Domain Siamese CNNs for Sparse Multispectral Disparity Estimation

David-Alexandre Beaupré
Guillaume-Alexandre Bilodeau

LITIV lab., Department of Computer and Software Engineering,
Polytechnique Montreal

25th International Conference on Pattern Recognition, Milan, Italy

10 - 15 January 2021

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Presentation overview I

Introduction - Disparity estimation

- The disparity d can be defined as an horizontal distance between a pixel p at coordinates (x_p, y_p) from the left image to the equivalent pixel q at (x_q, y_q) in the right image.
- With disparity information, we can determine the depth of the pixels in the scene which has applications in multiple domains such as robot navigation.

Figure: Illustration of the disparity for a multispectral image pair.

Introduction - Multispectral images

- When comparing different spectrums on which to apply stereo, we can see that the RGB-LWIR has the least similarities between both images.

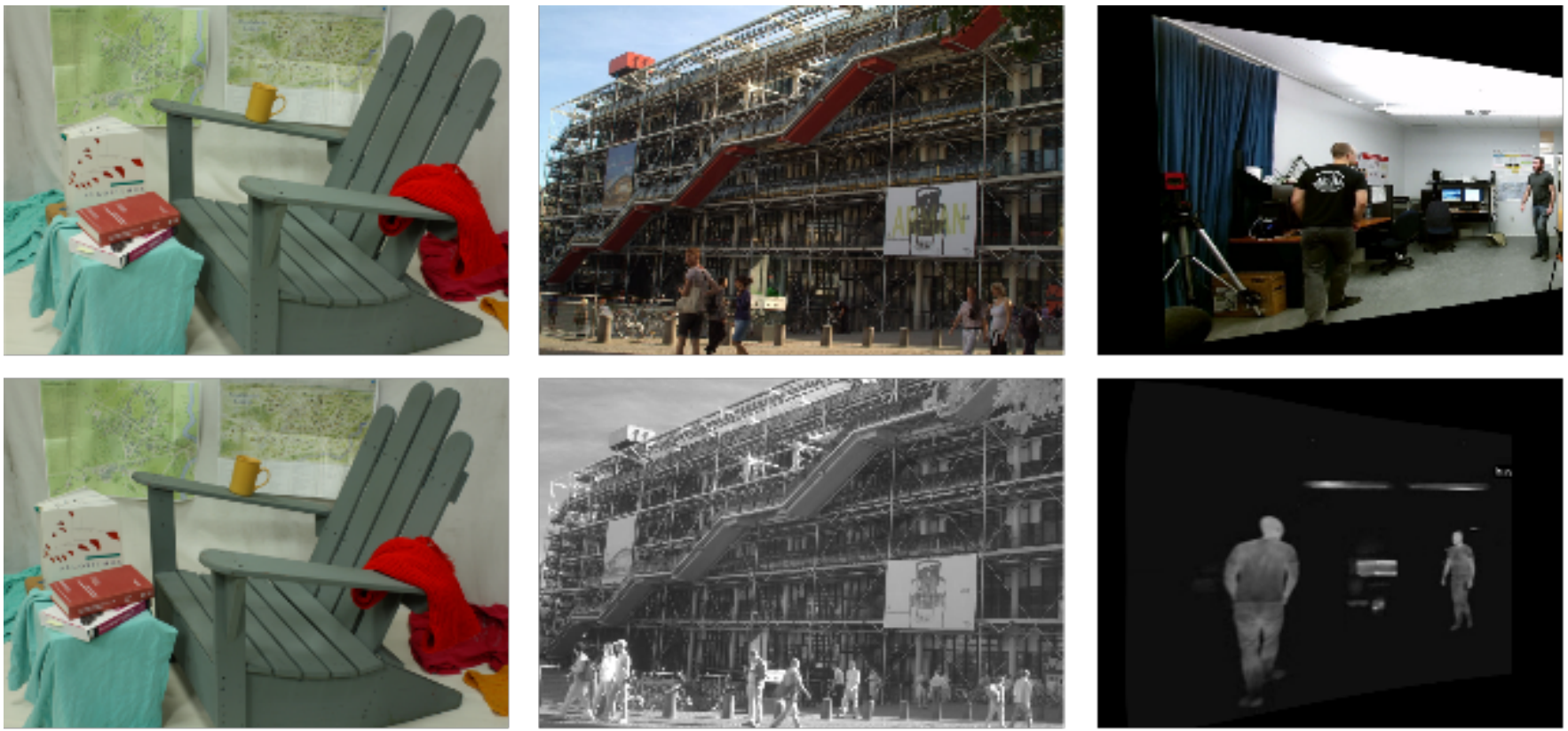

Figure: Comparison of image pairs: first colum: RGB-RGB, second column: RGB-NIR and third column: RGB-LWIR.

Related works

- In stereo matching most CNN approches fall into two categories: patch-based methods [1, 2] and end-to-end methods [3, 4]. Guo *et al.* [5] proposed an architecture to form the cost volume from correlation and concatenation, which leads to better performance compared to when only one is used. Kendall *et al.* [6] proposed the disparity regression that inspired us in our prediction process.
- In multispectral images, Bilodeau *et al.* [7] compared different classical descriptors for multispectral disparity estimation. Baruch *et al.* [8] proposed a siamese network where the branches do not share weights. They showed that this lead to favorable results when compared to weights being shared.

Proposed method

- Separate branches extract the RGB and LWIR features, which are then merged and forwarded to two classification heads.

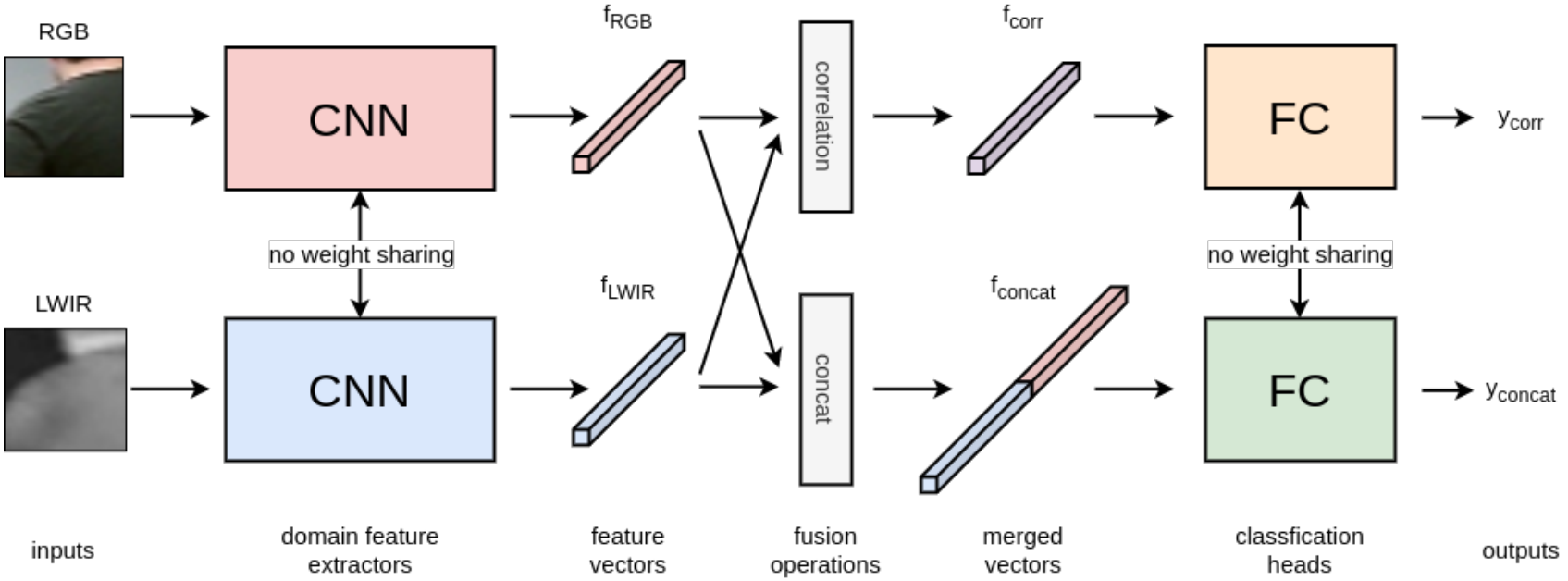

Figure: Detail of the proposed architecture.

Training and prediction

Training

- Creation of positive and negative patches so that the network learn a binary classification.

$$loss_{corr/concat} = -\frac{1}{N}\sum_{i=1}^{N} gt_i log(y_i),$$

Prediction

- Regress the disparity from the probability vector showing how likely two patches are the same.

$$\hat{d}_{corr/concat} = \sum_{d=0}^{disp_{max}} d \times p_d.$$

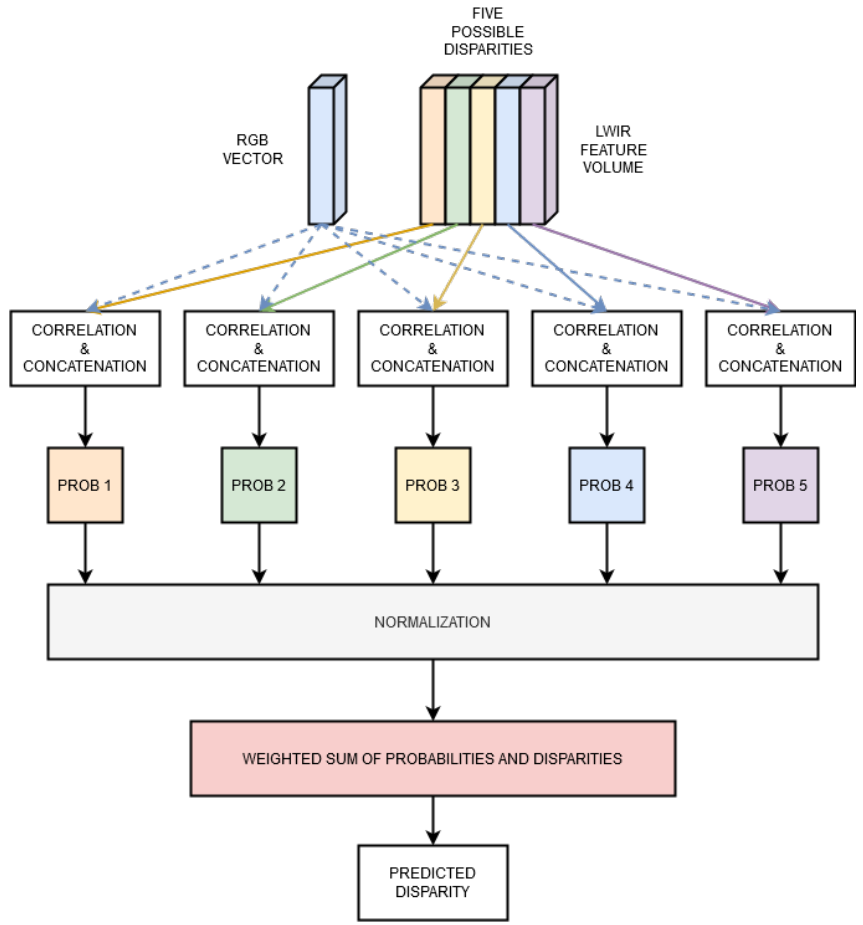

Figure: Illustration of the prediction process.

Datasets

- Two datasets were used to train and evaluate our method: the LITIV 2014 [7] and the LITIV 2018 [9] datasets.
- Each dataset is separated into three folds, where one is kept for testing, and the other ones are used for training and validation.

	Training		Validation	Testing
	LITIV 2018	LITIV 2014	LITIV 2014	LITIV 2014
Fold 1	218 230 (vid04 + vid07 + vid08)	240 167 (vid02 + vid03)	35 378 (vid02 + vid03)	101 433 (vid01)
Fold 2	218 230 (vid04 + vid07 + vid08)	291 720 (vid01 + vid03)	34 688 (vid01 + vid03)	76 001 (vid02)
Fold 3	218 230 (vid04 + vid07 + vid08)	320 648 (vid01 + vid02)	34 220 (vid01 + vid02)	61 771 (vid03)

Table: Data separation and number of ground-truth points for the LITIV 2014 dataset.

	Training		Validation	Testing
	LITIV 2014	LITIV 2018	LITIV 2018	LITIV 2018
Fold 1	478 410 (vid01 + vid02 + vid03)	109 620 (vid07 + vid08)	44 226 (vid07 + vid08)	32 192 (vid04)
Fold 2	478 410 (vid01 + vid02 + vid03)	91 904 (vid04 + vid08)	49 286 (vid04 + vid08)	38 520 (vid07)
Fold 3	478 410 (vid01 + vid02 + vid03)	99 858 (vid04 + vid07)	41 566 (vid04 + vid07)	38 403 (vid08)

Table: Data separation and number of ground-truth points for the LITIV 2018 dataset.

Comparison to other methods

	Correlation branch only			Concatenation branch only			Corr + Concat (proposed model)		
	≤ 1 px	≤ 3 px	≤ 5 px	≤ 1 px	≤ 3 px	≤ 5 px	≤ 1 px	≤ 3 px	≤ 5 px
Fold 1	0.524	0.859	0.984	0.551	0.894	0.981	**0.588**	**0.901**	**0.985**
Fold 2	0.454	0.854	0.978	0.472	0.897	0.985	**0.474**	**0.904**	**0.986**
Fold 3	0.541	0.875	0.982	0.558	0.895	0.982	**0.629**	**0.916**	**0.989**

Table: Ablation study of the proposed model. **Boldface**: best results.

	Fold 1		Fold 2		Fold 3		Overall	
	≤ 1 px	≤ 4 px	≤ 1 px	≤ 4 px	≤ 1 px	≤ 4 px	≤ 1 px	≤ 4 px
DASC Sliding Window [9]	0.104	0.265	0.086	0.236	0.121	0.289	0.104	0.263
Multispectral Cosegmentation [9]	0.253	0.562	0.236	0.531	0.307	0.678	0.265	0.590
Proposed Model	**0.480**	**0.943**	**0.446**	**0.877**	**0.406**	**0.972**	**0.442**	**0.930**

Table: Evaluation of our architecture on the LITIV 2018 dataset. **Boldface**: best results.

Method	≤ 3 px
Proposed Model (36×36)	**0.906**
Siamese CNNs [10] (37×37)	0.779
Mutual Information [7] (40×130)	*0.833*
Mutual Information [7] (20×130)	0.775
Mutual Information [7] (10×130)	0.649
Fast Retina Keypoint [7] (40×130)	0.641
Local Self-Similarity [7] (40×130)	0.734
Sum of Squared Differences [7] (40×130)	0.656

Table: Evaluation of our architecture on the LITIV 2014 dataset. **Boldface**: best results.

Conclusion

- We proposed a new CNN architecture able to estimate the disparity between images from the RGB and LWIR domains. Our CNN is made of two branches with no weight sharing, and each branch is responsible of extracting the features from the input images. The features are then merged with the correlation and concatenation operation, which leads to a more robust network able to determine if the inputs are similar or not.
- Experiments made on the LITIV 2014 [7] and LITIV 2018 [9] dataset show that our method outperforms previous methods (either classical descriptors or CNNs).

References I

[1] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research*, vol. 17, pp. 1–32, 2016.

[2] W. Luo, A. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "Segstereo: Exploiting semantic information for disparity estimation," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[4] X. Song, X. Zhao, H. Hu, and L. Fang, "Edgestereo: A context integrated residual pyramid network for stereo matching," in *Computer Vision – ACCV 2018*, C. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Cham: Springer International Publishing, 2019, pp. 20–35.

[5] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3273–3282.

References II

[6] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[7] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi, "Thermal-visible registration of human silhouettes: A similarity measure performance evaluation," *Infrared Physics & Technology*, vol. 64, no. C, pp. 79–86, 2014.

[8] E. B. Baruch and Y. Keller, "Multimodal matching using a hybrid convolutional neural network," *CoRR*, vol. abs/1810.12941, 2018. [Online]. Available: http://arxiv.org/abs/1810.12941

[9] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Online mutual foreground segmentation for multispectral stereo videos," *International Journal of Computer Vision*, Jan 2019. [Online]. Available: https://doi.org/10.1007/s11263-018-01141-5

[10] D.-A. Beaupre and G.-A. Bilodeau, "Siamese cnns for rgb-lwir disparity estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.