I spy with my little eye:
Learning Optimal Filters for Cross-Modal Stereo under Projected Patterns

Wei-Chen Chiu Ulf Blanke Mario Fritz
Max-Planck-Institute for Informatics, Saarbrcken, Germany
{walon, blanke, mfritz}@mpi-inf.mpg.de

Abstract

With the introduction of the Kinect as a gaming interfaces, its broad commercial accessibility and high quality depth sensor has attracted the attention not only from consumers but also from researchers in the robotics community. The active sensing technique of the Kinect produces robust depth maps for reliable human pose estimation. But for a broader range of applications in robotic perception, its active sensing approach fails under many operating conditions such like objects with specular and transparent surfaces.

Recently, an initial study has shown that part of the arising problems can be alleviated by complimenting the active sensing scheme with passive, cross–modal stereo between the Kinect's rgb and ir camera. However, the method is troubled by interference from the IR projector that is required for the active depth sensing method. We investigate these issues and conduct a more detailed study of the physical characteristics of the sensors as well as propose a more general method that learns optimal filters for cross–modal stereo under projected patterns. Our approach improves results over the baseline in a point-cloud-based object segmentation task without modifications of the kinect hardware and despite the interference by the projector.

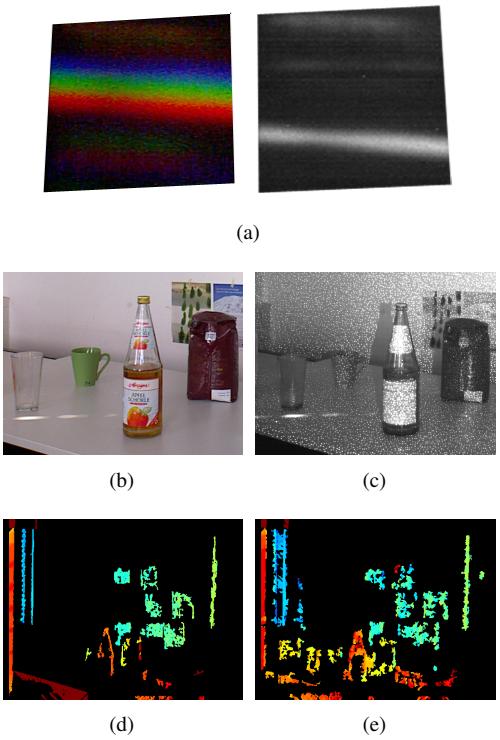


Figure 1: 1(a) Response of RGB camera (left) and IR camera (right). 1(b) and 1(c) Image pair obtained by Kinect with projected IR pattern. 1(d) Disparity map on unfiltered pairs. 1(e) Disparity map on patch-filtered image pairs.

1. Introduction

Despite the advance of elaborated global (e.g. [2]) and semi-global (e.g. [5]) stereo matching techniques, real-time stereo on standard hardware is still dominated by local method based on patch comparisons. The more surprising it is that we have seen very little work on improving the correspondences by a learning approach that would be better suited to a certain setting or conditions [10]. Yet the use of different pre-filters that are used by practitioners to improve the matching process are clear evidence that there is room for improvement over basic patch-based differences.

In our case the need for learning is even more apparent as we attempt cross-modal matching between the IR and RGB image obtained from the Kinect sensor. Such a system was recently proposed [1] which augments the active sensing strategy of the kinect by a passive stereo algorithm between the two available imagers. A very simple pixel based re-weighting scheme was proposed that produces an IR like image for improved depth estimates.

This paper identifies three issues and consequently improves over the previous work threefold:

First, we take a closer look at the sensor characteristics

of the kinect and realize that the overlap in the spectral response between the sensors is very small. This argues for a learning based approach that exploits smoothness and correlations in the BRDF function of the materials, as no satisfactory linear reconstruction of the IR channel is possible. Nevertheless we attempt such a reconstruction and add this as a baseline.

Second, as argued above we know from practical considerations that patch based stereo matching is often improved by pre-filter operations. This is a richer class than the pixel based weighting previously investigated. We propose a method for learning optimal filters for improving cross–modal stereo that is rich enough to capture channel-based weighting and filtering like sharpening, smoothing and edge detection.

Third, we realize that for the best results previously obtained [1] the IR projector had to be covered for capturing the IR-RGB pair. However, this is impracticable as it eliminates the active depth sensing scheme. We show that our learning based algorithm can achieve robustness of the stereo algorithm to these nuisances introduced by the projector – and in fact we are able to recover the performance previously only achieved with the covered projector.

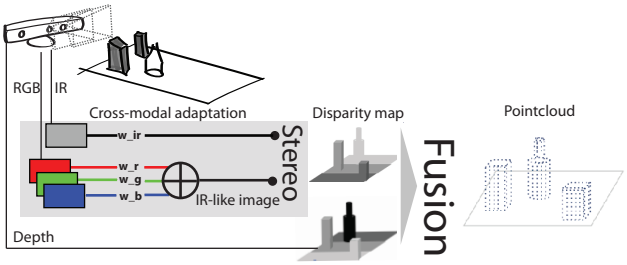
2. Related Work

Stereo vision has a far back reaching history and is studied well in the past decades. However, lighting conditions or specific material property such as transparency and specularity [7] still complicate stereo matching. In practice such variations are typically reduced by filtering techniques (e.g., laplacian of gaussians [9]), non-parametric matching costs (e.g., census [13]) or by hand tuning parameters for optimal matching. [6] provide a thorough comparison of several stereo matching techniques with respect to complex radiometric variations. They compare a large set of filters, and rank them according to performance and computational efficiency.

More recently the path of machine learning is taken to find automatically optimal models for stereo matching [10]. Also [6] propose to learn pixelwise cost based on mutual information from ground truth data. However, both approaches are global-based matching scheme and prohibit real time applications. Also the sensitivity to local changes [6] limits its applicability for matching across modalities that exhibit global as well as local variation.

Objects with transparent or specular surfaces have been also focused on as a detection task. [3] learns object models from data and [8] detect transparent objects employing a second time of flight camera. For increasing robustness of visual category recognition across different sources such as from the web, high quality DSLRs or webcams [12] the metric learning formulation proved to be a successful.

The Kinect depth estimate yields impressive results on



(a)

Figure 2: Diagrams for weighted fusion scheme.

(un)structured lambertian surfaces. Yet it completely fails on specular, transparent or reflective surfaces. This issue is addressed in previous work [1] by using a cross–modal stereo approach based on built–in RGB and IR sensor that complements the Kinect depth estimate.

In order to study the similarity between IR and RGB the authors of [1] investigate various fusion schemes and present a pixel-based optimization based on ground truth (see Fig 2). In contrast to their work, we replace the pixel based weighting and focus on learning patch based filters.

3. Capturing and Analyzing Sensor Characteristics of the Kinect

In [1] a mapping between IR and RGB was learned from patterns that were illuminated by environmental light. However, there was no justification given if there is any hope to actually recover the sensor response characteristic by a linear combination of the RGB channels. Therefore we provide here a first analysis of the sensor characteristics of the imagers in the Kinect.

To this end we capture diffracted light which is projected on a “white” surface. This allows us to determine the characteristics of the Kinect cameras by measuring their response to different wavelength (see Fig 1).

The setup is depicted in Fig. 3. Environmental light is

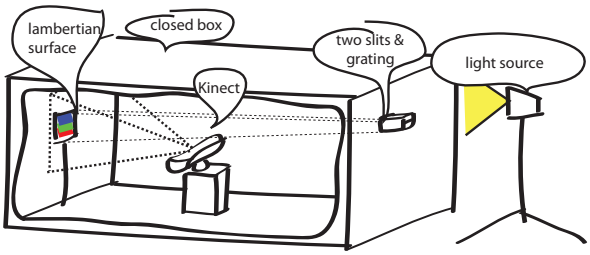


Figure 3: Schematic for experimental setup for reading sensor characteristics.

shielded so that we are only capturing the relevant wavelength. A special target that is almost perfectly lambertian ensures that the results are not corrupted by any specular effects. A light source is directed toward two small slits that serve as an aperture for selecting close to parallel light rays. This minimizes overlap between nearby wavelength on our target. Behind the slits a optical grating pattern causes diffraction which separates out the different wavelength. The light source is a 500 Watts Halogen light which – as a black-body-like radiator – emits light across the visual spectrum well into the infrared part, following roughly Planck's law [11]. We do not use a calibrated light source in this study and consider it of lesser importance as we are mostly interested in relative sensitivities under naturally occurring light.

After acquiring reference images in ambient light, we calibrate the images and calculate the response profile across wavelength. We do this for each RGB channel separately, as well for the IR-image. (See Fig 4 bottom). This gives us the sensitivity for each channel independently.

Having an estimate for the sensor response characteristics, we can now estimate a reconstruction of the IR sensor by a linear weighting of the RGB responses. Therefore we find the following least squares solution:

$$\min_{w} ||R_{ir} - [R_r R_g R_b]w||_2 \qquad (1)$$

where R_{ir} is the spectral response of the IR sensor and R_r, R_g, R_b are the responses of the red, blue and green channel respectively.

Results Fig 5 depicts the sensor readings we have obtained. The raw sensor data is plotted in pale colors, while the saturated colors show a gaussian fit. For each channel we subtract the minimum response in order to compensate for sensor noise and residual ambient light and then fit a gaussian mixture model with 3 modes as we observe 3 maxima of the diffraction pattern. The dominant mode is plotted per channel. There are 4 IR channels as we read the raw IR image from the Kinect that comes in a bayer pattern. We expected slightly different response characteristics for each channel, but they turn out to be almost identical. Furthermore we observe that the overlap between IR and RGB-channels is relatively small. The linear reconstruction of the IR channel from Equation 1 results in the cyan line shown in Fig 5. As the profile is very flat, we also show an amplified version. The low magnitude indicates that the reconstruction is not working well. The weights for the individual channels are as follows: $w_{red} = 0.0111, w_{green} = -0.0066, w_{blue} = 0.0022$. Obviously, the red channel has the highest weight, as it is closest to the infrared part. Interestingly, we get a negative weight for green which "pushes" the red channel further to the infrared part. The positive weight for blue again compensates partially for the introduced dip in the green to blue wavelength. This is also an interesting parallel to [1] where similar weights were obtained by training pixel correspondences without explicit knowledge of the spectral sensitivities.

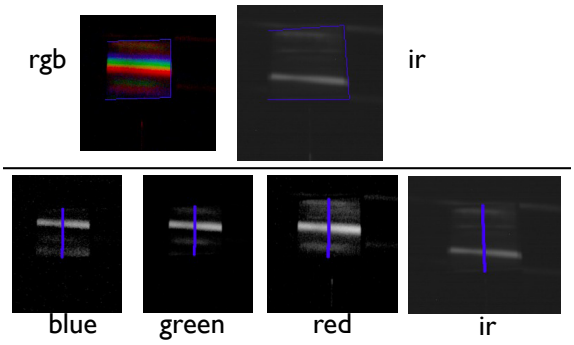


Figure 4: Spectrum from Experiment as in Fig 3.

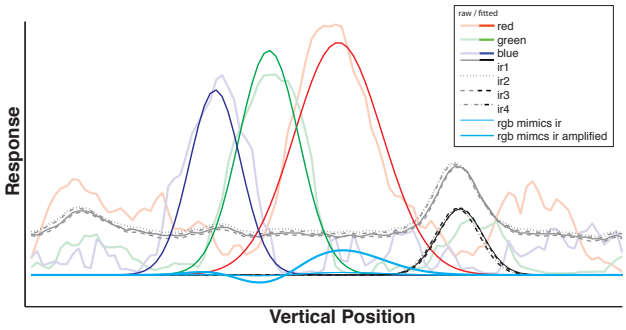


Figure 5: Spectrum from Experiment as in Fig 3. Light colored plots correspond to the response along blue lines through raw image data (Fig 4). Strong colored plots correspond to Gaussian fitted curves.

In summary, we have to conclude that the overlap of the IR and RGB sensitivity of the sensor is indeed smaller than expected, which seems very bad news for any cross–modal matching attempt. However, in practice we do often have light sources that cover a reasonable part of the spectrum – like the above used halogen lamp – and in addition typical materials also reflect in a relative broad and smooth spectrum. This gives a justification to learning-based approaches like the one in [1] that can exploit correlations and smoothness of BRDFs.

4. Methods

Our aim is to increase robustness as well as computational efficiency of cross–modal stereo under projected patterns by learned filters. A simplistic scheme that is exclusively based on pixel-wise re-weighting of IR and RGB seems to be too limited. A learning-based version of this linear scheme was attempted in [1] and we also derived a weighting based on spectral measurements in the previous

section.

As we want to stay in the roam of efficient patch–based stereo algorithm, we propose to extend the class of learned transformation to linear filters that leverage a pixel neighborhood in all channels to optimally preserve matches across modalities. These linear filters encompass smoothing, sharpening and edge detection methods that have been shown useful as prefilter in stereo algorithm and can potentially alleviate problems with the projected pattern.

The core idea is to collect corresponding pairs of patches between IR and RGB images into a set S for the training step. Then we use them to determine the weightings of each elements in the IR and RGB patches so that the corresponding patches have a smaller distance after the transformation. We employ an optimization framework to describe this problem.

We denote the $s–th$ corresponding pair of patches by $\{IR^s, C^s\} \in S$ where $C = \{r, g, b\}$ contains three color channels from the RGB image. With the assumption that the patch is in the size of $n \times n$, we would like to obtain the different weightings $\{w^{IR}_{i,j}, w^{C}_{i,j}\}$ for every pixels $\{IR^s_{i,j}, C^s_{i,j}\}$ of different positions (i, j) within IR and RGB patches $\{IR^s, C^s\}$. The resulting optimization problem reads:

$$\min_{w^{IR}, w^{C}} \sum_{s \in S} \left\| \sum_{i=1}^{n}\sum_{j=1}^{n} w^{IR}_{i,j} IR^s_{i,j} - \sum_{C=r,g,b} \sum_{i=1}^{n}\sum_{j=1}^{n} w^{C}_{i,j} C^s_{i,j} + b \right\|_1$$
$$\text{subject to} \sum_{C=r,g,b} \sum_{i=1}^{n}\sum_{j=1}^{n} w^{C}_{i,j} = 1. \tag{2}$$

where b is an offset. By applying these weightings for each color channels of RGB images and for IR images, we can transform the RGB images into "IR-like" images then use the same stereo matching algorithm to compute the disparity maps as usual. Note that this weighting procedure is the same as utilizing filters for images. We display an instance of our proposed filtering procedure in Figure 8.

5. Experiments

In order to evaluate the effectiveness of our approach we compare to the results from [1] in the same experimental setting. A clustering approach is used to segment objects in a table-top scenario. The dataset was deliberately designed to expose problems of the standard kinect depth sensing scheme on materials that are, e.g., specular or transparent. The groundtruth is given as 2d bounding boxes and the PASCAL matching criterion is used that requires the intersection over union between groundtruth and detection bounding box to be larger than 0.5. The dataset consist of 106 objects in 19 scenes.

5.1. Learning Filters

Given image pairs of IR and RGB images, our goal is to learn optimal adaptation between IR and RGB images using the Kinect hardware without any modifications. We manually collect thousands of corresponding pairs of 3×3 patches between low–resolution IR and RGB images under the influence of the IR–projector. The patches are distributed over normal and difficult regions including transparent, specular and reflective surfaces. To solve the optimization problem in Equation 2, we use cvx [4], a matlab–based toolbox for convex optimization.

The resulting filters w^r, w^g, w^b, and w^{IR} are as follows with the offset $b = -56.6978$:

$$w^r = \begin{bmatrix} 0.1451 & 0.1900 & 0.1228 \\ -0.0354 & 0.0089 & 0.1244 \\ 0.1788 & 0.0985 & 0.1809 \end{bmatrix}$$
$$w^g = \begin{bmatrix} 0.1844 & -0.0806 & 0.1249 \\ 0.1866 & -0.1393 & 0.1129 \\ 0.1981 & -0.0841 & 0.0841 \end{bmatrix}$$
$$w^b = \begin{bmatrix} -0.0702 & -0.0260 & -0.0098 \\ -0.1430 & -0.0915 & -0.0600 \\ -0.0984 & -0.0654 & -0.0365 \end{bmatrix} \tag{3}$$
$$w^{IR} = \begin{bmatrix} 0.0049 & 0.0961 & -0.0006 \\ 0.1532 & -1.0000 & 0.1084 \\ -0.0018 & 0.0741 & -0.0062 \end{bmatrix}$$

Visualizations are provided in Figure 8.

5.2. Evaluation

Our evaluation uses the same data as [1] and is consistent with their setting in order to ensure comparability. The stereo image pairs from the Kinect were obtained under two conditions. The first one is to cover the IR projector and the second one is under the normal situation with the IR projector.

From the evaluation of different fusion schemes under these two settings, the best results previously obtained are an average precision of 69% with IR–projector–on and 72.5% with IR–projector–off by using a late fusion strategy.

Comparing to the average precision 46% of the built-in Kinect depth estimate, the method can improve the Kinect depth over 20% but meet the practical issues as we mentioned in Section 4. In contrast, our method simply uses the early fusion scheme with applying the filters on IR–RGB images captured under IR–projector–on but still achieves an average precision of 71.5%.

The Precision–Recall curves of the late fusion scheme under IR–projector–off setting, our proposed method, and Kinect–only are plotted in Figure 6.

Our approach outperforms all their settings using IR–projector–on and is on par with their best result with modified hardware. Also note that our method shows strong

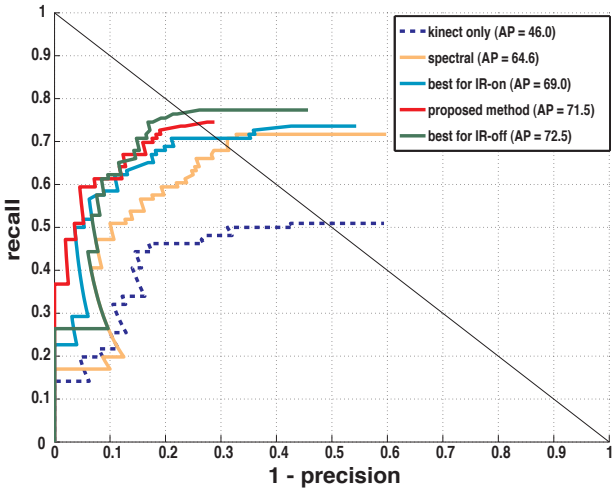
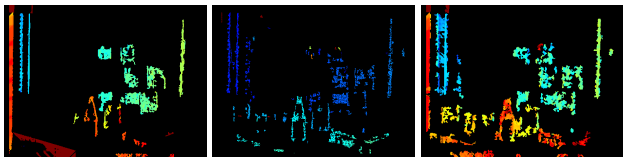


Figure 6: Precision–Recall curves of late fusion scheme under IR–projector–off setting, our proposed method and Kinect–only.



(a) (b) (c)

Figure 7: Disparity maps computed from (a) original image and (b) IR–RGB–image pair without IR–projector, and from (c) images filtered by our trained filters.

improvements in precision and produces the first false positives not until almost 40% recall which is 10% more than the competing methods. In Figure 7 we show examplary disparity maps computed from images in the IR–projector off case, and filtered images by our proposed method.

5.3. Discussion

In the middle column of Figure 8, we present the visualization of the filters obtained from optimization process. The filters of red channel and blue channel resemble smoothing operators and the filter of green channel, the smoothing seems to be applied along the y–axis while the x–axis direction resembles a Laplacian operator. The filter of IR channel basically computes a filter similar to a 2–dimensional Laplacian operator.

6. Conclusions

We have presented a method to optimize filters for improved stereo correspondence IR and RGB images that is robust to projected IR patterns. We have experimentally analyzed the spectral characteristics of the Kinect cameras in order to justify such an approach. Adapting RGB in frequency domain to mimic an IR image did not yield improved performance. The small overlap between RGB and IR seems prohibiting this approach. In contrast, learning several filters based on image patches allowed improved stereo vision across modalities. We conclude therefore, that our pre-filtered, cross-modal, SAD-based stereo vision algorithm profits most from combination in the spatial domain, rather than in the frequency domain. Our patch-based approach shows increased performance and improved robustness against IR-specific interference from the projector.

However, the Kinect hardware limitation still disallows us to capture RGB and IR simultaneously. This requires to switch both channels by software and limits the frame rate to 2-3fps. While sufficient for many applications, an increased frame rate is certainly preferable.

Upon publication, we will make the source code of our method available.

7. Acknowledgements

We would like to thank Ivo Ihrke for helpful discussions and suggestions.

References

[1] W. C. Chiu, U. Blanke, and M. Fritz. Improving the kinect by cross-modal stereo. In *BMVC11*. 1, 2, 3, 4

[2] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision*, 2006. 1

[3] M. Fritz, M. Black, G. Bradski, S. Karayev, and T. Darrell. An additive latent feature model for transparent object recognition. In *NIPS*, 2009. 2

[4] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. `http://cvxr.com/cvx`, Apr. 2011. 4

[5] H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *CVPR05*. 1

[6] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *TPAMI*, 31:1582–1599, 2009. 2

[7] I. Ihrke, K. N. Kutulakos, H. P. A. Lensch, M. Magnor, and W. Heidrich. State of the art in transparent and specular object reconstruction. In *STAR Proceedings of Eurographics*, 2008. 2

[8] U. Klank, D. Carton, and M. Beetz. Transparent object detection and reconstruction on a mobile platform. In *IEEE International Conference on Robotics and Automation*, 2011. 2

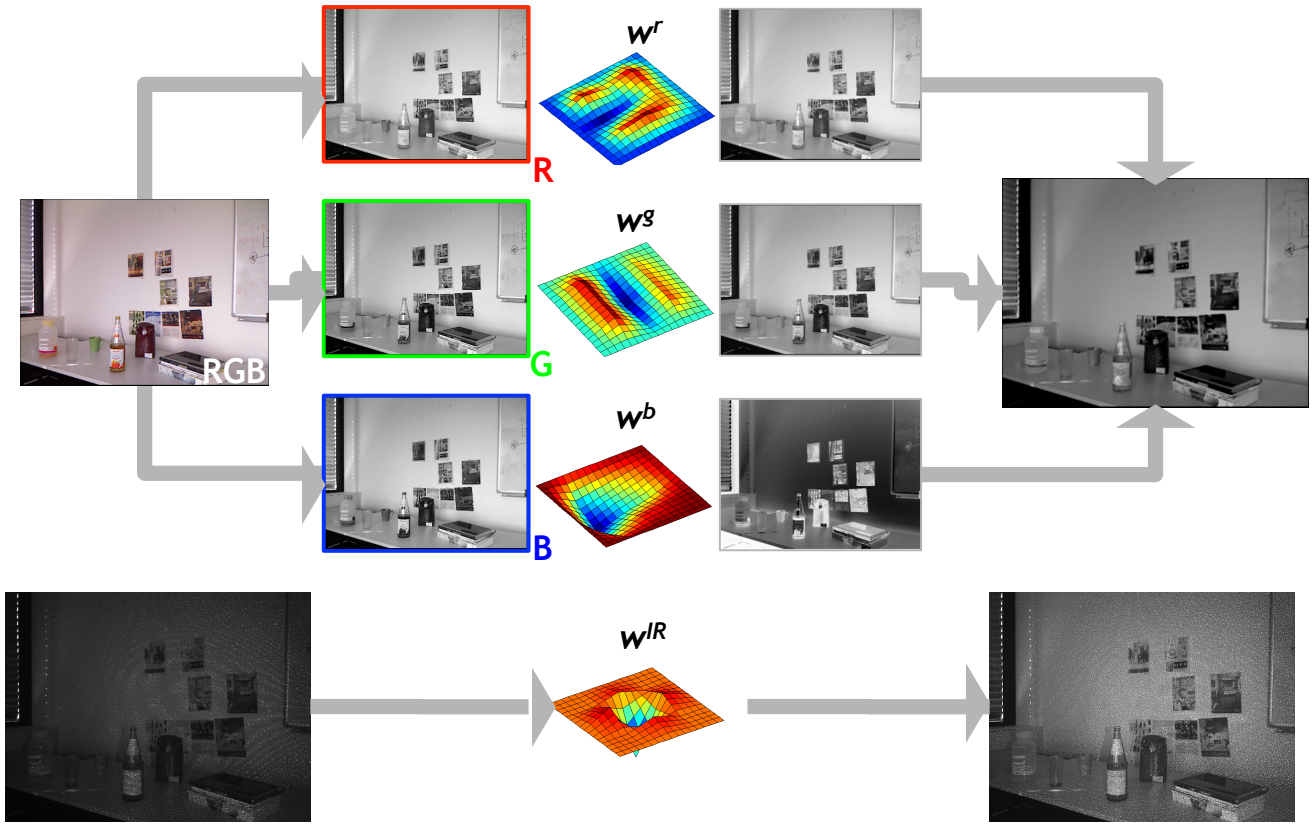


Figure 8: (Left) Incoming RGB and IR image with IR–projector–on from Kinect. (Top) color channels from RGB image: red, green, blue and learned filters w_R, w_G, w_B, w_{IR} from optimization. Note here we do the zero–padding and up–sample the filter for a better visualization. The resulting images filtered for each color channels. (Right) transformed RGB (summed filtered images for three color channels) and IR images.

[9] K. Konolige. Small vision systems: Hardware and implementation. In *ROBOTICS RESEARCH-INTERNATIONAL SYMPOSIUM-*, volume 8, pages 203–212. MIT PRESS, 1998. 2

[10] Y. Li and D. P. Huttenlocher. Learning for stereo vision using the structured support vector machine. In *Computer Vision and Pattern Recognition*, 2008. 1, 2

[11] M. Planck. On the law of distribution of energy in the normal spectrum. *Annalen der Physik*, 4(553):1, 1901. 3

[12] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 2

[13] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. *Computer Vision—ECCV'94*, pages 151–158, 1994. 2