Combining Visible and Near-Infrared Cues for Image Categorisation

Neda Salamati [1,2]
neda.salamati@epfl.ch
Diane Larlus[2]
diane.larlus@xrce.xerox.com
Gabriela Csurka[2]
gabriela.csurka@xrce.xerox.com

[1] École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
[2] Xerox Research Centre Europe (XRCE)
Meylan, France.

Motivation. Conventional digital cameras are sensitive to wavelengths from the ultraviolet to the near-infrared (NIR) domain (200-1100nm). NIR range is blocked by a component in the camera, so only the visible part of the spectrum is captured. Thus most of the consumer cameras would be able to capture NIR information if the NIR blocking filter would be removed. Visible and NIR representations of the same scene are very similar, in particular the shape of different objects in the scene is preserved. However, material dependency of the reflection in NIR results in some differences of contrast and intensity between these images (see Figure 1 for an illustration), therefore visible and NIR information are complementary. Recent studies [1, 3] have demonstrated the usefulness of this significant amount of potentially available information in different image processing and computer vision tasks.

Contribution. In this paper we consider the scene categorisation problem in the context of images for which both standard visible RGB channels and NIR information are available. Using an efficient local patch based image representation,

- Consistent with previous work, we confirm the observation that the combination of both colour and NIR cues can be useful for the image categorisation task, on a state-of-the art pipeline,
- We propose a thorough study on how to compute and best use texture and colour descriptors when NIR information is available,
- We investigate the complementarity between the different considered descriptors, and propose efficient ways to combine them.

(a) water

(b) old-building

Figure 1: Two scenes from the EPFL dataset. On the left: the conventional RGB image of the scene, on the right: its NIR counterpart.

Framework. In our study, we apply a well-performing and generic categorisation method [2] that works as follows. Image signatures are generated by encoding some low-level image descriptors using Fisher vectors. During training, signatures from all training images are used to train a discriminative classifier. This classifier is then applied to each testing image signature. In this study we use linear SVM as classifier.

The classifier is trained on two types of features; SIFT descriptors, which provide local texture information and colour descriptors (*COL*), which look directly at the intensity values in all the image channels.

Experiments. Experiments are conducted on a recently released [1] colour and NIR semantic categorisation dataset (see Figure 1 for some examples). Some of our observations also transpose to standard colour images, in this case, we show additional evidence on two other datasets, i.e., MIT-Scene 8 and PASCAL VOC 2007.

Images are composed of 4-channels (r, g, b, and NIR as n). As proposed in [1], we also generate an alternative 4-channel representation by projecting the pixel values from the r, g, b, n space into the decorrelated PCA space $p1$, $p2$, $p3$, $p4$. Thus, the descriptors ($D \in \{SIFT, COL\}$) can be computed on each channel ($\{D_i | i \in \{r,g,b,n,l,p1,p2,p3,p4\}\}$, where l is the luminance, or on any combination of these channels ($\{D_{i,j} | i,j \in \{r,g,b,n,p1,p2,p3,p4\} \& i \neq j\}$).

Main results. In analysing the performance of *SIFT* and *COL* descriptors individually, we observe that incorporating NIR information increases the classification accuracy (see Table 1 for details). The results also show that $SIFT_n$ performs comparably with $SIFT_{p1}$, while SIFT computed on any single visible colour channel performs worse than $SIFT_n$. For the *COL* descriptors, we observed the good performances obtained by these colour descriptors, which encode only simple statistics.

$SIFT_l$	$SIFT_n$	$SIFT_{p1}$	$COL_{r,g,b}$	$COL_{r,g,b,n}$
83.4 ± (2.6)	83.7 ± (2.4)	83.0 ± (2.3)	81.7 ± (2.8)	82.2 ± (1.6)

Table 1: Accuracy for *SIFT* and *COL* features on different channels.

To combine the *SIFT* and *COL* descriptors, we consider two different strategies. *Early fusion* (*EF*) is done at the descriptor level, and *late fusion* (*LF*) combines the features at the latest stage by averaging the classifier outputs obtained for both descriptors.

As another strategy of taking into account the information in all the channels, one can concatenate SIFT descriptors per channel over the full 4D space (multi-channel *SIFT*). The results show that colour information is better used when considered in a specific descriptor. The results obtained by the late fusion of independently trained *SIFT* and *COL* descriptors outperform the multi-channel SIFT descriptor both in the original space ($SIFT_{r,g,b}$) and in the PCA space ($SIFT_{p1,p2,p3,p4}$). All reported results significantly outperform the result on the EPFL dataset reported in [1]. This big difference is justified by the FV framework we used. The best results are obtained with $SIFT_n + COL_{p1,p2,p3,p4}$. $SIFT_n + COL_{p1,p2,p3}$ and $SIFT_{l,n} + COL_{r,g,b}$ lead to very similar performances.

Descriptor1	Descriptor2	Fusion type	Accuracy
$COL_{r,g,b}$	$SIFT_n$	EF	84.1 ± (2.6)
		LF	86.2 ± (2.0)
$COL_{r,g,b,n}$	$SIFT_n$	EF	84.1 ± (2.9)
		LF	86.5 ± (2.4)
$COL_{p1,p2,p3,p4}$	$SIFT_n$	EF	83.2 ± (3.0)
		LF	**87.9 ± (2.2)**
$COL_{r,g,b}$	$SIFT_l$	LF	84.5 ± (2.3)
Brown and Süsstrunk [1]			72.0 ± (2.9)

Table 2: Accuracy (mean ± std) for different fusions on EPFL dataset.

[1] M. Brown and S. Süsstrunk. Multispectral SIFT for scene category recognition. In *CVPR*, 2011.

[2] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[3] N. Salamati, C. Fredembach, and S. Süsstrunk. Material classification using color and NIR images. In *IS&T/SID CIC17*, 2009.