Learning Human Reconstruction from Synthetic Dataset

Akin Caliskan[1], Armin Mustafa[1], Evren Imre[2], and Adrian Hilton[1]

[1] Center for Vision, Speech and Signal Processing, University of Surrey

[2] Vicon Sensing Sytems Ltd, Oxford

Existing stereo reconstruction methods for narrow baseline image pairs [2, 4] give limited performance for wide baseline views. This paper proposes a framework to learn and estimate dense stereo for people from wide baseline image pairs. A synthetic people stereo patch dataset (S2P2) is introduced to learn wide baseline dense stereo matching for people. The proposed framework learns human specific features from synthetic data for patch match and adapts it to real data. In addition to patch match learning, a stereo constraint is introduced in the framework to solve wide baseline stereo reconstruction of humans. Quantitative and qualitative performance evaluation of the proposed approach against state-of-the-art methods demonstrates improved wide baseline stereo reconstruction on challenging datasets. We show that it is possible to learn stereo matching from synthetic people dataset and improve performance on real datasets

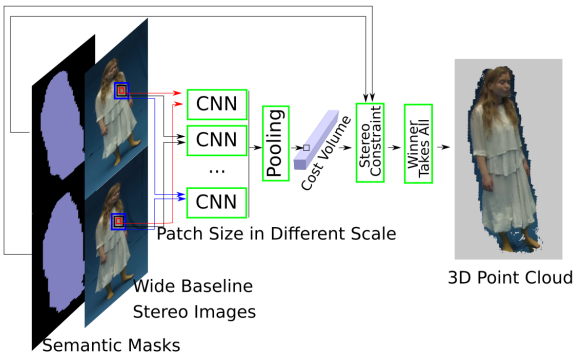

Figure 1: The proposed stereo reconstruction method.

Method: The proposed approach is motivated by requirement for high-quality 3D content (especially for humans) in augmented reality/virtual reality and autonomous driving. However, existing scanning technologies require advanced camera setups, and controlled studio capture environments, which are complex and costly solutions. We propose dense stereo reconstruction for humans from wide baseline image pairs to address this need. The proposed supervised learning based framework first learns stereo matching from a new synthetic human specific dataset *S2P2* for wide-baseline cameras and followed by stereo reconstruction refinement using semantic human constraint, as seen in Figure 1. Variation of human body surface for example folded clothing, hair, face details, makes it challenging to extract reliable stereo reconstruction from wide baseline image pairs. Given a wide baseline stereo pair of images of a person, we aim to obtain per-pixel dense correspondence for stereo reconstruction. The stereo pair of images are fed into a CNN module, which is trained on a human specific dataset to obtain the matching cost for each pixel. This generates a cost volume which is refined using a semantic stereo constraint to obtain the final depth map. We use a Siamese network architecture [1] as the backbone; because, it allows training of stereo matching between a pair of left and right image patches. The network consists of four consecutive 2D convolution layers and RELU (Rectified Linear Units) after each convolution layer. The computed feature vectors are fed into a fully-connected network (FCN) to estimate the similarity score between patches. Since we are solving a binary classification problem, we use binary cross entropy loss to train our network. During the training stage, we use a balanced number of positive and negative patches extracted from the *S2P2* dataset.

Results: The proposed method outperforms NCC [2] and Daisy [4] in all depth estimation metrics, which is illustrated in Table . NCC [3] and Daisy [4] generate local descriptors that are prone to fail in ambiguities, like repetitive textures, lack of textures, or lighting changes and large changes in shape. These failures can be resolved during post processing stage in wide baseline human stereo reconstruction methods [3]. As shown in Table , the proposed method gives approximately 25% RMSE error reduction for two camera baseline values compared to MC-CNN, which is the state of the art patch based stereo reconstruction method.

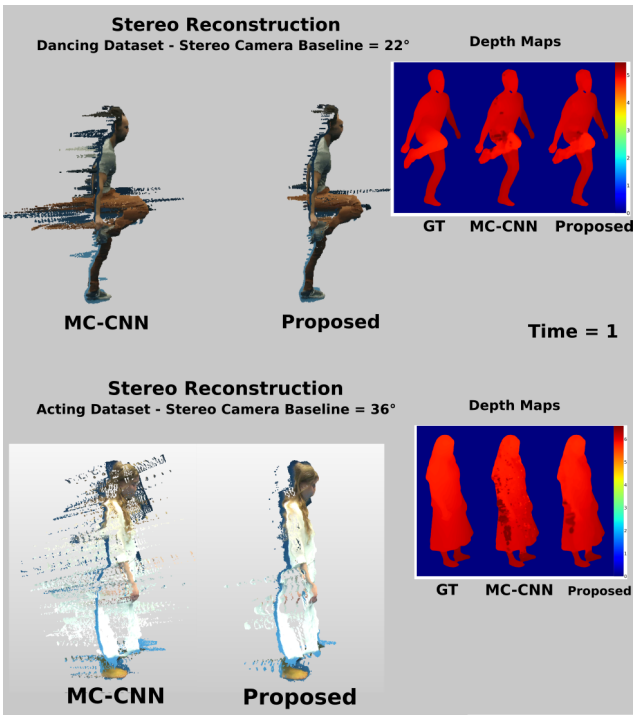

Figure 2: Point cloud stereo reconstruction results with depth map estimations for varying camera baselines.

Method	Camera Baseline ≈ 40°			
	Abs Rel	Squ Rel	RMSE	$RMSE_{log}$
Dataset:Acting				
NCC [2]	5.21	3.52	46.6	8.10
Daisy [4]	2.05	0.95	24.1	3.70
MCCNN [5]	1.42	0.43	16.1	2.57
Ours	**1.03**	**0.26**	**12.6**	**1.99**
Dataset:Dancing				
NCC [2]	6.08	2.51	35.3	7.29
Daisy [4]	2.55	0.88	20.6	3.89
MCCNN [5]	1.76	0.39	17.3	3.41
Ours	**1.71**	**0.33**	**15.3**	**3.01**

Table 1: Depth estimation error results for 2 datasets against four compared methods are listed in the table.

The reconstruction results are shown in Figure with corresponding depth maps for MC-CNN and the proposed method for different camera baselines. The proposed method which learns from human specific features is able to capture details of clothing and hair which are challenging to reconstruct in wide baseline stereo setups, which shows that learning from a human-specific dataset improves wide baseline stereo performance. More results are available in the project video: `https://youtu.be/dPPn_Hm0KFM`.

[1] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[2] Xiaoyan Hu and Philippos Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2121–2133, 2012.

[3] Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. General dynamic scene reconstruction from multiple view video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 900–908, 2015.

[4] Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–830, 2009.

[5] Jure Zbontar, Yann LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016.