QuPath/Groovy script which can be found here: https://github.com/KDM-Echo/IMVIP22-Magnification. Patients were divided into a 70:15:15 ratio for training, validation and independent test sets. The training, test and validation datasets are split at a patient level to avoid any potential data leakage.

2.3 Model and Training

The semantic segmentation model was built using the PyTorch framework (v1.11.0 + CUDA 11.6) through the segmentation-models-pytorch python package. The model consists of a UNet++ Architecture with an efficientnet-b0 encoder. UNet++ was chosen due to its regular use in digital pathology image segmentation, and the relatively lightweight nature of the model. The model was initiated with pretrained weights from ImageNet. The model was trained for 5 epochs. The Dice Loss function from the segmentation-models-pytorch was used. The optimiser was ADAM set at a learning rate of 0.001. A batch size of 8 was used as this maximised available memory. The model was trained on an Octane VI laptop with 32GB RAM, an intel Core i7-9700 CPU, and RTX 2070 8GB GPU. Models which achieved the highest Intersect over Union (IoU) score in the validation set were brought forward for analysis in the independent test cohort.

2.4 Model Evaluation

Models were evaluated by accuracy, precision, recall, Dice score, IoU score and Matthew's Coefficient in the independent patient set, both at a tile level, which were guaranteed to contain a TLS, and at a whole slide level for all four magnifications. The time taken to run inference on these independent, whole slide test images was also recorded.

3 Results

A summary of the validation results can be seen in Table 1. At a WSI level, the 4X magnification achieved the highest Dice, IOU score, pixel accuracy and Matthew's Coefficient score. In contrast, 1X and 20X models achieved the worst IOU and Matthew's Coefficient. At a WSI level, recall increased with magnification, with precision

decreasing at 10X and 20X magnification. The times taken to segment the 23 whole slide images in the test set were 61 seconds at 1X (22.8 slides/min) 19 minutes 44 seconds at 4X (1.16 slides/min), 105 minutes 16 seconds at 10X (0.22 slides per minute) and 534 minutes 56 seconds at 20x (23.25 min/slide).

Magnification	Accuracy	Dice	IOU	Recall	Specificity	Precision	Matthew's Coefficient
1X (Tile)	0.991	0.611	0.571	0.459	**0.998**	0.725	0.573
4X (Tile)	**0.956**	0.610	0.550	0.709	0.976	0.704	**0.682**
10X (Tile)	0.858	0.579	0.487	0.827	0.868	0.685	0.657
20X (Tile)	0.836	**0.816**	**0.735**	**0.930**	0.676	**0.830**	0.641
1X (WSI)	0.996	0.804	0.788	0.459	**0.999**	**0.700**	0.565
4X (WSI)	**0.997**	**0.935**	**0.931**	0.709	0.998	0.579	**0.639**
10X (WSI)	0.986	0.844	0.842	0.823	0.987	0.201	0.403
20X (WSI)	0.900	0.657	0.656	**0.903**	0.900	0.035	0.166

Table 1. Modelling results for the independent test set. Best result for each category is highlighted in bold. At a tile level, TLS are guaranteed to be in the image whilst the WSI level looks at the entire patient.

4 Discussion

4.1 Correct Magnification is Crucial in Histological Object Detection

In this study, we examined the influence of magnification on quality of complex semantic segmentation. When acting on the individual tiles that were guaranteed to have a TLS, it was found that the model trained at 20X magnification yielded the best Dice and IOU scores. In general, precision and recall both tend to increase with magnification when inferring on these tiles. This is likely due to the changes in negative to positive pixel ratio within the training set between magnifications (88:1 in 1x, 15.4:1 in 4x, 3.5:1 in 10x, 0.74:1 in 20x). With considerably more positive pixels to learn from, it is understandable why these trends exist, with less potential for false positives due to larger proportion of positive pixels in the ground truth images. In real-world scenarios however, these models will not have specifically curated images to evaluate. For this reason, we applied each model across the whole image for the independent test set. All models achieved marked increases in IoU scores however, this is likely due to the relatively large imbalance between non-TLS and TLS labelled pixels indicating that tile-based evaluation is important to assess the model's ability in the given segmentation task. Interestingly, the highest IoU scores were achieved in the 4X magnification (Table 1), a contrast to the best performing 20X images at the tile level. Where recall again tended to increase with magnification, precision appeared to fall, indicating increasing false positives. While this may be down to the increase positive pixel ratio as mentioned before, this may also be the result of a loss of architectural context. Figure 1 shows the same TLS at the different magnifications. These results make it clear that identifying the correct magnification per structure is imperative in model success. The average size of a TLS in our images was 0.21mm^2, close to the 10x tile size of 0.27mm^2, where the negative-positive pixel ratio increases drastically. It may also be advantageous to develop a model that utilises multiple magnifications, simulating the zooming in and out on a physical microscope [Rijthoven et all, 2021].

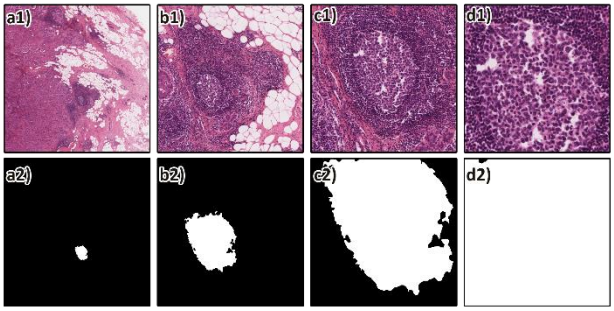


Figure 1. Raw image and corresponding mask of the same TLS at 1x (a) 4x (b), 10x (c) and 20x (d).

4.1 Making a Clinically Relevant Tool

It is important to take into account additional time and hardware costs for each approach. A higher magnification results in a greater number of tiles to examine meaning a longer inference time. For a tool to be clinically relevant, it should at least match the time taken for a pathologist to manually examine the slide in the same manner. Inference at 20X magnification took approximately 9 hours for 23 whole slide images compared to with 20 minutes at 4X for

 ISBN 978-0-9934207-7-1

the same images. For context, review and digital annotation of these images took approximately 5-10 minutes per image. Ultimately there remains somewhat of a knowledge and understanding gap between digital pathology image analysis and the more traditional pathology aspects, meaning the majority of DIA AI research is not translated to clinical use [Steiner, 2020]. It is unreasonable to expect pathologists to develop a background in coding to utilise these deep learning techniques, just as it is to expect computer scientists to become experts in cancer pathology. The development of robust AI models is a collective endeavour and efforts must be made to bridge the knowledge gap between the two disciplines. The development of open-source, graphical user interfaces (GUI) offering a code-free, point-and-click experience, may be a potential remedy. By providing a GUI we can remove the intimidation of coding, and by keeping the software open source, we provide a pathway for community driven development.

5 Conclusion

In this study, we demonstrate the importance of selecting the correct magnification and appropriate context to perform semantic segmentation of digital pathology images. When developing deep learning models for DIA, it is important to understand how changing magnification can impact training data and tissue architectural context. We demonstrate why it is important to extrapolate models to whole patient images for evaluation to reflect real world scenarios and how this can influence optimal model selection. The results show (Figure 1 and Table 1) that increased metrics do not necessarily translate to a robust clinically relevant model. Finally, we show it is important to consider time as a factor, as these tools ultimately need to be designed to improve times to diagnosis as well as impact on patient experience.

References

[Bankhead et al. 2017] Bankhead P., Loughrey M., Fernández J. et al (2017) QuPath: Open source software for digital pathology image analysis. *Scientific Reports (7: 16878)*

[Browning et al., 2020] Browning L., Colling R., Rakha E. et al (2020) Digital pathology and artificial intelligence will be key to supporting clinical and academic cellular pathology through COVID-19 and future crises: the PathLAKE consortium perspective. *Journal of Clinical Pathology (74: 443-447)*

[CRUK, 2015] CRUK 1 in 2 people in the UK will get cancer, *Accessed July 15th 2022* (https://news.cancerresearchuk.org/2015/02/04/1-in-2-people-in-the-uk-will-get-cancer/)

[George et al., 2020**]** George J., Gkousis E., Feast A. et al. (2020). Estimating the cost of growing the NHS cancer workforce in England by 2029, *CRUK*

[Lara et al., 2021] Lara H., Li Z., Abels E., et al. (2021) Quantitative Image Analysis for Tissue Biomarker Use: A White Paper From the Digital Pathology *Association Appl Immunohistochem Mol Morphol (29(7):479-493)*

[Naizi et al., 2019] Niazi M. et al. (2019) Digital pathology and artificial intelligence, *Lancet Oncol (20: 253-261)*

[Rijthoven et all, 2021] Rijthoven M. et al. (2021) HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images *Medical Image analysis (68(9):101890)*

[Salto-Tellez et al, 2019] Salto-Tellez M., Maxwell P., Hamilton P. (2019) Artificial intelligence-the third revolution in pathology *Histopathology (3:372-376)*

[Schumacher et al, 2022] Shumacher, T., Thommen, D. Tertiary Lymphoid Structures in Cancer *Science (375(6576): eabf9419)*

[Steiner et al, 2020] Steiner D., Chen P., Mermel C. (2020) Closing the translation gap: AI applications in digital pathology *Biochimica et Biophysica Acta – Reviews on Cancer (1875(1):188452)*

Sign2Speech: A Novel Sign Language to Speech Synthesis Pipeline

Dan Bigioi[1*], Théo Morales[2*], Ayushi Pandey[2*], Frank Fowley[3*], Peter Corcoran[1] and Julie Carson-Berndsen[3]

[1] *National University of Ireland Galway*
[2] *Trinity College Dublin, Ireland*
[3] *University College Dublin, Ireland*
[*] *Each author contributed equally to this paper*

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/CTDH7516

Abstract

The lack of assistive Sign Language technologies for members of the Deaf community has impeded their access to public information, and curtailed their civil rights and social inclusion. In this paper, we introduce a novel proof-of-concept method for end-to-end Sign Language to speech translation without an intermediate text representation.We propose an LSTM-based method to generate speech from hand pose, where the latter can be obtained from applying an off-the-shelf pose predictor to fingerspelling videos. We train our model using a custom dataset of synthetically generated signs annotated with speech labels, and test on a real world dataset of fingerspelling signs. Our generated output resembles real-world data sufficiently on quantitative measurements. This indicates that our techniques can be used to generate speech from signs, without reliance on text. The use of synthetic datasets further reduces the reliance on real-world, annotated data. However, results can be further improved using hybrid datasets, combining real-world and synthetic data. Our code and datasets are available at https://github.com/DanBigioi/Sign2Speech.

Keywords: Sign Language, Speech Synthesis, Machine Learning, Recurrent Neural Networks, Machine Translation

1 Introduction

Sign Language is the first language of an estimated 70 million deaf people in the world, according to the World Federation of the Deaf. Among the Deaf Community, deafness is not considered to be a disability but is the identity of a linguistic and cultural minority group. However, many Deaf persons have faced severe difficulties in accessing public information in their preferred language, which has negatively impacted the realisation of their civil rights and liberties [CIB, 2017]. The potential benefits of developing such assistive technologies include increased social integration and a heightened sense of autonomy and personal well-being.

Presently, most synthetic speech generation pipelines follow a text-to-speech framework. However, annotating Sign Language to text can prove cost-inefficient and error-prone. Moreover, a two-step sign-to-text-to-speech conversion can introduce a lag in real-time inference of the speech. In this paper, we introduce a proof-of-concept prototype for the direct translation of Sign Language to speech via deep learning techniques. Our novel contributions are twofold:

- We propose a method to generate speech from hand pose, where the latter can be obtained from an off-the-shelf pose predictor.
- We design and implement a proof-of-concept system for Irish Sign Language (ISL) to Speech and assess the feasibility of ISL to speech without intermediary text representation. We focus mainly on translating fingerspelling signs to speech as part of this initial work.

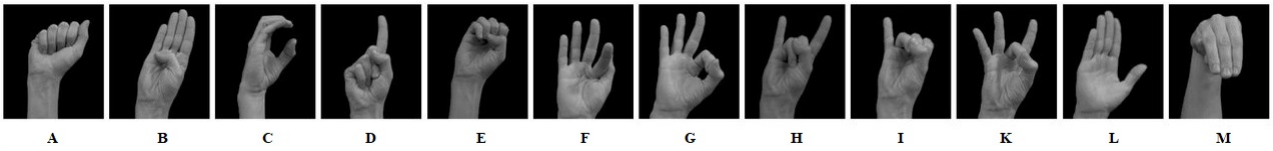


Figure 1: **The ISL Fingerspelling Alphabet** – Static signs examples. (Source: Irish Deaf Society)

2 Background

Despite recent research into Sign Language processing, there is no practical, automated solution for the interpretation of Sign Language [Dyzel et al., 2020]. The main challenges in Sign Language processing are: 1) the sparsity of suitable datasets for Deep Learning models; 2) the linguistic complexity of Sign Languages; and 3) the Computer Vision difficulties due to occlusion and perspective issues as well as subject fluency and variability.

As a highly expressive visual-gestural language, the articulators in Sign Language are the hands, arms, head, body and face. The manual features, or those involving hand movements include handshape, orientation, movement and location. Non-manual features, including facial expressions, mouth shapes, and body and head movements [Leeson et al., 2020], add further complexity to the visual medium. As preliminary work in this direction, we have focused the present work on the manual features of the static ISL fingerspelling alphabet.

Of the 81 fundamental handshapes described in [Leeson et al., 2020], there are 26 fingerspelling signs that form parallel to the English alphabet. These include 23 stationary letter signs and 3 moving letter signs (“J”, “X” and “Z”) in the ISL fingerspelling alphabet. The stationary signs are depicted in Figure 1. These signs are not only used for spelling out proper names and abbreviations, but also encode lexical items (“M” articulates “mother”).

Fingerspelling presents an important test-case to computer vision because of the high frequency of fingerspelled alphabets in connected Sign Language. Secondly, the speed and dexterity of native signers can make their segmentation a harder problem.

3 Methodology

3.1 Datasets

To train the proposed model, a dataset of ISL fingerspelling labelled with audio targets is necessary. We use the synthetic data generation framework proposed by [Fowley and Ventresque, 2021] to produce training datasets for our models, and generate Mel spectrograms from audio recordings. The sequence training dataset includes 9,360 synthetically generated 30-frame sequences, with 360 sample clips for each of the 26 letters. Each sequence represents an animation from a neutral pose in frame 0 to the letter pose in frame 29. Rather than using RGB frames of hand signs as input, we extract hand-pose key-points and use these as our training features. Each sequence has an associated audio label.

Our models were tested on the ISL-HS corpus of ISL fingerspelling signs [Oliveira et al., 2017] recorded by six native signers. The dataset is available as 468 short videos of static ISL alphabet letters. To ensure generalizability on an unseen test dataset, we included signer-specific variations to each sign. Additionally, signer fluency and dexterity was varied by introducing random finger bone rotations.

3.2 Network Architecture & Training Details

Network Architectures To model the sequential features from the extracted hand poses, we use a bidirectional Long Short-Term Memory (LSTM) to regress the Mel spectrogram coefficients. Fig. 2 depicts a high-level overview of the pipeline.

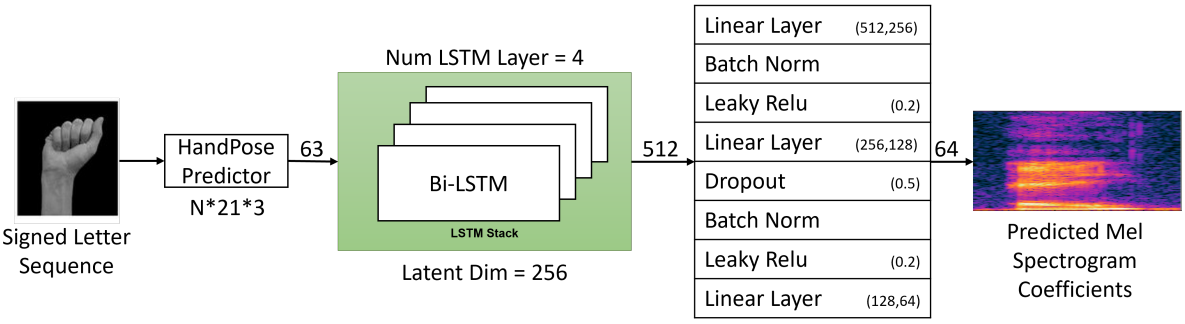


Figure 2: **Long Short-Term Memory (LSTM)-based network architecture** – The proposed architecture decodes temporal features from and predicts Mel spectogram coefficients in an autoregresive manner. The keypoints are predicted by an off-the-shelf method.

Training Details The LSTM-Based architecture was trained for a total of 100 epochs, with a batch size of 64, the mean squared error loss function, and a learning rate of 0.001. 70% of the synthetic sequence dataset was used for training, 20% for validation, and 10% was withheld for testing. Additionally, the LSTM-model was further evaluated on the real world fingerspelling ISL-HS corpus. [Oliveira et al., 2017].

4 Results

Our models were tested on the ISL-HS corpus of ISL fingerspelling signs [Oliveira et al., 2017]. Since the purpose of this experiment is to demonstrate the feasibility of generating speech signals from fingerspelling sequences, we do not evaluate the naturalness and quality of produced audio waveforms. Instead, the results are evaluated with the mean squared error, displayed in Table 1. Presently, the audio sounds "robotic" in nature as we are inverting the generated mel spectrogram using the Griffin-Lim reconstruction algorithm. For better quality speech, the use of a neural vocoder is recommended. Fig. 3 below depicts samples outputted by our LSTM-model.

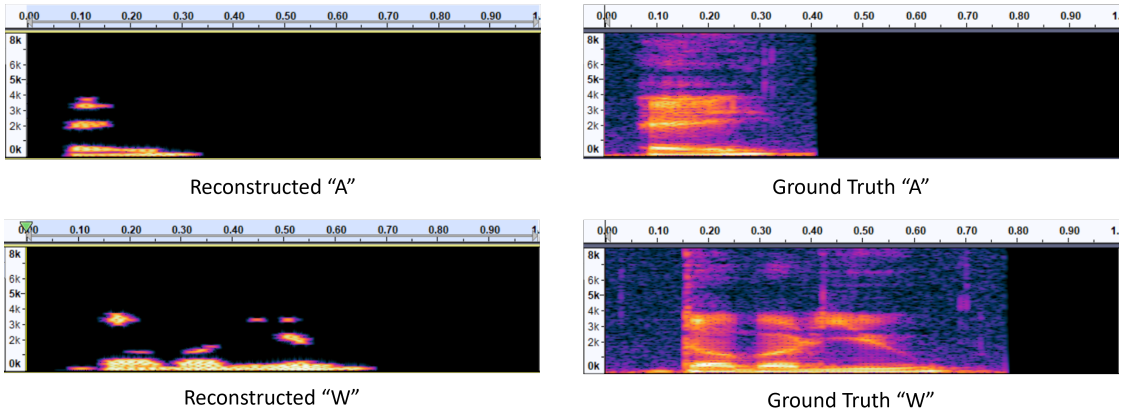


Figure 3: **Spectrograms of generated sounds versus ground truth** – The X-axis represents the time step and the y-axis encodes the Mel coefficients. The reconstructions don’t encode the noise but are close to the ground-truth signals.

While most poses that are input into the network clearly sound like their intended letters, there are a few that are unintelligible or sound different from the desired ground truth. We suspect that this is due to the limited size of our dataset, which does not capture enough variability that is present in real world data. It can also be attributed to a class overlap, where poses gesturally similar to each other (e.g, P and E) result in similar

LETTER	A	B	C	D	E	F	G	H	I	J	K	L	M
MSE	47	41	61	64	137	50	75	51	479	55	69	66	49

LETTER	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
MSE	76	81	116	44	47	72	37	94	88	81	130	86	69

Table 1: **Mean Squared Error (MSE) of letters from real-world ISL-HS corpus** – Most letters show a similar MSE value, although a few letters such as I, E, X, and P are challenging for the network. P and E being nearly identical sounds, they are more difficult to differentiate for the model.

sounding audio output. A potential solution to this is to build a hybrid dataset that combines real world data with synthetic data.

5 Conclusion & Future Work

Through this work, we show that modern deep-learning based techniques can prove feasible for sign-to-speech translation. We also show that synthetic data can alleviate the reliance on labelled, real-world data. As future work, we will experiment with creating hybrid datasets for training, and using vocoders for generating natural sounding speech. Also, following from our initial designs on fingerspelling signs, we will also enhance this pipeline to generate full words and sentences directly from sign. In future, we plan to expand this work to handle more complex sign sequences, and to generate full words and sentences from sign rather than just fingerspelling signs.

Acknowledgments This work has the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, and the ADAPT Centre (Grant 13/RC/2106). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission

References

[CIB, 2017] CIB (2017). *Information provision and access to public and social services for the Deaf Community*. Government of Ireland.

[Dyzel et al., 2020] Dyzel, V., Oosterom-Calo, R., Worm, M., and Sterkenburg, P. S. (2020). Assistive technology to promote communication and social interaction for people with deafblindness: A systematic review. *Frontiers in Education*, 5:164.

[Fowley and Ventresque, 2021] Fowley, F. and Ventresque, A. (2021). Sign language fingerspelling recognition using synthetic data. *Proc. 29th Irish Conferenfe on Artificial Intelligence and Cognitive Science*, pages 1–6.

[Leeson et al., 2020] Leeson, L., Sheridan, S., Cannon, K., Murphy, T., Newman, H., and Veldheer, H. (2020). Hands in motion: Learning to fingerspell in irish sign language. *TEANGA the Journal of the Irish Association for Applied Linguistics*, 11:120–141.

[Oliveira et al., 2017] Oliveira, M., Chatbri, H., Little, S., Ferstl, Y., Oconnor, N. E., and Sutherland, A. (2017). Irish Sign Language recognition using principal component analysis and Convolutional Neural Networks. *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*.

Geometrically reconstructing confocal microscopy images for modelling the retinal microvasculature as a 3D cylindrical network

Evan P. Troendle[1,*], Peter Barabas[1], and Tim M. Curtis[1]

[1]*Wellcome-Wolfson Institute for Experimental Medicine, Queen's University Belfast, Belfast, BT9 7BL, Northern Ireland, United Kingdom*

[*]*Corresponding author: Evan P. Troendle, E.Troendle@qub.ac.uk*

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/KTXE9847

Abstract

Microvascular networks can be modelled as a network of connected cylinders. Presently, however, there are limited approaches with which to recover these networks from biomedical images. We have therefore developed and implemented computer algorithms to geometrically reconstruct three-dimensional (3D) retinal microvascular networks from micrometre-scale imagery, resulting in a concise representation of two endpoints and radius for each cylinder detected within a delimited text file. This format is suitable for a variety of purposes, including efficient simulations of molecular delivery. Here, we detail a semi-automated pipeline consisting of the detection of retinal microvascular volumes within 3D imaging datasets, the enhancement and analysis of these volumes for reconstruction, and the geometric construction algorithm itself, which converts voxel data into representative 3D cylindrical objects.

Keywords: Image Processing, Retina, Vasculature, 3D Reconstruction, Shape Recovery

1 Introduction

Measuring three-dimensional (3D) blood vessel volumes is an essential step in analysing healthy capillary network architectures as well as deviations from them during pathological processes. From the clinical perspective, quantitative analysis of blood vessel topography in 3D offers a baseline to compare against and provides a basis for understanding blood flow dynamics or communication between blood vessel and tissue volumes [Secomb, 2008]. Further, reconstruction of complex, realistic vascular networks allows for efficient computer simulations that physically model the transport of molecules from the vasculature into tissues [Troendle et al., 2018]. Efficient representations of 3D vascular structures are paramount to identify the structural features that impact extravasation and molecular delivery within heterogeneous microvasculatures to advance systemic drug delivery and development strategies.

Retinal microvasculatures are valuable to investigate scientifically as they are readily accessible, yet highly dense and complex microvascular systems that can be adequately imaged in full at high resolution *ex vivo* [Prahst et al., 2020]. The study of retinal microvascular complications in most mammals can be used to better understand disease progression in humans [Curtis et al., 2009] as their anatomical features and physiology recapitulate many features found within human retinas [Chang, 2013]. At present, therapeutics for advanced retinal vascular diseases are primarily administered intravitreally [Virgili et al., 2018, Solomon et al., 2019]. However, a better understanding of systemic pharmaceutical delivery to the retina could assist in the development of less-invasive drug delivery approaches.

In this study, we describe a sequence of semi-automated image processing techniques that can be used to geometrically reconstruct the retinal microvasculature in 3D as a network of connected cylinders. These techniques allow for further 3D quantification of the retinal microvasculature to better understand and model retinal diseases and therapies. Furthermore, this methodology should be applicable to the study of any microvascular bed.

2 Methods

2.1 Preparing the retina for microscopy

Due to biological and technical limitations at present, it is only feasible to experimentally acquire microvascular imagery in sufficient resolution for 3D reconstruction *ex vivo*. Present *in vivo* techniques are of insufficient resolution to image the entire retinal microvasculature at the capillary level [Roorda et al., 2006, Yu et al., 2015]. For this study, retinal tissue was harvested from the left eye of a streptozotocin (STZ)-induced diabetic male C57BI6J mouse model aged 19 weeks. STZ was dissolved in 0.1M citrate buffer pH 4.5 at 10 mg/ml and injected intraperitoneally once per day on five consecutive days at a dose of 50 mg STZ / kg body weight. The eye was fixed for 1 hour in 4% paraformaldehyde solution and stored in phosphate-buffered saline (PBS) until immunohistochemistry was performed to prepare the blood vessels for acquisition using confocal scanning laser microscopy.

For detecting blood vessels, biotinylated isolectin B4 from the plant *Griffonia simplicifolia* was used, as it specifically binds to the α-galactose residues of glycoproteins, which are typically present on endothelial cells [Kirkeby and Moe, 2001]. While the lectin serves as the primary agent for recognizing the target molecules, the secondary in this case was streptavidin conjugated to the fluorophore Alexa Fluor™ 568. The murine retina was permeabilised using a solution of TritonX-100 [Johnson, 2013] in PBS[T8787-100ML, Merck] (0.5% v/v), followed by incubation with isolectin B4 [L2140-1MG, Merck] (1:200) in PBS+TritonX-100 for 4 days at 4 °C on a rocker. Subsequently, PBS was used to wash the retinal samples, using a fast PBS wash (10 exchanges) followed by another wash with 30 minute incubation in refreshed PBS on rocker at 4 °C over ten washing iterations; the total washing time amounted to 6.5 hrs. Streptavidin, Alexa Fluor™ 568 conjugate [S11226, Thermo Fisher Scientific] (1:200) in PBS+TritonX-100 was incubated overnight at 4 °C on a rocker, followed by the same wash steps that were used after the primary step. Finally, the retina was flat-mounted under dim red light in VECTASHIELD H-1000 [H-1000, Vector Laboratories] and protected under a coverslip.

Animal work was approved by the Queen's University of Belfast Animal Welfare and Ethical Review Body (AWERB). Work adhered to Department of Health, Social Services and Public Safety (DHSSPS) project license PPL2814.

2.2 Scanning confocal microscopy for the acquisition of the retina

A Leica TCS SP8 laser scanning confocal inverted microscopy system [Leica Microsystems] equipped with a 20× air objective [Leica 20x HC PL APO CS2; NA = 0.75] was used to image the murine retinal sample from the vitreal surface through the inner retina. The confocal pinhole was set to 1 Airy unit, the scanning speed was 400 Hz, and 512 × 512 pixels (px.) were captured in each imaging frame across a field of 8 × 9 laser patterning tiles encompassing the entire flat-mount preparation. The z-step was set to 1 µm with 75 steps in total, resulting in a tissue depth of 75 µm; the duration of the image acquisition took 2 hours.

Within the LasX Leica software, x-y images belonging to the same optical plane were stitched with 10% overlap for all z-slices, resulting in aggregated images that best depicted independent focal planes. This yielded a dataset of 75 3890 × 4343 px. images, with each pixel corresponding to voxels of spatial extent 1.14 µm × 1.14 µm × 986.6 nm. Image data were exported to an uncompressed 3D Tag Image File Format (TIFF) for input into subsequent processing steps for 3D geometric reconstruction. The maximum intensity z-axis projection of the acquired murine flat-mounted retina is shown in Figure 1.

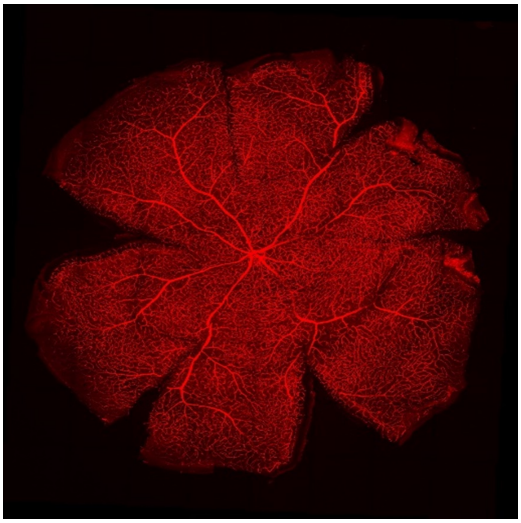


Figure 1: Maximum intensity z-axis projection of the murine flat-mounted retina.

 ISBN 978-0-9934207-7-1

3 Results

3.1 Semi-automated detection and refinement of microvascular volumes in 3D

For the remainder of this study, image processing algorithms utilised the OpenCV C++ API [Bradski, 2000]. To detect microvascular endothelia from the confocal microscopy images, noise was firstly reduced by applying a 2D median filter [Huang et al., 1979] (5×5 px. kernel) over each image slice. Next, a Laplacian of Gaussian [Lindeberg, 2015] filter ($\sigma = 2.0$ px ≈ 2.28 µm) was utilised to further reduce noise and detect edge-like features corresponding to the endothelial boundaries. Following this step, a global intensity threshold was applied to detect the endothelial edges as a binary (i.e., bi-partite-valued) mask.

As the isolectin B4 staining is selective to visualising endothelia, the vessel lumens emitted low to non-detectable fluorescence levels and were not initially included in the binary mask following the application of the global intensity threshold. To recover these regions as part of the total microvascular volume, a filling approach was necessary. Traditional recursive [Newman and Sproull, 1979] or iterative scan-line-based [Vučković et al., 2019] region filling algorithms in 2D and 3D were not suitable for these experimentally-derived images due to the presence of small gaps in the endothelial boundaries that persisted after enhancement (see Figure 2). This necessitated the development of a novel region-filling approach as described below.

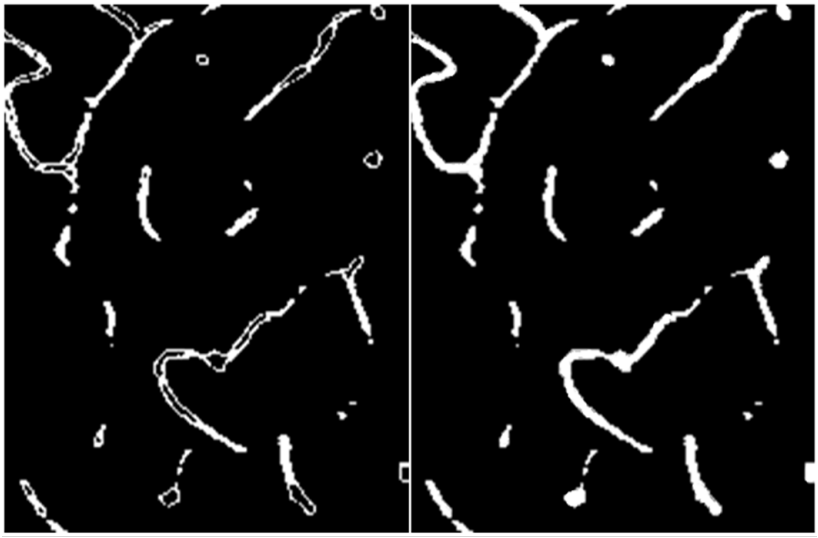


Figure 2: Detection and filling of microvascular volumes from their endothelial boundaries. Left: Boundary pixels of the microvascular endothelia were detected for a subset of the 3D imaging data; note the existence of many small gaps precluding conventional region filling techniques. Right: The resultant data after filling according to the algorithm described in this manuscript.

To fill the microvascular volumes, first-order x-, y-, and z- image intensity (I) partial derivatives (i.e., $\partial I/\partial x$, $\partial I/\partial y$, and $\partial I/\partial z$) were initially assessed for every pixel in the original microscopy data (Figure 1). For each boundary coordinate in the binary mask (Figure 2), an orthogonal unit vector was constructed from the image intensity derivative components. Then, a potential linear filling operation propagated from each boundary coordinate in the direction of the orthogonal vector and was accepted should the voxel march [Amanatides and Woo, 1987] hit another boundary pixel with an orthogonal vector pointing in approximately the opposite direction (e.g., <5° from parallel). This approach can be implemented in 2D as well as 3D, by applying the algorithm within each 2D slice and thus omitting the z- components, resulting in a filled binary microvascular volume.

Following filling, the 3D Euclidean distance transform [Felzenszwalb and Huttenlocher, 2012] and non-maximum suppression [Marin et al., 2015] was applied on the microvascular volume to estimate the vessel centrelines and radii. A depiction of the 3D Euclidean distance transform applied to the microvascular volume as detected and filled by the 2D region filling algorithm is shown in Figure 3.

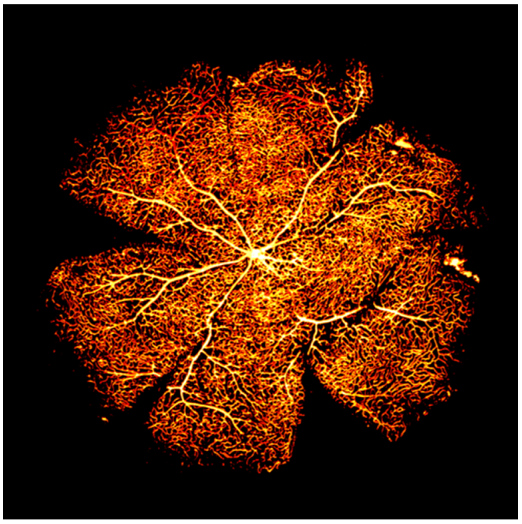


Figure 3: Maximum intensity z-axis projection of the 3D Euclidean distance transform applied to the detected and filled microvascular voxels within the murine flat-mounted retina. Vessel diameters detected ranged from 1.14 to 42.18 µm.

3.2 3D reconstructions of the microvasculature

3.2.1 Surface reconstructions

From the binary mask of the filled microvascular volumes (e.g., Figure 2), a surface mesh of the microvasculature was prepared using isosurface reconstruction via marching cubes [Lorensen and Cline, 1987]. The 3D surface reconstruction of the murine retinal flat-mount is shown in Figure 4.

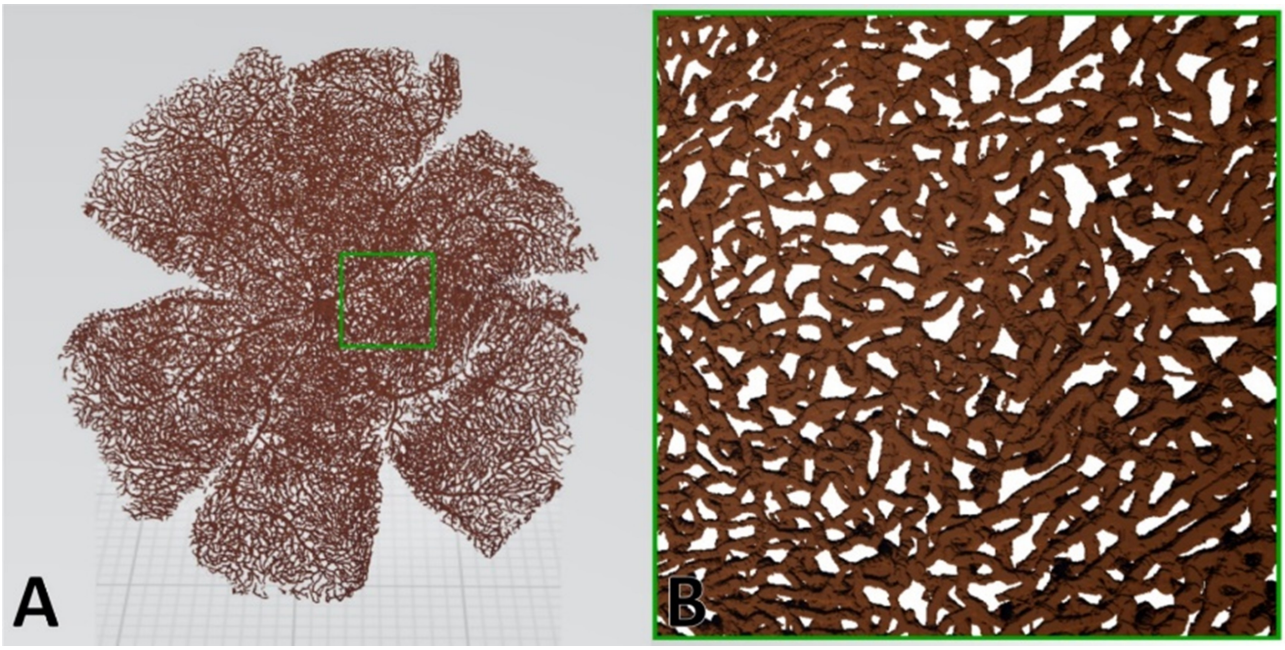


Figure 4: 3D surface mesh of the entire murine retinal flat-mount. A: Isosurface reconstruction of the full binarised imaging dataset allows the representation of microvascular surfaces. B: A zoomed-in region reveals intricate interwoven retinal microvascular layers present in 3D.

3.2.2 Geometric reconstructions

As geometric reconstructions are a computationally expensive task, the application of the algorithm was restricted to a quarter of the retinal flat-mount. To geometrically encode the retinal microvasculature as a series of interconnected cylinders, spherical objects were first detected by marching a series of voxelised spheres in

 ISBN 978-0-9934207-7-1

descending size about the binary mask of voxels detected in the microvasculature, accepting a candidate sphere as present should ≥90% of its volume intersect the mask. Regions within the mask occupied by successfully placed spheres were then excluded from the search space until no further potential candidates could be found. The smallest spheres detected (R = 1 px. ≈ 1.14 µm) were omitted from further processing as smaller objects were more likely to represent false-positive detections due to remaining noise as determined by expert inspection. The result of this approach is reflected in Figure 5.

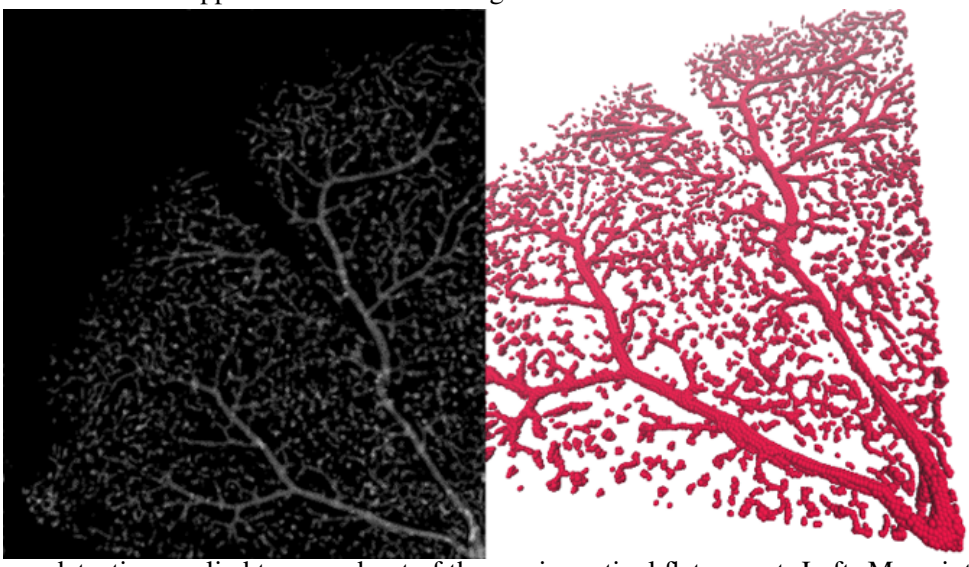


Figure 5: Sphere detection applied to a quadrant of the murine retinal flat-mount. Left: Mean intensity projection denoting spherical detection; Right: The corresponding sphere detection as visualised in 3D.

Next, all possible cylinders were assigned by marching [Amanatides and Woo, 1987] between the centres of all pairs of spheres, tagging a candidate cylinder should a straight 3D line be contained entirely within the detected microvascular volume binary voxel mask. The radius of each cylinder was set to the mean of the two sphere radii. The result of this approach is shown in Figure 6.

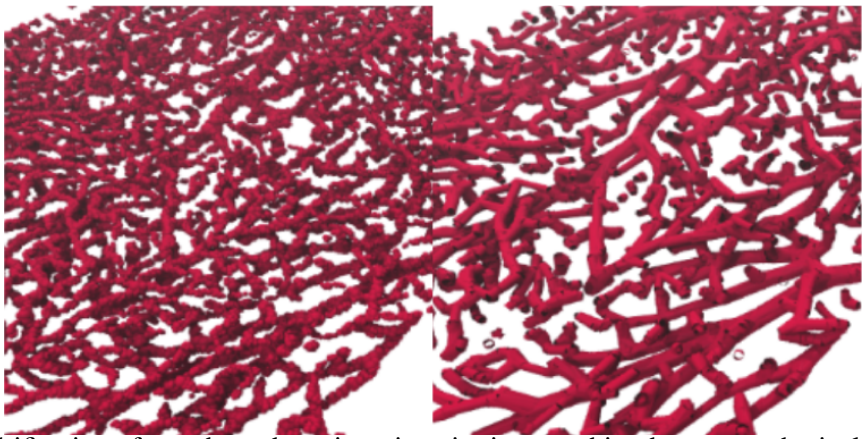


Figure 6: Cylindrification after sphere detection via pairwise marching between spherical nodes constrained within the binary voxel mask. Left: A region of spheres as detected in Figure 5. Right: The resulting cylindrical network following pairwise marching between spherical nodes constrained to the binary voxel mask.

As cylinders are over-detected from the pairwise node marching approach, firstly, trivial cylinders that provide no additional information (i.e., those which are fully contained within others) are removed entirely using rapid point in cylinder detection calculations [Barbier and Galin, 2004]. Next, remaining pairs of cylinders that overlap partially [Ketchel and Larochelle, 2005] are each divided into three cylinders (i.e., pre-overlap, overlap, and post-overlap, relative to each pair) until no colliding cylinders remain apart from end-to-end connections.

 ISBN 978-0-9934207-7-1

Branch points must then be resolved through splitting cylinders at the point of bifurcation according to the centreline detection derived from the distance transform, maintaining appropriate connectivity. Lastly, relatively collinear cylinders that touch end-to-end may be consolidated into lengthier cylinders by evaluating the scalar product of the cylindrical axes to further simplify the vascular network. The memory to store the 3D geometric reconstruction of the retinal microvascular network quadrant as described, which was initially composed of ~209 million 1 byte voxels of input imaging data, is now represented by N = 98,389 cylinders, which are each represented by 7 32-bit floating point numbers in the form of two endpoint x-, y-, and z- coordinates with radius or diameter. This corresponds to a reduction in computer memory utilisation by a factor of ~100×.

4 Discussion

The primary purpose for undertaking this work was to retrieve microvasculatures from experimental imagery as input for particle-based simulations of molecular transport [Troendle et al., 2018]. The work presented in this paper provides the necessary platform to perform such simulations within real microvascular networks in the near future. Computer simulations of therapeutic compounds moving within the retinal microvasculature are likely to provide new insights into molecular transport within the retina and avenues for developing improved drug delivery approaches. Additionally, retinal microvascular surface meshes (see Figure 4) may be refined [Garland and Heckbert, 1997] and used for computational fluid dynamics studies [Lu et al., 2016] or as a basis for the development of mimetic *in vitro* microfluidic platforms as derived from retinal imagery.

This study provides a new approach for extracting and cylindrifying microvascular networks from biomedical imagery. Kelch et al. analysed the 3D microvascular topography of a murine lymph node by combining confocal microscopy with image processing [Kelch et al., 2015]. Similar to this study, they developed an alternative algorithm to fill incomplete microvascular volumes. Their goal was to obtain detailed topographic spatial analyses of their microvasculature in 3D, but they did not test the utility of their algorithms for geometric reconstructions. Another protocol for the 3D reconstruction of microvascular networks was recently released by Wächli et al., which utilises corrosion casting, scanning electron microscopy, synchrotron radiation and desktop microcomputed tomography (μCT) imaging, and computational network analysis [Wälchli et al., 2021]. While this work could also be adapted to output 3D cylindrical network models, it does not in its presently published form. Nonetheless, a prevailing limitation within this application area is that the specific image processing parameters and steps needed to perform microvascular geometric reconstructions are dependent upon the features of the biological sample, its preparation for imaging, and the image acquisition modality itself.

The algorithms used in this study for microvascular refinement and geometric detection may facilitate the development of machine learning algorithms to improve predictions of disease risk. For example, the algorithm by Cheung et al. can detect increased probability of cardiac arrests via assessing retinal vessel diameters [Cheung et al., 2021]. Topographic information extracted from 3D cylindrical network models may be used to enhance their predictions. Further, supervised machine learning algorithms can be developed to accelerate the recovery of 3D cylindrical objects from vascular imagery by employing the methods used here to generate labelled target data for training machine learning models.

As medical imaging technology further advances, it may be possible to apply these methods to high resolution 3D *in vivo* microvascular datasets. With the recent development of adaptive optics ocular coherence tomography angiography (AO-OCTA) [Camino et al., 2020], the reconstruction of the human retinal microvasculature may soon be possible *in vivo*.

5 Conclusions

This manuscript describes semi-automated image processing algorithms to geometrically reconstruct the retinal microvasculature in 3D as represented by a network of interconnected cylinders. This work can be further extended by improving computational efficiency of the reconstruction protocols and by assessing the accuracy of this method systematically across a wider array of retinal samples.

 ISBN 978-0-9934207-7-1

Acknowledgments

We would like to thank Northern Ireland Health and Social Care R&D Division (STL/4748/13), and the Medical Research Council (MC_PC_15026) for supporting this research.

Author contributions

EPT & TMC conceived the study. PB conducted the animal research, prepared the retina, and acquired the input imagery. EPT designed and performed the image processing and 3D geometric reconstruction algorithms. EPT drafted the manuscript and produced the graphical figures. EPT, PB, & TMC critically evaluated the manuscript.

References

[Amanatides and Woo, 1987] Amanatides, J. and Woo, A. (1987). A fast voxel traversal algorithm for ray tracing. In *EG1987 Proceedings (Technical Papers)*.

[Barbier and Galin, 2004] Barbier, A. and Galin, E. (2004). Fast distance computation between a point and cylinders, cones, line-swept spheres and cone-spheres. *Journal of Graphics Tools*, 9:11–19.

[Bradski, 2000] Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

[Camino et al., 2020] Camino, A., Zang, P., Athwal, A., Ni, S., Jia, Y., Huang, D., and Jian, Y. (2020). Sensorless adaptive-optics optical coherence tomographic angiography. *Biomed Opt Express*, 11(7):3952–3967.

[Chang, 2013] Chang, B. (2013). Mouse models for studies of retinal degeneration and diseases. *Methods in molecular biology (Clifton, N.J.)*, 935:27–39.

[Cheung et al., 2021] Cheung, C. Y., Xu, D., Cheng, C.-Y., Sabanayagam, C., Tham, Y.-C., Yu, M., Rim, T. H., Chai, C. Y., Gopinath, B., Mitchell, P., Poulton, R., Moffitt, T. E., Caspi, A., Yam, J. C., Tham, C. C., Jonas, J. B., Wang, Y. X., Song, S. J., Burrell, L. M., Farouque, O., Li, L. J., Tan, G., Ting, D. S. W., Hsu, W., Lee, M. L., and Wong, T. Y. (2021). A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. *Nature Biomedical Engineering*, 5(6):498–508.

[Curtis et al., 2009] Curtis, T. M., Gardiner, T. A., and Stitt, A. W. (2009). Microvascular lesions of diabetic retinopathy: clues towards understanding pathogenesis? *Eye*, 23(7):1496–1508.

[Felzenszwalb and Huttenlocher, 2012] Felzenszwalb, P. F. and Huttenlocher, D. P. (2012). Distance transforms of sampled functions. *Theory of Computing*, 8(1):415–428.

[Garland and Heckbert, 1997] Garland, M. and Heckbert, P. S. (1997). Surface simplification using quadric error metrics. In *SIGGRAPH 1997*, page 209–216. Association for Computing Machinery, Inc.

[Huang et al., 1979] Huang, T., Yang, G., and Tang, G. (1979). A fast two-dimensional median filtering algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(1):13–18.

[Johnson, 2013] Johnson, M. (2013). Detergents: Triton x-100, tween-20, and more. *Materials and Methods*, 3:163.

[Kelch et al., 2015] Kelch, I. D., Bogle, G., Sands, G. B., Phillips, A. R. J., LeGrice, I. J., and Rod Dunbar, P. (2015). Organ-wide 3d-imaging and topological analysis of the continuous microvascular network in a murine lymph node. *Scientific Reports*, 5(1):16534.

[Ketchel and Larochelle, 2005] Ketchel, J. S. and Larochelle, P. M. (2005). Collision detection of cylindrical rigid bodies using line geometry. In *Proceedings of IDETC/CIE 2005*, pages 811–825. ASME.

[Kirkeby and Moe, 2001] Kirkeby, S. and Moe, D. (2001). Binding of Griffonia simplicifolia 1 isolectin B4 (GS1 B4) to α-galactose antigens. *Immunology & Cell Biology*, 79(2):121–127.

[Lindeberg, 2015] Lindeberg, T. (2015). Image matching using generalized scale-space interest points. *Journal of Mathematical Imaging and Vision*, 52(1):3–36.

[Lorensen and Cline, 1987] Lorensen, W. E. and Cline, H. E. (1987). Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1987*, pages 163–169. Association for Computing Machinery, Inc.

[Lu et al., 2016] Lu, Y., Bernabeu, M. O., Lammer, J., Cai, C. C., Jones, M. L., Franco, C. A., Aiello, L. P., and Sun, J. K. (2016). Computational fluid dynamics assisted characterization of parafoveal hemodynamics in normal and diabetic eyes using adaptive optics scanning laser ophthalmoscopy. *Biomedical Optics Express*, 7(12):4958–4973.

[Marin et al., 2015] Marin, D., Zhong, Y., Drangova, M., and Boykov, Y. (2015). Thin structure estimation with curvature regularization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 397–405.

[Newman and Sproull, 1979] Newman, W. M. and Sproull, R. F. (1979). *Interactive Raster Graphics: Filling Areas*, book section 17, page 253. McGraw-Hill, Inc., 2nd edition.

[Prahst et al., 2020] Prahst, C., Ashrafzadeh, P., Mead, T., Figueiredo, A., Chang, K., Richardson, D., Venkatraman, L., Richards, M., Russo, A. M., Harrington, K., Ouarne, M., Pena, A., Chen, D. F., Claesson-Welsh, L., Cho, K. S., Franco, C., and Bentley, K. (2020). Mouse retinal cell behaviour in space and time using light sheet fluorescence microscopy. *eLife*, 9.

[Roorda et al., 2006] Roorda, A., Garcia, C. A., Martin, J. A., Poonja, S., Queener, H., Romero-Borja, F., Sepulveda, R., Venkateswaran, K., and Zhang, Y. (2006). What can adaptive optics do for a scanning laser ophthalmoscope? *Bulletin of the Belgian Society of Ophthalmology*, (302):231–44.

[Secomb, 2008] Secomb, T. W. (2008). Theoretical models for regulation of blood flow. *Microcirculation*, 15(8):765–775.

[Solomon et al., 2019] Solomon, S. D., Lindsley, K., Vedula, S. S., Krzystolik, M. G., and Hawkins, B. S. (2019). Anti-vascular endothelial growth factor for neovascular age-related macular degeneration. *Cochrane Database of Systematic Reviews*, (8).

[Troendle et al., 2018] Troendle, E. P., Khan, A., Searson, P. C., and Ulmschneider, M. B. (2018). Predicting drug delivery efficiency into tumor tissues through molecular simulation of transport in complex vascular networks. *Journal of Controlled Release*, 292:221–234.

[Virgili et al., 2018] Virgili, G., Parravano, M., Menchini, F., and Evans, J. R. (2018). Anti-vascular endothelial growth factor for diabetic macular oedema. *Cochrane Database of Systematic Reviews*, (10).

[Vučković et al., 2019] Vučković, V., Arizanović, B., and Le Blond, S. (2019). Generalized n-way iterative scanline fill algorithm for real-time applications. *Journal of Real-Time Image Processing*, 16(6):2213–2231.

[Wälchli et al., 2021] Wälchli, T., Bisschop, J., Miettinen, A., Ulmann-Schuler, A., Hintermüller, C., Meyer, E. P., Krucker, T., Wälchli, R., Monnier, P. P., Carmeliet, P., Vogel, J., and Stampanoni, M. (2021). Hierarchical imaging and computational analysis of three-dimensional vascular network architecture in the entire postnatal and adult mouse brain. *Nature Protocols*, 16(10):4564–4610.

[Yu et al., 2015] Yu, P. K., Balaratnasingam, C., Xu, J., Morgan, W. H., Mammo, Z., Han, S., Mackenzie, P., Merkur, A., Kirker, A., Albiani, D., Sarunic, M. V., and Yu, D.-Y. (2015). Label-free density measurements of radial peripapillary capillaries in the human retina. *PLOS ONE*, 10(8):e0135151.

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/EVLM5779

Deep Multi-Task Networks For Occluded Pedestrian Pose Estimation

Arindam Das[1,2], Sudip Das[3], Ganesh Sistu[4], Jonathan Horgan[4],
Ujjwal Bhattacharya[3], Edward Jones[2], Martin Glavin[2], and Ciarán Eising[2,5]

[1]*Detection Vision Systems, Valeo India,* [2]*National University of Ireland, Galway,*
[3]*Indian Statistical Institute, Kolkata, India,* [4]*Valeo Vision Systems, Ireland,*
[5]*University of Limerick, Ireland*

Abstract

Most of the existing works on pedestrian pose estimation do not consider estimating the pose of an occluded pedestrian, as the annotations of the occluded parts are not available in relevant automotive datasets. For example, CityPersons, a well-known dataset for pedestrian detection in automotive scenes does not provide pose annotations, whereas MS-COCO, a non-automotive dataset, contains human pose estimation. In this work, we propose a multi-task framework to extract pedestrian features through detection and instance segmentation tasks performed separately on these two distributions. Thereafter, an encoder learns pose specific features using an unsupervised instance-level domain adaptation method for the pedestrian instances from both distributions. The proposed framework has improved state-of-the-art performances of pose estimation, pedestrian detection, and instance segmentation.

Keywords: Pedestrian Pose Estimation, Unsupervised Domain Adaptation, Multi-task Learning

1 Introduction

Human Pose Estimation (HPE) is a widely deployed computer vision application [Chen et al., 2020], along with semantic segmentation [Das. et al., 2019], and object tracking and trajectory prediction [Sridevi et al., 2021]. Addressing HPE for occluded pedestrians was first reported in ClueNet [Kishore et al., 2019] and later *person-to-person* occlusion was investigated in [Das et al., 2020]. In this paper, we propose a novel *end-to-end two stage fully convolutional network* for the purpose of estimating pose of occluded pedestrians, in the context of a Multi-Task Learning (MTL) architecture where object detection and segmentation are performed. The proposed framework can 1) learn pedestrian pose estimation (PPE) from a dataset that does not contain pose annotations and 2) accurately estimate the pose of the occluded parts of the human body without the annotations of the same occluded parts. To address the issue of the non-availability of pose annotations during training in the CityPersons [Zhang et al., 2017] dataset, we consider a related dataset (MS-COCO [Lin et al., 2014]) to get adequate supervision specific to PPE during training. To achieve this, we apply unsupervised instance level domain adaptation on each pedestrian into the target domain. Here, the dataset with human pose annotations acts as the source domain, whereas the samples with only pedestrian detection and segmentation annotations become the target domain in domain adaptation. The main contributions of this study are: 1) the proposal of a two-stage end-to-end fully convolutional network to perform occluded PPE, 2) preserving the information of detection and segmentation in an MTL architecture, and 3) in achieving state-of-the-art performance on pedestrian detection, instance segmentation and PPE tasks respectively.

2 Methodology

2.1 Distribution Specific Multi-task Learning

Two distinct multi-task learning (MTL) networks are setup for two different data distributions - CityPersons and MS COCO. Each MTL network aims to extract domain specific features of human instances to perform human

 ISBN 978-0-9934207-7-1

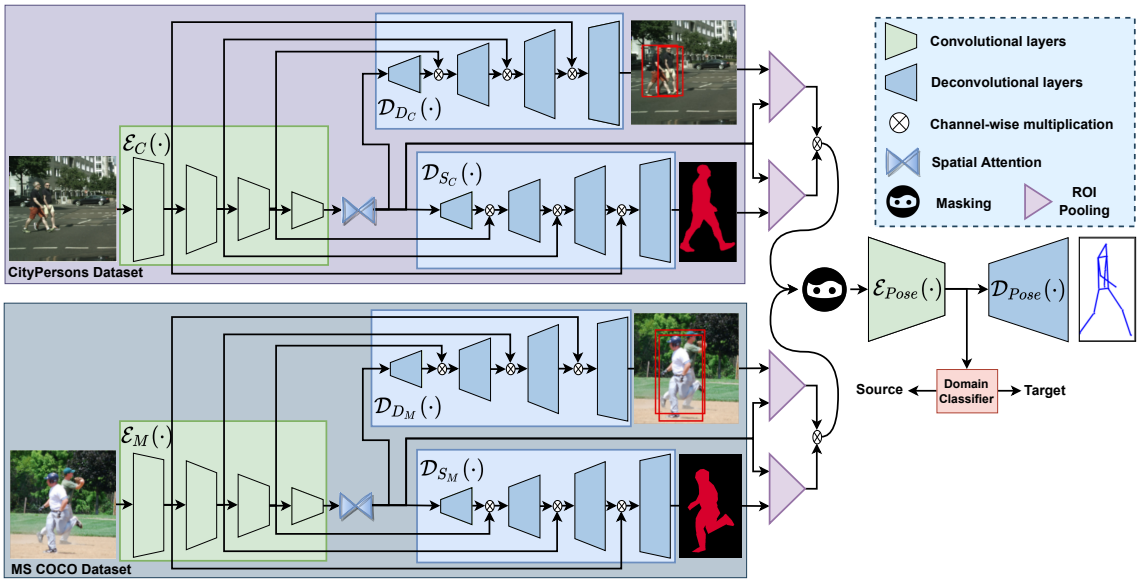


Figure 1: Our proposed MTL architecture for end-to-end PPE.

detection and instance segmentation respectively. Our proposed framework is illustrated in Figure 1. Feature Pyramid Network (FPN) style encoder is used in each distribution specific MTL instance. We denote these as $\mathscr{E}_C(\cdot)$ and $\mathscr{E}_M(\cdot)$ respectively. We use a set of deconvolution layers to perform instance segmentation ($\mathscr{D}_{S_C}(\cdot)$ and $\mathscr{D}_{S_M}(\cdot)$) and detection ($\mathscr{D}_{D_C}(\cdot)$ and $\mathscr{D}_{D_M}(\cdot)$) in two instances of MTL. Two independent MTLs are necessary here as the distribution of the datasets considered in this work is not close. One of the main advantages of having distribution specific MTL is to exploit these two domains independently to generate strong distribution specific features. We use spatial attention in between encoder and task specific decoders. This basically applies global average pooling on feature maps to generate vector and that is multiplied with the same feature map to generate most attentive area from the features. This technique helps to preserve the spatial consistency in feature space.

2.2 Distribution Invariant PPE

As the pedestrian dataset does not provide pose annotations, we apply instance level domain adaptation from another dataset that contains the details of human pose. Pedestrian instances from the detection and segmentation decoder are projected back to the last layer of encoded feature maps using ROI pooling, that are masked and fed as input to the encoder ($\mathscr{E}_{Pose}(\cdot)$) for PPE. The motivation to apply *instance-level* domain adaption is to minimize the domain shift between two different distributions of the same category. In this unsupervised domain adaptation setup, when a sample from source domain passes through the pose encoder, then the weights are updated in $\mathscr{E}_{Pose}(\cdot)$, domain classifier ($\mathscr{D}_C(\cdot)$) and $\mathscr{D}_{Pose}(\cdot)$ but $\mathscr{D}_{Pose}(\cdot)$ is not updated for the input from target domain. This means the features extracted by $\mathscr{E}_{Pose}(\cdot)$ are fed to $\mathscr{D}_C(\cdot)$ to determine the actual source distribution (i.e., CityPersons or MS COCO). In this adversarial learning setup, $\mathscr{D}_C(\cdot)$ promoted the accurate classification of the input domain, while $\mathscr{E}_{Pose}(\cdot)$ encourages the generation of better domain invariant features specific to human instances.

To train the proposed framework in end-to-end fashion, we express overall MTL loss function, $\mathscr{L}_{total}$ as the simple weighted combination of other losses from the detectors ($\mathscr{L}_{det_c}$, $\mathscr{L}_{det_m}$), instance segmentation ($\mathscr{L}_{seg_c}$, $\mathscr{L}_{seg_m}$), domain classification ($\mathscr{L}_{dc}$) and PPE ($\mathscr{L}_{pe}$) as,

$$\mathscr{L}_{total} = \mathscr{L}_{det_c} + \mathscr{L}_{det_m} + \alpha\mathscr{L}_{seg_c} + \alpha\mathscr{L}_{seg_m} + \beta\mathscr{L}_{dc} + \gamma\mathscr{L}_{pe} \tag{1}$$

where $\mathscr{L}_{det_c}$, $\mathscr{L}_{seg_c}$ are the losses of detection and segmentation tasks trained on CityPersons dataset. Likewise, losses $\mathscr{L}_{det_m}$ and $\mathscr{L}_{seg_m}$ are obtained from an MTL trained on MS COCO. α, β and γ are the weights corresponding to instance segmentation, domain classifier and PPE losses respectively. For the instance segmentation task, we use binary cross-entropy loss. For object detection task, we adopt the loss from [Dasgupta et al., 2022]. We follow the loss function for PPE and $\mathscr{L}_{dc}$ as described in [Kishore et al., 2019] and [Kocabas et al., 2018].

 ISBN 978-0-9934207-7-1

Model	AP	AP_{50}	AP_{75}	AP_M	AP_L
MultiPoseNet [Kocabas et al., 2018]	69.6	86.3	76.6	65.0	76.3
CFN [Huang et al., 2017]	72.6	86.7	69.7	78.3	79.0
ClueNet[Kishore et al., 2019]	73.9	89.6	**78.2**	70.9	79.1
[Das et al., 2020]	74.2	89.9	74.9	79.3	76.6
Ours	**75.7**	**90.3**	76.3	**80.7**	**79.5**

Table 1: Comparison using AP for PPE task on MS COCO dataset.

Model	Occlusion %					
	20%	**30%**	**40%**	**50%**	**60%**	**70%**
ClueNet [Kishore et al., 2019]	88.06	83.93	79.8	73.4	64.0	58.8
[Das et al., 2020]	90.3	84.31	81.2	74.06	64.9	59.1
Ours	**92.0**	**85.9**	**82.4**	**75.3**	**65.7**	**59.3**

Table 2: Comparative study of PPE with SOTA models having different occlusion percentages on MS-COCO dataset.

Model	R	HO	R+HO
Faster RCNN	15.52	64.83	41.45
Tao et al. [Song et al., 2018]	14.4	52.0	34.24
OR-CNN [Zhang et al., 2018]	**11.0**	51.0	36.11
ClueNet [Kishore et al., 2019]	11.87	47.68	30.84
[Das et al., 2020]	13.29	46.07	29.13
Ours	12.01	**44.7**	**27.8**

Table 3: MR based comparison of SOTA models for pedestrian detection on CityPersons.

Model	Training Data	Person	Rider
Mask-RCNN	CityPersons + COCO	34.8	27.0
PANet [Liu et al., 2018]	CityPersons + COCO	41.5	33.6
[Das et al., 2020]	CityPersons + COCO	42.1	33.9
Ours	CityPersons + COCO	**42.7**	**34.7**

Table 4: Comparison using IoU for instance segmentation on CityPersons.

3 Experimentation Details

3.1 Datasets and Implementation Details

The proposed method is trained and evaluated on the two publicly available datasets already mentioned - CityPersons and MS COCO. Annotations of detection and instance segmentation specific to human instances are used from both datasets, along with pose information from MS COCO. Out of 17 available key points, we use 13 (as discussed in [Kishore et al., 2019]). Two MTL networks specific to detection and instance segmentation tasks were pre-trained on CityPersons and MS COCO before initiating the second stage of training for PPE. Curriculum Learning for Mask and Predict strategy was used to gradually increase the masking percentage as the training progresses. Momentum Optimizer with 0.9 and an initial learning rate of 0.01 was used. After each $15k$ iterations, the learning rate is reduced by a factor of 10. The weights α, β and γ are set to 0.5, 1, 1 respectively. Data augmentation methods, such as applying random flipping, blurring, brightness, etc., are added to make the proposed framework more robust. Training was completed on two Nvidia Tesla P6 GPUs with batch size set to 1.

3.2 Results

Table 1 shows the performance of PPE of the proposed model and compares with recently published methods. Since the annotations of the occluded parts of the pedestrians are not available, we created such occluded pedestrians by masking (with different percentages) random parts of fully visible pedestrians [Das et al., 2020] and compared with existing techniques as presented in Table 2. The proposed approach has clearly improved the existing benchmark in estimating pose for the occluded pedestrians. Comparison study using Miss Rate (MR) on the validation set of CityPersons for pedestrian detection task with a few existing methods is shown in Table 3 where a few specific scenarios such as Reasonable (R) with [.65, inf] visibility, Heavy Occlusion (HO) with [.20, .65] visibility and Reasonable+Heavy occlusion (R + HO) with [.20, inf] visibility are considered. Fig. 2 shows the qualitative results on CityPersons and MS COCO dataset for PPE of the proposed model. Table 3.2 provides the results of instance segmentation of the proposed framework that slightly improves the SOTA performance for both categories - person and rider. As part of ablation study, different standard backbone encoders are tested using average precision metric for PPE and among these ResNeXt-101 outperformed other encoders as presented in Table 3.2.

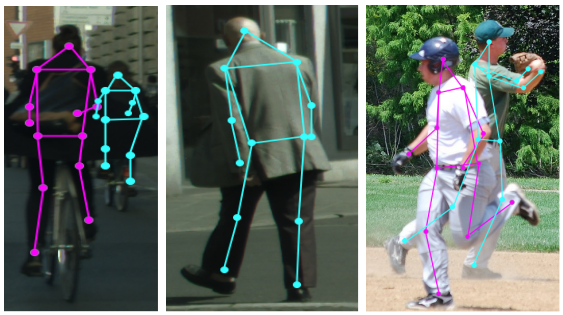


Figure 2: Sample outputs of the pose estimation.

 ISBN 978-0-9934207-7-1

Backbone	AP	AP_{50}	AP_{75}	AP_M	AP_L
ResNet-50	69.7	86.3	71.94	64.2	71.1
ResNet-101	71.4	87.8	72.1	76.1	74.3
ResNeXt-101	**75.7**	**90.3**	**76.3**	**80.7**	**79.5**

Table 5: Ablation study on different backbone encoders for PPE on test data of MS-COCO.

4 Conclusion

In this work, an end-to-end two stage network is developed that is trained in an unsupervised manner to accurately estimate the pose of pedestrians regardless the level of occlusion. We apply unsupervised domain adaptation at instance level to reduce the distribution gap between two set of features obtained from two distinct MTL setup. Experimental results demonstrate the robustness of the proposed strategy and provide strong confirmation as it improved respective state-of-the-art results on PPE, pedestrian detection, and instance segmentation.

References

[Chen et al., 2020] Chen, Y., Tian, Y., and He, M. (2020). Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897.

[Das. et al., 2019] Das., A., Kandan., S., Yogamani., S., and Křížek., P. (2019). Design of real-time semantic segmentation decoder for automated driving. In *VISAPP*.

[Das et al., 2020] Das, S., Kishore, P. S. R., and Bhattacharya, U. (2020). An end-to-end framework for pose estimation of occluded pedestrians. In *2020 IEEE International Conference on Image Processing (ICIP)*.

[Dasgupta et al., 2022] Dasgupta, K., Das, A., Das, S., Bhattacharya, U., and Yogamani, S. (2022). Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*.

[Huang et al., 2017] Huang, S., Gong, M., and Tao, D. (2017). A coarse-fine network for keypoint localization. In *ICCV*.

[Kishore et al., 2019] Kishore, P. S. R., Das, S., Mukherjee, P. S., and Bhattacharya, U. (2019). Cluenet: A deep framework for occluded pedestrian pose estimation. In *BMVC*.

[Kocabas et al., 2018] Kocabas, M., Karagoz, S., and Akbas, E. (2018). Multiposenet: Fast multi-person pose estimation using pose residual network. In *ECCV*.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *ECCV*.

[Liu et al., 2018] Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). Path aggregation network for instance segmentation. In *CVPR*.

[Song et al., 2018] Song, T., Sun, L., Xie, D., Sun, H., and Pu, S. (2018). Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *ECCV*.

[Sridevi et al., 2021] Sridevi, M., Sugirtha, T., Rashed, H., Kiran, B. R., and Yogamani, S. (2021). Object detection, tracking and trajectory prediction for autonomous driving. *Autonomous Driving and Advanced Driver-Assistance Systems (ADAS): Applications, Development, Legal Issues, and Testing*, page 105.

[Zhang et al., 2017] Zhang, S., Benenson, R., and Schiele, B. (2017). Citypersons: A diverse dataset for pedestrian detection. In *CVPR*.

[Zhang et al., 2018] Zhang, S., Wen, L., Bian, X., Lei, Z., and Li, S. Z. (2018). Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *ECCV*.

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/POYA9239

Reality Analagous Synthetic Dataset Generation with Daylight Variance for Deep Learning Classification

Thomas Lee, Susan Mckeever and Jane Courtney

TU Dublin, School of Electrical & Electronic Engineering

Abstract

For the implementation of Autonomously navigating Unmanned Air Vehicles (UAV) in the real world, it must be shown that safe navigation is possible in all real world scenarios. In the case of UAVs powered by Deep Learning algorithms, this is a difficult task to achieve, as the weak point of any trained network is the reduction in predictive capacity when presented with unfamiliar input data. It is possible to train for more use cases, however more data is required for this, requiring time and manpower to acquire. In this work, a potential solution to the manpower issues of exponentially scaling dataset size and complexity is presented, through the generation of artificial image datasets that are based off of a 3D scanned recreation of a physical space and populated with 3D scanned objects of a specific class. This simulation is then used to generate image samples that iterates temporally resulting in a slice-able dataset that contains time varied components of the same class.

Keywords: Simulation, Drones, Deep Learning, Autonomous Aerial Vehicles

1 Introduction

Figure 1: 'FIDIM' Scene view (wireframe enabled)

Over the years, there has been a notable increase in the activity within the UAV autonomy research space [Lee et al., 2021]. Many projects have a specific focus on neural network based solutions [Rojas-Perez and Martinez-Carranza, 2020]. These projects aim to infer a safe solution from a learned model to autonomous UAV tasks rather than engineer one. These tasks include image classification, object detection, and image segmentation in order for the drone to detect and identify obstacles, safe movement vectors, or unsafe situations. These tasks often require access to training datasets that are of considerable size and variation to achieve good generalisation [Deng et al., 2009], but also have specific elements through which quality training for that task can be assured [Udacity, 2016]. The challenge of training for autonomous UAV navigation is a matter of data availability, due to the high degree of variability in even common environments such as urban, suburban, forested, or rural areas it is not feasible to manually gather a dataset on all variations. Furthermore, such environments are defined arbitrarily and do not have binary transitions, but gradually shift from one environment to the other. This issue is compounded when considering other environmental factors such as light level or weather variation. It is possible to generate artificial data to make up for a lack of samples and doing so is considered to prevent overfitting [Shorten and Khoshgoftaar, 2019]. This is typically done through the augmentation of an existing, manually collected dataset [Takahashi et al., 2020]. However, these solutions are not robust, and vary

 ISBN 978-0-9934207-7-1

from simple image processing transforms with some examples being flipping, cropping, and noise injection on existing data within the set to the synthesis of entirely samples based on the output of trained adversarial networks. in both cases, images are modified but new image subjects are not created. Augmentation can generate samples but cannot generate new sample content including information that does not already exist within the original set, simulated data with new sample content is used in reinforcement learning to create unknown situations which can assist in generalisation, however this process is time consuming and costly, requiring a simulation to be run over the course of the model's training. The information in Section 2 introduces the background concepts and previous work which lead to the creation of the simulator, while the method outlined in Section 3 details the construction of the simulated dataset. A goal of this project is to generate not just photo realistic image samples with unique content, but the same image content under unique environmental contexts, in the case of this project a cyclical time axis is introduced to mimic a Day-Night cycle which is then analysed using a pretrained YOLOv3 network in Section 4 [Redmon and Farhadi, 2018].

1.1 Autonomous Navigation

Autonomous navigation tasks tend to be complex, incorporating elements of physics calculation, state estimation [Al-Sharman et al., 2020], perception [Yang and Wang, 2020], and decision making in various degrees. Many projects aim to achieve these tasks through a monocular camera and the utilization of a Convolutional Neural Network to allow for the prediction of task-related data from input images. However, Neural Networks have accuracy issues regarding operation in environments which the autonomous navigator is not trained for [Loquercio et al., 2018] which may not cripple autonomous UAV projects mechanically but could legislatively, given that the EASA Special Condition guidelines [European Union Aviation Safety Agency, 2020] state that: "Certification of light UAS with highly integrated systems will be fundamentally based on a safety assessment that includes thrust/lift/power systems and also interaction with structures". These accuracy issues are considered to be the trade-off of performance in a Deep Learned solution based on how specific the task is [Alshehri et al., 2019]. Consider a specific task such as the identification of a specific brand of item in an image. Generalisation of that same task, such as identifying objects of a similar type (ie: other brands of that item) is likely to result in a reduction of the model's capability to identify the original brand, if even by a small amount. The reduction of task-specific performance in generalised models is exacerbated with perception based problems such as collision avoidance in UAVs [Loquercio et al., 2020]. There are many variables that bring additional complexity to the task of interpreting the image beyond what the subject is, for example: When and where the image was taken are two dimensions with massive impact on the resulting image, and the perception of said image makes up the majority of modern Deep Learned autonomous navigation solutions.

2 Background

Since UAV research in the physical space is costly (both in time and budget), projects often test the UAV design in simulation during the initial stages of project development, only moving to physical experimentation for final validation stages. A review of recent literature [Loquercio et al., 2020] identified two common simulation interfaces which are used throughout the UAV research space. Gazebo is an open source robotics simulator which can seamlessly interface with ROS, a software framework for the development and operation of robotics. An alternative to Gazebo is AirSim, an open source robotics simulator developed by Microsoft. Although a simulator like Gazebo supports high fidelity rendering the main focus of the simulator is not visual quality but realistic device control through ROS and the accurate simulation of physical forces. This can lead to discrepancies between the simulation results and physical experiment results in image based tasks like Computer Vision and more modern Deep Learning RCNN tasks. Given that these networks are trained on image inputs, it is proposed that the visual quality is more valuable than the physical accuracy in the context of image data generation. By using a modern graphical rendering engine designed for real time interaction, it is possible to develop simulations that are more analogous to the physical world in visual aspects [Wang et al., 2021]. A 2019 survey on data augmentation [Shorten and Khoshgoftaar, 2019] presents a taxonomy on several methods of data augmentation

for machine learning applications, these methods include color space manipulation, noise injection, translations and similar image processing tasks for the use of expanding a limited dataset. While these approaches are helpful, even more complex solutions are noted to simply prevent overfitting [Takahashi et al., 2020], they do not increase the amount of information actually contained within the dataset.

2.1 Trained Autonomous Features

Previous work sought to categorise the functions in DNN based autonomous UAV projects [Lee et al., 2021], and organize them into a usable taxonomy. Based on the results of that literature review, the most common function projects sought to train was that of autonomous collision avoidance. Typically through CNN based, ResNet optimised strategies [Palossi et al., 2018]. It was determined that [Loquercio et al., 2018] was the most notable and uses steering angle and crash probability captured from ground based vehicles to train an autonomous collision avoidance policy in a quad-rotor helicopter UAV. Additionally, many projects intend to utilise transfer learning of generic image classification models such as the "ImageNet" dataset [Deng et al., 2009] for the function of Object Distinction, such as locating civilians in a collapsed building [Hartawan et al., 2019], or anomalies in natural environments [Yong and Yeong, 2018]. While not as popular in the UAV space, Classification as a task is an immensely popular application of Deep Learning in general, and in certain contexts Classification can serve as the basis for a Collision Avoidance solution. One of the key conclusions drawn from [Lee et al., 2021] was noting a distinct lack or papers that approached classifying, distinguishing. or generalising multiple different environments in a neural network. Environmental Distinction is considered by the author to be a key feature in the development path of the next generation of autonomous navigation.

2.2 Environment Specificity

One of the fundamental issues with using a neural network to solve a general task is the restriction placed on the network by its predefined complexity. When a network structure is defined in code it is also bounded to that structure which has a maximum mathematical complexity with which it can reach. Because of these limits, it is logical to assume that, given that the complexity of a task scales exponentially with the number of dimensions being considered on input, trying to generalize a complex task (for many use cases) using a bounded network can result in a decrease of specific (single use case) performance due to reaching that limit of the network's complexity as it is defined. This trade-off can be mitigated in design, through careful architecture decisions in the network or by tailoring the data being fed into the network into a more general context (converting an RGB image into grey scale before using it for training is an example of this). This is not usually an issue for rudimentary tasks, but when considering different environments in a CNN this becomes a much more distinct problem. A model's predictive capability is primarily dependant on the data being fed into it, specific datasets are often gathered or annotated manually by experts for use in the training of that task [Deng et al., 2009], this process is typically done in sessions as befits the researcher's schedule and as a result of this, uncontrollable environment parameters can change from session to session. This naturally biases the information to these uncontrollable parameters such as time, location, and weather and can lead to a less accurate, state-based representation of a dynamic area. While it is theoretically possible to sample the entire spectrum of an area's change and create a dataset containing all reasonable permutations of environmental parameters, it is unfeasible to do so manually, given that the addition of just one dimension leads to an exponential gain in the amount of data to be collected.

2.3 Simulating Data

Though traditional thinking would have it that in order to model decisions for use in the physical world, the dataset used for training must contain physical world data. However, artificial data has already been used as a means of base network training for adjustment later using real data [Vierling et al., 2020], and simulation as a means of generating datasets for learning tasks has been used for tasks such as Image Segmentation [Richter et al., 2016]. [Perri et al., 2022] presents a novel benefit and rationale for the use of simulation for data

Figure 2: Various bike assets used for dataset generation

Figure 3: Single position/rotation sliced by time axis in synthetic dataset

generation, the main concept being that using a simulation allows for greater control aspects than is possible in the traditional gathering process. Notably, that it is possible during simulation to simply pause the simulation and acquire all the necessary samples instantaneously (from a runtime perspective), without having to account for minor deviations in time or position due to hardware or human limitations. This effect also applies between samples, rather than wait for the sampling hardware to relocate to a new position as would be required in a physical sampling session, it is possible to move the camera and any attached sampling hardware to the next position immediately without having to ensure that the position is correct or make readjustments.

3 Method

Previous work [Lee et al., 2022] detailed the method regarding the acquisition of the 3D Scene and development of the simulator used for this experiment. To summarise, a 3D Scan of a physical location and a simulated quadcopter UAV camera is imported into a prototype simulator, referred to developmentally as "Fine, I'll Do It Myself" (FIDIM). FIDIM is being developed for the generation of visually realistic synthetic image data for Deep Learning using the Unity 3D Game engine which is common in Game Development as a framework due to the engines ease of use and robust package library. This scene was augmented using a small library of 3D Scanned bicycle assets (5 in total) to be used as classification subjects for the position/rotation log (see Fig. 2). Bicycle objects were chosen due the availability of models and "Bicycle" being a classifiable label by the base Yolov3 network used for the analysis in Section 4.

 ISBN 978-0-9934207-7-1

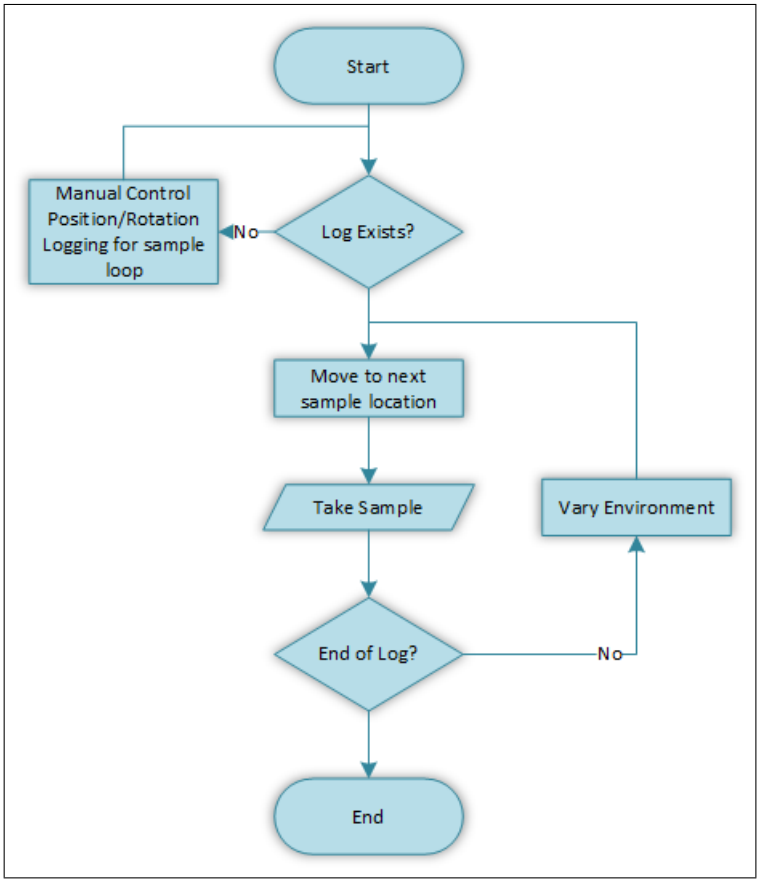


Figure 4: Automated sampling strategy flowchart, with environmental variance

3.1 Time Axis Simulation

Exhibiting different time effects in a visually realistic manner is key to the value of this dataset. Since the average light level in an given outdoor scene is primarily dictated by the time of day, with the season, and location having less significant impact. By using sufficiently realistic light simulation combined with adjustments to the texture lighting and skybox colour the transition from day to night can be simulated as seen in Fig. 3. From this simulated axis, 6 time configurations were selected. Further development into time axis simulation will account for geographic variation, seasonal change, and solar angle but were considered unnecessary for the initial simulation, as the time configurations are only relative to the simulations initial parameters at this time.

3.2 Sampling

The strategy for the generation of synthetic data samples is outlined in Fig. 4. After manual collection, the simulation will automatically generate the iterated dataset once the logs have been created and the chosen time variations are set. Additionally, each image that is generated contains the time iteration number and position number embedded in the filename for later reference, it is possible that other useful outputs can be contained in samples this way such as virtual lidar values or similar sensor data.

4 Results

For the first round of sampling, the virtual UAV was used to manually define a position and rotation log containing 411 entries was iterated over 6 time settings to create a dataset which contains a total of 2466 images. Each

 ISBN 978-0-9934207-7-1

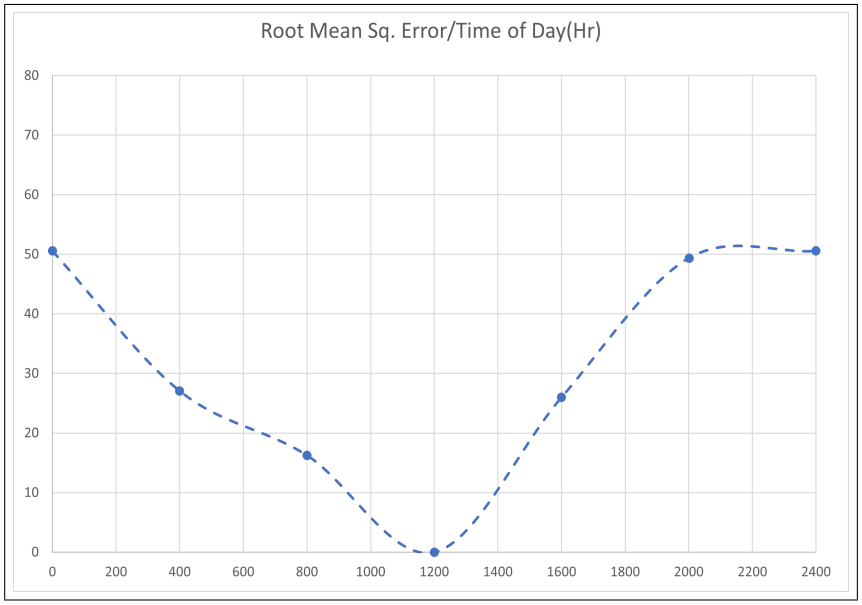


Figure 5: Linear graph of RMSE by Time of Day, represented as a percentage of deviation from ideal conditions, using base Yolov3 network classification

new position in the log creates 6 unique images related to that position, this allows for the data to be sliced in ways that are unfeasible in traditional sampling. These samples were then classified using the YOLOv3 pretrained network [Redmon and Farhadi, 2018], note that the pretrained network was used as is and no training was done using the simulated data. The highest average light in the simulation occurs at 1200hrs (noon) which is used as an ideal comparison for the rest of the time configurations. Average Root Mean Squared Error(RMSE) is then calculated by comparing the classification results from a given time variation sample to the ideal time sample.

4.1 Comments

For initial sampling, four equidistant points on the time axis; 6 hours, 12 hours, 18 hours and 24 hours respectively. However, these time configurations were too distant to provide adequate insight. In order to generate more informative samples, the number of configurations was increased to ensure a 4 hour gap between points along the time axis (see Fig. 5). It is important to note that since the graph is represented linearly but is cyclical in reality, ’0’ and ’2400’ on the x axis are the same point and thus have the identical Average RMSE, additionally, noon was expected to yield an RMSE of 0 since it is the point of comparison from which the RMSE of the 5 other time configurations are calculated.

5 Conclusion

This work sought to investigate the use of custom, photo realistic simulation and synthetic data generation as method of measuring deviation in the classification probability of the commonly used YOLOv3 pretrained network. The dataset, which is generated across a simulated time axis, when evaluated showed an increase in the Average RMSE of the network response peaking at 50.56% at the darkest point on the time axis being 2400hrs (simulated midnight), which is to be expected based on how the average light level (and thus visibility) in a scene changes as the cyclical time axis progresses from day into night. Since object visibility is critical to the task of classification, daylight level is shown to be of high impact to the performance of the classification network even when the change in simulated time is quite minimal.

 ISBN 978-0-9934207-7-1

5.1 Future Work

This project is an initial part of a series of investigations on the creation of a spatially linked area which is then simulated in a modern 3D engine that has been purpose-built for the task of creating reality-analogous synthetic image samples. Future development is aimed at generating simulated data that iterates over other environmental axes such as weather, the creation of a locally accessible "Digital Twin Area" to serve as a test-bed for Autonomous Navigation Projects, and increasing the overall quality of the simulation. Additionally, further analysis into methods of verifying trained network performance, such as Uncertainty estimation is being considered.

6 Acknowledgments

This project was funded by the ADVANCE CRT PhD Program.

References

[Al-Sharman et al., 2020] Al-Sharman, M. K., Zweiri, Y., Jaradat, M. A. K., Al-Husari, R., Gan, D., and Seneviratne, L. D. (2020). Deep-learning-based neural network training for state estimation enhancement: Application to attitude estimation. *IEEE Transactions on Instrumentation and Measurement*, 69(1):24–34.

[Alshehri et al., 2019] Alshehri, A., Member, S., Bazi, Y., and Member, S. (2019). Deep Attention Neural Network for Multi-Label Classification in Unmanned Aerial Vehicle Imagery. *IEEE Access*, 7:119873–119880.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, and Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, number June, pages 248–255. IEEE.

[European Union Aviation Safety Agency, 2020] European Union Aviation Safety Agency (2020). Special Condition for Light Unmanned Aircraft Systems - Medium Risk INTRODUCTORY. Technical report, EASA.

[Hartawan et al., 2019] Hartawan, D. R., Purboyo, T. W., and Setianingsih, C. (2019). Disaster victims detection system using convolutional neural network (CNN) method. In *Proceedings - 2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2019*, pages 105–111.

[Lee et al., 2021] Lee, T., Mckeever, S., and Courtney, J. (2021). Flying Free: A Research Overview of Deep Learning in Drone Navigation Autonomy. *Drones*, 5(2):52.

[Lee et al., 2022] Lee, T., Mckeever, S., and Courtney, J. (2022). Generating Reality-Analogous Datasets for Autonomous UAV Navigation using Digital Twin Areas. In *2022 33rd Irish Signals and Systems Conference (ISSC)*, pages 1–6. IEEE.

[Loquercio et al., 2020] Loquercio, A., Kaufmann, E., Ranftl, R., Dosovitskiy, A., Koltun, V., and Scaramuzza, D. (2020). Deep Drone Racing: From Simulation to Reality With Domain Randomization. *IEEE Transactions on Robotics*, 36(1):1–14.

[Loquercio et al., 2018] Loquercio, A., Maqueda, A. I., Del-Blanco, C. R., and Scaramuzza, D. (2018). DroNet: Learning to Fly by Driving. *IEEE Robotics and Automation Letters*, 3(2):1088–1095.

[Palossi et al., 2018] Palossi, D., Loquercio, A., Conti, F., Flamand, E., Scaramuzza, D., and Benini, L. (2018). A 64mW DNN-based Visual Navigation Engine for Autonomous Nano-Drones. *IEEE Internet of Things Journal*, 6(5):8357–8371.

[Perri et al., 2022] Perri, D., Simonetti, M., and Gervasi, O. (2022). Synthetic data generation to speed-up the object recognition pipeline. *Electronics (Switzerland)*, 11(1):1–19.

[Redmon and Farhadi, 2018] Redmon, J. and Farhadi, A. (2018). YOLOv3: An Incremental Improvement. *CoRR*, 1804.02767.

[Richter et al., 2016] Richter, S. R., Vineet, V., Roth, S., and Koltun, V. (2016). Playing for data: Ground truth from computer games. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9906 LNCS:102–118.

[Rojas-Perez and Martinez-Carranza, 2020] Rojas-Perez, L. O. and Martinez-Carranza, J. (2020). DeepPilot: A CNN for Autonomous Drone Racing. *Sensors*, 20(16):4524.

[Shorten and Khoshgoftaar, 2019] Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60.

[Takahashi et al., 2020] Takahashi, R., Matsubara, T., and Uehara, K. (2020). Data Augmentation Using Random Image Cropping and Patching for Deep CNNs. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2917–2931.

[Udacity, 2016] Udacity (2016). An Open Source Self-Driving Car.

[Vierling et al., 2020] Vierling, A., Sutjaritvorakul, T., and Berns, K. (2020). Dataset Generation Using a Simulated World. In *International Conference on Robotics in Alpe-Adria Danube Region*, pages 505–513.

[Wang et al., 2021] Wang, Z., Han, K., and Tiwari, P. (2021). Digital twin simulation of connected and automated vehicles with the unity game engine. *Proceedings 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence, DTPI 2021*, pages 180–183.

[Yang and Wang, 2020] Yang, R. and Wang, X. (2020). UAV Landmark Detection Based on Convolutional Neural Network. *2nd IEEE Eurasia Conference on IOT, Communication and Engineering 2020, ECICE 2020*, 0(1):5–8.

[Yong and Yeong, 2018] Yong, S. P. and Yeong, Y. C. (2018). Human Object Detection in Forest with Deep Learning based on Drone's Vision. In *2018 4th International Conference on Computer and Information Sciences: Revolutionising Digital Landscape for Sustainable Smart Society, ICCOINS 2018 - Proceedings*, pages 1–5. IEEE.

A Comparison of Feature Extraction Methods Applied to Thermal Sensor Binary Image Data to Classify Bed Occupancy

Rebecca Hand, Ian Cleland and Chris Nugent

School of Computing, Ulster University

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/QLZV1440

Abstract

Low-resolution thermal sensing technology is suitable for sleep monitoring due to being light invariant and privacy preserving. Feature extraction is a critical step in facilitating robust detection and tracking, therefore this paper compares a blob analysis approach of extracting statistical features to several common feature descriptor algorithm approaches (SURF and KAZE). The features are extracted from thermal binary image data for the purpose of detecting bed occupancy. Four common machine learning models (SVM, KNN, DT and NB) were trained and evaluated using a leave-one-subject-out validation method. The SVM trained with feature descriptor data achieved the highest accuracy of 0.961.

Keywords: Thermal Sensor, Image Processing, Bed Occupancy Detection, Feature Description, Classification

1 Introduction

From an accurate bed occupancy metric further sleep quantifying metrics can be generated such as, the total time spent in bed and the number and timing of bed exits [Gilakjani et al., 2018]. Many researchers have developed sleep monitoring systems based on vision sensor data. A near-infrared camera-based solution to determine sleep quality, by detecting body movements, proved more accurate to actigraphy [Liao and Yang, 2009]. Nevertheless, the high-resolution images are intrusive and introduce privacy concerns. A low-resolution visual sensor was developed to determine sleep activity based on detected motion patterns while preserving privacy [Eldib et al., 2015]. This system was limited by lighting conditions: overestimating sleep duration by around 20%, and missing bathroom visits. The application of low-resolution thermal sensing to sleep monitoring offers advantages over other vision sensors, as they operate irrespective of lighting conditions and preserve privacy [Kido et al., 2009].

Vision sensors produce image data from which features are extracted for object recognition or scene classification. There are several approaches to features extraction, including: blob analysis; feature detector and descriptor algorithms. Blob analysis algorithms, such as MATLAB's 'regionprops', is the traditional approach to feature extraction from thermal sensor binary image data, extracts statistical measures from blobs such as the centroid or area [Burns et al, 2019]. Feature-detector algorithms such as SURF (speeded up robust features) and KAZE are used to detect key points (i.e., blobs) within an image, returning their location for processing by feature descriptor algorithms. Feature descriptor algorithms compute feature vectors of encoded identifiable numeric information which represent the detected key point and surrounding pixels. This paper compares the performance of four classification models to classify bed occupancy from several feature permutations.

2 Related Work

Thermal sensors have been applied to several areas of human activity recognition, such as fall detection [Kido et al., 2009], sedentary behaviour monitoring [Synnott et al, 2000] and assessing activities of daily living [Burns et al., 2019]. These low-resolution thermal solutions rely on blob statistical features. The following sections will briefly describe blob statistical features and feature detectors and descriptors.

Blob Statistical Features: In the thermal systems presented in [Shetty et al., 2017] and [Kuki et al, 2012] human movements were tracked by determining the centroid of the heat source blob. Synnott et al. [Synnott et al, 2016] monitored sedentary behaviour using ten hand crafted metrics based upon the presence of a blob within binary images. Example metrics include total occupancy time and mean occupancy time. A thermal sensor and ultrasonic solution to recognising bedside events, such as bed entry, detected human location from four hand crafted metrics based upon the blob area and max temperature [Asbjørn and Jim, 2017]. Like the approach in this paper, [Burns et al., 2019] used the 'regionprops' function to extract statistical features from thermal binary image blobs. The statistical features (e.g., bounding box and area) were used to recognize daily kitchen activities.

Feature Detectors and Descriptors: A study evaluating the use of popular feature detector and descriptor algorithms across a wide range of RGB images, binary images and transformed images, determined the SIFT-based (scale-invariant feature transform) descriptor, such as SURF, was the best performing [Mikolajczyk and Schmid, 2003]. The SURF detector and descriptor algorithm has been used for object detection, recognition and tracking in RGB image data. Guo and Wang [Guo and Wang, 2019] employed a SURF-based algorithm to recognise clothing accessories, while Loussaief and Abdelkrim [Loussaief and Abdelkrim, 2018] and Horak et al. [Horak et al., 2017] employed SURF-based algorithms to recognise vehicle number plates and traffic stop signs.

While SURF-based algorithms have been widely applied to RGB images, they are also capable of detecting and extracting feature points from greyscale and binary images. A SURF-based algorithm was used to detect signature forgery within binary images [Okawa, 2016]. In this study, the performance of the SURF-based algorithm was compared to the performance of a KAZE-based algorithm. Although KAZE is a relatively newer and less popular approach to blob feature detector and extraction, the KAZE-based algorithm outperformed the SURF-based algorithm achieving an equal error rate of 1.6%, compared to 5.5% [Okawa, 2016].

To our knowledge, feature descriptor algorithms have not been applied to thermal binary images. This paper will investigate the application of SURF and KAZE to detect bed occupancy. Initial investigations revealed each feature detector returned different key point locations. SURF predominantly located the centre of the head while KAZE predominantly located limbs or shoulders. The SURF and KAZE approaches will be compared to the traditional approach to feature extraction from thermal binary images, that is, blob analysis.

3 Materials and Methods

Thermal data was collected using a Heimann HTPA 32x32 Infrared Thermopile Array Sensor installed onto a frame providing a bird's eye view of a bed and surrounding area. Four examples of bed occupancy data (with target durations of 1, 3, 5 and 10 minutes) were provided from five participants (4: Female,1: Male). The data collection protocol (detailed in [Hand et al., 2022]) resulted in 55,529 frames. The temperature data was reshaped, cropped, re-scaled and visualized as a greyscale image before binarization to identify heat sources. Figure 1 presents an overview of the data pre-processing. The classes were largely imbalanced with 45,522 in bed frames, (i.e., >50% of the human heat blob was within the bed location) and 10,007 not in bed frames (i.e., <50% of the human heat blob was within the bed location).

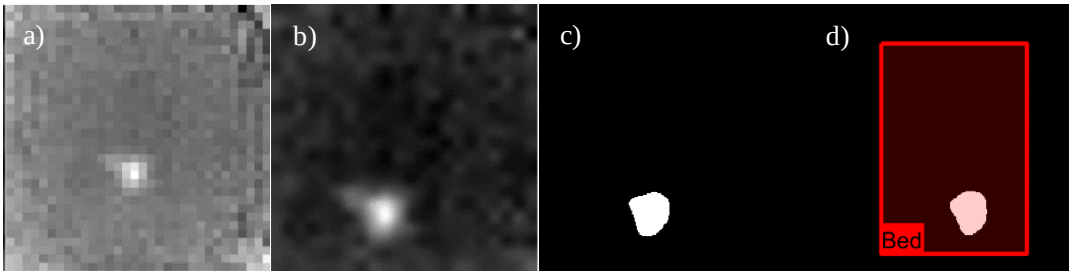


Figure 1: Pre-processed thermal data: a) a 32x32 greyscale image; b) the 220x220 greyscale image; c) the binary image; and d) with the bed location superimposed for labelling.

4 Implementation

Implementation was undertaken using MATLAB. Table 1 summarizes the feature data resulting from each approach applied in this paper. The blob analysis approach returns a 17- dimensional feature vector and the feature detector and descriptor algorithms (SURF, KAZE) return a 64- dimensional feature vector. Each permutation of the SURF and KAZE detector and descriptor algorithms are evaluated. The labelled feature vectors are used to train and test four classification models (i.e., Decision Tree (DT), k-Nearest Neighbor (KNN(k=5)), Support Vector Machines (SVM), and Naïve Bayes (NB)) using the Classification Learner app. The hyperparameter options were disabled. Each classification model was trained and tested using a leave-one-subject out validation method. Performance measures including accuracy and F-measure were derived from the overall summed confusion matrix for each classification model. The performance of each classification model was evaluated using the feature data resulting from each approach to feature extraction.

Approach	**Feature Data**
SURF Descriptor	**64**-dimensional vector, sums Haar wavelets responses over a 4 × 4 rectangular grid, in dominant orientation, Gaussian weighting
KAZE Descriptor	**64**-dimensional vector, sums overlap of Haar wavelets responses over a 4 × 4 rectangular grid, in dominant orientation, two step Gaussian weighting
BLOB Analysis	**17**-dimensional vector, Area, X Centroid, Y Centroid. X Bounding Box, Y Bounding Box, H Bounding Box, W Bounding Box, Major Length, Minimum Length, Eccentricity, Orientation, Convex Area, Circularity, Filled Area, Equivalent Diameter, Solidity, Extent

Table 1: A description of the feature vector data extracted using each approach.

5 Results

The performance of each classification model in terms of classifying bed occupancy is presented in Table 2.

Model	**Accuracy (SEM)**					**F-measure**				
	BLOB	SURF_ SURF	KAZE_ KAZE	KAZE_ SURF	SURF_ KAZE	BLOB	SURF_ SURF	KAZE_ KAZE	KAZE_ SURF	SURF_ KAZE
DT	0.951 (1.19)	0.898 (3.62)	0.899 (4.4)	0.882 (3.4)	0.926 (2.3)	0.970	0.938	0.938	0.927	0.954
KNN	0.916 (4.42)	0.873 (3.12)	0.882 (5.32)	0.873 (3.95)	0.94 (1.65)	0.947	0.921	0.925	0.920	0.963
SVM	**0.953** (1.67)	0.92 (3.1)	0.914 (4.03)	0.902 (2.61)	**0.961** (0.92)	**0.972**	0.915	0.956	0.940	**0.977**
NB	0.903 (4.57)	0.573 (15.33)	0.896 (2.4)	0.908 (2.52)	0.767 (14.18)	0.940	0.924	0.936	0.944	0.845

Table 2: Accuracy (with Standard Error of Mean) and F-measure comparisons of the classification models to detect bed occupancy. The best performing model for each feature extracting method is highlighted in bold.

For the blob analysis feature extraction method, the SVM model achieved the highest accuracy of 0.953 and F-measure of 0.972, closely followed by the DT model achieving 0.951 accuracy and 0.970 F-measure. The DT and NB models resulted in a high number of false positives while the KNN and SVM models resulted in a high number of false negatives. For the feature detector and descriptor method, SURF_KAZE SVM model achieved the highest accuracy of 0.961 and the highest F-measure of 0.977. Interestingly, the DT, KNN and SVM

 ISBN 978-0-9934207-7-1

performed almost equally between the SURF_SURF data and KAZE_KAZE data. Also, the NB models did not perform well when used with the SURF detector algorithm, illustrated through the low accuracy and high SEM.

6 Conclusions

This study illustrates the potential of using feature detector and feature descriptor data to classify binary thermal images. Based on the current study, the best model to detect bed occupancy was determined to be the SURF_KAZE SVM model due to achieving the highest accuracy (lowest SEM) and F-measure of all models. Nevertheless, with the BLOB SVM model performed comparably and considering the difference in feature dimensionality, it can be concluded that both approaches to feature extraction - blob analysis and feature description - are valid approaches to use in future studies. A thermal sensing approach to detecting bed occupancy is fundamentally challenged by the accumulation of residual heat resulting in false positive readings [Hand et al, 2022]. In future work the background subtraction algorithm must identify and remove residual heat.

References

[Asbjørn and Jim, 2017] Asbjørn D. and Jim T. (2017), "*Recognizing Bedside Events Using Thermal and Ultrasonic Readings*", Sensors (Vol. 6), 2017.

[Burns et al., 2019] Burns M., Morrow P., Nugent C., and McClean S. (2019), "*Fusing Thermopile Infrared Sensor Data for Single Component Activity Recognition within a Smart Environment*", JSAN (Vol. 8), 2019.

[Eldib et al., 2015] Eldib M., Deboeverie F., Philips W., and Aghajan H. (2015), "*Sleep Analysis for Elderly Care Using a Low-Resolution Visual Sensor Network*", LNIP (Vol. 9277), 2015.

[Gilakjani et al., 2018] Gilakjani, S., Bouchard M., Goubran R., and Knoefel F. (2018), "*Long-Term Sleep Assessment by Unobtrusive Pressure Sensor Arrays*", ICBSP, 2018.

[Guo and Wang, 2019] Guo J. and Wang X, (2019) "*Image Classification Based on SURF and KNN*", ICIS, 2019.

[Hand et al., 2022] Hand R., Cleland I. and Nugent C. (2022), "*Detecting Bed Occupancy Using Thermal Sensing Technology: A Feasibility Study*", LNICST, (Vol. 431) 2022.

[Horak et al., 2017] Horak K., Klecka J., Bostik O., and Davidek D. (2017), "*Classification of SURF Image Features by Selected Machine Learning Algorithms*", TSP, 2017.

[Kido et al., 2009] Kido S., Miyasaka T., Tanaka T., Shimizu T., and Saga T. (2009), "*Fall Detection in Toilet Rooms Using Thermal Imaging Sensors*", SII, 2009.

[Kuki et al., 2012] Kuki M., Nakajima H., Tsuchiya N., and Hata Y. (2012), "*Human Movement Trajectory Recording for Home Alone by Thermopile Array Sensor*", SMC, 2012.

[Liao and Yang, 2009] Liao W. and Yang C (2009), "*Video-based Activity and Movement Pattern Analysis in Overnight Sleep Studies*", ICPR, 2009.

[Loussaief and Abdelkrim, 2018] Loussaief S. and Abdelkrim A. (2018), "*Machine Learning Framework For Image Classification*", SETIT, 2018.

[Mikolajczyk and Schmid, 2003] Mikolajczyk K. and Schmid C. (2003), "*A Performance Evaluation of Local Descriptors*", CVPR, 2003.

[Owaka, 2016] Okawa M. (2016), "*Vector of Locally Aggregated Descriptors with KAZE Features for Offline Signature Verification*", GCCE, 2016.

[Shetty et al., 2017] Shetty A., Disha, Shubha B., and Suryanarayana K. (2017), "*Detection and Tracking of a Human using the Infrared Thermopile Array Sensor - 'Grid-EYE'*", ICICT, 2017.

[Synnott et al., 2016] Synnott, J., Rafferty J., and Nugent C. (2016), "*Detection of Workplace Sedentary Behaviour Using Thermal Sensors*", EMBC, 2016.

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/UAOV9084

Recurrent Super-Resolution Method for Enhancing Low Quality Thermal Facial Data

David O'Callaghan*[1], Cian Ryan[1], Waseem Shariff[1,2], Muhammad Ali Farooq[1,2], Joseph Lemley[1], and Peter Corcoran[2]

[1]*Xperi Corporation, Galway, Ireland*
[2]*School of Engineering, National University of Ireland, Galway*

Abstract

The process of obtaining high-resolution images from single or multiple low-resolution images of the same scene is of great interest for real-world image and signal processing applications. This study is about exploring the potential usage of deep learning based image super-resolution algorithms on thermal data for producing high quality thermal imaging results for in-cabin vehicular driver monitoring systems. In this work we have proposed and developed a novel multi-image super-resolution recurrent neural network to enhance the resolution and improve the quality of low-resolution thermal imaging data captured from uncooled thermal cameras. The end-to-end fully convolutional neural network is trained from scratch on newly acquired thermal data of 30 different subjects in indoor environmental conditions. The effectiveness of the thermally tuned super-resolution network is validated quantitatively as well as qualitatively on test data of 6 distinct subjects. The network was able to achieve a mean peak signal to noise ratio of 39.24 on the validation dataset for 4x super-resolution, outperforming bicubic interpolation both quantitatively and qualitatively.

Keywords: Super-resolution, Deep learning, Thermal imaging, LWIR

1 Introduction

Super-resolution (SR) is used to increase the details of a low-resolution (LR) image by applying statistical approaches, and various optimization techniques on either a sequence of images in a collective manner or a single image to generate high-resolution image. In this article, we have focused on developing an imaging pipeline to output super-enhanced thermal images for in-cabin driver monitoring systems. Primarily, with the advancements in safe and autonomous systems for vehicular technology, it is essential to monitor the driver's activity for enabling enhanced security and safety features. This can be achieved by face localization, facial landmark detection, drowsiness and fatigue detection. In this work, particular focus is on developing a thermal super-resolution algorithm for enhancing the existing thermal imaging data captured from Video Graphics Array (VGA) uncooled prototype Long-wave Infrared (LWIR) camera developed under the Heliaus project [Heliaus, 2022]. Given recent developments in microbolometer technology, uncooled thermal imaging sensors are now less expensive and can be used for wide range real-world applications [Farooq et al., 2021]. This sensor has added benefits as they can detect the thermal emissivity of objects and work independently of illumination conditions (i.e., a thermal camera can operate independently in any circumstance, including day and night), making them a more reliable source of data for vehicular in-cabin applications. The proposed SR algorithm integrates recurrent-based multi-image super-resolution (MISR) and single-image super-resolution (SISR), thus taking benefits from both the methods to produce optimal results.

*david.ocallaghan@xperi.com

 ISBN 978-0-9934207-7-1

2 Background

Image super-resolution is a popular research topic in the scientific community. And, given the challenges of thermal vision, super-resolution of thermal images is a difficult research problem. A number of useful thermal datasets have been released in recent years [Wang et al., 2010, Espinosa-Duró et al., 2013]. It is currently crucial to develop an effective neural network algorithm to super-resolve these images using available datasets. The deep back-projection network (DBPN) [Haris et al., 2018] was proposed for super-resolution of single visible images. Further improvements include up-sampling layers with a recurrent network [Haris et al., 2019], an unsupervised approach using CycleGAN [Rivadeneira et al., 2020], Spatio-Temporal feature fusion deep neural network [Zhang et al., 2021], with different up-sampling and asymmetrical residual learning in the network [Patel et al., 2021]. Considering thermal image super-resolution is an open research topic, a CVPR workshop also included a thermal super-resolution challenge. The challenge was tackled by 9 separate teams, each with a unique solution [Rivadeneira et al., 2021]. Recently a new approach was proposed to this problem that is based on a generative adversarial network with channel filtering mainly focused on super-resolution for power equipment [Haris et al., 2018]. This network outperformed the state-of-the-art in the reconstruction of thermal imaging images (of power equipment) by having a higher peak signal to noise ratio (PSNR) and structural similarity (SSIM) to the enhanced image.

3 Methodology

3.1 Data collection

The acquired thermal SR dataset was created after a series of data acquisitions which were conducted as part of the Heliaus project [Heliaus, 2022] in a driving simulator. The dataset consisted thermal images of 36 individuals (7 female and 29 male) using a VGA quality 640×480 uncooled thermal camera based on microbolometer technology developed under the Heliaus project. The mean age of the individuals was 37.7 years and the standard deviation was 11.3 years. On average, 543 sequences of thermal images (each with a length of 10 frames) were extracted from the recording of each subject. This resulted in a dataset of 19,558 thermal image sequences. The data of 30 subjects were used for training and 6 subjects were used for testing. The effective train-test split ratio was 0.83:0.17.

3.2 Neural Network Architecture

The proposed novel super-resolution architecture is illustrated in Figure 1. In the left image, low-resolution images are combined with latent space feature maps computed at the previous time-step and the previous SR network output. Information is propagated through time in a recurrent manner. The right image illustrates the individual components of the unified cell. This comprises of multi-image, single-image, residual and reconstruction components. The feature space outputs of the multi-image and single-image components are combined through element-wise subtraction and further processed, the results of which are added (element-wise) to the single image output to generate a super-resolved image. The network is fully convolutional and can take any input size. A given trained network can then up-sample with a fixed scale; i.e., 2x or 4x.

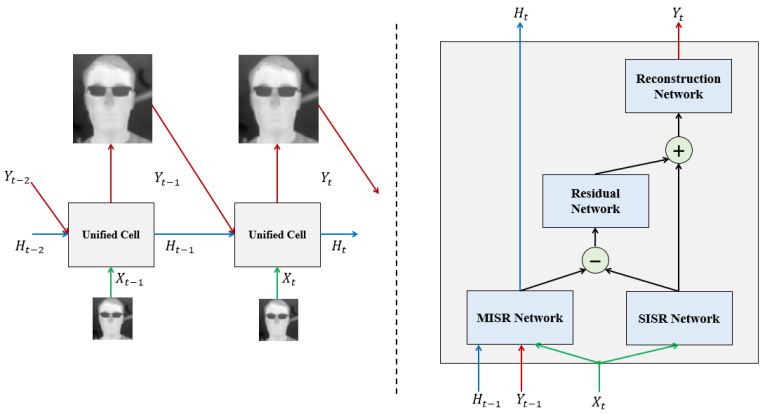


Figure 1: The unrolled recurrent network (left) and components of the unified cell (right) of our super-resolution network.

The proposed network integrates recurrent-based MISR and SISR, drawing benefits from both the methods to super-resolve thermal images. The single image network can be replaced by any state-of-the-art network but,

currently, we have adopted the DBPN [Haris et al., 2018]. The MISR and Residual components include series of residual blocks. That is, convolutional layers with residual connections between block input and output. The reconstruction component is simply a convolutional layer that takes a combination of the outputs of the Residual component and SISR component to produce the super-resolved frame.

The complexity of our neural network architecture was analysed using 3 different metrics: The number of parameters, the number of multiply-accumulates (MACs) and the number of floating-point operations (FLOPs). The 4x version of the network had 5,151,645 parameters. For an 80×80 input thermal image, this network performs 285.45 GMACs and 570.91 GFLOPs.

3.3 Training

The network was trained using PyTorch,the deep learning framework. The sequence length of each sample passed to the network during training was randomised between 1 and 10 to allow the recurrent network to generalise to varying lengths of input. The same 128×128 random crop was taken from each HR frame in the sequence and the LR frames were created by down-sampling by a factor of 4 and applying Gaussian blur and additive noise to simulate degradation due to the camera resolution. The ADAM optimizer was used with a learning rate of 10^{-4} reduced by a factor of 0.5 every 25 epochs for a total of 100 epochs. Mean absolute error was used as the loss function and a weight decay of 10^{-4} was applied. No form of pre-training was done. Figure 2 shows the loss and PSNR progression on the training and validation data during training.

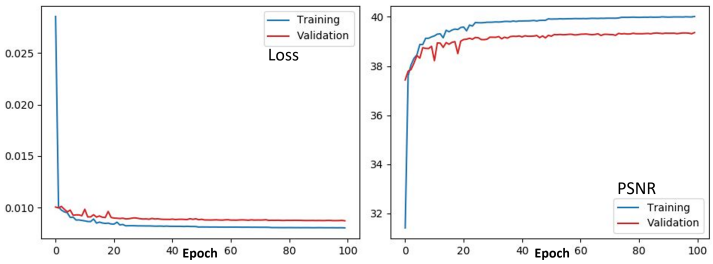


Figure 2: Progression of mean absolute error loss and PSNR during training.

4 Experimental Results

The performance of the proposed super-resolution network trained on our collected thermal dataset was evaluated using PSNR and SSIM as quantitative metrics. The mean values of these metrics on our validation dataset using our trained SR network and bicubic interpolation are shown in Table 4.

	PSNR	SSIM
SR	39.235 ± 2.639	0.901 ± 0.051
Bicubic	37.705 ± 1.952	0.879 ± 0.045

Table 1: Performance on validation dataset.

The proposed SR network outperforms bicubic interpolation significantly in both of these metrics and also does so visually. Qualitative examples of our SR networks performance can be visualized in Figure 3.

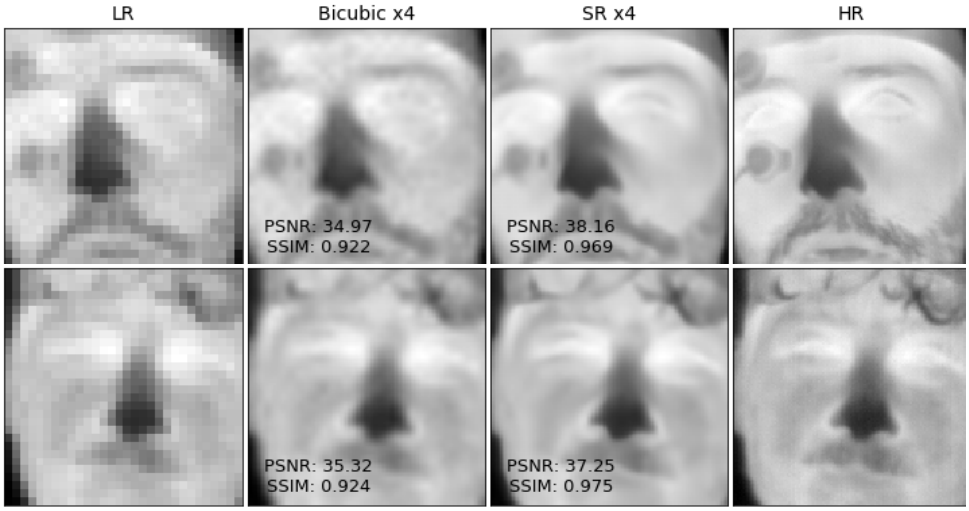


Figure 3: Examples of SR network performance on sample images from our validation set.

 ISBN 978-0-9934207-7-1

5 Conclusions

In this study, we propose a novel multi-image super-resolution architecture for in-cabin driver monitoring systems that exhibits exemplary performance on a locally acquired thermal facial dataset acquired from an uncooled thermal camera. The proposed network integrates recurrent-based MISR and SISR. The trained network (PSNR: 39.24, SSIM: 0.90) outperforms bicubic interpolation (PSNR: 37.71, SSIM: 0.88) both quantitatively and qualitatively. As part of future work, we plan to optimise the network to further accelerate the inference speed thus making it more efficient and computationally less expensive for real-time deployment on low-power edge computing devices (such as NVIDIA-Jetson). We also plan to retrain the network on a larger and more diverse dataset.

References

[Espinosa-Duró et al., 2013] Espinosa-Duró, V., Faundez-Zanuy, M., and Mekyska, J. (2013). A new face database simultaneously acquired in visible, near-infrared and thermal spectrums. *Cognitive Computation*, 5(1):119–135.

[Farooq et al., 2021] Farooq, M. A., Corcoran, P., Rotariu, C., and Shariff, W. (2021). Object detection in thermal spectrum for advanced driver-assistance systems (adas). *IEEE Access*, 9:156465–156481.

[Haris et al., 2018] Haris, M., Shakhnarovich, G., and Ukita, N. (2018). Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673.

[Haris et al., 2019] Haris, M., Shakhnarovich, G., and Ukita, N. (2019). Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3897–3906.

[Heliaus, 2022] Heliaus (2022). Heliaus - thermal vision augmented awareness. `https://www.heliaus.eu/`. Accessed: 2022-06-22.

[Patel et al., 2021] Patel, H. M., Chudasama, V. M., Prajapati, K., Upla, K. P., Raja, K., Ramachandra, R., and Busch, C. (2021). Thermisrnet: an efficient thermal image super-resolution network. *Optical Engineering*, 60(7):073101.

[Rivadeneira et al., 2020] Rivadeneira, R. E., Sappa, A. D., and Vintimilla, B. X. (2020). Thermal image super-resolution: A novel unsupervised approach. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics*, pages 495–506. Springer.

[Rivadeneira et al., 2021] Rivadeneira, R. E., Sappa, A. D., Vintimilla, B. X., Nathan, S., Kansal, P., Mehri, A., Ardakani, P. B., Dalal, A., Akula, A., Sharma, D., et al. (2021). Thermal image super-resolution challenge-pbvs 2021. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4359–4367.

[Wang et al., 2010] Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., Chen, F., and Wang, X. (2010). A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia*, 12(7):682–691.

[Zhang et al., 2021] Zhang, W., Sui, X., Gu, G., Chen, Q., and Cao, H. (2021). Infrared thermal imaging super-resolution via multiscale spatio-temporal feature fusion network. *IEEE Sensors Journal*, 21(17):19176–19185.

Beyond Social Distancing: Application of real-world coordinates in a multi-camera system with privacy protection

Frances Ryan, Feiyan Hu, Julia Dietlmeier, Noel E. O'Connor, Kevin McGuinness

Insight SFI Research Centre for Data Analytics, Dublin City University, Dublin, Ireland

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/RTNS4233

Abstract

In this paper, we develop a privacy-preserving framework to detect and track pedestrians and project to their real-world coordinates facilitating social distancing detection. The transform is calculated using social distancing markers or floor tiles visible in the camera view, without an extensive calibration process. We select a lightweight detection model to process CCTV videos and perform tracking within-camera. The features collected during within-camera tracking are then used to associate passenger trajectories across multiple cameras. We demonstrate and analyze results qualitatively for both social distancing detection and multi-camera tracking on real-world data captured in a busy airport in Dublin, Ireland.

Keywords: Social distancing, people detection, cross-camera tracking, coordinate calibration

1 Introduction

During pandemics, social distancing helps slow the spread of the virus and as Covid-19 has proven will be an important tool in the fight against future pandemics. Additionally, identifying and managing areas with regular overcrowding is vital in crowded public places to ensure pedestrian safety and monitor wait or queue times.

A key element of social distancing detection is transforming pedestrian bounding boxes to real-world coordinates. This usually requires a precomputed/learnt transform matrix. Researchers have proposed different ways to infer this transform in the past year. The most popular way is to find three or more pairs of points in both image and real-world coordinates; a perspective matrix is computed using these pairs of points. However, most of the time it is challenging to specify real-world 2D coordinates. [Yang et al., 2021] tried to find real-world reference points using the floor plan of public buildings such as a train station, while others have tried using information about detected pedestrians height to estimate size of reference objects and real-world coordinates [Cong et al., 2020]. There are also some researchers that do not compute the distance between every pedestrian pair. Instead, for each pedestrian, a circular violation zone is established, and pedestrians appearing within the zone are marked as a violating group. [Punn et al., 2020] used the estimated depth from the camera to estimate the violation area for each pedestrian. [Aghaei et al., 2021] assume camera roll and pan angle to be 0, and project the image to the real-world and then use torso size instead of whole body height to compute the area of the violation zone.

In our work, we compute the perspective transform matrix by taking advantage of the commonly used social distancing markers on the floor. There are some benefits to this choice of reference points: 1) all stickers are on the same real-world plane; 2) the relative distance between stickers is largely consistent; 3) markers such as these have become commonplace during the pandemic.

The main contributions of this paper are as follows:

- A proof of concept using social distancing markers/floor tiles to compute the perspective matrix.
- Extension of a single camera to a multi-camera system by trajectory matching for cross-camera tracking.
- Demonstration and evaluation of both functionalities on challenging data from a busy airport.

 ISBN 978-0-9934207-7-1

Figure 1 gives an overview of the proposed system. Frames from multiple views are passed through the face and pedestrian detection module. The face boxes can be blurred, hence anonymizing the image data. Pedestrian boxes are input into the tracking module. The bottom-centre points(approximate foot location) are transformed to a top-down view, where person distances can be calculated. Points can further be transformed to world coordinates to plot in the existing GIS system. Within-camera tracks and appearance features can be passed to the cross-camera trajectory association module to assign final identities. To achieve a balance between speed and accuracy, we used the YOLOv5x[1] detector, trained on the CrowdHuman dataset [Shao et al., 2018]). DeepSort [Wojke et al., 2017] is used for within-camera tracking. To extract good ReID features, we use the method proposed in [Jia et al., 2019] to tackle domain shift using instance normalization in early layers and feature normalization in deep layers.

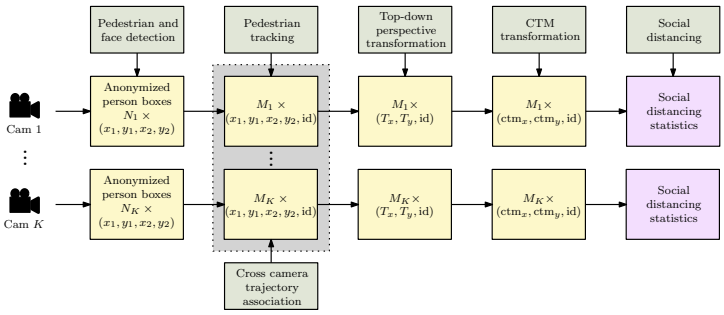


Figure 1: System overview.

2 System architecture

2.1 Transformation to Real-World Coordinates

The perspective transform matrix to top-down view is computed from 4 pairs of points. Source points – image coordinates describing the location of 4 distancing markers (or other fixed points, e.g. the corners of floor tiles) that form a square or rectangle – and target points. Transforming between image coordinates and top-down view with multiple reference points forms a system of linear equations that is solved by Gaussian elimination with the optimal pivot element to estimate the 3×3 perspective transform matrix. The transformation from top-down to world coordinates is a problem similar to rigid body movement, thus the Customised Transverse Mercator (CTM) transformation matrix can be computed by solving an Orthogonal Procrustes problem [Schönemann, 1966]. The problem can be formulated as the following optimization with the constraint that the rotation matrix R is orthonormal:

$$\begin{aligned} \min_{\sigma,R,\vec{t}} \|RA - \hat{B}\|^2, \quad \sigma\hat{B} = B - \vec{t}\cdot\mathbb{I} \\ \text{s.t.} \quad R^\top R = RR^\top = \mathbb{I} \in \mathbb{R}^{3\times 3}, \end{aligned} \tag{1}$$

where σ is a scaling factor, R is a rotation matrix, and $\vec{t}$ is a translation vector. A and B are matrices whose columns contain the corresponding points in top-down and CTM coordinates. $\hat{B}$ is shifted and scaled back from B such that the scale of $\hat{B}$ is the same as A and a pair of corresponding points in A and $\hat{B}$ are aligned as origin points. To solve the optimization, σ is estimated first as average scale factors between corresponding edges. The edge between two nodes is calculated using the L^2 norm. $\vec{t} = B[:,1] - \sigma A[:,1]$ is estimated using σ as a vector formed from a pair of corresponding points in B to A. Finally, we solve the Orthogonal Procrustes problem by singular vector decomposition as $UDV^\top = \text{svd}(\hat{B}A^\top)$. The rotation matrix is calculated as $R = UJV^\top$, where J is an identity matrix. In the case of $\det(UV^\top) = -1$, the value on diagonal of J that corresponds to the smallest value of D is set to -1.

2.2 Multi-Target Cross-Camera Re-Identification

In the airport, cameras can be selected such that the following conditions are satisfied: pedestrians 1) first appear in a set of non-overlapping 'query' cameras at various times; 2) travel through a known topology of subsequent

[1] https://github.com/ultralytics/yolov5

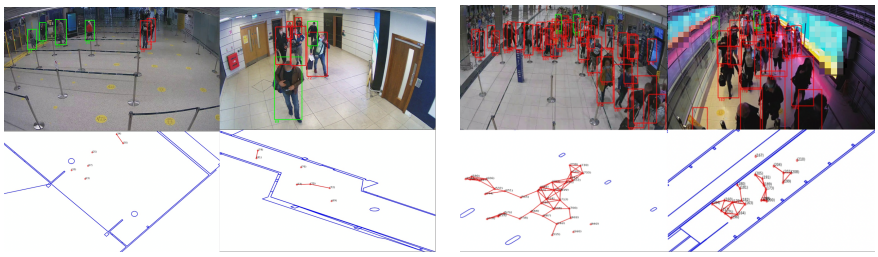
'gallery' cameras; 3) can only reappear in the next camera after having left the previous camera. Within-camera trajectories are post-processed to exclude those that are very short, that matched with low-confidence or very small bounding boxes. The trajectories from both query and gallery cameras are then input into the cross-camera trajectory association module. The within-camera appearance features collected for each trajectory are averaged such that there is one feature vector per identity from all cameras. Initially, all trajectories from the query cameras are associated with the next camera according to the network topology. This is done by calculating the cosine distance between all averaged feature vectors and associated using the Hungarian algorithm [Kuhn, 1955]. Time constraints [Dietlmeier et al., 2022] are applied to remove the possibility of matching with target identities appearing outside of pre-configured time constraints, e.g. before a query identity left the query camera. A threshold is set such that if the distance exceeds this, identities should not be associated. Unmatched query trajectories are still carried forward for potential association with a trajectory in the next camera. If, on the other hand, the match distance is within the threshold, the trajectories are appended and carried forward for association in further cameras.

3 Experiments and Results

Figure 2: (left) distances are correctly detected (right) the pair are falsely found to be within $2m$.

Social distancing. We demonstrate the system performing the social-distance monitoring function across various areas within the airport. Detected people far from the camera are removed by post-processing based on bounding-box location, due to inaccuracy caused by the perspective transform in those locations. The Euclidean distances between pairs of people are calculated and assessed against the configurable 'social distance.'. Figure 3a shows social-distance monitoring can be effective for sparse and medium crowd density.

Overestimation of social distance. Figure 2 demonstrates failures when the bounding box is cut-off due to a person exiting a scene, since this results in inaccurate estimations of foot location for a person and, hence, the transformed position is incorrect. This could result in overestimation of social distancing violations in this area. However, this can be prevented by assigning a specific region of interest within certain camera views to carry out the analysis, thereby excluding areas where the bounding box may be unreliable.



(a) (b)

Figure 3: a) Successful examples; b) Failure cases of social distancing detection. For each, the top image shows the camera view with bounding boxes in red indicating people within $2m$ of another while the bottom image plots people in the CTM coordinate space – blue lines outline airport walls.

Underestimation of social distance. Figure 3b shows failures may occur in areas with high crowd density in the far background. Single person bounding boxes may be detected among a crowd and the person is mistakenly marked as exceeding a $2m$ distance, resulting in underestimation of violations in these areas. In some areas false positive detection is caused by reflections or people moving on the other side of glass panels, these can easily be filtered out based on location since the cameras are fixed. Nonetheless, we have found it is feasible to use floor markers to estimate the distance between pedestrians.

Multi-camera tracking. Figure 4 shows examples of tracking passengers cross-camera – disembarking the aircraft, entering and later exiting the immigration hall. The green arrow in the leftmost image in each group indicates the initial query and successful tracking through the remaining cameras. Incorrect matches are indicated with a red arrow. In general and as expected, people wearing distinctive clothing or bags are easily tracked across cameras even in the presence of reasonably high crowd density, as demonstrated by examples in Figure 4a.

 ISBN 978-0-9934207-7-1

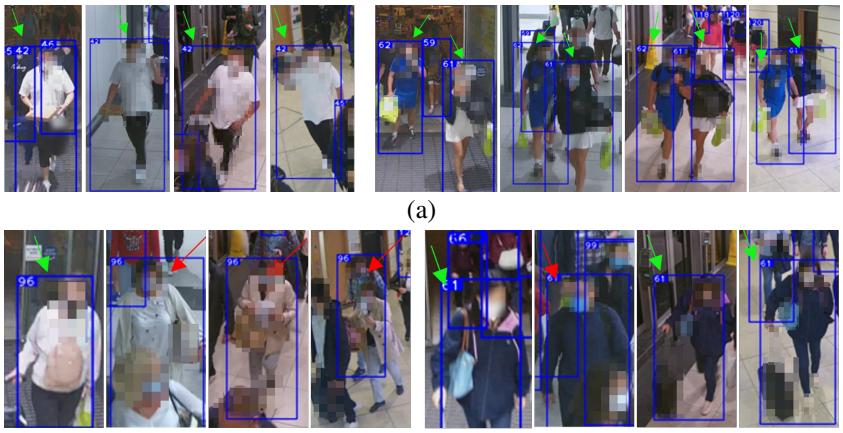


Figure 4: a) Successful examples b) failure cases of tracking passengers.

Identity switches may occur between people wearing similar clothing as shown in the examples in Figure 4b. These mismatches can be due to people changing items of clothing, e.g. putting on a jacket, between query camera and subsequent cameras, or in cases where dense crowds cause occlusions entering an area. In certain cases, the match might be incorrect in a camera where there are dense crowds, but the correct match is found in other cameras where the crowd is more dispersed. Such failures could be minimized by selecting cameras where crowds are more likely to be dispersed or using overlapping cameras, where available, to tackle severe occlusion. The convolutional neural network features generalize well to the real-world data and our colleagues in the airport confirm that the tracking in world coordinates is helpful for operations. For social distancing detection, quantitative evaluation is challenging because it is time consuming and error-prone to obtain accurate ground-truth and annotation for multi-camera tracking data.

4 Conclusion

We have proposed an approach to computer vision based social distancing detection for multi-camera system in a challenging real-world environment. We showed an effective way to calculate perspective transform using distancing markers present in the image as reference points. Furthermore, we demonstrated the transformation from top-down coordinates to CTM coordinates by using a matrix that is computed by solving an Orthogonal Procrustes problem, extending the system to use trajectory-based association for cross-camera tracking. In future work, we hope to use the developed framework to assist in annotating airport data to further evaluate the system and investigate the scaling of a cross-camera tracking system across a larger area.

References

[Aghaei et al., 2021] Aghaei, M., Bustreo, M., Wang, Y., Bailo, G., Morerio, P., and Del Bue, A. (2021). Single image human proxemics estimation for visual social distancing. In *WACV*, pages 2785–2795.

[Cong et al., 2020] Cong, C., Yang, Z., Song, Y., and Pagnucco, M. (2020). Towards enforcing social distancing regulations with occlusion-aware crowd detection. In *ICARCV*, pages 297–302. IEEE.

[Dietlmeier et al., 2022] Dietlmeier, J., Hu, F., Ryan, F., O'Connor, N. E., and McGuinness, K. (2022). Improving person re-identification with temporal constraints. In *WACV Workshop*, pages 540–549.

[Jia et al., 2019] Jia, J., Ruan, Q., and Hospedales, T. (2019). Frustratingly easy person re-identification: Generalizing person Re-ID in practice. In *BMVC*, pages 141.1–141.14.

[Kuhn, 1955] Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

[Punn et al., 2020] Punn, N. S., Sonbhadra, S. K., Agarwal, S., and Rai, G. (2020). Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and deepsort techniques. *arXiv preprint arXiv:2005.01385*.

[Schönemann, 1966] Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31:1–10.

[Shao et al., 2018] Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., and Sun, J. (2018). CrowdHuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*.

[Wojke et al., 2017] Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *ICIP*.

[Yang et al., 2021] Yang, D., Yurtsever, E., Renganathan, V., Redmill, K. A., and Özgüner, Ü. (2021). A vision-based social distancing and critical density detection system for COVID-19. *Sensors*, 21:4608.

Acoustic Source Localization Using Straight Line Approximations

Swarnadeep Bagchi, Ruairí de Fréin

Technological University Dublin

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/LJRB7078

Abstract

The short paper extends an acoustic signal delay estimation method to general anechoic scenario using image processing techniques. The technique proposed in this paper localizes acoustic speech sources by creating a matrix of phase versus frequency histograms, where the same phases are stacked in appropriate bins. With larger delays and multiple sources coexisting in the same matrix, it becomes cluttered with activated bins. This results in high intensity spots on the spectrogram, making source discrimination difficult. In this paper, we have employed morphological filtering, chain-coding and straight line approximations to ignore noise and enhance the target signal features. Lastly, Hough transform is used for the source localization. The resulting estimates are accurate and invariant to the sampling-rate and shall have application in acoustic source separation.

Keywords: Shape Representation, Kinesics, Delay Estimation, Array Signal Processing

1 Introduction

We wish to extend a delay estimation technique, called the tiled-Elevatogram [de Fréin, 2017] to general anechoic mixtures. In terms of large delay estimation, this technique is more efficient than many existing methods [Chen et al., 2004]. A stereo mixing scenario consist of J discrete time sources, $s_1[n], s_2[n], \ldots, s_J[n]$. The Elevatogram technique performs delay estimation of a single source, $s_j[n]$. It assumes that $s_j[n]$ is physically close to any one sensor and uses the mixing model: $x_1[n] = s_j[n]$ and $x_2[n] = \sum_{j=1}^{J} a_j s_j[n-\delta_j]$. The delay δ_j is measured in samples. In this paper, we evaluate the the candidature of this method in normal anechoic condition given as:

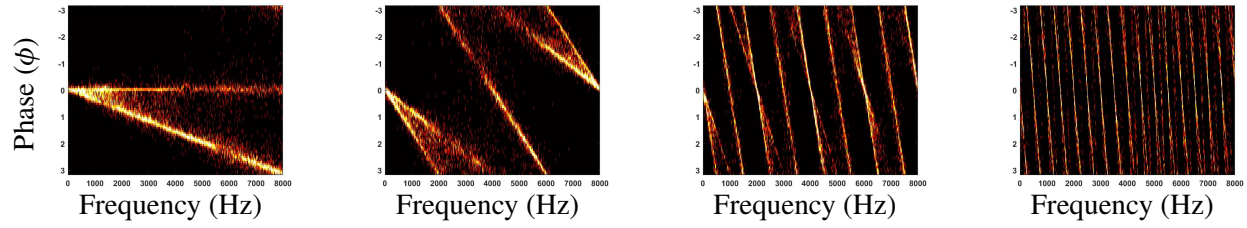
$$x_1[n] = \sum_{j=1}^{J} s_j[n] \tag{1}$$

$$x_2[n] = \sum_{j=1}^{J} a_j s_j[n-\delta_j] \tag{2}$$

1.1 The Elevatogram Technique

The transform of choice for Time-Frequency (TF) representation is synchronized short-Time Fourier transform (sSTFT) [de Fréin and Rickard, 2011]. It provides the mapping of $x_1[n]$ as: $\mathbf{X}_1 : x_1[n] \in \mathbb{R} \mapsto \mathbf{X}_1[k,\tau] \in \mathbb{C}$, where k and τ are the discrete frequency and time indices, respectively, where $1 \leq k \leq K$, and $1 \leq \tau \leq T$. Similarly, $x_2[n]$ is represented as $\mathbf{X}_2[k,\tau]$. We have considered a two source scenario. The Elevatogram multiplies the two TF representations element-wise, $\hat{\mathbf{X}} = \mathbf{X}_1 \odot \overline{\mathbf{X}}_2$, where $\mathbf{X} \in \mathbb{C}^{K \times T}$ and $\overline{\mathbf{X}}_\mathbf{2}$ is the complex conjugate of $\mathbf{X}_2$. The phase of matrix $\hat{\mathbf{X}}$ is defined as $\angle\hat{\mathbf{X}}$. It quantizes a phase range where each level in the range is denoted by $\hat{\phi}$. These levels are the phase bins. The range is divided into L uniformly spaced levels. The levels are spaced with a gap of Δ. For each discrete frequency, k, in the TF representation, a phase histogram is constructed. This categorizes the phase content by stacking the same phase measurement in relevant bins. A set of TF bins, I_k,

 ISBN 978-0-9934207-7-1



(a) $\delta_1 = 0, \delta_2 = 1$ (b) $\delta_1 = 2, \delta_2 = 4$ (c) $\delta_1 = 16, \delta_2 = 8$ (d) $\delta_1 = 64, \delta_2 = 32$

Figure 1: The delays, δ_j are in samples. The slanting lines are observed for real speech utterances, where each line corresponds to a source, $s_j[n]$. The larger the delay, the more often the lines get phase wrapped. They indicate the energy concentration of a particular phase bin as a function of frequency.

contributing to a particular phase bin, $\hat{\phi}$, is given as: $\{I_k := [k, \tau] : |\phi_{k,\tau} - \hat{\phi}| \leq \Delta\}$. This process is repeated for all available frequencies K. The result is a phase-frequency matrix denoted by $\mathbf{P} \in \mathbb{R}^{L \times K}$, called the Elevatogram (Figure 1). Then a source can be localized in delays using:

$$\delta_{est} = -\frac{K}{L} \tan(\phi). \tag{3}$$

The two bright lines correspond to the two sources, $s_1[n]$ and $s_2[n]$. Given a sampling rate of $F_s = 16$ kHz for real speech utterances, Figure 1(a) depicts a line which horizontally bisects the plane in two halves corresponding to a source incurring a delay, $\delta_1 = 0$ samples. The other line which slants to 8 kHz corresponds to $\delta_2 = 1$ samples. This phenomenon depicts that the sources get phase wrapped at 8 kHz. A delay of one sample causes a source to get phase wrapped at $\frac{F_2}{2}$ kHz location on the frequency axis. In Figure 1(b), the sources with delays $\delta_1 = 2$ and $\delta_2 = 4$ samples get phase wrapped at 4 kHz and 2 kHz, respectively. It is also observed here that source $s_2[n]$ get the first jump at 2 kHz and the next time at 6 kHz. In this case, the 2 kHz is the fundamental phase wrap location. Likewise, two sources incurring delays of $\delta_1 = 16$ and $\delta_2 = 8$ samples having their fundamental jumps at 1 kHz and 0.5 kHz, respectively (Figure 1(c)). Doubling of delays cause the sources to be phase wrapped at fundamental locations decreased by a factor of 2 (Table 1). The parallel lines on the Elevatogram are indeed a single line, but broken into parts owing to phase wrapping. The larger the delay, more frequently the sources undergo the phase jumps. Now, inspired by the classical Hough transform [Bhuyan, 2019], the objective of the Elevatogram technique is to find a set of most significant collinear points forming a straight line. This is done by the voting procedure, parameterizing the elevatogram matrix, $\mathbf{P} \in \mathbb{R}^{L \times K}$, by distance-angle, $\rho - \phi$, to determine the accumulator cell value that receives the highest vote. With increased delays and multiple constituent sources in the mixture, it becomes difficult to identify the set of collinear points.

Delays (samples)	Frequency (kHz)
1	$\frac{F_s}{2}$
2	$\frac{F_s}{4}$
4	$\frac{F_s}{8}$
8	$\frac{F_s}{16}$
16	$\frac{F_s}{32}$
32	$\frac{F_s}{64}$
64	$\frac{F_s}{128}$

Table 1: Doubling of the delay causes a source to be wrapped at the fundamental frequency location, which decreases by a factor of 2.

This paper proposes a strategy to inpaint the set of most significant collinear points on the Elevatogram.

2 Method

Pre-processing of data includes skeletonizing the elevatogram, $\mathbf{P}$. We smooth it using a 2×2 average mask. Performing morphological filtering on it, we thin down the lines (Figure 2). Lastly, applying spur operation empirically, lines branching from the delay-lines, shorter than 5 pixels in length are removed. The portions excluded contain no significant information. The minimized skeleton is of few pixel thick. These lines slant diagonally from left to right as shown in the last diagram of Figure 2. We call these as *delay-kinesics*. Owing to increase in delays and less power, a_j, of the sources, these lines tend to get mishaped. Our goal is retrieving these delay-kinesics by inpainting these minimized lines on the Elevatogram. Using chain coding,

 ISBN 978-0-9934207-7-1

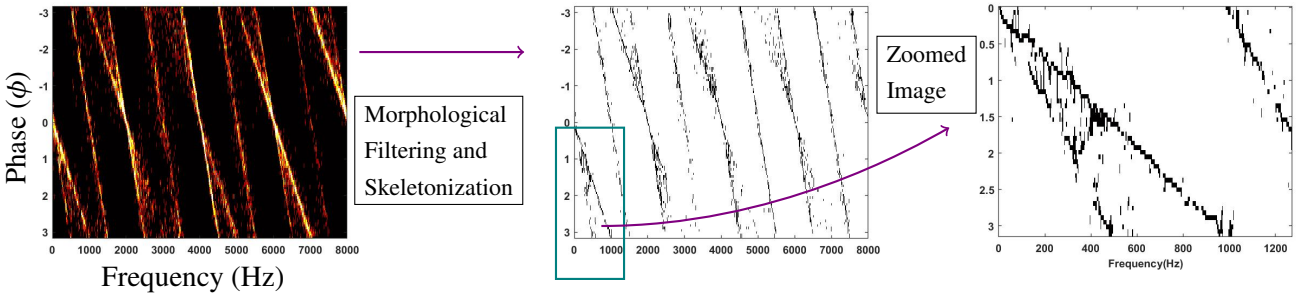


Figure 2: The middle figure is the grayscaled version of the Elevatogram (LHS figure) after morphological filtering. In the RHS figure, we zoom in the portion marked in green on the middle figure. Here, the two sources get phase wrapped first at locations 1000 Hz and 500 Hz. These lines are known as delay-kinesics.

we find the direction of the delay kinesics. We consider the Elevatogram to be a part of the fourth quadrant of a Cartesian coordinate system. Our move starts from the origin, traversing down to the boundary of the matrix along the slanted line. Inspired by Directional Freeman Chain Code of Eight Directions (DFCCE) [Bhuyan, 2019], we divide each state of our move, from '0' to '7' directions. The next move can be in any one of these directions (Figure 3). If the total number of moves required to reach the edge of the Elevatogram is M, and each level has seven Degrees of Freedom (DOF), this makes the aggregate available DOFs equal to $7 \times M$. Our proposed approach quantitizes the Elevatogram by this quantity resulting in limited resolution for delay estimation. The drawback of the quantization method is that two straight lines located very close to one another as they reach the frequency axis, shall be difficult to be discriminated. Our proposed method shall consider these two as a single line resulting in erroneous source localization. For achieving a straight line approximation, we start at any random coordinate near the origin that contains a "dark" pixel. Then we search for a neighbouring "dark" pixel within a sub-matrix of dimensions 2×2. Searching within a small enclosure ensures that the lines bifurcate. This is done in parallel. Each crawl along the line directions corresponds to a unique delay-kinesic. If we are unable to find a neighbouring dark pixel within this sub-matrix, we increase its size to maximum of 5×5. If we are unable to find it within this, then we assume that no significant straight line exists in that direction. We change our direction of search. This implies that the image pre-possessing steps do not split a line by more than 5 pixels. Once we reach the boundary of the elevatogram, the corresponding pixel is connected to the origin using a straight line approximation. Discrimination between the two sources is based upon the assumption that they are at least apart by 40 pixels horizontally at $\phi = 2$ rads on Figure 3. Our objective is to construct an inpainted Elevatogram. We observe in the original Elevatogram, Figure 1(c), that a set of parallel lines correspond to a source. Once we derive Figure 4(a), we calculate the slope, m, pertaining to the two sources in it. The subsequent parallel line starts where the previous line ends. The coordinates of this line is computed using the formula: $m = \frac{y_2 - y_1}{x_2 - x_1}$. The only unknown is y_2. We notice that $x_1 = 1$ and $x_2 = L$. The subsequent parallel lines are calculated in the same procedure and then they are inpainted in a separate matrix as depicted in Figure 4(b). We thicken each line correspondin to a source in the inpainted Elevatogram by 8 pixels.

3 Experiments and Results

Experiments are conducted using real speech utterances from the TIMIT database. They are sampled at a rate of F_s = 16 kHz, [Garofolo, 1993]. A K-sample FFT Hamming window is used where $K = 2048$. The number of phase quantization levels are $L = 100$. Keeping the window length higher is to obtain a significant differential between the two lines as they fall on the frequency axis. In Figure 3, the lines located at 500 Hz and 1 kHz could not have been appropriately ascertained if K was small. It would have looked liked a contorted phosphorescence of high intensity pixels. A straight line approximation of the

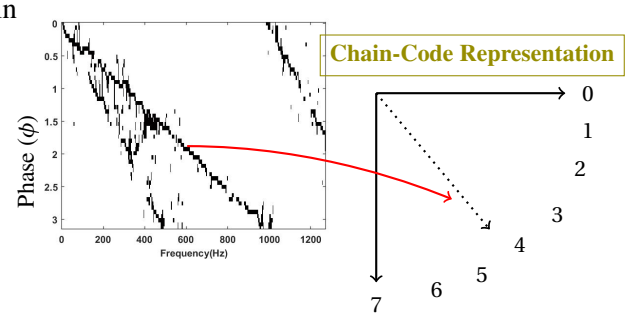
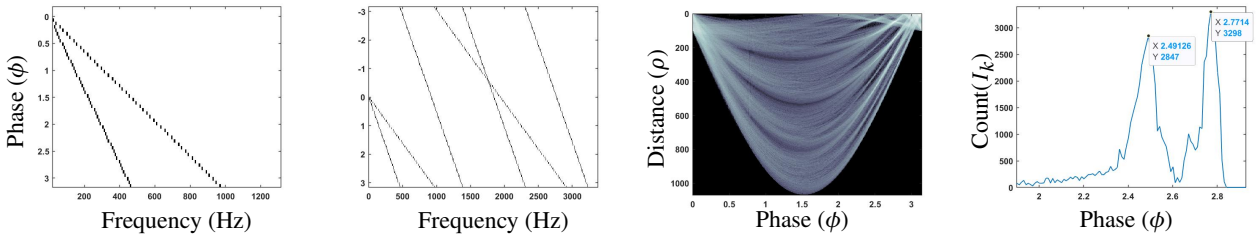


Figure 3: Each code in $0-7$ corresponds to a particular direction.

 ISBN 978-0-9934207-7-1

sources is given in Figure 4. In our mixture, we have considered actual delays as $\delta_1 = 16$ and $\delta_1 = 8$. We perform the traditional Hough transform for the inpainted image (Figure 4(b)). The prominent peaks are at $\phi = 2.49$ and $\phi = 2.77$ rads (Figure 4(c) and Figure 4(d)). Now, substituting these values in Eqn. 3, we get $\delta_{est_1} = -\frac{2048}{100}\tan(2.49) = 15.6$ samples and $\delta_{est_2} = -\frac{2048}{100}\tan(2.77) = 7.94$ samples. The error variance between actual δ and estimated δ_{est} are 0.4 and 0.06 samples, respectively.



(a) Delay Kinesic (b) Inpainted Elevatogram (c) Hough's Transform (d) Peaks Location

Figure 4: From the delay-kinesics in (a), we develop the inpainted Elevatogram (b). Hough Transform of (b) yields the accumulator matrix in (c). Two most prominent peaks are observed in (d) at $\phi = 2.49$ and 2.77 rads.

4 Conclusion and Discussions

This paper deals in acoustic source localization using image processing techniques. We have performed the same experiments by down-sampling the speech utterances from a sampling rate of $F_s = 16$ kHz to $F_s = 4$ kHz. The location of the fundamental phase jump remains the same and is unchanged with the sampling rate. They are invariant to the attenuation coefficients of the source signals. Our technique can be applied to speech signals of low sampling rates. The limitation of the chain-coded Elevatogram compared to the original tiled-Elevatogram is that it is quantizes a set of delays. This results in a limited resolution for delay estimation. Unlike the tiled-Elevatogram, our chain-code approach performs unsatisfactorily for high delay estimates as the lines get so steep that it is impossible to differentiate the sources.

Acknowledgments

This paper has emanated from research supported in part by a Grant from Science Foundation Ireland under Grant number 18/CRT/6222 and Grant Number 15/SIRG/3459.

References

[Bhuyan, 2019] Bhuyan, M. K. (2019). *Computer vision and image processing: Fundamentals and applications*. CRC Press.

[Chen et al., 2004] Chen, J., Huang, Y., and Benesty, J. (2004). Time delay estimation. *Audio signal processing for next-generation multimedia communication systems*, pages 197–227.

[de Fréin, 2017] de Fréin, R. (2017). Tiled time delay estimation in mobile cloud computing environments. In *IEEE ISSPIT*, pages 282–287.

[de Fréin and Rickard, 2011] de Fréin, R. and Rickard, S. T. (2011). The synchronized short-time-Fourier-transform: Properties and definitions for multichannel source separation. *IEEE Trans. Sig. Proc.*, 59(1):91–103.

[Garofolo, 1993] Garofolo, J. S. (1993). TIMIT acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*.

Integrating feature attribution methods into the loss function of deep learning classifiers

James Callanan[1], Carles Garcia-Cabrera[2], Niamh Belton[1], Gennady Roshchupkin[3], Kathleen M Curran[1]

[1] *University College Dublin,* [2] *Dublin City University,* [3] *Erasmus University Rotterdam*

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/OMXA8857

Abstract

Feature attribution methods are typically used post-training to judge if a deep learning classifier is using meaningful concepts in an input image when making classifications. In this study, we propose using feature attribution methods to give a classifier automated feedback throughout the training process via a novel loss function. We call such a loss function, a heatmap loss function. Heatmap loss functions enable us to incentivize a model to rely on relevant sections of the input image when making classifications. Two groups of models were trained, one group with a heatmap loss function and the other using categorical cross entropy (CCE). Models trained with the heatmap loss function were capable of achieving equivalent classification accuracies on a test dataset of synthesised cardiac MRI slices. Moreover, HiResCAM heatmaps suggest that these models relied to a greater extent on regions of the MRI slices within the heart. A further experiment demonstrated how heatmap loss functions can be used to prevent deep learning classifiers from using non-causal concepts that disproportionately co-occur with images of a certain class when making classifications. This suggests that heatmap loss functions could be used to prevent models from learning dataset biases by directing where the model should be looking when making classifications.

Keywords: Loss function, Dataset bias, Grad-CAM, HiResCAM, Deep learning

1 Introduction

Many feature attribution methods are differentiable with respect to the network's weights and biases. This makes it possible to integrate them into a model's loss function. Models were successfully trained with both Grad-CAM (Selvaraju et al., 2017) and HiResCAM (Draelos and Carin, 2020) integrated into their loss functions. However, the heatmap loss function used in the experiments below consisted of a weighted sum of a HiResCAM component and a mean squared error (MSE) component. The HiResCAM component served to disincentivize the classifier from relying on irrelevant portions of images when making classifications and the MSE component acted to incentivize the model to make correct class classifications. A training, validation and testing dataset of synthetic cardiac MRI slices were generated along with their corresponding segmentation masks. The areas of the MRI slices outside of the heart were deemed irrelevant for cardiac disease classifications. Consequently, the HiResCAM component of the loss function was set equal to the sum of the HiResCAM heatmap values that lay outside of the heart. Many other metrics have been proposed to evaluate the degree of overlap between feature attribution maps and segmentation masks in segmentation problems such as Dice (1945). There is potential for these to be adapted for use in a heatmap loss function.

 ISBN 978-0-9934207-7-1

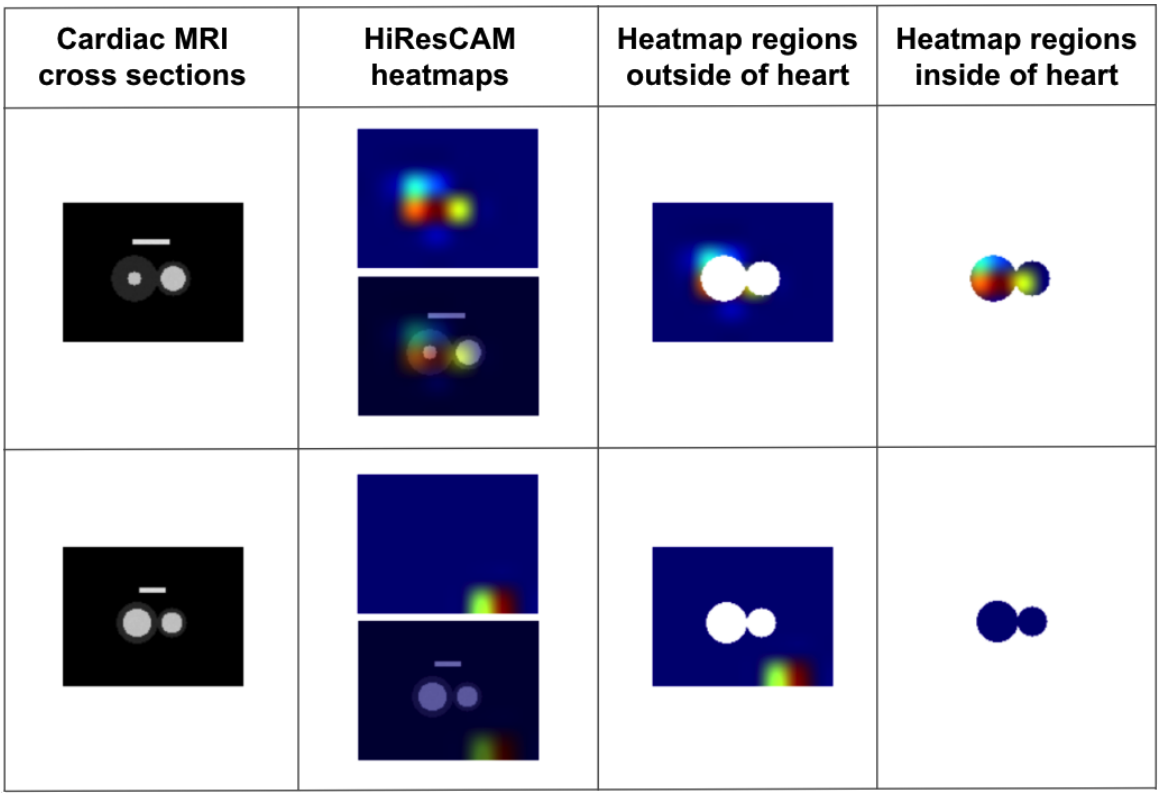


Figure 1: **Visual explanation of the heatmap loss component:** The heatmap loss component is calculated by summing the portion of the HiResCAM heatmap that lies outside of the heart (shown in column 3 above). This definition is imperfect as both MRI slices above incur equivalent heatmap losses despite the classifier relying on more information within the heart in the top MRI slice. Thus, there is likely scope to define a better heatmap loss function.

2 Methods and Results

Balanced datasets of MRI slices through the center of the heart were generated. These slices were meant to mimic short axis cardiac MRI cross sections through the center of the heart. Four classes of cardiac MRIs were generated, these included; normal, hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM) and arrhythmogenic right ventricular cardiomyopathy (ARV). Attempts were made to make the synthetic datasets representative of a real world cardiac dataset by injecting noise and taking into account disease biomarkers, aetiology and sex prevalence. Below are sixteen sample MRI slices taken from the training dataset used in the first experiment, these correspond to sixteen different ‘patients’.

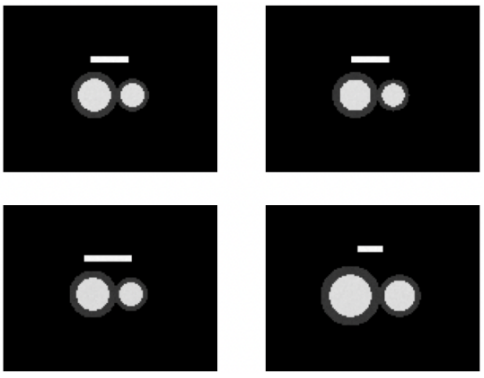


Figure 2: Exp 1: Normal MRI slices

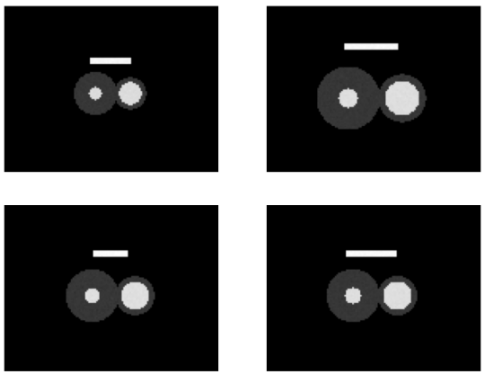


Figure 3: Exp 1: HCM MRI slices

 ISBN 978-0-9934207-7-1

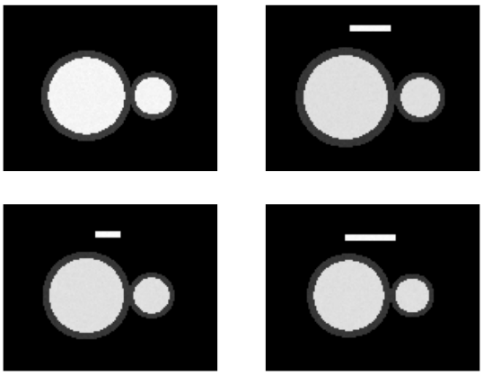


Figure 4: Exp 1: DCM MRI slices

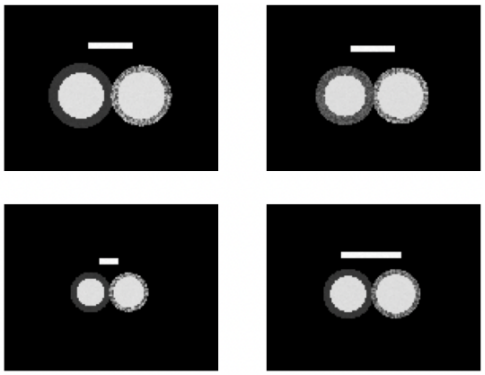


Figure 5: Exp 1: ARV MRI slices

In the first experiment, ~30 models were trained with a heatmap loss function and ~30 models with CCE. All models had identical VGG16 inspired architectures. Learning rates were varied using Keras-Tuner within a range that exhibited good convergence of training and validation loss and accuracies. Of the ~60 models trained, models with a classification accuracy of less than 95% on a validation dataset were discarded. This meant 23 models trained with a heatmap loss function and 25 models trained with CCE remained. HiResCAM-heart overlap metrics were computed on a test dataset comprising of 300 MRIs for models in both groups. The results of which can be seen in Figure 6. A Shapiro-Wilkes test confirmed the distributions of the overlap metrics were not normally distributed among groups. Consequently, a two-sided Mann-Whitney U-test ($\alpha = 0.05$) was performed to test for a statistically significant difference between the group's overlap metrics. The models trained with the heatmap loss function were found to have systematically higher degrees of heatmap-heart overlap, with a p-value $\approx 1 \times 10^{-9}$.

A second experiment was carried out to test whether models would rely on knowledge of the patient's sex when making classifications. It was hypothesized that a model may base classifications off knowledge of a patient's sex because many cardiac diseases occur disproportionately among the sexes. For example, for every female case of ARV there are ~2.7 male cases. The datasets used in these experiments were designed to mirror real world differences in disease prevalence among the sexes. In this experiment, all systematic differences between the MRIs of males and females were removed (i.e. size and body fat's sex dependence). However, a label was included in the bottom corner of male patient's MRIs to distinguish them. This enabled us to separate the concept of sex from the heart. Thus, we could test for a model's reliance on sex when making classifications by calculating the degree of overlap between the HiResCAM heatmap and sex label. Approximately 50 models were trained with both loss functions, after discarding those with a validation accuracy <95%, 22 models trained with the heatmap loss function and 23 models trained with the CCE loss function remained. HiResCAM-heart overlaps as well as HiResCAM sex label overlaps were computed on a test dataset of 300 MRIs. Statistically significant differences were found in the distributions of the heatmap-heart and heatmap-sex label overlaps among both groups. The models trained with the heatmap loss function had higher degrees of heatmap-heart overlap (p-value $\approx 1 \times 10^{-8}$) as can be seen in in Figure 7. These models also had lower degrees of heatmap-sex label overlap (p-value $\approx 1 \times 10^{-7}$).

Perfect classification accuracy was achieved by models in groups across all experiments on an unseen testing dataset.

 ISBN 978-0-9934207-7-1

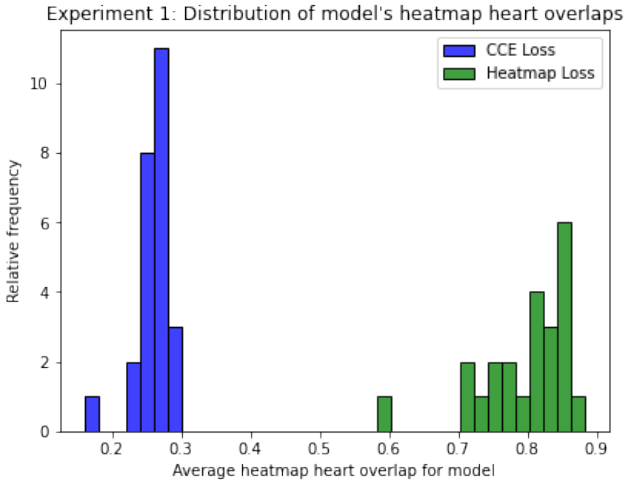


Figure 6: Exp 1: Distribution of overlaps

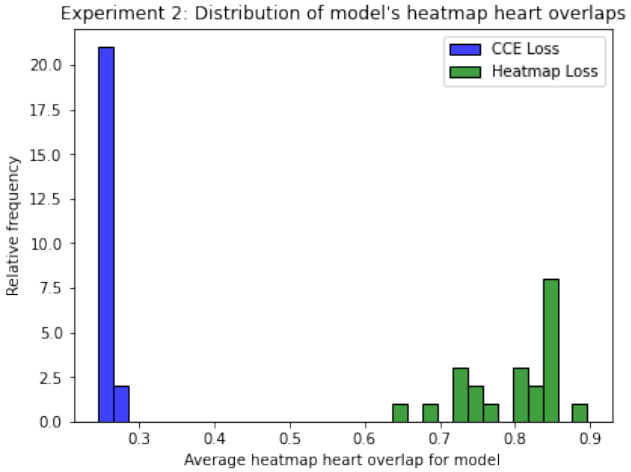


Figure 7: Exp 2 (sex label): Distribution of overlaps

3 Discussion and Conclusion

We have demonstrated that a model trained with a heatmap loss function can achieve high classification accuracies on a synthetic dataset. However, we envision this loss function being used to overcome an issue that affects existing high performing classifiers, the issue of learned bias (as outlined by Kim et al. (2018), Selvaraju et al. (2017) and Ghorbani et al. (2019)). Eliminating learned bias using techniques such as data augmentation, oversampling and undersampling is likely infeasible, if not impossible. The disproportional co-occurrence of non-causal concepts within images of a given class seems inevitable, especially when the set of possible concepts that a classifier can detect is extremely large. Thus, we believe heatmap loss functions warrant further investigation. Future research should test the feasability of heatmap loss functions on a real world dataset. Ideally the chosen dataset would be large enough to train deep learning classifiers using conventional loss functions and would contain both classification labels and segmentation masks. Moreover, several obstacles need to be investigated further such as; the intrinsic limitations of the feature attribution methods used, the requirements of regions to be separable from the object being classified and the increased training times required when using a heatmap loss function. The code associated with this project along with a more in depth discussion can be found at https://jamescallanan.github.io/HeatmapLossFunction.

References

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Draelos, R. L. and Carin, L. (2020). Hirescam: Faithful location representation in visual attention for explainable 3d medical image classification. *arXiv preprint arXiv:2011.08891*.

Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. (2019). Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Distance measurement between smartphones within an ad-hoc camera array, using audible PRBS

Pádraic McEvoy, Damon Berry, Ted Burke

Biomedical Research Group, Technological University Dublin

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/KEBP4512

Abstract

An approach for measuring the distance between two smartphones is presented in this paper. The method uses each smartphone's microphone(s) and speaker(s) to concurrently emit and record audio in order to calculate the sound propagation delay and hence distance. Each device in turn emits a different audible pseudo-random binary sequence (PRBS) - specifically, a maximum length sequence (MLS). Each device captures both emitted signals in one continuous recording. The propagation delay between the devices is calculated by comparing their respective recordings, and in particular the temporal positions of the emitted signals within each recording. Each device emits one of the signals, records both signals, and then sends its recording to a master device for analysis, which is performed by a custom web application and is therefore independent of operating system. A mean error of 32.29 mm was found in initial testing, which was conducted using Samsung Galaxy A10 devices running Android 10. The key innovation in this method is that it requires no clock time synchronisation between devices because the distance is determined by comparing inter-transmission delays in the two recordings. Potential future improvements are discussed, including how to take into account the exact locations of each phone's microphone and speaker to increase accuracy.

Keywords: ad-hoc camera array, smartphone, calibration, PRBS, audio.

1 Introduction

The increasing availability and hardware capabilities of smartphones make them excellent candidates for use within various sensor arrays, particularly within camera arrays, as described in this study. It is common knowledge that the speed and propagation delay of a signal within a medium can be utilised to determine how far the signal has travelled. In this paper, we describe a method for measuring the distance between two smartphones using audio signals that are broadcast by each device and captured by the nearby partner device(s). The difference in propagation delays acquired at each smartphone is used to compute the distance between them. The method requires that each smartphone within the array be capable of producing and recording audio, that the devices are within audible range of one another, and that each device can connect to the internet. No time synchronisation between device clocks is required because all timing is relative to the devices in the array and synchronisation information is present in the audio. Each smartphone emits its own maximum length (MLS) pseudo-random binary sequence (PRBS) signal. The cross-correlation of each recorded audio with template copies of the PRBS signals results in correlation peaks at the moments that each signal arrives at a specific device. The variation in the timing of arrival of the audio PRBS signals reveals the distances between devices within the array and, due to the unique PRBS signatures, identifies each device.

2 Related work and problem statement

A number of previous studies have estimated the distance between and/or position of smartphones [Orujov et al., 2018, Chen et al., 2019, Dinh et al., 2020, Li et al., 2020], each showing varying accuracy ranging between 0.3 and 0.8 metres for distances from 0 and 20 metres. There have also been investigations into the use of smartphones as distributed camera arrays for various applications [Latimer et al., 2014, Wang et al., 2015]. These systems use additional hardware or software, such as additional static receiver nodes throughout the area of investigation.

Consider the case of a multi-camera array where the location of each recording device is uncertain. Many multi-camera array applications require the user to carefully position the cameras, which can take time. A system that detects relative locations using the device's built-in hardware without the need for software or hardware modifications would be beneficial. If the absolute distance between each pair of recording devices can be measured, it is possible to estimate the relative positions of these devices. In this case, it is assumed that all of the smartphones can concurrently record and transmit sound and that they are within audible range of one another. Clearly, the exact position of a smartphone's speaker, microphone, and camera vary depending on the model, but this was ignored in this experiment.

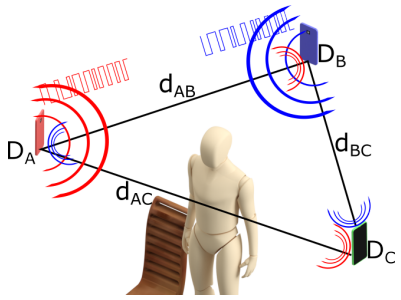


Figure 1: Ad-hoc smartphone camera array (positions initially unknown). The distance between devices may be calculated with the proposed approach. Absolute distances can be used to determine relative positions.

3 Theory and proposed approach

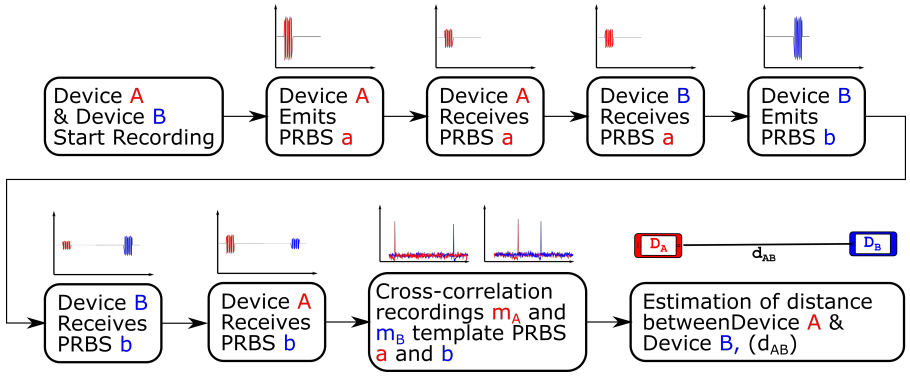


Figure 2: Block diagram showing how the approach uses the difference in received propagation times from two or more devices to measure the distance between them. The difference in propagation times is determined by cross-correlation between received audio and template copies of each emitted audible PRBS signal.

This section explains how the audio recorded by each smartphone in a pair can be used to calculate the distance between them. The average speed of sound in air, c, is $343\,\mathrm{m\,s^{-1}}$. The sampling frequency used on each smartphone's microphone, f_s, is 48 kHz and the sampling period, T_s, is $\frac{1}{f_s}$. Ideally, the microphone and speaker within each smartphone would be positioned very close to each other, but this varies with smartphone make and model. A known source of error in this investigation was that the midpoint between speaker and microphone was assumed to be the point from which sound was emitted and at which incoming sound was recorded. A different PRBS was assigned to each device. Each used PRBS was a maximum length sequence (MLS), generated by an 8-bit Fibonacci linear feedback shift register (LFSR) with taps set by various primitive polynomials of degree 8, resulting in a sequence 255 bits long. A bit repeat factor, $\delta = 3$, was applied, which effectively set the LSFR clock frequency to 16 kHz and increased the length of each sequence to 765 bits, when sampled at f_s.

$$p(n) = MLS_8\left(\left\lfloor \frac{n}{3} \right\rfloor\right) \tag{1}$$

So,
$$a(n) = MLS_{8[8,4,3,2]}\left(\left\lfloor \frac{n}{3} \right\rfloor\right) \tag{2}$$
$$b(n) = MLS_{8[8,5,3,1]}\left(\left\lfloor \frac{n}{3} \right\rfloor\right) \tag{3}$$

The period of the generated PRBS is
$$T_{prbs} = \frac{2^n - 1}{f_{LFSR}} = 15.938\,\mathrm{ms} \tag{4}$$

The assigned PRBS is played in a three-loop segment, in order to generate a steady-state response, and so

$$3NT_s = 47.8125\,\mathrm{ms} \tag{5}$$

 ISBN 978-0-9934207-7-1

The raw audio is captured as discrete-time 32-bit pulse-code modulation (PCM) signals from each device and sent to the web server for logging. Each of these signals include time-shifted and attenuated duplicates of the PRBS audio streams with added noise. The PCM signals are cross-correlated with pure models of each PRBS to calculate time lags $l_{A(a)}$, $l_{A(b)}$, $l_{B(a)}$ and $l_{B(b)}$. The resulting cross-correlation signals, $g_{Aa}(l)$, $g_{Ab}(l)$, $g_{Ba}(l)$ and $g_{Bb}(l)$ are defined as

$$g_{Dp}(l) = \sum_{n=0}^{N-1} m_D(n+l)\,p(n) \tag{6}$$

Where $D \in \{A, B\}$, shown in Figure 1, and $p \in \{a, b\}$. A prominent peak is observed in each $g_{Dp}(l)$ at $l_{D(p)}$ representing the propagation time between each speaker and microphone, shown in Figure 3. The largest magnitude value in each one of the cross-correlation signals' lag values, measured in samples, is found to be as follows

$$l_{D(p)} = \arg\max_{l} g_{Dp}(l) \tag{7}$$

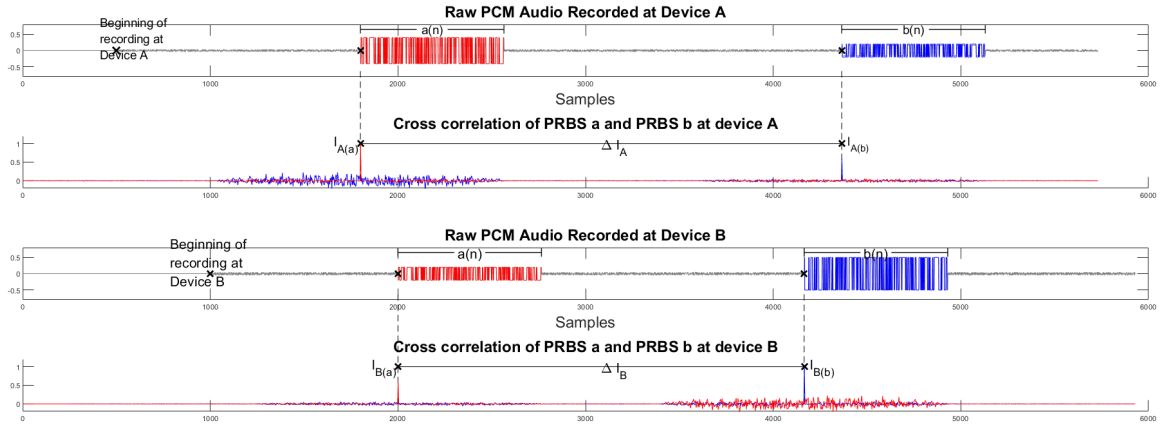


Figure 3: Representation of the raw audio recorded at each device and corresponding cross-correlation peaks. Notice how each device begins recording at a different time, showing that the approach described is clock time independent.

And so, as seen in Figure 3

$$\Delta l_A = l_{A(b)} - l_{A(a)} \tag{8}$$

$$\Delta l_B = l_{B(b)} - l_{B(a)} \tag{9}$$

$$\tau_A = \Delta l_A T_s \tag{10}$$

$$\tau_B = \Delta l_B T_s \tag{11}$$

Where Δl_A is the discrete time difference between the instants $l_{A(a)}$ and $l_{A(b)}$. The distance between device A and device B can be determined by:

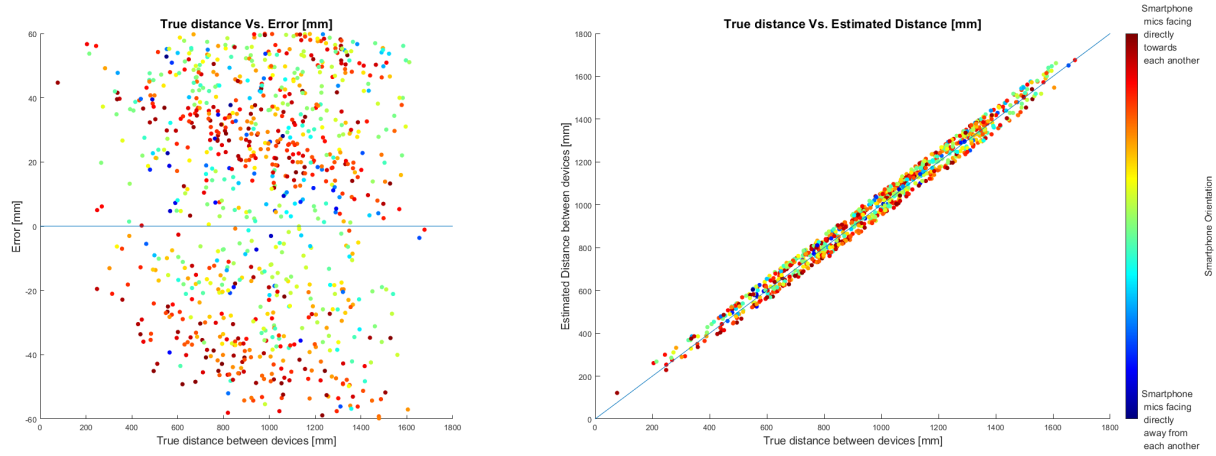
$$d_{AB} = \left(\frac{\tau_A - \tau_B}{2}\right) c \tag{12}$$

Where c is the propagation speed of the signal (average speed of sound in air, $343 m/s$).

4 Experiment and results

To test the accuracy of the method described, an experiment was conducted within a lab setting. A specifically developed web application was opened on three Samsung Galaxy A10 smartphones and a laptop. Three robots received commands from the web application, running on the laptop, via Bluetooth to move the phones to random locations within the investigation area. All smartphones sequentially emitted their assigned audible PRBS signals, while capturing continuous audio. This enabled the capture of a wide range of data in a short space of time. The area of investigation was limited to 1.8 x 1.8 metres by the equipment used. Images were gathered by a camera placed above the investigation area, with colour point detection used to determine ground truth distance information.

 ISBN 978-0-9934207-7-1



(a) True distance Vs Error showing the orientation of each smartphones microphone

(b) True distance Vs Estimated distance, showing the orientation of each smartphones microphone

Figure 4: Results from 895 sets of recordings showing errors between 0.1 - 59.5mm. (mean error of 32.29mm)

5 Conclusions, limitations and future work

Figure 4 shows the results from the experiment described above. Distance errors between 0.1 and 59.5 mm can be observed. A colour map is used to visualise the orientation of each smartphones microphone, relative to its partner device. The authors believe there are several potential sources of inaccuracy, which will be further investigated. It may be possible to reduce inaccuracies that occur due to cross-correlation and the audio sampling frequency used by sub-sample measurement. Information collected from sensors, such as inertial measurement units (IMU) can determine the orientation of the smartphones, giving insight into the difference in location between embedded microphone, speaker and camera allowing for accuracy improvements.

Current results suggest that the technique performs well in a real-world setting. This could allow smartphones to be used within ad-hoc multi-camera arrays, even where the smartphones are either occluded or not within view of other cameras within the array. Further study is needed for calibration of the smartphone microphone and speaker position variation, sampling frequency or sub-sample measurement to impact future accuracy. Planned future work also includes investigation into integrating the system with machine vision methods using smartphone cameras.

References

[Chen et al., 2019] Chen, P., Liu, F., Gao, S., Li, P., Yang, X., and Niu, Q. (2019). Smartphone-based indoor fingerprinting localization using channel state information. *IEEE Access*, 7:180609–180619.

[Dinh et al., 2020] Dinh, T.-M. T., Duong, N.-S., and Sandrasegaran, K. (2020). Smartphone-based indoor positioning using ble ibeacon and reliable lightweight fingerprint map. *IEEE Sensors Jnl*, 20(17):10283–94.

[Latimer et al., 2014] Latimer, R., Holloway, J., Veeraraghavan, A., and Sabharwal, A. (2014). Socialsync: Sub-frame synchronization in a smartphone camera network. In *European Conference on Computer Vision*, pages 561–575. Springer.

[Li et al., 2020] Li, P., Yang, X., Yin, Y., Gao, S., and Niu, Q. (2020). Smartphone-based indoor localization with integrated fingerprint signal. *IEEE Access*, 8:33178–33187.

[Orujov et al., 2018] Orujov, F., Maskeliūnas, R., Damaševičius, R., Wei, W., and Li, Y. (2018). Smartphone based intelligent indoor positioning using fuzzy logic. *Future Generation Computer Systems*, 89:335–348.

[Wang et al., 2015] Wang, Y., Wang, J., and Chang, S.-F. (2015). Camswarm: Instantaneous smartphone camera arrays for collaborative photography. *arXiv preprint arXiv:1507.01148*.

Triple Loss based Satellite Image Localisation for Aerial Platforms

Eduardo A. Avila H., Tim McCarthy, John McDonald

National University of Ireland Maynooth

2022 | 24th Irish Machine Vision and Image Processing Conference (IMVIP) | DOI: 10.56541/PJFN5642

Abstract

We present a vision-based technique for aerial platform localisation using satellite imagery. Our approach applies a modified VGG16 network in conjunction with a triplet loss to encode aerial views as discriminative scene embeddings. The platform is localised by comparing the encodding of its current view with a database of pre-encoded embeddings using a cosine similarity metric. Recent image based localisation research has shown potential for such learned embeddings, however, to ensure reliable matching they require dense sampling of views of the environment, thereby limiting their operational area. In contrast, the combination of our proposed architecture in conjunction with the triplet loss shows robustness over greater spatial shifts, reducing the need for dense sampling. We demonstrate these improvements through comparison with a state-of-the-art approach using simulated ground truth sequences derived from a real-world satellite dataset covering a 1.5km × 1km region in Karslruhe.

Keywords: Satellite, Localisation, Machine Learning,Computer Vision, Triplet Networks

1 Introduction

Localisation of mobile platforms has been dominated by Global Navigation Satellite Systems (GNSS) such as the Global Positioning System (GPS) [Kaplan, 2017]. Although, advancements in electronics have allowed such systems to become ubiquitous, their dependency on satellite availability results in an accuracy that varies as a function of location. Furthermore, these systems are also susceptible to noise, multi-path effects and signal jamming.

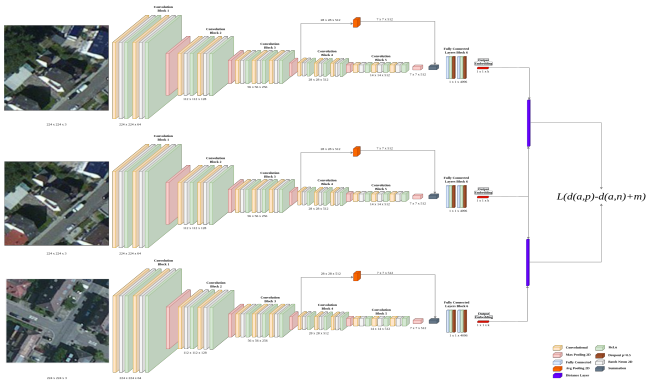


Figure 1: Network architecture visualisation consisting of three modified VGG16 CNN branchs (shared weights). Top branch takes as input a positive pair, middle branch takes the anchor satellite image and bottom branch a negative pair image. The outputs are used to calculate the similarity metric between the image pairs.

Such issues have motivated researchers to develop alternative approaches for localisation using exteroceptive sensors such as optical sensors [Gianpaolo Conte, 2008, Bay et al., 2008] or LiDAR [Shan, 2018]. These approaches require algorithms for solving the data association problem, involving relating onboard sensor measurements to an a-priori global model of the environment. Recent work in this area has explored the potential of machine learning techniques to compute compact representations of complex real-world sensor data [Kim, 2017, Sixing Hu, 2018, Regmi, 2019, Bianchi, 2021].

In this paper we propose a localisation technique based on a modified VGG-16 network architecture trained using a triplet loss, where a cosine distance is employed as the similarity metric between images. Once trained, a localisation likelihood is predicted over extended areas leveraging the use of a pre-encoded database. A high-level view of our approach is shown in Fig. 1.

 ISBN 978-0-9934207-7-1

2 State of the Art

Localisation is a central problem in mobile robotics as platforms require an accurate estimate of their pose for robust planning and navigation. Although the problem has been addressed for sensor types ranging from sonar to LiDAR, given their low cost and wide availability, current research in the area is dominated by visual sensors. Classic computer vision techniques have been employed to solve localisation of platforms over extended areas. Algorithms such as SIFT [Lowe, 1999], SURF [Bay et al., 2008] or SfM [Sattler et al., 2011] [Schonberger and Frahm, 2016] have been employed for image based geo-location through feature extraction and image matching techniques [Torii et al., 2011]. However, these approaches can be computationally expensive, and lack robustness to extreme variations in appearance, for example, lighting and weather changes.

To overcome these challenges recently researchers have focused on the potential of deep learning approaches to compute discriminative image representations suitable for localisation tasks. Convolutional neural networks (CNNs) have been explored to geo-locate ground [Vo Nam N, 2016] [Kim, 2017] [Sixing Hu, 2018] and aerial platforms posing the problem as an image matching task [Bianchi, 2021] [Regmi, 2019]. The combination between triplet loss and CNN architectures has also been explored in re-identification problems [Hermans et al., 2017]. Many techniques utilise a pre-defined database of image encodings to reduce the online computational requirements [Hays, 2008] [Bianchi, 2021]. Unfortunately, these approaches are still sensitive to spatial shifts in the input resulting in a requirement for dense sampling of the environment. We aim to address this problem by increasing matching robustness over larger spatial shifts thereby improving the localisation robustness and reducing required sampling densities.

3 Approach & Experimental Results

In our approach we estimate the location likelihood of a platform using a modified version of VGG16 presented by [Kim, 2017]. The network is trained with a triplet margin loss [Veit et al., 2017] [Ren, 2019] [Hermans et al., 2017] based on a cosine distance between an anchor, positive and negative triple as shown in Eq.(1) and Eq.(2), respectively.

$$L(\mathbf{E}_a, \mathbf{E}_p, \mathbf{E}_n) = \max(d(\mathbf{E}_{a_i}, \mathbf{E}_{p_i}) - d(\mathbf{E}_{a_i}, \mathbf{E}_{n_i}) + m, 0) \tag{1}$$

$$d(\mathbf{E}_a, \mathbf{E}_v) = 1 - \frac{\mathbf{E}_a \cdot \mathbf{E}_v}{\max(|\mathbf{E}_a|_2 \cdot |\mathbf{E}_v|_2, \epsilon)} \tag{2}$$

Here, $\mathbf{E}_i$ is the embedding for image i, and the indices a, p, and n refer to the anchor, positive, and negative images, respectively. The use of the triplet margin loss helps the network learn a discriminative feature space, where the distance between embeddings for similar images is minimised through the postive contribution of the distance between the positive pairs in the loss. Similarly the loss maximimises the distance between dissimilar embeddings through the negative contribution from non-matching pairs. The network architecture is composed of five convolution blocks as shown in Figure 1. An additional 2D batch normalization layer is added before every ReLu layer in the convolution blocks. Finally, a dropout layer is added after the ReLu layers in the fully connected block.

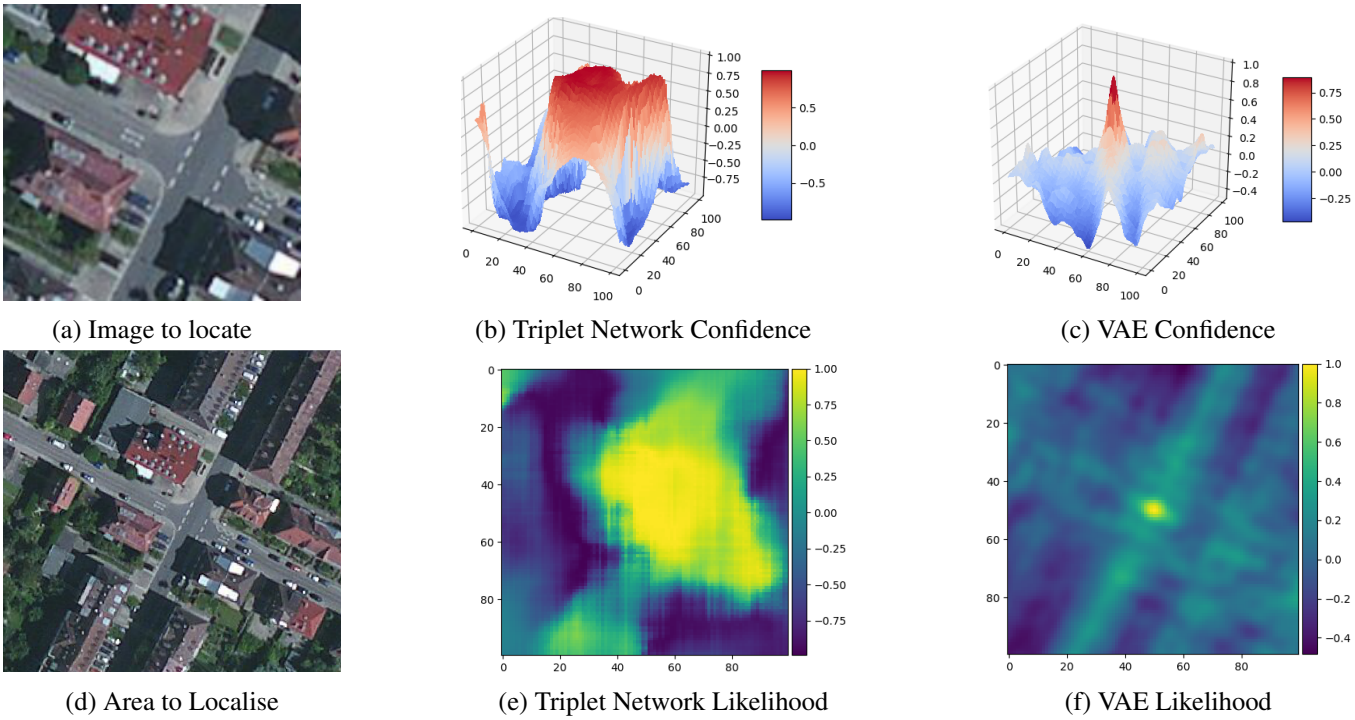
To train the network we utilise satellite imagery taken over Karlsruhe, Germany, at a ground sampling density (GSD) of 10cm. Separate training (2km × 2km) and testing (1.5km × 1.3km) areas are extracted. For training, ~114K triples are generated from a set of sub-images randomly sampled in both position and orientation over the training area. Once trained, the test area of 1.5km by 1.3km is sampled every 5m vertically and horizontally. For each sample a 50m by 50m satellite image is pre-encoded using the resultant network. To geo-locate an image it is first encoded with the network and then the resulting embedding is compared against the pre-computed database applying the cosine similarity, thereby calculating the localisation likelihood.

The localisation confidence of our methodology is compared against the one calculated with the Variational Auto Encoder (VAE) proposed by [Bianchi, 2021]. Here, the VAE architecture from [Bianchi, 2021] which used gray scale input images, is extended to encode RGB images. The localisation confidence is calculated over a 100m × 100m area with 1m shifts as shown in Figure 2, and over a 500m × 500m area with 5m shifts, as shown in Figure 3.

The results demonstrate a narrow peak exhibited by the VAE, showing its capable of matching only with its own embedding, not matching against other regions with similar morphology. Figure 2 exposes a dense sampling strategy required by the VAE to ensure that the sub-image itself is contained within the pre-computed database. In

 ISBN 978-0-9934207-7-1

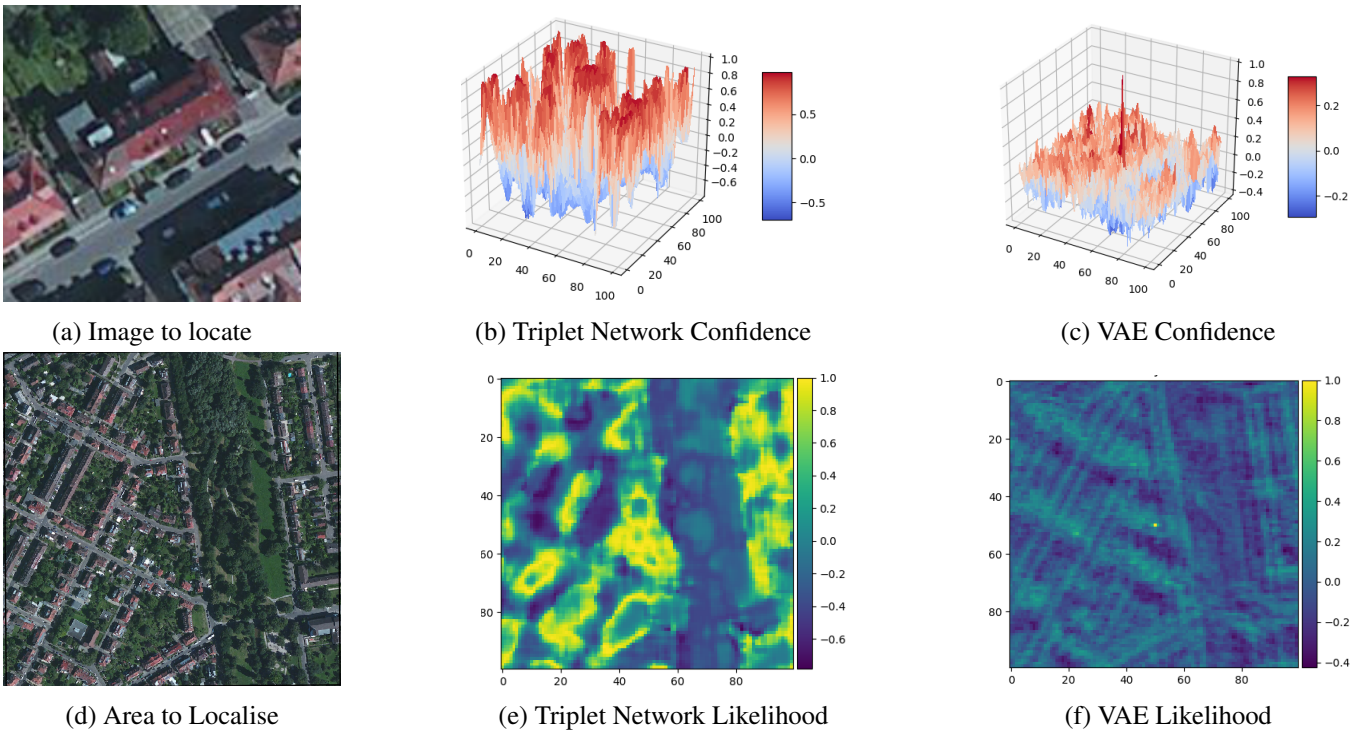
contrast with our approach a high similarity signal is output over a larger area, ensuring a strong match over larger spatial shifts, thereby allowing a significantly reduced sampling density in the pre-computed database. In Figure 3, the proposed network shows strong matches in areas of similar morphology (i.e. sub-images showing roads and buildings), whereas we see weak responses in the central strip of grass and trees.



(a) Image to locate (b) Triplet Network Confidence (c) VAE Confidence

(d) Area to Localise (e) Triplet Network Likelihood (f) VAE Likelihood

Figure 2: Localization likelihood calculated over a 100m^2 region. Samples were collected every 1m.



(a) Image to locate (b) Triplet Network Confidence (c) VAE Confidence

(d) Area to Localise (e) Triplet Network Likelihood (f) VAE Likelihood

Figure 3: Localization likelihood calculated over a 500m^2 region. Samples were collected every 5m.

 ISBN 978-0-9934207-7-1

4 Conclusion

A vision-based technique for aerial platform localisation using satellite imagery is proposed. We demonstrate the possibility of aerial localisation over large areas when applying a triplet loss function to learn image morphology. In future work we intend to integrate the output of our network within a particle filtering framework, similar to the work of [Kim, 2017].

Acknowledgments

Material supported by U-Flyte 17/SPP/3460, funded under the Science Foundation Ireland

References

[Bay et al., 2008] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359.

[Bianchi, 2021] Bianchi, B. T. D. (2021). Uav localization using autoencoded satellite images. *RA-L*, 6(2).

[Gianpaolo Conte, 2008] Gianpaolo Conte, P. D. (2008). An integrated uav navigation system based on aerial image matching. pages 1–10. IEEE.

[Hays, 2008] Hays, E. A. A. (2008). Im2gps: estimating geographic information from a single image. In *2008 IEEE CVPR*, pages 1–8. IEEE.

[Hermans et al., 2017] Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv:1703.07737*.

[Kaplan, 2017] Kaplan, Elliott D, H. C. (2017). *Understanding GPS/GNSS: Principles and applications*. Artech.

[Kim, 2017] Kim, W. M. R. (2017). Satellite image-based localization via learned embeddings. In *ICRA*. IEEE.

[Lowe, 1999] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150–1157. IEEE.

[Regmi, 2019] Regmi, S. M. (2019). Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 470–479.

[Ren, 2019] Ren, Chen Zhiyong, X. S. (2019). Triplet based embedding distance and similarity learning for text-independent speaker verification. In *APSIPA ASC*, pages 558–562. IEEE.

[Sattler et al., 2011] Sattler, T., Leibe, B., and Kobbelt, L. (2011). Fast image-based localization using direct 2d-to-3d matching. pages 667–674. IEEE.

[Schonberger and Frahm, 2016] Schonberger, J. L. and Frahm, J.-M. (2016). Structure-from-motion revisited. In *IEEE CVPR*, pages 4104–4113.

[Shan, 2018] Shan, B. E. (2018). Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In *IEEE/RSJ IROS*, pages 4758–4765. IEEE.

[Sixing Hu, 2018] Sixing Hu, Mengdan Feng, e. a. (2018). Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. pages 7258–7267. IEEE.

[Torii et al., 2011] Torii, A., Sivic, J., and Pajdla, T. (2011). Visual localization by linear combination of image descriptors. pages 102–109. IEEE.

[Veit et al., 2017] Veit, A., Belongie, S., and Karaletsos, T. (2017). Conditional similarity networks. In *IEEE CVPR*, pages 830–838.

[Vo Nam N, 2016] Vo Nam N, H. J. (2016). Localizing and orienting street views using overhead imagery. In *ECCV*, pages 494–509.

Published by the Irish Pattern Recognition & Classification Society
iprcs.org
ISBN 978-0-9934207-7-1 DOI 10.56541/iprcs_imvip2022