Cross-Spectral Stereo Correspondence using Dense Flow Fields

Naveen Onkarappa[1], Cristhian A. Aguilera-Carrasco[1]
Boris X. Vintimilla[2] and Angel D. Sappa[1,2]

[1] *Computer Vision Center, Universitat Autònoma de Barcelona,*
08193 Bellaterra, Barcelona, Spain

[2] *CIDIS-FIEC, Escuela Superior Politécnica del Litoral (ESPOL),*
Campus Gustavo Galindo, Km 30.5 vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador

{*naveen, caguilera, asappa*}*@cvc.uab.es, boris.vintimilla@espol.edu.ec*

Keywords: Cross-Spectral Stereo Correspondence, Dense Optical Flow, Infrared and Visible Spectrum.

Abstract: This manuscript addresses the cross-spectral stereo correspondence problem. It proposes the usage of a dense flow field based representation instead of the original cross-spectral images, which have a low correlation. In this way, working in the flow field space, classical cost functions can be used as similarity measures. Preliminary experimental results on urban environments have been obtained showing the validity of the proposed approach.

1 INTRODUCTION

The coexistence of cameras working at different spectral bands is increasing (e.g., (Torabi et al., 2011), (Barrera et al., 2012)). For instance, in video surveillance applications long wavelength infrared sensors (LWIR), also referred in the literature to as thermal sensors, complement visual systems at night (Snidaro et al., 2004). The same happens in the driving assistance domain, where thermal information helps detecting pedestrians (Krotosky and Trivedi, 2007). Working on such a cross-spectral domain, we propose to explore the possibility of finding correspondences between the images through a cost function based on the use of dense flow fields.

The multimodal correspondence problem between LWIR and visible spectrum (VS) images has been recently studied in several works for different applications such as image registration or stereovision. The main challenge lies on the low correlation between LWIR and VS. A way to overcome this lack of correlation is by means of the use of prior knowledge of the objects in the scene; so the correspondence search is limited to those regions of interest (ROI) such as human silhouettes (Krotosky and Trivedi, 2008), faces (Socolinsky and Wolff, 2009), or contrasted objects (Yang and Chen, 2011) (e.g., hot or cold objects on a smooth background). Although these ROI based approaches have shown attractive results, the main problem remains unsolved.

A more general solution has been presented in (Barrera et al., 2013). In that work the authors propose a matching cost function for a multispectral stereo rig based on mutual and gradient information in a scale space representation. This cost function is later on used in a minimization scheme, which is able to extract dense disparity maps assuming a piecewise planar representation. Although interesting results have been presented, the main drawback lies on the use of an expensive minimization scheme and on the assumption of a piecewise planar representation. In order to avoid the low correlation problem, (Pistarelli et al., 2013) propose to search for correspondences in the Hough space; hence, edges from the given cross-spectral images are extracted and represented in the Hough space. These edges correspond to predominant geometries in the scene. Only matchings over edges can be obtained with this approach.

Similarly to the previous approaches, in the current work we propose not to search for correspondences in the image space, where there is a low correlation between LWIR and VS, but in a space where both images have the same representation. The proposed work is intended to be used in dynamic scenarios where there is a relative motion between the cross-spectral stereo rig and the objects contained in the scene. In our particular case the cameras are used for driving assistance and they are placed on the top of our vehicle (see Fig. 1). The proposed approach is based on the fact that since both cameras are rigidly attached to the same rig, and images accordingly calibrated and rectified, the optical flow information can be used for the correspondence search. Cross-spectral video sequences of urban scenarios are used for the evaluation.

The manuscript is organized as follow. Section 2 gives details about the cross-spectral stereo rig. The proposed approach is introduced in Section 3. Experimental results on urban scenarios are discussed in Section 4. It should be noticed that current manuscript is intended to show how cross-spectral correspondence problem can be tackled by representing the given images in another domain. More deep and rigorous validations would be required to show the advantage of the proposed approach. Finally, conclusions and future work are given in Section 5.

2 SYSTEM SETUP

The stereo head used in the current work consists of a pair of cameras separated by a baseline of about 12 cm and a non verged geometry. The images provided by the cross-spectral stereo head are calibrated and rectified using (Bouguet, 2010); a process similar to the one presented in (Barrera et al., 2012) is followed. It consists of using a thin aluminium metallized paper, with black and white squares that is observed by both cameras. The LWIR camera (Gobi-640-GigE from Xenics) provides images up to 50*fps* with a resolution of 640×480 pixels. The visible spectrum camera is an ACE from Basler with a resolution of 658×492 pixels. Both cameras are synchronized using an external trigger. Camera focal lengths were set so that pixels in both images contain similar amount of information from the given scene. Figure 1 shows an image of the stereo rig together with a commercial stereo camera (Bumblebee XB3 from Point Grey), which will be used in a future work for validating the cross-spectral stereo correspondence. The whole platform is placed on the roof of a vehicle for driving assistance applications.

3 PROPOSED APPROACH

This section details the two steps of the proposed approach. First, an overview of the variational formulation used for the optical flow estimation is provided. Then, cost functions that could be used for finding correspondences are detailed.

3.1 Dense Optical Flow

Several variational approaches have been proposed for optical flow estimation in recent years[1]. Variational formulations involve a data term and a regularization term. The data term formulates the assumption matching characteristics, typically the intensity of the pixel; it is also called brightness constancy assumption (BCA) (Horn and Schunk, 1981). The BCA can be formulated as: $I_1(\boldsymbol{x}+\boldsymbol{u}) - I_0(\boldsymbol{x}) = 0$, where I_0 and I_1 are two consecutive images, $\boldsymbol{x} = (x,y)$ is the pixel location within the image space $\Omega \subseteq \mathbf{R}^2$; $\boldsymbol{u} = (u(\boldsymbol{x}), v(\boldsymbol{x}))$ is the two-dimensional flow vector. Linearizing the above equation using first-order Taylor expansion we get BCA as: $(I_x u + I_y v + I_t)^2 = 0$, where subscripts denote the partial derivatives. Using only BCA does not provide enough information to infer meaningful flow fields, making the problem ill-posed. In order to solve this ill-posed problem a regularization term is needed. In (Horn and Schunk, 1981) a regularization term, based on the assumption that resulting flow field is globally smooth all over the image, is proposed. Combining BCA and homogeneous regularization in a single variational framework and squaring both constraints yields the following energy function:

$$E(\boldsymbol{u}) = \int_{\Omega} \{ \underbrace{(I_x u + I_y v + I_t)^2}_{Data\ Term} + \alpha \underbrace{(|\nabla u|^2 + |\nabla v|^2)}_{Regularization} \} \, d\boldsymbol{x}, \tag{1}$$

where α is the regularization weight. This energy function is minimized for flow vectors using corresponding Euler-Lagrange equations.

Since the current work is intended for driving scenarios, we decided to use the Laplacian of flow components, instead of their gradients, as penalization in the regularization term. This strategy has been recently presented in (Onkarappa and Sappa, 2013) showing interesting results. Hence, the energy function becomes:

$$E(\boldsymbol{u}) = \int_{\Omega} \{ \underbrace{(I_x u + I_x v + I_t)^2}_{Data\ Term} + \alpha \underbrace{(|\triangle u|^2 + |\triangle v|^2)}_{Regularization} \} \, d\boldsymbol{x}. \tag{2}$$

The resulting flow field is depicted using the classical optical flow color map, where color indicates the

[1] http://vision.middlebury.edu/flow/

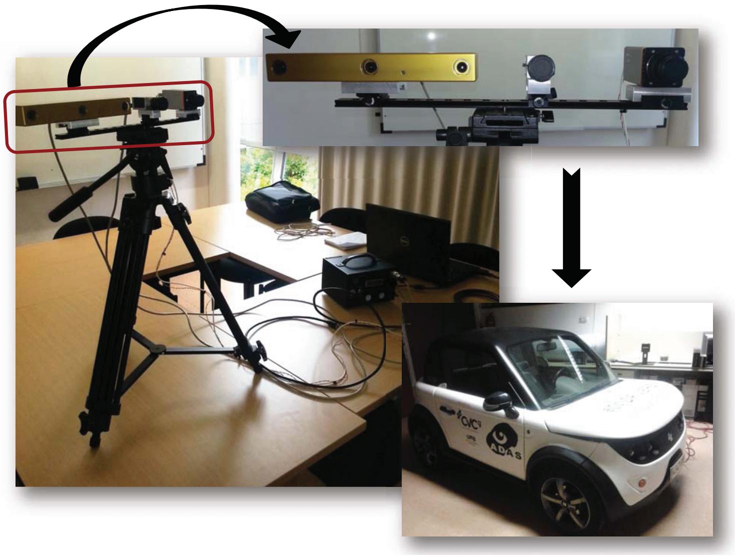

Figure 1: Cross-spectral stereo rig together with a commercial system to be used as a ground truth.

direction and intensity indicates the magnitude (Fig. 2(*right*) shows optical flow depicted using color map coding). Hereinafter, I_{VS} will refer to the color coded flow field of visible spectrum sequence, while I_{LWIR} corresponds to the color coded flow field of the LWIR spectrum sequence.

3.2 Correspondence Search

Once every pixel in both images (VS and LWIR) is associated with a vector representing the relative motion between the camera and the point in the 3D space, correspondences are computed. This section just presents three similarity measures that could be used as correspondence functions in the flow field domain: sum of absolute differences (SAD), sum of squared differences (SSD), and normalized cross-correlation (NCC) (Faugeras et al., 1993). During last decades there have been several improvements looking for invariance to photometric distortions or for producing a disparity map with sub-pixel accuracy (e.g., (Shimizu and Okutomi, 2001), (Scharstein and Szeliski, 2002), (Psarakis and Evangelidis, 2005)).

In the current work, since the cross-spectral matching is not implemented at the photometric level, but at a color level corresponding to the flow field representations, we propose to evaluate SAD, SSD and NCC—their corresponding expressions are presented in Equations (3), (4) and (5), where d represents the disparity value, N the correlation window and $\overline{N}$ the mean value of the elements in that correlation window. In all the cases the images I_{VS} and I_{LWIR} representing the flow fields are considered. At the moment just a winner-take-all strategy is considered for assigning the best matches, in the future other strategies will be implemented to improve results.

4 EXPERIMENTAL RESULTS

The proposed approach has been tested using cross-spectral image pairs from the on-board stereo rig presented in Section 2. The main intention of current work is to evaluate the validity of the proposed scheme, instead of the accuracy of obtained results that will need larger data sets with different scenarios. Figure 2 shows a pair of VS images and a pair of LWIR images (left column) together with their corresponding dense flow field (right column). From these flow fields the three cost functions presented in Equations (3), (4) and (5) have been evaluated. The corresponding sparse disparity maps obtained using the different cost functions are similar.

In all the cases sparse representations were obtained since only those correspondence values with a cost smaller than a given threshold were considered as valid. We can conclude that moving from the cross-spectral domain to another representation, common to both inputs, can be useful for finding correspondences. Similar conclusion was presented in the literature by mapping the given images to the Hough space (e.g., (Pistarelli et al., 2013)). The main problem lies on the tradeoff between the dense optical flow representation and the accuracy of those flow vectors.

$$C_{SAD_{(x,y,d)}} = \sum_{(i,j)\in N_{x,y}} |I_{VS}(i,j) - I_{LWIR}(i+d,j)| \tag{3}$$

$$C_{SSD_{(x,y,d)}} = \sum_{(i,j)\in N_{x,y}} (I_{VS}(i,j) - I_{LWIR}(i+d,j))^2 \tag{4}$$

$$C_{NCC_{(x,y,d)}} = \frac{\sum_{(i,j)\in N_{x,y}} (I_{VS}(i,j) - \overline{I_{VS}(N_{x,y})}) \cdot (I_{LWIR}(i+d,j) - \overline{I_{LWIR}(N_{x,y})})}{\sqrt{\sum_{(i,j)\in N_{x,y}} (I_{VS}(i,j) - \overline{I_{VS}(N_{x,y})})^2 \cdot (I_{LWIR}(i+d,j) - \overline{I_{LWIR}(N_{x,y})})^2}} \tag{5}$$

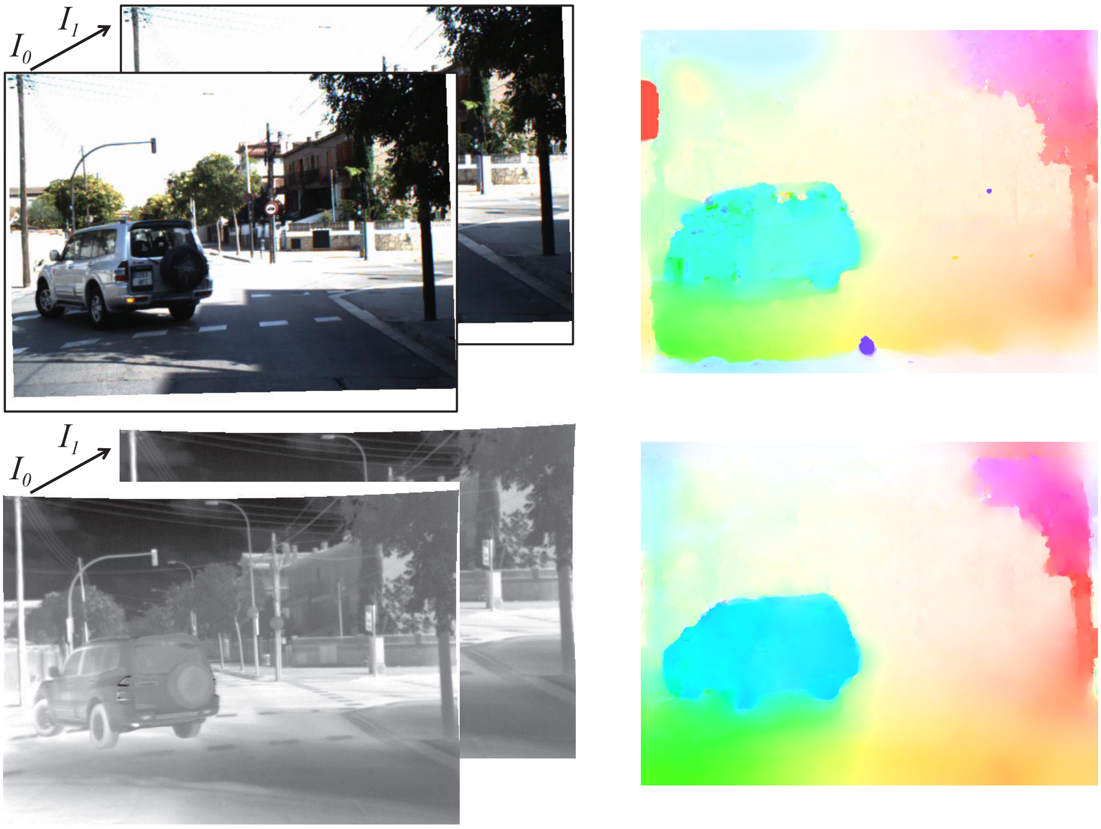

Figure 2: $(left)$ Two pairs of cross-spectral images (VS and LWIR). $(right)$ Dense flow fields used for the correspondence search.

The dense optical flow representation is obtained by the regularization term, which somehow smooths the results. This smoothing indirectly affect the accuracy during the correspondence search.

From the obtained results it is not clear which is the best correspondence function, all of them have a similar result. Another problem to be considered for future work is the lack of ground truth disparity value, although we expect to use the Bumblebee camera (at the left side in our stereo rig, see Fig. 1) temporal synchronization problems need to be solved first (Bumblebee work up to 15*fps*).

5 CONCLUSIONS AND FUTURE WORK

This manuscript presents an approach to tackle the low correlation between cross-spectral stereo images by representing the given inputs in a common space. It is based on the use of dense optical flow in the cost function formulation. As a preliminary study, we present some conclusions using different similarity measures. More deep validations should be performed to find the best similarity measure together with the best matching criteria (currently a winner-takes-all is considered). Additionally, ground truth values will be considered for such a validation. As mentioned in Section 1, current manuscript is intended to show preliminary results from our approach and how the proposed scheme can represent an alternative to overcome the lack of correlation, or to improve the results from cross-spectral correspondence search.

ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish Government under Research Project TIN2011-25606 and PROMETEO Project of the ”Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación de la República del Ecuador”. Cristhian A. Aguilera-Carrasco was supported by a grant from ”Universitat Autònoma de Barcelona”.

REFERENCES

Barrera, F., Lumbreras, F., and Sappa, A. D. (2012). Multimodal stereo vision system: 3d data extraction and algorithm evaluation. *J. Sel. Topics Signal Processing*, 6(5):437–446.

Barrera, F., Lumbreras, F., and Sappa, A. D. (2013). Multispectral piecewise planar stereo using manhattan-world assumption. *Pattern Recognition Letters*, 34(1):52–61.

Bouguet, J.-Y. (2010). Camera calibration toolbox for matlab.

Faugeras, O., Hotz, B., Mathieu, H., Vieville, T., Zhang, Z., Fua, P., Theron, E., Moll, L., Berry, G., Vuillemin, J., Bertin, P., and Proy, C. (1993). Real time correlation-based stereo: algorithm, implementations and applications. In *INRIA Technical Report, No. 2013, August 1993*, page 56.

Horn, B. K. P. and Schunk, B. G. (1981). Determining optical flow. *Artificial Intelligence*, 17:185–203.

Krotosky, S. J. and Trivedi, M. M. (2007). On color-, infrared-, and multimodal-stereo approaches to pedestrian detection. *IEEE Trans. on Intelligent Transportation Systems*, 8(4):619–629.

Krotosky, S. J. and Trivedi, M. M. (2008). Person surveillance using visual and infrared imagery. *IEEE Trans. on Circuits and Systems for Video Technology*, 18(8).

Onkarappa, N. and Sappa, A. D. (2013). Laplacian derivative based regularization for optical flow estimation in driving scenario. In *15th International Conference on Computer Analysis of Images and Patterns, York, UK, August 27-29, 2013*, pages 483–490.

Pistarelli, M. D., Sappa, A. D., and Toledo, R. (2013). Multispectral stereo image correspondence. In *15th International Conference on Computer Analysis of Images and Patterns, York, UK, August 27-29, 2013*, pages 217–224.

Psarakis, E. Z. and Evangelidis, G. D. (2005). An enhanced correlation-based method for stereo correspondence with sub-pixel accuracy. In *10th IEEE International Conference on Computer Vision, 17-20 October 2005, Beijing, China*, pages 907–912.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42.

Shimizu, M. and Okutomi, M. (2001). Precise sub-pixel estimation on area-based matching. In *International Conference on Computer Vision*, pages 90–97.

Snidaro, L., Foresti, G. L., Niu, R., and Varshney, P. K. (2004). Sensor fusion for video surveillance. In *Electrical Engineering and Computer Science*.

Socolinsky, D. A. and Wolff, L. B. (2009). Face recognition in low-light environments using fusion of thermal infrared and intensified imagery. In Hammoud, R. I., editor, *Augmented Vision Perception in Infrared*, Advances in Pattern Recognition, pages 197–211. Springer London.

Torabi, A., Najafianrazavi, M., and Bilodeau, G. (2011). A comparative evaluation of multimodal dense stereo correspondence measures. In *IEEE Int'l Symp. on Robotic and Sensors Environments*, volume 1, pages 143–148.

Yang, R. and Chen, Y. (2011). Design of a 3-D infrared imaging system using structured light. *IEEE Trans. on Instrumentation and Measurement*, 60(2):608 –617.