Semantic Image Segmentation Using Visible and Near-Infrared Channels*

Neda Salamati[1,2], Diane Larlus[2], Gabriela Csurka[2], and Sabine Süsstrunk[1]

[1] IVRG, IC, École Polytechnique Fédérale de Lausanne, Switzerland
[2] Xerox Research Centre Europe, 6, Chemin de Maupertuis, Meylan, France

Abstract. Recent progress in computational photography has shown that we can acquire physical information beyond visible (RGB) image representations. In particular, we can acquire near-infrared (NIR) cues with only slight modification to any standard digital camera. In this paper, we study whether this extra channel can improve semantic image segmentation. Based on a state-of-the-art segmentation framework and a novel manually segmented image database that contains 4-channel images (RGB+NIR), we study how to best incorporate the specific characteristics of the NIR response. We show that it leads to improved performances for 7 classes out of 10 in the proposed dataset and discuss the results with respect to the physical properties of the NIR response.

1 Introduction

Semantically segmenting a scene given an image is one of the eminent goals in computer vision. While we have seen a lot of progress in recent years using sophisticated image descriptors [1,2] and better machine learning techniques [3,4], segmentation still remains challenging. Whereas humans have no difficulties performing semantic image interpretation, machine vision systems still struggle mainly because of the ambiguity of the influence of light and surface reflectance on a given pixel value. For example, a dark pixel can either result from a dark surface reflectance under normal lighting conditions or a light surface reflectance under shadow. Decoding the contributions of light and reflectance from an image is an ill-posed problem [5]. To solve it, we either need to make assumptions about the world, or to capture more information.

In this paper, we study semantic segmentation using the latter approach. Specifically, we propose to use near-infrared (NIR) images in addition to visible (RGB) images as input. Silicon sensors of digital cameras are naturally sensitive in the NIR wavelengths range (750-1100nm). By removing the NIR blocking filter affixed to the sensor, digital cameras can capture both RGB and NIR images [6]. RGB and NIR cues have been successfully combined in many applications like dehazing [7], dark flash photography [8], and scene categorization [9,10].

We believe that the intrinsic properties of NIR images make them relevant for semantic segmentation. First, due to the NIR radiation being adjacent to the visible spectrum, NIR images share many characteristics with visible images. In

* This work was supported by the Swiss National Science Foundation under grant number 200021-124796/1 and Xerox foundation.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 461–471, 2012.

particular, the shapes of objects in the scene are preserved, *i.e.*, borders of physical objects in the visible images match the borders in the NIR image, which is necessary for segmentation. Second, the intensity values in the NIR images are more consistent across a single material, and consequently across a given class region, due to the unique reflectances of certain natural and man-made composites to NIR radiation [11]. For instance, vegetation is consistently "bright", and sky and water are "dark". Third, texture in NIR images is more intrinsic to the material. This is partly due to the transparency of most colorants and dyes in NIR; texture introduced by (color) patterns on a surface is less dominant in NIR. Additionally, there is generally less haze present in NIR images [7]. Consequently in landscape scene images, distant regions appear sharper (see Figure 1).

These properties of NIR images have been used by the remote sensing and military communities for years to detect and classify natural and/or man-made objects [12]. However, in this paper we approach semantic image segmentation from a different point of view. First, as opposed to the aerial photography, we address images in typical street and landscape photography. Second, most remote sensing applications use true hyper-spectral capture, with several bands in the NIR and even the IR. Our framework only uses a single channel that integrates all NIR radiation, and that can be captured by a standard sensor of any digital camera. This is in-line with the recent developments in computational photography, where different camera set-ups are proposed to concurrently capture three visible (RGB) and one NIR channel, either on two sensors with a beam splitter [13] or on a single sensor [14]. To this end, we apply a state-of-the-art segmentation framework to the task. Our proposed system is based on a Conditional Random Field (CRF) model [15], where we exploit different possibilities of combining the visible and NIR information in the recognition part and in the regularization part of the model.

The contributions of this paper are three-fold. First, we study how to best use NIR in a CRF based segmentation framework, exploring different options for both recognition and regularization parts of the model. Second, we provide a new semantic segmentation dataset that contains images having both visible and NIR channels (RGB+NIR) and pixel-level annotations. Finally, we discuss the results obtained with both cues for different classes and connect our observations with material characteristics and other properties of NIR radiation.

1.1 Previous Work

Semantic Image Segmentation is the process of partitioning an image into regions, where each region corresponds to a semantic class within a predefined list. The appearance of these classes are learned using labeled images. Methods usually fuse two sub-tasks: a recognition part, responsible for the labeling, and a regularization part that enforces neighboring pixels to belong to the same class. The recognition part is based on the local appearance that is considered at the pixel [1], or at the patch level [3]. Different features are used to describe the local appearance, among them texture (filter banks), color statistics, and SIFT [16]. The low-level features are often transformed into higher-level features, such as the Bag-of-Visual-Words [17] or the Fisher Vector (FV)

representations [18], before feeding them into a classifier. In our work, we follow [2] for local representation and use FVs for recognition. The local consistency is usually enforced by pairwise constraints between neighboring pixels. The local appearance and the local consistency are often combined using Markov random fields (MRF) [3] or conditional random fields (CRF) [1]. In this paper, we use a CRF model.

NIR. The spectral signature of different materials in the NIR part of the spectrum is the foundation for most of remote sensing applications. In such tasks, to achieve a successful classification, data with high spectral resolution is required [12]. On the contrary, [19] exploits the material-based low-level segmentation task using the 4-channel images that can be potentially captured by any digital camera. Incorporating this freely available data, we are going one step further and not only segment the scene but also semantically label every region of the image. For that purpose, we explore a framework using supervised classifiers that learn the relation between visible and NIR information given a class.

The rest of this article is organized as follows. The CRF model we used is described in Section 2. The various experiments are described in Section 3. Finally, a discussion is proposed in Section 4.

2 Model

Semantic segmentation is formulated as a discrete labeling problem that assigns each pixel $i \in \{1, ..., N\}$ to a label from a fixed set Ψ. Given the observations $\mathbf{x} = \{x_1, ..., x_N\}$, the task is to estimate a set of random variables $\mathbf{Y} = \{Y_1, ..., Y_N\}$, taking values in Ψ^N. We employ a CRF that considers the posterior distribution to define the Gibbs energy: $E(y) = -\log P(Y = y \mid X = x) - \log Z$, where Z is a normalization constant. The maximum a posteriori (MAP) labeling y^* of the random field is defined as:

$$y^* = \text{argmax}_{y \in \Psi^N} P(Y = y \mid x) = \text{argmin}_{y \in \Psi^N} E(y) \tag{1}$$

The labeling is formulated as a pairwise CRF, whose energy can be written as:

$$E(y) = E_{unary}(y) + \lambda E_{pair}(y), \tag{2}$$

and is composed of a unary and a pairwise term. As in [20], we assign a weight λ to E_{pair} that models the trade-off between recognition and regularization.

The Unary Term is responsible for the recognition part of the model and uses the probability for each pixel to belong to a class. E_{unary} is considered as the cost of assigning labels y to observations x, and is defined as: $E_{unary} = \sum_i -\log(p(Y_i = y_i|\mathbf{x}))$. We used a patch based representation, since patches contain more information than pixels. Low-level descriptors are computed for each patch, and transformed into Fisher Vectors (FV) [18]. FVs computed on the patches of the training images and their labels are used to train linear SVM classifiers. For a test image, FV representations of patches are given to the

Fig. 1. Examples of 4 channel images (visible = RGB, left and NIR, middle) and their ground truth (right) from our dataset

classifiers, a score is inferred for each patch, and for each class. The scores can be transformed into probabilities at the pixel level [2], yielding probability maps.

The Pairwise Term regularizes the pixel labeling, neighboring pixels are encouraged to share labels. We used a 4-neighbor system ν (each pixel is connected to its 4 direct neighbors). We relax the regularization constraints along the edges, using a contrast sensitive Potts model: $E_{pair} = \sum_{(i,j)\in\nu} \delta_{y_i,y_j} \exp(-\beta||p_i - p_j||^2)$ where δ_{y_i,y_j} is the Kronecker delta and $\beta = \frac{1}{2<\|p_i - p_j\|^2>}$ as in [20]. This potential penalizes disagreeing neighboring pixel labels, and the penalty is lower were pixel values change. That way, borders between regions are encouraged to follow edges. The pixel value p_i can be considered in the visible domain ($p_i = \{r_i, g_i, b_i\}$), in the NIR domain ($p_i = n_i$) or in both (4 dimensions).

Model Inference is carried out by the multi-label graph optimization library of [21,22] using α-expansion.

3 Experiments

First we present our dataset and the implementation details (Sec. 3.1). We then compare different descriptors for the model recognition part (Sec. 3.2), and the regularization part (Sec. 3.3) is then studied for the most promising ones.

3.1 Proposed Dataset and Technical Details

Our dataset is based on a previously released scene dataset [9], where images are composed of 3 visible (RGB) channels and a NIR channel. To the best of our knowledge, that is the only set of diverse natural images for which both visible and NIR channels have been recorded. The original dataset consists of 477 images, divided into 8 outdoor and 1 indoor scenes. We discarded the indoor and old building classes, whose appearance is too different from the other classes[1].

The remaining 370 images were manually segmented and annotated at the pixel level with the following classes: *Building, Cloud, Grass, Road, Rock, Sky, Snow, Soil, Tree, Water*. We followed the MSRC annotation style [1], pixels are labeled as one of these classes or as *void* class. *Void* corresponds to pixels whose class is not defined as part of our classes of interest, or are too ambiguous to be labeled (see Figure 1).

We extract patches of size 32×32 on a regular grid (every 10 pixels) at 5 different scales. To extract different scales the images are resized by factors of

[1] Keeping these scene classes would have resulted in images with only (or mostly) void regions, according to our 10 pre-selected semantic classes.

Table 1. Left: evaluation (average of per-class and overall accuracies) of the segmentation for different local descriptors and their combinations. Right: p-value for the paired t-test, for the overall accuracy per image of different strategy pairs.

Descriptor	Per-class	Overall
COL_{rgb}	71.78	77.05
COL_{rgbn}	74.00	79.47
COL_{p1234}	72.50	77.09
$SIFT_l$	65.62	72.89
$SIFT_n$	66.18	73.82
$SIFT_{p1}$	65.80	73.07
$SIFT_{rgb}$	72.94	78.68
$SIFT_{rgbn}$	74.77	81.70
$SIFT_{p1234}$	75.09	81.77
$COL_{rgb} + SIFT_l$	77.36	82.55
$COL_{rgbn} + SIFT_n$	**78.88**	**84.26**
$COL_{p1234} + SIFT_n$	77.57	82.85

Strategy A	Strategy B	p-value
COL_{rgb}	COL_{rgbn}	3.10^{-4}
$COL_{rgb} + SIFT_l$	$COL_{rgbn} + SIFT_n$	1.10^{-4}
$SIFT_{p1234}$	$COL_{rgbn} + SIFT_n$	3.10^{-6}
$SIFT_{rgbn}$	$COL_{rgbn} + SIFT_n$	5.10^{-7}
$SIFT_{rgbn}$	$SIFT_{p1234}$	7.10^{-7}
$SIFT_n$	$SIFT_l$	9.10^{-2}

1, 0.7, 0.5, 0.35, and 0.25. We consider two different features. The SIFT feature ($SIFT$) [16] encodes texture using histograms of oriented gradients for each bin of a 4×4 grid covering the patch. The color feature (COL) encodes the intensity values in each image channel using mean and standard deviation in each bin of the same grid covering the patch. Low-level descriptors are computed for each patch and their dimensionality is reduced by PCA to 96. A visual codebook with 128 Gaussians is built in the projected space, and each patch is transformed into a FV. By using the same PCA dimension and the same codebook size, the FV representation of all descriptors has the same dimensionality.

We randomly split our dataset into 5 sets of images (5 folds) and define 5 sets of experiments. For each experiment, one fold is used for validation, one is used for testing, and the remaining images are used for training the model. Results for the 5 test-folds are grouped and evaluated at once, producing a single score for the dataset for each evaluation measure.

We consider two measures that evaluate the segmentation as a pixel level categorization problem. The first one (**overall**) is the overall accuracy (*i.e.*, the number of correctly classified pixels divided by the total number of pixels), the second one (**per-class**) is the average of the per-class accuracy (*i.e.*, average over the classes of the ratio between true positives and positives). The pixels labeled as *void* in the ground truth are not considered for evaluation.

3.2 Descriptor Comparison

This set of experiments evaluates the recognition part of our model. As described in Section 2, each pixel is associated with a probability of belonging to each of the classes. We produce a semantic segmentation by assigning pixels to their most likely label, $y^* = \text{argmax}_{y \in L} P(Y = y|\mathbf{x})$. This is equivalent to our full model, using $\lambda = 0$. In other words, only the unary term is considered here.

First, we compare different COL and $SIFT$ features. COL extracts statistics over RGB channels, consequently we call it COL_{rgb}. The standard SIFT, computed on the luminance channel (visible image) is called $SIFT_l$. The color descriptor can be extended to 4-channel images (RGB+NIR), defining COL_{rgbn}.

We also consider the alternative 4-D color space proposed by [9], and introduce COL_{p1234}, that concatenates COL features computed on each of the 4 alternative channels $p1$, $p2$, $p3$, and $p4$, that are obtained from PCA applied to RGBN. We propose to examine $SIFT_n$, a SIFT descriptor computed on the NIR image. The $SIFT_{p1}$ descriptor, computed only on the first channel ($p1$) of the alternative color space, is also considered.

As color and texture are complementary, we look at different ways to combine them. The first one, proposed initially for visible images by [23], is a multispectral SIFT, $SIFT_{rgb}$, that concatenates $SIFT$ descriptors computed on the R, G, and B channels. This can be extended in a straightforward manner to 4-dimensional images, defining $SIFT_{rgbn}$, and $SIFT_{p1234}$ respectively. To combine color and texture, we also consider combinations of $SIFT$ and COL descriptors, by averaging the relevant probability maps. Results are reported for $COL_{rgb} + SIFT_l$ that contains only visible information, and for $COL_{rgbn} + SIFT_n$ and $COL_{p1234} + SIFT_n$ that also include NIR.

Table 1 (left) compares the segmentation accuracy obtained by these descriptors and their combination. In order to get an intuition whether results are significantly different, we computed statistical significance using the paired t-test on overall results per image for the most interesting pairs of descriptors. Results are reported in Table 1 (right). A p-value smaller than 0.05 means the descriptors are statistically different from each other, with a 5% confidence level.

From Table 1, we can make the following observations. First, COL descriptors using NIR information outperform the visible only COL_{rgb}. The original 4-D color space (COL_{rgbn}) performs better than the de-correlated space (COL_{p1234}). For the SIFT descriptor, $SIFT_n$ does slightly better than $SIFT_l$, as we expected due to the material dependency of the NIR response, but the difference is not significantly different on this dataset, at the 5% confidence level. The same applies when comparing $SIFT_n$ and $SIFT_{p1}$.

The best single descriptor is $SIFT_{p1234}$, as already shown for image classification in [9]. This descriptor encodes texture for the different color channels, visible and NIR. Still, this best descriptor is outperformed by the late fusion of COL and single-channel $SIFT$ descriptors.

As main conclusions, the best visible only approach for recognition is $COL_{rgb} + SIFT_l$. We will use this descriptor as our visible-only baseline for the rest of the paper. According to our study, the best way to include the NIR information for local recognition is through $COL_{rgbn} + SIFT_n$. It outperforms the best visible method ($COL_{rgb} + SIFT_l$) by almost 2% (+1.71 overall accuracy).

3.3 Graph Model

In the previous section, we studied the recognition part of our model, and acknowledged the gain obtained using NIR together with the 3 visible channels to build local descriptors. For the regularization part, we use the previously obtained probability maps per class, and apply them in the full model described in Section 2. The resulting energy function is optimized using the library of [21,22]. λ is fixed to 5 for all the experiments.

Table 2. Results for the full CRF model

Descriptor	Pairwise	Per-class	Overall
$COL_{rgb} + SIFT_l$	VIS	78.45	83.82
	NIR	78.48	83.71
	$VIS + NIR$	78.58	83.94
$COL_{rgbn} + SIFT_n$	VIS	80.10	85.50
	NIR	80.01	85.38
	$VIS + NIR$	**80.30**	**85.72**
$COL_{p1234} + SIFT_n$	VIS	78.68	84.21
	NIR	78.76	84.20
	$VIS + NIR$	78.87	84.34
$SIFT_{p1234}$	VIS	76.68	83.41
	NIR	76.76	83.37
	$VIS + NIR$	76.81	83.50

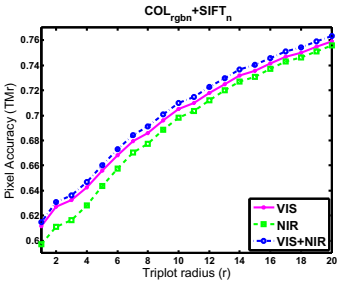

Fig. 2. Trimap accuracy, for varying radius (size of the band around the border considered for evaluation)

We focus on the most promising combinations of descriptors from the recognition part: $COL_{rgbn} + SIFT_n$, $COL_{p1234} + SIFT_n$, and $SIFT_{p1234}$. We compare them to our visible only baseline $COL_{rgb} + SIFT_l$. For the pairwise potential, the Potts model is extracted on the visible image (VIS), the NIR one (NIR) and the full 4-channel image ($VIS + NIR$). This means that the pixel intensity difference is computed for pixel values p_i being of dimension 3, 1 and 4 respectively. Results are presented in Table 2.

First, we note that regularization always improves segmentation accuracy (comparing Table 1 and 2), this improvement is modest (between 1% and 1.7%).

Second, we observe that visible only pixel information (VIS) for the pairwise potential is comparable to the NIR Potts model (none of the difference is statistically significant). Nevertheless, combination of both ($VIS+NIR$) is always slightly better than any of the individual pairwise models, and this is statistically different at the 5% confidence level.

To better understand the role of the regularization, we also consider a third evaluation measure, *trimap accuracy* [24] that considers the overall classification accuracy for pixels in a narrow band around borders between two regions in the ground truth. Results for the $COL_{rgbn} + SIFT_n$ descriptors, and the different pairwise potentials are shown in Figure 2[2]. These results support our previous claim that the 4D pixel representation leads to better results[3].

The best results using visible only information were obtained by the full model, with $COL_{rgb} + SIFT_l$ as descriptors for recognition and a regularization using the RGB image. This is our visible baseline, referred to as VB. The best results for RGB+NIR images were obtained by $COL_{rgbn} + SIFT_n$ and $VIS + NIR$ for regularization. Our best strategy is referred as BS in the following discussion.

4 Result Analysis and Discussion

In this section, we analyze and compare the segmentation results obtained with the two strategies VB and BS based on the confusion matrices reported in Table 3 and sample images in Figure 3.

[2] Curves for the other descriptors look similar, and have been omitted.
[3] Statistical significance is obtained for this measure as well.

Table 3. Confusion matrices for i) descriptor $COL_{rgbn} + SIFT_n$ and 4-dimentional pairwise (BS) on top and ii) $COL_{rgb} + SIFT_l$ with visible-only pairwise (VB) below

		Tree	Grass	Soil	Build.	Road	Rock	Snow	Water	Sky	Cloud
VISIBLE + NIR	Tree	**94.3**	2.4	0.3	1.1	0.1	1.1	0.1	0.1	0.2	0.3
	Grass	13.4	**81.0**	1.0	0.4	0.7	2.4	0.0	1.0	0.0	0.1
	Soil	10.5	9.3	**64.6**	1.0	6.3	5.6	1.1	1.3	0.1	0.0
	Build.	4.7	0.4	0.4	**89.6**	2.9	0.6	0.0	0.2	0.2	1.1
	Road	1.1	1.4	1.5	8.7	**84.2**	0.5	1.7	0.8	0.0	0.0
	Rock	23.4	1.9	2.2	1.4	0.3	**64.3**	4.5	1.0	0.4	0.6
	Snow	0.8	0.0	1.6	1.4	4.2	20.0	**69.6**	0.6	0.0	1.8
	Water	4.4	0.5	1.5	4.9	0.2	1.6	0.5	**83.2**	0.9	2.4
	Sky	1.5	0.0	0.0	2.0	0.2	0.1	0.0	0.1	**78.6**	17.5
	Cloud	0.9	0.0	0.0	0.4	0.0	0.6	0.2	0.1	4.3	**93.5**
VISIBLE ONLY BASELINE	Tree	**91.8**	3.7	0.5	2.0	0.2	0.9	0.1	0.1	0.2	0.4
	Grass	12.6	**80.6**	0.9	0.7	0.7	3.2	0.0	1.2	0.0	0.0
	Soil	11.9	10.7	**60.5**	0.3	6.9	5.5	1.8	2.2	0.1	0.0
	Build.	3.6	0.4	0.7	**89.8**	2.9	0.3	0.5	0.4	0.3	1.1
	Road	1.4	1.5	1.6	8.0	**83.0**	1.0	2.3	1.1	0.0	0.0
	Rock	21.2	1.3	1.7	2.1	0.1	**66.3**	5.2	1.4	0.4	0.3
	Snow	0.2	0.0	0.1	1.3	2.1	20.0	**72.7**	0.0	0.0	3.7
	Water	5.8	2.0	1.2	6.6	1.8	2.9	0.5	**73.4**	0.5	5.3
	Sky	1.6	0.0	0.0	1.8	0.2	0.1	0.1	0.5	**73.5**	22.3
	Cloud	0.6	0.0	0.0	0.7	0.0	0.3	0.8	0.1	4.5	**92.9**

From Table 3, we note that the almost 2% overall difference between BS and VB relates to an improvement for 7 classes out of 10. The largest improvement is observed for classes *water* (+9.8%), *sky* (+5.1%), *soil* (+4.1%). *Trees*, *clouds*, and *grass* are improved by 2.5%, 0.6%, and 0.4%, respectively. On the other hand, the performance on classes *building*, *rock* and *snow* were slightly lower (less than 2% except for the class *snow*).

Haze Effect. First, we observe the benefit of NIR in the presence of haze. As stated by Rayleigh's law, the light scattered from small particles ($< \lambda/10$) is inversely proportional to the wavelength λ ($1/\lambda^4$) [6]. Particles in the air (haze) satisfy this condition and are scattered more in the short-wavelength range of the spectrum. Thus, when images are captured in the NIR, atmospheric haze is less visible and the sky becomes darker (see image 5 in Figure 3). The "haze transparency" characteristic of NIR results in sharper images for distant objects (see images 3 or 5). In particular, vegetation and mountains at a distance in the visible image is smoothed and bluish, which may affect classification results. The sharper and haze-free appearance in NIR helps classification and leads to better segmentation, such as for the class *rock* in image 2.

Border Accuracy. For some images, we observed that borders are more precisely detected while incorporating NIR in the pairwise potential. This can be explained by the material dependency of NIR responses that may reduce wrong edges due to clutter, or may result in more contrasted edges between classes. This information, used in the regularization part of our model, helps to align borders between regions with a change of material (see images 3, 4, 5, and 6).

Water. The class *water* exhibits the largest improvement. Since water absorbs radiation in the NIR, this class appears very dark and becomes very distinctive.

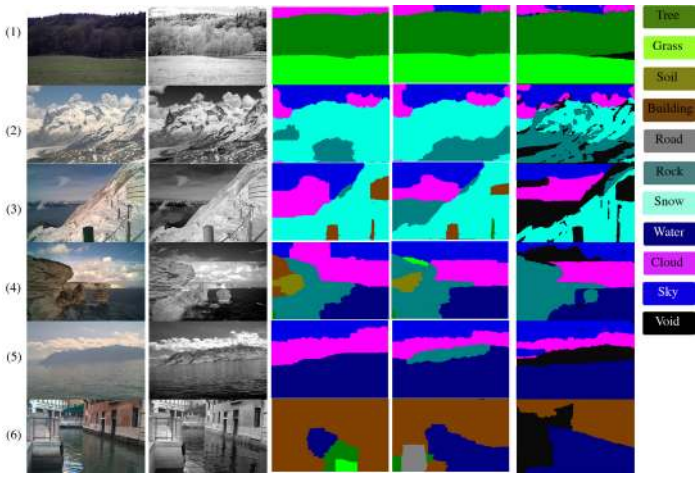

Fig. 3. Visible and NIR images (1st and 2nd column), segmentation results for visible only ($COL_{rgb}+SIFT_l$ and pairwise on VIS, 3rd column) and NIR+visible ($COL_{rgbn}+SIFTn$ and pairwise on $VIS+NIR$ 4th column) methods. Ground truth (5th column)

Even if in visible images, the blue color can be confused with other classes (sky, far away mountains), it has a unique 4-D appearance that lead to 10% improvement compared to the visible baseline. Errors due to reflection are also reduced, for instance in the image 6 in Figure 3.

Clouds and Sky. The classes *cloud* and *sky* are better segmented in the BS scenario, and, more importantly, are less confused. Sky is dark in NIR, due to Rayleigh's scattering mentioned above, while clouds remain white. As clouds are formed from particles larger than $\lambda/10$, Mie scattering [6] that is independent of wavelength applies. Thus, the contrast between these two classes are higher in NIR images, allowing more accurate segmentation (see images 2, 5 of Figure 3).

Tree and Grass. Vegetation is better predicted when NIR information is present due to the unique value of chlorophyll in NIR. Both *tree* and *grass* accuracies are improved. However, as both contain chlorophyll, classes *tree* and *grass* have similar responses in NIR and becomes more confused when NIR is present.

Building. The accuracy of the class *building* stays approximately the same for VB and BS. This class is not made of a homogeneous material, hence, material based information does not bring any advantage.

Conclusions. In this paper, we presented a framework for semantic image segmentation using the RGB and NIR information captured by any ordinary digital camera. Segmentation was formulated using a CRF model, and we studied how to incorporate the NIR cue, either in the recognition or in the regularization parts of our model, and showed that integrating NIR along with conventional RGB images improves the segmentation results. In particular, we observed that

the overall improvement is due to a large improvement for some classes whose response in the NIR domain is particularly discriminant, as water, sky or cloud. The use of this potentially free additional information is a promising direction to improve semantic segmentation, which we plan to test on a broader range of classes.

References

1. Shotton, J., Winn, J.M., Rother, C., Criminisi, A.: *TextonBoost*: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 1–15. Springer, Heidelberg (2006)
2. Csurka, G., Perronnin, F.: An efficient approach to semantic segmentation. IJCV 95 (2011)
3. Verbeek, J., Triggs, B.: Region classification with markov field aspects models. In: CVPR (2007)
4. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.S.: Graph Cut Based Inference with Co-occurrence Statistics. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 239–253. Springer, Heidelberg (2010)
5. Finlayson, G.D., Drew, M.S., Funt, B.V.: Color constancy: generalized diagonal transforms suffice. Journal of the Optical Society of America 11, 3011–3019 (1994)
6. Fredembach, C., Süsstrunk, S.: Colouring the Near-infrared. In: CIC (2008)
7. Schaul, L., Fredembach, C., Süsstrunk, S.: Color image dehazing using the Near-infrared. In: ICIP (2009)
8. Krishnan, D., Fergus, F.: Dark flash photography. In: SIGGRAPH (2009)
9. Brown, M., Süsstrunk, S.: Multispectral SIFT for scene category recognition. In: CVPR (2011)
10. Salamati, N., Larlus, D., Csurka, G.: Combining visible and Near-infrared cues for image categorisation. In: BMVC (2011)
11. Salamati, N., Fredembach, C., Süsstrunk, S.: Material classification using color and NIR images. In: CIC (2009)
12. Zhou, W., Huang, G., Troy, A., Cadenasso, M.L.: Object-based land cover classification of shaded areas in high spatial resolution imagery of urban areas: A comparison study. Remote Sensing of Environment 113, 1769–1777 (2009)
13. Zhang, X., Sim, T., Miao, X.: Enhancing photographs with NIR images. In: CVPR (2008)
14. Kermani, Z., Lu, Y., Süsstrunk, S.: Correlation-based joint acquisition and demosaicing of visible and Near-infrared images. In: ICIP (2011)
15. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: ICML (2001)
16. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
17. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV SLCV Workshop (2004)
18. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR (2007)
19. Salamati, N., Süsstrunk, S.: Material-based object segmentation using Near-infrared Information. In: CIC (2010)
20. Rother, C., Kolmogorov, V., Blake, A.: "GrabCut": interactive foreground extraction using iterated graph cuts. In: SIGGRAPH (2004)

21. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE TPAMI 26, 1124–1137 (2004)
22. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via Graph Cuts? IEEE TPAMI 26, 147–159 (2004)
23. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE TPAMI 32, 1582–1596 (2010)
24. Kohli, P., Ladický, L., Torr, P.H.S.: Robust higher order potentials for enforcing label consistency. IJCV 82, 302–324 (2009)