Multimodal Stereo Vision Using Mutual Information with Adaptive Windowing

Mustafa Yaman
Computer Eng., Middle East Technical University
Ankara, Turkey
mustafa.yaman@ceng.metu.edu.tr

Sinan Kalkan
Computer Eng., Middle East Technical University
Ankara, Turkey
skalkan@ceng.metu.edu.tr

Abstract

This paper proposes a method for computing disparity maps from a multimodal stereovision system composed of an infrared and a visible camera pair. The method uses mutual information (MI) as the basic similarity measure where a segmentation-based adaptive windowing mechanism is proposed for greatly enhancing the results. On several datasets, we show that (i) our proposal improves the quality of existing MI formulation, and (ii) our method can provide depth comparable to the quality of Kinect depth data.

1. Introduction

Stereovision [1],[2] deals with computing depth information in a scene by finding the projections of 3D points in images of the scene captured from two or more cameras. Finding which pixels in the different images are the corresponding projections of the same 3D point is the most crucial part in stereovision. Although there are many approaches in the literature for the correspondence problem in unimodal stereo, they are not directly applicable to multimodal stereovision since they depend on pixel intensities, which cannot be used in multimodal stereovision as images are captured by multimodal cameras, such as for an infrared/thermal camera vs. visible camera pair.

Although there are plenty of local (e.g. [7],[8]) or global [9], dense [5],[9] or sparse [3],[6] approaches available in the literature for unimodal stereo (see also [3],[4],[5] for reviews), there are not many studies on multi-modal stereo vision (except for [10]-[18]). All these studies use Mutual Information (MI) as the basis for computing the similarity measures. Egnal [10] is, up to our knowledge, the first to use mutual information (MI) for stereo image pairs that were unimodal but red / blue filtered or differently lighted, but also multimodal (an Near-IR and Visible/NearIR couple). The results were promising and revealed the power of MI compared to standard correlation based methods especially on images with different spectral characteristics for the same scene, although still had low accuracy / quality. Fookes et al. extended the MI-based approach with adaptive windowing [11] and integrated prior probabilities using a 2D match surface [12]. However, their methods were only tested on synthetically altered unimodal images, which do not actually include different segmentation / edge characteristics that multimodal images may have. Similarly, Krotosky and Trivedi [13]-[15] used MI successfully for pedestrian detection / person tracking using a multimodal (infrared / visible camera pair) stereovision system. They constrained stereo correspondence within region of interests (ROI) including people's bodies, and proposed a disparity voting method for computing the final depth information for the corresponding regions.

In a very recent work on multi-modal stereovision, Campo et. al. [16] propose an MI-based method where the similarity measures were extended using the gradient information. They implemented a multimodal stereo head (thermal vs. visible) and a database also. The 3D depth results presented in their work are yet quite sparse for the scenes tested however claimed promising due to the challenge of trying to match two separate spectral bands.

Recently, LSS (local self similarity), originally proposed for image template matching [19] is tried as a thermal-visible stereo correspondence measure [17], since it is already shown to outperform MI in template matching, and has some advantages like better handling textured / colored regions as long as they have a similar spatial layout. They implemented an ROI based image matching by tracking people in the scene and compared with MI based similarity descriptors, and showed that LSS measures outperform MI and HoG (Histogram of Oriented Gradients) [18], however this measure is not yet tested for a dense disparity / depth map calculation.

1.2 This study and the contributions

We propose a new MI-based multimodal stereovision framework whose novelty is a new adaptive windowing method MI formulation. We determine the adaptively sized windows by the segmentation of the images, which help generating a robust correlation surface when computing joint probabilities to compute the joint entropy and the MI similarity metric. Our results are not quantitatively comparable to existing MI-based methods since they have used different sets of images; however, by visual evaluation, it is possible to say that there is significant improvement in obtained disparity/depth maps. Besides, using synthetically altered images of Middlebury stereo image database [20] (where the left images are replaced with the synthetically altered versions of these images (similar to [12]), it was possible to compute the statistics of test results to gain more knowledge of the performance of our method. In addition, by using the Kinect device having an IR and RGB camera along with an IR projector, evaluation of the method has been possible for a real multimodal camera system since it has a built in depth computation, which was not performed

before as we know of for such applications.

2. Method

Our method (Figure 1) takes a rectified multimodal image pair and follows these steps:

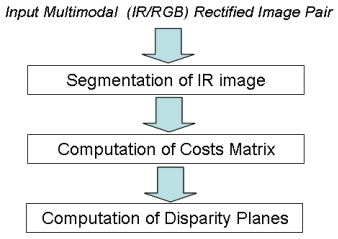


Figure 1. Overview of the Method

Step1-Segmentation of the IR Image: We use the mean shift segmentation algorithm [23] for segmenting the IR image. With this step, we get non-overlapping segments representing homogenous regions in the IR image. We assume that each segment corresponds to a planar surface in the scene (a common assumption, see, e.g., [24]). The reason for segmenting only the IR image is that the surfaces in IR images are also common in the RGB images but the reverse is not true. RGB images contain more detailed and textured surfaces which do not exist in the IR images in our datasets.

Step2-Computation of Cost Matrix: The cost matrix computation step is the most important step containing our contributions in this study (see Alg. 1). The inputs to the algorithm are the left (IR) image L, the right (RGB) image R and the left segmentation S_l.

Algorithm 1: Cost Matrix Computation

Inputs: L, R, S_l,
Compute $P_{prior}(L,R)$
for $r=[0:height)$ **do**:
  for $c = [0, width)$ **do**:
    for $d = [0, d_{max}]$ **do**:
      $C(r,c,d) = - M(W_L(r,c), W_R(r,c-d), S_l, P_{prior})$
    end for
  end for
end for

The algorithm first computes joint prior probabilities for all corresponding pixel intensities in left and right images without considering any disparities, in a straightforward fashion:

$$P_{prior}(I_l,I_r) = h(I_l,I_r) \Big/ \sum_{l,r} h(I_l,I_r) \quad (1)$$

where I_l, I_r are the intensity of corresponding pixels. Prior probabilities are computed using h, the 2D histogram of corresponding pixel intensities.

Next, we compute the cost matrix for all pixels by computing MI (negative of the MI measure is used) using the proposed adaptive windowing scheme as:

$$W_L(r,c) = L(r_{\min}:r_{\max}, c_{\min}:c_{\max}) \quad , \quad (2)$$

$$\begin{aligned} c_{\min} &= c - \delta c_l - \omega, c_{\max} = c + \delta c_r + \omega \\ r_{\min} &= r - \delta r, r_{\max} = r + \delta r \end{aligned} \quad , \quad (3)$$

where δc_l and δc_r are distances to the border of the segment which the current pixel (r,c) belongs to, and the window is enlarged by ω, the assumed thickness of discontinuity at the images on the segment border (Figure 2). δr similarly provides the window size in vertical direction and it is currently a user configured parameter ($\leq$ *5 pixels*) determined experimentally. We did not consider the segment borders in the vertical direction since the segment plane may not be a fronto-planar surface and may confuse the cost calculation. We applied the same window to the right image by moving the window for each candidate disparity d.

$$W_R(r,c-d) = R(r_{\min}:r_{\max}, c_{\min}-d:c_{\max}-d) \quad (4)$$

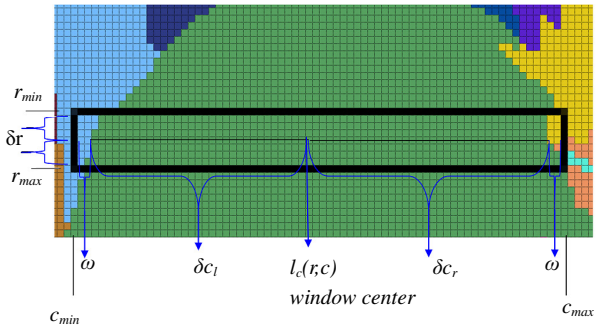


Figure 2. Adaptive window calculation

After we determine the adaptive windows to be matched, we compute MI between the two windows (W_L, W_R) using the segment information and the prior probabilities as:

$$M(W_L,W_R,S_l,P_{prior}) = \sum_W P(I_l,I_r) \ln \frac{P(I_l,I_r)}{(P(I_l)P(I_r))} \quad (5)$$

where joint probabilities are computed using the adaptive correlation surface that we developed (P_w), which is incorporated with prior probabilities [12] (P_{prior}) as below:

$$P(I_l,I_r) = \lambda P_w(I_l,I_r) + (1-\lambda) P_{prior}(I_l,I_r) \quad (6)$$

The correlation surface enabling joint probability calculation is another key contribution of ours for the MI cost calculation, where the joint histogram is calculated by considering pixels within the current segment in the window and the pixels nearby the edge of the segment as:

$$P_w(I_l,I_r,S_i) = h_w(I_l,I_r,S_i) \Big/ \sum_w h_w(I_l,I_r,S_i) \quad (7)$$

$$h_w(I_l,I_r,S_i) = \sum_w T(I_l,I_r,S_i) \quad (8)$$

$$T(I_l,I_r,S_i) = \begin{cases} k & if\ S_i(l) = S_i(l_c)\ \&\ L1(l,S_i(l_c)) > \xi \\ k/(\lambda L1(l,S_i(l_c))) & elseif\ \ L1(l,S_i(l_c)) \leq \xi \\ 0 & otherwise \end{cases} \quad (9)$$

The *L1 distance* term (Figure 3) in Eq. 9 incorporates the pixels near the segment borders to MI calculation with some penalty due to possible occlusions around borders and this way we managed to consider both the segment and the edges excluding other segments within the rectangular window in MI measure computation.

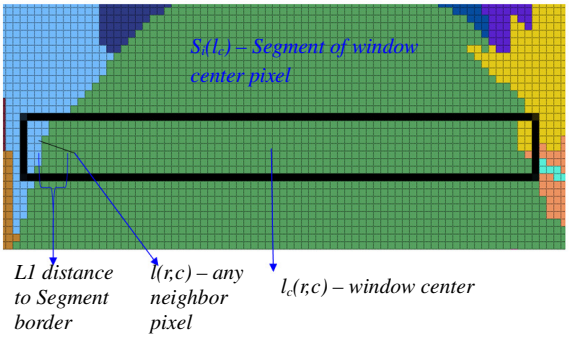


Figure 3. Adaptive MI computation surface using segmentation

Step3-Computation of Disparity Planes: In this step, we first compute the WTA ("winner take all") disparities for all pixels by selecting the disparity with minimum cost for all pixels in the cost matrix. Later, we fit planes to WTA disparities in a segment using RANSAC [22].

However, we perform this iteratively by splitting segments if the outliers of the computed plane constitute regions of size greater than a designated threshold. This way, we reduce the dependency of the performance of the algorithm on the initial segmentation.

3. Results & Discussion

Regarding the Middlebury images [20], it is possible to compute the performance statistics with the ground truth provided for unimodal stereovision as percentage of bad pixels having the disparity error greater than a designated threshold. In Table 1 & Table 2, we generated the percentage of bad disparities (*disparity error > 1 pixel*) for a set of Middlebury images [20], although, the images are synthetically altered *(cos (I*π/255))* for the left images. Table 1 includes the Winner-Take-All (WTA) results and Table 2 includes the final disparity planes fitted to segments. In each table, we also provide the statistics computed when no adaptive windowing is used but rather, a regular rectangular window extracted from the neighborhood of corresponding pixels are used for MI measure computation.

Table 1. Results on Synt. Altered Middlebury Images for WTA Disparity Selection

Image*	Adap.	Bad (all)	Bad (nocc)	Bad (disc)
Tsukuba	No	17.7%	16.1%	24.7%
	Yes	**6.6%**	**5.6%**	**16.7%**
Venus	No	26.0%	24.8%	40.7%
	Yes	**10.5%**	**9.7%**	**20.0%**
Teddy	No	43.3%	36.9%	45.1%
	Yes	**36.2%**	**29.9%**	**36.6%**
Cones	No	35.8%	27.8%	40.2%
	Yes	**28.3%**	**20.0%**	**30.7%**

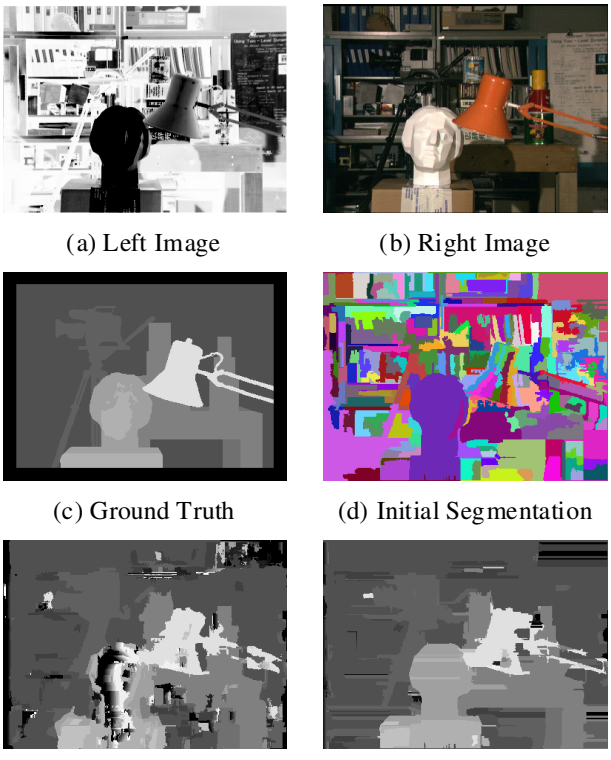
**none-adaptive method window size=11, adaptive method vertical window size=11*

Table 2. Results on Synt. Altered Middlebury Images for Disparity Plane Computation

Image*	Adap.	Bad (all)	Bad (nocc)	Bad (disc)
Tsukuba	No	16.8%	15.4%	24.9%
	Yes	**6.2%**	**5.4%**	**16.7%**
Venus	No	24.9%	23.6%	39.1%
	Yes	**11.8%**	**11.1%**	**20.1%**
Teddy	No	43.4%	37.0%	44.5%
	Yes	**36.1%**	**30.0%**	**37.5%**
Cones	No	34.5%	26.5%	39.0%
	Yes	**28.0%**	**19.9%**	**30.5%**

**none-adaptive method window size=11, adaptive method vertical window size=11*

In Figure 4, we provide an example Middlebury image pair, along with adaptive and non-adaptive window disparity plane results with the initial and final segmentations after the disparity plane computation step.



(a) Left Image (b) Right Image

(c) Ground Truth (d) Initial Segmentation

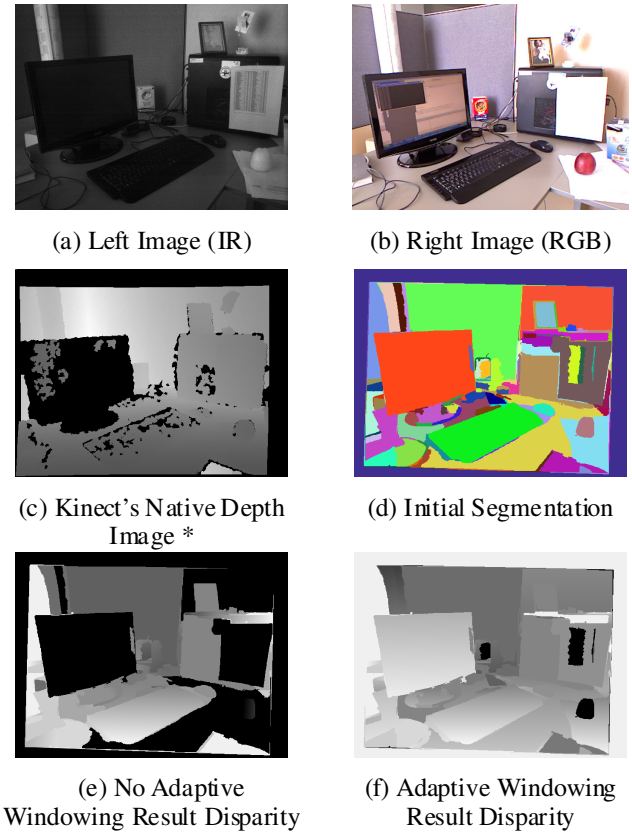
(e) No Adaptive Windowing Result (f) Adaptive Windowing Result

Figure 4. Results on a synthetically altered Middlebury image (see text for details).

As can be observed from both visual and statistical results, significant progress is achieved by the method when compared to a non-adaptive local window MI calculation scheme for multi-modal stereovision.

In Figure 5, we provide sample results from Kinect data, including Kinect's native depth image and our none-adaptive and adaptive method result disparity images for visual comparison. We see that our method again improves the disparity map compared to the fixed-window MI method. Moreover, we observe that the

disparity map generated by our method can be used to improve the quality of the depth calculated by Kinect, especially on edges and non fronto-planar surfaces.



(a) Left Image (IR) (b) Right Image (RGB)

(c) Kinect's Native Depth Image * (d) Initial Segmentation

(e) No Adaptive Windowing Result Disparity (f) Adaptive Windowing Result Disparity

Figure 5. Results on a Kinect captured image pair (**brighter pixels has more depth)*

4. Conclusion

In this paper, we proposed a multi-modal stereovision framework using novel adaptive-window based MI similarity. On synthetically altered Middlebury database and a set of Kinect-captured RGB-IR image pairs, we show that our method can improve over MI-based similarity. We have also developed an energy-based optimization method and observed similar improvements (results not provided).

References

[1] R. Hartley and A. Zisserman: "Multiple View Geometry in Computer Vision," Cambridge Univ. Press, 2000.

[2] R. Szeliski, "Computer Vision: Algorithms and Applications," Springer, 2010.

[3] U.R. Dhond, et. al.: "Structure from Stereo—A Review," IEEE Trans. Systems, Man, and Cybernetics, vol. 19, pp. 1489-1510, 1989.

[4] Z. B. Myron et. al.: "Advances in Computational Stereo," PAMI, vol.25, no.8, pp.993-1008,2003.

[5] D. Scharstein and R. Szeliski: "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," IJCV, vol. 47, no. 1, pp. 7-42, 2002.

[6] V. Venkateswar and R. Chellappa: "Hierarchical Stereo and Motion Correspondence Using Feature Groupings," IJCV, vol. 15, pp. 245-269, 1995.

[7] M. J. Hannah: "Computer Matching of Areas in Stereo Images," Ph.D. Thesis, Stanford University, 1974.

[8] K. Ambrosch et. al.: "Flexible Hardware-Based Stereo Matching," EURASIP Journal on Embedded Systems, 2008.

[9] C. Cassisa: "Local vs global energy minimization methods: application to stereo matching," Proc. of the Int. Progress in Informatics and Computing, 2010.

[10] G. Egnal: "Mutual information as a stereo correspondence measure," Technical Report MS-CIS-00-20, Uni. of Pennsylvania, 2000.

[11] C. Fookes et. al.: "A new stereo image matching technique using mutual information," Int. Conf. on Computer, Graphics and Imaging, 2001.

[12] C. Fookes et. al.: "Multi-spectral stereo image matching using Mutual Information," 2nd Int. Symposium on 3D Data Processing, Visualization, and Transmission, 2004.

[13] S.J. Krotosky, and M.M. Trivedi: "Multimodal Stereo Image Registration for Pedestrian Detection," Proc. IEEE Conference on Intelligent Transportation Systems, 2006.

[14] S.J. Krotosky, and M.M. Trivedi: "Registration of Multimodal Stereo Images using Disparity Voting from Correspondence Windows," Proc. IEEE International Conference on Advanced Video and Signal based Surveillance, 2006.

[15] S.J. Krotosky, and M.M. Trivedi: "Mutual Information Based Registration of Multimodal Stereo Videos for Person Tracking," Computer Vision and Image Understanding, Special Issue on Advances in Vision Algorithms and Systems Beyond the Visible Spectrum, vol.106, pp.2-3, 2007.

[16] F.B. Campo et. al.: "Multimodal Stereo Vision System: 3D Data Extraction and Algorithm Evaluation," IEEE Journal of Selected Topics In Signal Processing, vol. 6, no.5, 2012.

[17] A. Torabi and G.-A. Bilodeau: "Local self-similarity as a dense stereo correspondence measure for thermal-visible video registration," IEEE CVPR Workshops, 2011.

[18] A. Torabi et.al: "A comparative evaluation of multimodal dense stereo correspondence measures," IEEE Int. Symposium on Robotic and Sensors Environments (ROSE), pp. 143-148, 2011.

[19] E. Shechtman and M. Irani: "Matching local self-similarities across images and videos," CVPR, 2007.

[20] The Middlebury Stereo Vision Page, The Datasets: http://vision.middlebury.edu/stereo/data/

[21] R. Szeliski et. al.: "A Comparative Study of Energy Minimization Methods for Markov Random Fields", ECCV, 2006.

[22] M.A. Fischler and R.C. Bolles: "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," Comm. ACM, vol. 24, pp. 381-395, 1981.

[23] D. Comanicu and P. Meer: "Mean shift: A robust approach toward feature space analysis," PAMI, vol.24, pp.603–619, 2002.

[24] Z. Wang and Z. Zheng: "A region based stereo matching algorithm using cooperative optimization", CVPR, 2008.