Valeria Olyunina 17309442 olyuninv@tcd.ie

Supervisor: Matthew Moynihan, Prof. Aljosa Smolic

**Multi-view camera synthesis for improved 3D reconstruction**

**Introduction:**

In the entertainment industry, Virtual reality (VR) and Augmented Reality (AR) are gaining popularity. However, most content available for AR/ VR is synthetic, created by artists and designers (Zhang 2004; Pages *et al.* 2018). Free Viewpoint Video (FVV) and performance capture can offer a VR/AR experience with the spatio-temporal fidelity of a live performance. However, the quality FVV and multi-view performance capture is dependant on the technology used to capture and process the input videos. Usually FVV is reconstructed from multiple camera recordings using techniques like SfM (Structure from Motion) and SfS (Shape-from-Silhouettes). For accuracy, this video is usually captured in green screen studios for easy background removal. Current efforts aim to improve the methodologies for 3D reconstruction from multiple camera views, both in studio design and “in the wild” capture scenarios (see (Pages *et al.* 2018) for FVV from mobile phone recordings). The quality of the SfM reconstruction suffers from holes where occlusions occur (S. Li *et al.* 2018), inaccuracies from lack of reference points in sparse camera setup, lack of texture, transparent or reflective features, due to camera lens, noise, camera angle (Wenger 2016). One suggested approach for improving the accuracy in case of sparse reconstructions of photogrammetry-based methods - is to provide more points of view for feature extraction via generating synthetic views (synthetic cameras). The latter can be achieved with IBR (image-based rendering) techniques, one example of which is video interpolation. Image Based Rendering (IBR) techniques can create new images directly from the existing set of images without doing a full 3D reconstruction (Zhang 2004). Recently, neural networks were shown to have better performance at computer vision tasks than previously designed procedural approaches (King 2016). This is due, in part to deep-learning being able to extract and combine tens of thousands of features from images, where a human approach would only find dozens. The deep-learning networks have been used for classification, segmentation and creation of new images/ video. IBR was previously combined with deep-learning to create arbitrary points of view when given a collection of images (Flynn *et al.* 2015). (Niklaus and Liu 2018) suggest a new video interpolation technique based on deep-learning that is context-aware and produces a high-quality interpolated frame from two temporal frames.

**Research Question:**

This MSc dissertation will check if a methodology similar to (Niklaus and Liu 2018) can be applied to frame interpolation in multi-view spatial context, rather than temporal - to create additional synthetic cameras for better 3D reconstruction. The *aim of this research* is to create an effective frame interpolation technique for multi-view synthesis using deep-learning that would improve the quality of the existing 3D model reconstruction and FVV based on human motion videos.

Several useful applications can come out from the creation of quality 3D videos of natural scenes. Apart from the use in Augmented, Virtual reality and 3D Television, the created 3D content will be annotated and so it can be used for further deep-learning training for video recognition and video prediction. This dissertation will combine several interesting areas of Computer vision, 3D reconstructions and Augmented reality fields.

The *keywords* for the research topic and methodology are:

Multi-view synthesis, Multi-View Video, Free-Viewpoint video (FVV), video frame interpolation, image-based rendering (IBR), photogrammetry, video prediction, deep-learning, neural networks, 3D reconstruction, depth estimation, structure from motion, shape from silhouette, optical flow.

For *abbreviations and definitions* relevant to the research – please see Glossary in the Appendix 1.

The primary *sources of literature* will be: Journal articles in Elsevier (Journal of visual communication and Image representation, Pattern recognition), PAMI (IEEE Transactions on Pattern Analysis and Machine Intelligence), IEEE Transactions on Image Processing, IJCV (International Journal on Computer Vision) and International conferences: ICLR (International Conference on Learning Representations), CVPR (Computer Vision and Pattern Recognition), ICCV ([International Conference on Computer Vision](http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=79303&copyownerid=105620)), NIPS (Neural Information Processing Systems), FUSION (International Conference on Information Fusion). The following websites: IEEE Xplore digital library, Web of Science, arXiv, Google Scholar. Background literature as foundation for the review of existing algorithms for deep learning, 3D reconstruction and IBR: (Szeliski 2010); (Bhatti 2012); (Goodfellow *et al.* 2016);(Nielsen 2015); (Hartley and Zisserman 2004).

*Literature search strategy:*

* Research 3D reconstruction techniques, in particular - MVS and SfM. SfS is also of interest.
* Research traditional IBR / video interpolation techniques – as this are combined with deep-learning
* Research deep-learning architectures, in particular – convolutional networks architectures
* Research data and data formats
* Research specific to video interpolation using deep-learning – study in detail articles that have done similar research

**Methodology**

First, this research will test the methodology proposed by (Niklaus *et al.* 2017; Niklaus and Liu 2018) articles on a different set of data. These articles employ a combination of optical flow and 2 neural networks for video interpolation. In contrast to temporal interpolation between two frames of a single camera - this research will focus on the interpolation of inputs from multiview cameras at a single point in time in application to videos of human motion. (Niklaus *et al.* 2017) article provides the original implementation code in Torch and Lua programming language, but re-implementation is available in PyTorch and Python by (Kartašev *et al.* 2018). The methodology described is as follows:

* The algorithm works on 2 input frames (time 0, time 1)
* Both frames are convolved with a CNN neural network (based on conv1 layer of ResNet-18). As the result - each pixel in the input frame is assigned a contextual vector that describes it’s 7x7 neighbourhood.
* The frames are estimated for bi-directional optical flow, then warped to correspond to time t, between time 0 and 1. The above contextual information is also warped to correspond to time t.
* The resultant 2 optical flow images and 2 contextual information images are provided to GridNet CNN architecture that performs multi-resolution convolution-deconvolution to produce a single output image – interpolated frame at time t.

There are several questions that still need to be answered on methodology:

1. How to process colour images – should these be split into 3 channels or just converted to luminance value?
2. What data should be used to train the neural network?

* The first CNN does not require training data as it uses the first output of pre-trained ResNet network for feature extraction.
* The second CNN will need to be trained with data specific to this research. The network needs to deal primary with human motion videos – walking, sitting down, gestures, and maybe dancing. Some videos can be recorded in FVV Studio at V-Sense. However this may not be sufficient, as thousands of input videos are usually required, and overfitting will occur from using a single setting.
* CG simulations can be used to record multi-view video. This website suggests a suitable implementation in Blender - <https://mmspg.epfl.ch/MultiviewDatabase>.
* Also, some databases exist with multi-view data of human motion. For example, OU-ISIR Gait Database.

1. What format should the data be in when provided to neural network? What data pre-processing is going to be used?

* As an example, (Niklaus and Liu 2018) used 50,000 short videos – each consisting of 3 frames. The short videos were selected from 3,000 YouTube videos that had the highest Optical flow between the frames. The flow averaged at 4 pixels (very low for multi-view setup), but the highest flow was 41 pixel.
* In addition to formatting the data correctly, *data augmentation* techniques (cropping, blurring etc) will be applied to increase the amount of data available to deep learning.

1. Tuning the neural network – what hyperparameters will be tuned?

* There are a few hyperparameters applicable to tuning the neural network: learning rate, batch size, number of hidden layers etc. See (Lau 2017) for complete overview.

If above does not produce a good result - alternative architectures/ neural networks will be considered:

1. Alternative Feedforward architectures – different CNN architectures.
2. Recurrent architectures (for example, RNNs) – these are applicable to sequence modelling in video as output of the network is provided back to itself for further processing (Goodfellow *et al.* 2016).
3. CapsuleNets – recently developed by (Hinton *et al.* 2018) . This network type is designed to preserve positional and spatial information and can be applied in combination with CNNs. Hinton et al. suggest that the network performs better than CNN and requires less data, albeit it may take longer to train.
4. GANs are recently developed for generating new images, so will suit our application. The principle of the network is student -teacher setup. The student tries to create images that the teacher will believe to have come from the natural video, while the teacher tries to reject the images as being computer generated and suggests a way to make the images more believable. Original article on GANs by (Goodfellow *et al.* 2014).

Also, the network can be constrained with additional inputs - for example, depth estimation. Depth estimation with deep-learning was researched by (Garg *et al.* 2016) and (Agrawal *et al.* 2015).

**Related research**

Both 3D reconstruction and stereo correspondence are extensively researched areas of computer vision since last century. (Szeliski 2010) provides an overview of methodologies in Chapters 11 and 12, including techniques for reconstructing shape from one image – from shading, texture and focus, to applicable methodologies for stereo reconstructions from 2 images, to multi-view stereo.

Simple frame interpolation is a basic computer vision technique and is deployed in video equipment. Recently image-based modelling and rendering techniques (IBR/ IBMR) can be used for producing new images based on a set of provided images or light rays in the space directly, with unknown or limited amount of known geometry (Zhang 2004). The IBR techniques for generating new images based on dense motion estimation and phase-based methods have been deployed for video interpolation (Niklaus *et al.* 2017). However, these techniques do not perform well in challenging scenarios with e.g. lighting changes and motion blur (Meyer *et al.* 2018), and also are less applicable in videos with large interframe changes (Niklaus *et al.* 2017) – which is the case for multi-view videos.

Apart from (Niklaus *et al.* 2017; Niklaus and Liu 2018) whose approach was described above - several other authors suggested using neural networks for video-frame interpolation. (Meyer *et al.* 2018) proposes to use convolutional neural network to directly estimate phase decomposition of the intermediate frame, without Optical flow. They compare the results to the earlier implementation by Niklaus (Niklaus *et al.* 2017). They suggest that their method is better and is good for coping with larger motion. (Z. Zhang *et al.*) propose to use recurrent convolutional layers (RCL). They compare their results to their own earlier implementation (Z. Liu *et al.*), (Niklaus *et al.* 2017) – previous implementation - and (S. Meyer *et al.*) – previous implementation. Several researches also attempted to use GANs for video interpolation - (C. Li *et al.* ; Samsonov 2017).

Video interpolation is closely related to video extrapolation and video prediction. (Zhou *et al.* 2016) attempts to synthesize new views from an input frame using trained CNN. He trains a network based on car images – and creates believable images of the car from the requested point of view. (Flynn *et al.* 2015) were the first to suggest the use of CNNs for creating new viewpoints.

As an alternative to using deep-learning to produce synthetic video frames, deep-learning can be applied to produce 3D structure directly. (Huang *et al.* 2018) successfully implemented CNN-based reconstruction from as few as 3 images that output the likelyhood of the points location. They however only trained the network on synthetic samples that were in ideal setting (no background). (Garg *et al.* 2016) trained unsupervised CNN network to predict depth from a single image, basing the training on stereo pair of images.

**Evaluation criteria and method**

Quantitative evaluation of our method is possible - using synthetic ground-truth data. The proposed algorithm will be applied to images obtained from a CG model – then the output frames will be provided as additional input to the existing 3D reconstruction pipeline. Because the ground truth for a CG model is readily available as mesh information – digital estimate of the quality of 3D reconstruction will be obtained – with and without the additional frames.

Another approach is to use 3D reconstruction obtained with a lot of cameras on a real video scenario as ground truth. The approach similar to the above can be applied to compare the resultant 3D reconstruction with/ without the interpolated frames.

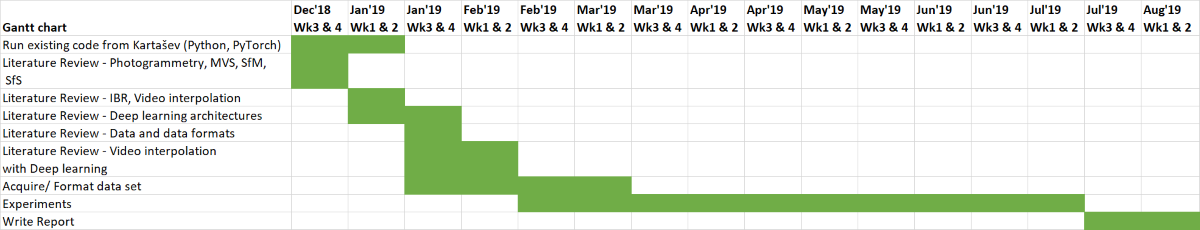
Benchmark evaluation can include <http://vision.middlebury.edu/stereo/data/> - which provide stereo dataset with ground truth and the results can be compared to other researchers - based on 2 images only.

*Ethics:* Permission from the ethics committee is not necessary. This dissertation will use only publicly available datasets (with no personal data or GDPR concerns) and will not be conducting a subjective quality study.

*Potential benefits of this research* include a novel technique for creation of synthesised views for multi-view stereo. These can be applied in 3D reconstruction and 360 video. Potentially, if the implemented approach differs from (Niklaus and Liu 2018) – this can be applied to video interpolation in general. Ultimately, there is a cost saving to using less cameras for the recording and fewer cameras make everyday application more feasible, i.e. using mobile phone cameras.

*Research Milestones*

|  |  |
| --- | --- |
| Milestone 1 - Test sample code (Kartašev *et al.* 2018) | January 15th |
| Milestone 2 - Complete Literature Review | February 15th |
| Milestone 3 - Prepare Dataset/ Test set/ Format data | March 15th |
| Milestone 4 - Experiments completed | July 19th |
| Milestone 5 - Report completed | August 15th |



*Proposed Table of contents:* Introduction, Aims, Background and literature review, Methodology, Results and Discussion, Conclusion.

This research will aim towards a *paper* publication.

The following *skills* are required for this research*:* PyTorch, Python, C++ and possibly cuda. The first two skills will need to be improved by the author of this dissertation. Also, if cuda is required - this will be added to the list. Blender will be needed for creating 3D models for training and test – the author of this dissertation has knowledge of this.

The proposed research is not intended to be a continuation of existing work nor is it intended to be a collaboration with any external organisations beyond Trinity College Dublin and V-Sense. This is a new project which aims to expand on existing research and implement it in a novel application beyond the scope of its original design.

Potential problems that may arise:

1) The proposed method may not produce good result as it is developed for video interpolation, where the range of motion is much smaller than in multi-view stereo setup. This research will look at other methods proposed in “Methodology” section.

2) Deep-learning requires a lot of training time – may run out of time. Will need to start running the experiments earlier than April.

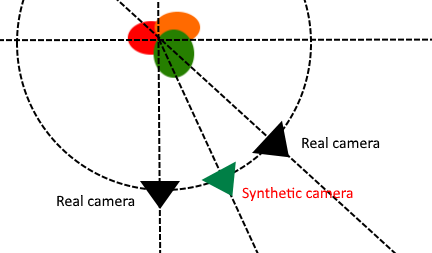


Fig.1 Synthetic camera illustration

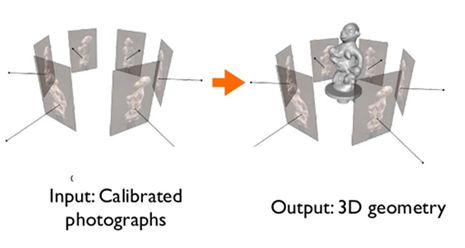


Fig.2. 3D Reconstruction illustration. Source: <http://www.cs.sfu.ca/~furukawa/>

**References:**

Agrawal, P., Carreira, J. and Malik, J. (2015) 'Learning to See by Moving', available.

Bhatti, A. (2012) *Current Advancements in Stereo Vision*, InTech.

C. Li, D. Gu, X. Ma, K. Yang, S. Liu and F. Jiang 'Video Frame Interpolation Based on Multi-scale Convolutional Network and Adversarial Training', in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, 553-560, available: <http://dx.doi.org/10.1109/dsc.2018.00089>.

Flynn, J., Neulander, I., Philbin, J. and Snavely, N. (2015) 'DeepStereo: Learning to Predict New  Views from the World’s Imagery', *CoRR abs/1506.06825*.

Garg, R., Kumar, V.B.G., Carneiro, G. and Reid, I. (2016) 'Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue', available: <https://arxiv.org/pdf/1603.04992.pdf>.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*, MIT Press.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) 'Generative Adversarial Networks', available: <https://arxiv.org/abs/1406.2661>.

Hartley, R. and Zisserman, A. (2004) *Multiple View Geometry in Computer Vision*,2 ed.

Hinton, G., Sabour, S. and Frosst, N. (2018) 'Matrix capsules with EM routing', in *ICLR 2018 Conference*.

Kartašev, M., Rapisarda, C. and Fay, D. (2018) 'Implementing Adaptive Separable Convolution for Video Frame Interpolation', *arXiv*, available: <https://arxiv.org/abs/1809.07759>.

King, D. (2016) *Easily Create High Quality Object Detectors with Deep Learning*, available:<http://blog.dlib.net/2016/10/easily-create-high-quality-object.html>

Lau, S. (2017) *A Walkthrough of Convolutional Neural Network — Hyperparameter Tuning*, available:<https://towardsdatascience.com/a-walkthrough-of-convolutional-neural-network-7f474f91d7bd>

Li, S. , Zhu, C. and Sun, M. (2018) 'Hole Filling With Multiple Reference Views in DIBR View Synthesis', *IEEE Transactions on Multimedia*, 20(8), 1948-1959, available: <http://dx.doi.org/10.1109/tmm.2018.2791810>.

Liu, Z. , Yeh, R. A., Tang, X. , Liu, Y. and Agarwala, A. 'Video Frame Synthesis Using Deep Voxel Flow', in *2017 IEEE International Conference on Computer Vision (ICCV)*, 4473-4481, available: <http://dx.doi.org/10.1109/iccv.2017.478>.

Meyer, S., Djelouah, A., McWilliams, B., Sorkine-Hornung, A., Gross, M. and Schroers, C. (2018) 'PhaseNet for Video Frame Interpolation', *arXiv*, available: <https://arxiv.org/abs/1804.00884>.

Meyer, S. , Wang, O., Zimmer, H. , Grosse, M. and Sorkine-Hornung, A.'Phase-based frame interpolation for video', in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1410-1418, available: <http://dx.doi.org/10.1109/cvpr.2015.7298747>.

Nielsen, M.A. (2015) *Neural Networks and Deep Learning*, Determination Press, available:<http://neuralnetworksanddeeplearning.com/>.

Niklaus, S. and Liu, F. (2018) 'Context-aware Synthesis for Video Frame Interpolation', *CoRR*, abs/1803.10967.

Niklaus, S., Mai, L. and Liu, F. (2017) 'Video Frame Interpolation via Adaptive Convolution', *CoRR*, abs/1703.07514.

Pages, R., Amplianitis, K., Monaghan, D., Ondrej, J. and Smolic, A. (2018) 'Affordable content creation for free-viewpoint video and VR/AR applications', *Journal of Visual Communication and Image Representation*, 53, 192-201, available: <http://dx.doi.org/10.1016/j.jvcir.2018.03.012>.

Samsonov, V. (2017) 'Deep frame interpolation', available: <https://arxiv.org/pdf/1706.01159.pdf>.

Sanz, P.R., Mezcua, B.R. and Pena, J., M. Sánchez (2012) 'Depth Estimation - An Introduction', available: <http://dx.doi.org/10.5772/45904>.

Seitz, S.M., Curless, B., Diebel, J., Scharstein, D. and Szeliski, R. (2006) 'A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms', *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Simon Niklaus, F.L. (2018) 'Context-aware Synthesis for Video Frame Interpolation', *CoRR*, abs/1803.10967.

Simon Niklaus, L.M., Feng Liu (2017) 'Video Frame Interpolation via Adaptive Convolution', *CoRR*, abs/1703.07514.

Smolic, A. (2011) '3D video and free-viewpoint video—from capture to display', *Elsevier. Pattern Recognition*, 44, 1958–1968.

Szeliski, R. (2010) *Computer Vision: Algorithms and Applications*, Springer.

Wenger, S.M.B. (2016) *Evaluation of SfM against traditional stereophotogrammetry and Lidar techniques for DSM creation in various land cover areas*, unpublished thesis, Stellenbosch University.

Zhang, C.a.C.T. (2004) 'A Survey on Image-Based Rendering - Representation, Sampling and Compression', *Signal Processing: Image Communication*, 19, 1-28.

Zhang, Z. , Song, L. , Xie, R. and Chen, L. 'Video Frame Interpolation Using Recurrent Convolutional Layers', in *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, 1-6, available: <http://dx.doi.org/10.1109/BigMM.2018.8499065>.

Zhou, T., Tulsiani, S., Sun, W., Malik, J. and Efros, A.A. (2016) 'View Synthesis by Appearance Flow', available: <https://arxiv.org/abs/1605.03557>.

**Glossary:**

AR – Augmented reality

CG – Computer Graphics

Depth estimation - set of techniques and algorithms aiming to obtain a representation of the spatial structure of a scene.

Disparity map - refers to the apparent pixel difference or motion between a pair of stereo image.

FVT – Free-viewpoint television. System for viewing natural video, allowing the user to interactively control the viewpoint and generate new views of a dynamic scene from any 3D position

FVV - Free-viewpoint video. Same as FVT.

IBMR - image-based rendering

IBMR - image-based modeling and rendering. Methods rely primarily on the original or trained set of images to produce new, virtual views, rather than 3D model.

MS-SfS – Multi-source Shape-from-Silhouette

MVS – Multi-view stereo. Aims to reconstruct disparity maps from a collection of images with known camera poses and calibration, possibly estimated using Structure from Motion (SfM) algorithms.

Optical flow - independent flow estimation for each pixel.

Photogrammetry - science of making measurements from photographs, especially for recovering the exact positions of surface points.

SfM – Structure from Motion. Photogrammetric range imaging technique for estimating 3D structure from 2D image sequences that may be coupled with local motion signals.

SfS – Shape-from-Silhouette. Shape reconstruction method which constructs a 3D shape estimate of an object using silhouette images of the object (for example, where silhouettes are obtained by object segmentation).

View Synthesis - aims to create new views of a specific subject starting from a number of pictures taken from given point of views.

VR – Virtual Reality