OmniFlow: Human Omnidirectional Optical Flow

Roman Seidel, André Apitzsch, Gangolf Hirtz

Chemnitz University of Technology

Faculty of Electrical Engineering and Information Technology

09126 Chemnitz, Germany

{roman.seidel,andre.apitzsch,g.hirtz}@etit.tu-chemnitz.de

Abstract

Optical flow is the motion of a pixel between at least two consecutive video frames and can be estimated through an end-to-end trainable convolutional neural network. To this end, large training datasets are required to improve the accuracy of optical flow estimation. Our paper presents OmniFlow: a new synthetic omnidirectional human optical flow dataset. Based on a rendering engine we create a naturalistic 3D indoor environment with textured rooms, characters, actions, objects, illumination and motion blur where all components of the environment are shuffled during the data capturing process. The simulation has as output rendered images of household activities and the corresponding forward and backward optical flow. To verify the data for training volumetric correspondence networks for optical flow estimation we train different subsets of the data and test on OmniFlow with and without Test-Time-Augmentation. As a result we have generated 23,653 image pairs and corresponding forward and backward optical flow. Our dataset can be downloaded from: https://mytuc.org/byfs

1. Introduction

In the last decade, large-scale synthetic training and benchmark datasets have driven innovation in computer vision, have accelerated the development of learning-based approaches and demonstrated a way of quantitative evaluation without capturing real-world data. Especially in optical flow estimation this plays an important role due to the high necessary effort to collect and label real-world data. Optical flow is the motion of a pixel between at least two consecutive video frames and can be estimated through an end-to-end trainable convolutional neural network (CNN) [4, 8].

Assuming an indoor scenario with fisheye cameras with a field of view (FOV) of greater than or equal to 180° a whole room can be captured with only one sensor. Fisheye cameras follow the omnidirectional camera model and the question arises whether to formulate their distortions implicitly in the model architecture or explicitly generate synthetic omnidirectional data.

Perspective human optical flow cannot be used to determine human motions in omnidirectional images due to unsuitable layer architectures or missing data. However, our data-driven approach generates a dataset with human optical flow in omnidirectional images that contains distortions and invariances against angle of view. Our dataset contains various indoor household activities such as *sitting down and standing up*, *walking* or *falling down*. Human optical flow for omnidirectional images can be used for computer vision tasks such as motion estimation, indoor navigation of robots and tracking of persons with indoor surveillance systems.

2. Related Work

Synthetic optical flow datasets for training or benchmarking CNNs are extensively available and firstly initiated by the Middlebury [1] and the MPI Sintel dataset [2]. In parallel, with a new correlation-based network architecture datasets with synthetic foreground and real-world random background images were created by [4, 8], namely FlyingChairs and FlyingThings. Aodha *et al.* [10] created a system for easily producing synthetic ground truth data for both optical flow and descriptor-matching, where the pipeline lacks realistic scenarios and has no humans included. With the goal of benchmarking algorithms in the perception of the environment of autonomous cars the KITTI Benchmark Suite [6] and – for multi object tracking – VirtualKITTI [3,5] were created. Both have optical flow ground truth; KITTI from a lidar sensor and virtual KITTI through a 3D rendering engine. A tabular overview of optical flow training and benchmark datasets based on perspective views is shown in Table 1.

Single and multi-human optical flow datasets in perspective front views were investigated in [16, 17] from which we differ in terms of camera geometry, body models and output image and flow resolution. A dataset which focuses on the application of crowd analysis was created in [19] where

Dataset	Synthetic/ Natural	Training/ Benchmark	#Frames	Resolution	Moving/Static camera	Person/ Non-person
Driving [13]	S	T	4,392	960×540	M	N
FlyingChairs [4]	S	T	21,818	512×384	M	N
FlyingThings3D [13]	S	T	22,872	960×540	M	N
HD1K [9]	N	B	1,083	$2,560 \times 1,080$	M	P
KITTI 2012 [6]	N	B	194	$1,242 \times 375$	M	(P)
KITTI 2015 [15]	N	B	200	$1,242 \times 375$	M	(P)
Middlebury [1]	S/N	B	8	640×480	SC	N
Monkaa [13]	S	T	8,591	960×540	M	N
SceneNet RGB-D [14]	S	T	$\sim 5,000,000$	320×240	M	N
Sintel [2]	S	B/T	1,064	$1,024 \times 436$	M	(P)
SplitSphere [7]	S	B	unknown	512×512	SC	N
UCL [11]	S	B	4	640×480	SC	N
UCL (extended) [12]	S	B	20	640×480	SC	N
Virtual KITTI [5]	S	B/T	21,260	$1,242 \times 375$	M	P
Virtual KITTI 2 [3]	S	B/T	21,260+2,126	$1,242 \times 375$	M	P
OmniFlow (ours)	S	B/T	23,653	$2,048 \times 2,048$	Static per scene	P

Table 1. Comparison of optical flow benchmark and training datasets.

an omnidirectional synthetic dataset which contains bounding boxes, segmentation masks and depth maps in indoor scenarios [18] lacks optical flow.

3. OmniFlow: Human Omnidirectional Optical Flow

This section describes the dataset creation pipeline of human omnidirectional optical flow. Our data is created by the rendering engine *Blender*, which we used to model a 3D indoor environment with randomly placed animated humans in various rooms, objects and a virtual camera with an omnidirectional camera geometry. The whole dataset creation pipeline is shown in Figure 1. The setup of these animated scenes contains a moving human from various camera locations with static background modeled as 3D indoor environment. We have generated 321 scenes each with 75 images and corresponding forward and backward flow and a total dataset size of 23,653 frames containing 18,921 frames for training and 2,366 frames each for validation and testing.

Dataset Creation Pipeline. To counter the gap between synthetic and real-world image data domain randomization [21] is implemented in our simulation. To this end, randomly set simulation parameters affect the building process of the entire scene. Various skeletons were used by the simulation separately for each scene while person models are rigged to these skeletons. Further simulation parameters are rooms with different textures and random illumination size and energy from two area lights. Additionally, objects such as tables, chairs or plants are placed inside each room. To achieve various viewing angles of the camera with respect to the character the camera location is variable, where the extent of the room is used for potential camera viewpoints.

Rendering Engine. The optical flow of OmniFlow is created by *vector pass* of Blender's Cycles rendering engine which exclusively delivers an omnidirectional camera, as shown in Figure 1. Given three consecutive image frames $t, t+1$ and $t+2$ the *vector pass* contains motion between these frames in forward (images t and $t+1$) and backward direction (images $t+1$ and $t+2$).

Camera. For rendering the 3D environment to 2D images, a virtual camera with the fisheye-equidistant camera model is randomly placed in a 4×4 m square in the center of the room. The camera is tracked to the hip bone of the animation to make sure that the human is centered with respect to the camera image. To cover the whole scene, the FOV is $180°$ without considering the sensor type and sensor size.

Rooms. On the basis of [18] we use four different rooms from ProBuilder, a world building tool in the Unity Editor with random textures[1] which are changed with every scene. We import the geometries and the corresponding textures for walls, ground and furniture to Blender. The rooms itself have a spatial expansion of 20×20 m and a height of 4 m.

Animations. We include various animations from the Motionbuilder-friendly BVH conversion release of CMU's Motion Capture Database[2] where frame-1 T-poses are included. Examples of animations from the CMU Motion Capture database are activities of the humans, e.g. *sitting down and standing up, falling, tooth brushing, brooming* and can be adapted to every other daily activity.

Characters. The characters were generated by *MakeHuman* and contain 16 user generated textures with upper body clothes, lower body clothes, shoes and body extensions (e.g. hairstyles). All characters were set to T-pose (base animation

---

[1] https://www.cc0textures.com
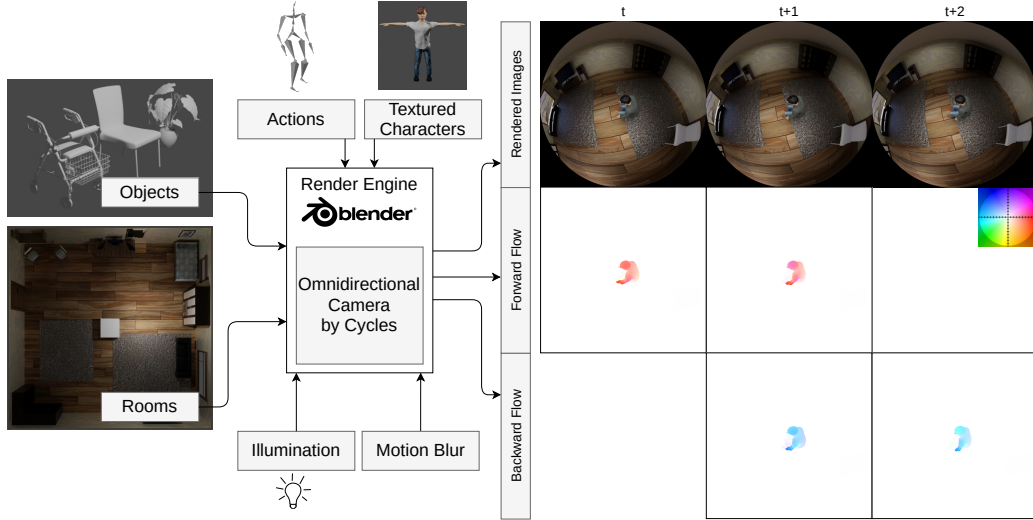[2] http://mocap.cs.cmu.edu/ and http://cgspeed.com

Figure 1. OmniFlow dataset creation pipeline. We rig textured characters to actions, place both in randomly selected rooms and add objects, illumination and (to some scenes) motion blur to generate high resolution image sequences with corresponding forward and backward flow.

pose), rigged to CMU's hybrid dataset and randomized in terms of weight, height, age and gender.

Objects. Natural objects, such as potted plants, chairs and tables are included in the scene to make the indoor environment realistic and evoke occlusions of objects with the animated humans.

Illumination. For indoor lighting we randomly set two light sources in each scene. An area-light directly above the character and a ceiling light with a distance of 2.5 to 7.5 meters to the character to get a realistic lightning scenario. The light varies for each scene in terms of energy and position in the above specified limits. Naturalistic outdoor illumination was realized by High Dynamic Range Images (HDRI) from HDRIHeaven[3]. With realistic day, dusk, dawn and night HDRIs we were able to simulate different times of day. Both, indoor lights and outdoor illumination were selected randomly in our scenes.

Motion Blur. Since fast-moving objects are appearing not constantly sharp during their motion, we applied motion blur to the human activity and switch it randomly on and off in each scene.

4. Evaluation

Our dataset is evaluated on a test set of OmniFlow of 10% of the whole dataset. We train a correspondence network for optical flow and fine-tune on five subsets 1k, 5k, 10k, 15k and 20k of OmniFlow with a pretrained model on FlyingChairs and FlyingThings. As long there is no further omnidirectional optical flow dataset for testing available we use test-time augmentation (TTA) [20] with three standard
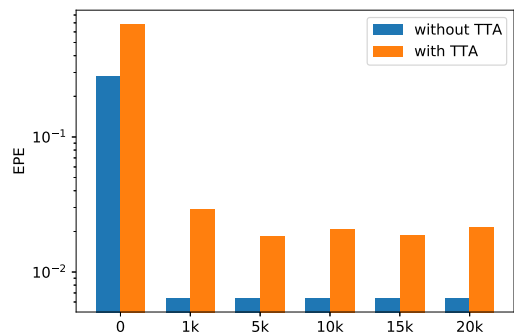

Figure 2. Determination of a *sufficient* amount of data for training RAFT. 0 means that we test a on FlyingChairs and FlyingThings pretrained model on test split of OmniFlow. Furthermore, we fine-tune RAFT on our data on selected subsets of 1k, 5k, 10k, 15k and 20k data with and without test-time augmentation.

augmentation methods cropping, scaling and horizonal flipping. Results on OmniFlow test set are shown in Figure 2.

Recurrent All Pairs Field Transforms. Recurrent All Pairs Field Transforms for Optical Flow (RAFT) consists of a per-pixel feature encoder for extracting features from both input images I_1, I_2 and a context encoder that extracts features only from I_1. A correlation layer generates a 4D correlation tensor for all pairs of pixels including subsequent pooling to produce lower resolution volumes. The update operator based on a gated activation unit (GRU-block) where fully connected layers are replaced by convolutions. The input feature map of the update operator is the concatenation of correlation, flow and context features. In general, we follow the training schedule of RAFT but resize our OmniFlow

---

data to a input resolution of 512×512 px and use a learning rate of 1e-6 and a weight decay of 1e-5.

Our Observations. We figure out that 5,000 image pairs of OmniFlow are sufficient to fine-tune RAFT with synthetic omnidirectional human optical flow data. Nevertheless we provide approx. 20,000 image pairs where CNNs based on correlation layers needs more data for training.

5. Conclusion

In this paper, we create OmniFlow: a new omnidirectional human optical flow dataset. With a 3D rendering engine, namely *Blender* we generate a naturalistic 3D indoor environment with textured rooms, characters, actions, objects, illumination and motion blur. We evaluate our data with TTA and have explored that the amount of 5k images is sufficient to fine-tune RAFT on FlyingChairs and FlyingThings. Our next steps are the investigation in other network architectures for correlation-based or semi-supervised optical flow CNNs and the interpretation of optical flow as fine-grained human activities.

References

[1] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1):1–31, 2011. 1, 2

[2] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012. 1, 2

[3] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv:2001.10773 [cs, eess]*, Jan. 2020. arXiv: 2001.10773. 1, 2

[4] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 1, 2

[5] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016. 1, 2

[6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1, 2

[7] Frédéric Huguet and Frédéric Devernay. A Variational Method for Scene Flow Estimation from Stereo Sequences. In *Proc. Intl. Conf. on Computer Vision*, Rio de Janeiro, Brasil, Oct. 2007. IEEE. 2

[8] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of Optical Flow Estimation With Deep Networks. pages 2462–2470, 2017. 1

[9] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–28, 2016. 2

[10] Oisin Mac Aodha, Gabriel J Brostow, and Marc Pollefeys. Segmenting video into classes of algorithm-suitability. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1054–1061. IEEE, 2010. 1

[11] O. Mac Aodha, G. J. Brostow, and M. Pollefeys. Segmenting video into classes of algorithm-suitability. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1054–1061, 2010. 2

[12] Oisin Mac Aodha, Ahmad Humayun, Marc Pollefeys, and Gabriel J Brostow. Learning a confidence measure for optical flow. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1107–1120, 2012. 2

[13] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 2

[14] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2678–2687, 2017. 2

[15] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 2

[16] Anurag Ranjan, David T. Hoffmann, Dimitrios Tzionas, Siyu Tang, Javier Romero, and Michael J. Black. Learning Multi-human Optical Flow. *International Journal of Computer Vision*, 128(4):873–890, Apr. 2020. 1

[17] Anurag Ranjan, Javier Romero, and Michael J. Black. Learning Human Optical Flow. *arXiv:1806.05666 [cs]*, July 2018. arXiv: 1806.05666. 1

[18] Tobias Scheck, Roman Seidel, and Gangolf Hirtz. Learning from THEODORE: A Synthetic Omnidirectional Top-View Indoor Dataset for Deep Transfer Learning. pages 943–952, 2020. 2

[19] G. Schröder, T. Senst, E. Bochinski, and T. Sikora. Optical Flow Dataset and Benchmark for Visual Crowd Analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018. 1

[20] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. When and Why Test-Time Augmentation Works. *arXiv:2011.11156 [cs]*, Nov. 2020. arXiv: 2011.11156. 3

[21] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training Deep Networks With Synthetic Data: Bridging the Reality Gap by Domain Randomization. pages 969–977, 2018. 2