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
Meylan, France.

Abstract

Standard digital cameras are sensitive to radiation in the near-infrared domain, but this additional cue is in general discarded. In this paper, we consider the scene categorisation problem in the context of images where both standard visible RGB channels and near infrared information are available. Using efficient local patch-based Fisher Vector image representations, we show based on thorough experimental studies the benefit of using this new type of data. We investigate which image descriptors are relevant, and how to best combine them. In particular, our experiments show that when combining texture and colour information, computed on visible and near-infrared channels, late fusion is the best performing strategy and outperforms the state-of-the-art categorisation methods on RGB-only data.

1 Introduction

Traditional photo cameras record radiation in a range that matches the spectrum of light visible to the human eye. Yet, the silicon sensor captures radiation in a larger range, in particular in the near-infrared (NIR). NIR radiation is artificially suppressed using a built-in filter, called hot mirror, to avoid possible artifacts produced by non-visible radiation. In general, computer vision algorithms process only images produced in that way. Would methods perform better if they had access to a larger spectrum of radiation than what provided by conventional camera? Recent studies [2, 16] have positively answered this question for some standard computer vision tasks. They have efficiently combined NIR and colour cues and it was shown that NIR radiation complements the visible light in such tasks.

In this paper, we similarly look at the NIR information captured jointly with visible information by standard digital cameras and further analyse the role of this channel in the image categorisation problem. Our paper contributes an extensive study on how this additional information can be efficiently integrated. Experiments are conducted on a recently released [2] Colour+NIR semantic categorisation dataset. When relevant, some of our observations are also confirmed on two additional visible-only datasets.

In our study, we apply a well-performing and generic categorisation method that works as follows. Texture or colour local descriptors are encoded using Fisher Vectors [14] to

BMVC 2011 http://dx.doi.org/10.5244/C.25.49

produce image signatures that are then fed into discriminative classifiers. We show evidence that the categorisation method we use is highly competitive, for both Colour+NIR images and standard datasets, making it a suitable framework for our study.

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

NIR radiation has been used in many different fields, such as astrophysics and medical imaging, often requiring some specific equipments. Infrared (IR) has also been considered more recently in the context of some classical computer vision problems. Li *et al.* [10] have used a LBP operation on NIR images combined with a classifier for face detection. Zhang *et al.* [21] have enhanced pedestrian tracking using IR. Unlike our work, the latter study considers radiation in the far-infrared band. In computational photography, Schaul *et al.* [17] have used NIR information for haze removal. Salamati *et al.* [16] have used colour and NIR cues for material classification. Closer to our work, Brown and Süsstrunk [2] have proposed a new local descriptor fusing visible and NIR information for scene categorisation.

Scene recognition is a core task of computer vision. Various methods have been proposed [5, 9, 12, 19, 20] using different types of descriptors. The recent study of Xiao *et al.* [20] shows evidence in favour of patch based local descriptors for colour images, which was also shown by Brown and Süsstrunk [2] in the context of Colour+NIR images.

The paper is organised as follows. Section 2 presents the datasets, and describes the nature and specificity of our data. The categorisation method is described in section 3, and in particular we elaborate on the local descriptors used in our experiments. Section 4 describes the conducted study. Finally, we draw conclusions in section 5.

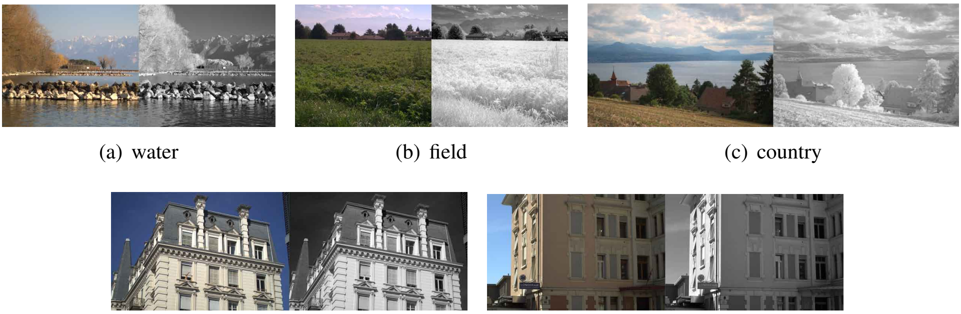
2 Datasets

Our experimental study is applied on a recently released dataset [2], where images are composed of visible and infrared channels.

Some of our observations could transpose to standard colour images, in this case, we show additional evidence on two other datasets, composed only of traditional visible (RGB) images. The first one is also a scene dataset (MIT-Scene 8), while the other one is a standard object classification dataset (PASCAL VOC 2007).

Near Infrared Dataset. Our experiments on combined visible and NIR information use the **EPFL scene classification dataset (EPFL)** [2] that consists of 477 images, divided into 9 scene categories (country, field, forest, mountain, old building, street, urban, water). Although the number of images in this dataset is limited, it contains challenging classes. There is some overlap between object categories. For example the classes in Figure 1(d) and 1(e) can be confused with each other, or Figure 1(c) can be confused with the class "water". Also, this is the only available benchmark dataset for categorisation that contains both RGB and NIR channels for each image.

As described in [2], the dataset was acquired as follows. To capture the NIR channel, the NIR blocking filter in the digital camera was removed. As such modified cameras do not allow for a joint acquisition, two separate exposures were captured, a visible only (with NIR



(a) water

(b) field

(c) country

(d) old-building

(e) urban

Figure 1: Some images of the EPFL dataset are displayed as pairs. On the left: the conventional RGB image of the scene, on the right: its NIR counterpart.

blocking filter) and a NIR one (without NIR filter but with a filter that suppresses visible light). The possible movements between the two shots are corrected using an alignment algorithm. At the end, images composed of 4 channels are produced. More details on the transformation between NIR sensor responses and a single NIR channel can be found in [6]. Examples of the RGB and NIR channels on some images are displayed as pairs in Figure 1. On the left, a conventional RGB image of a natural scene is shown, while the right side shows the NIR representation of the same scene. Note that NIR images have similar appearances, although there are some differences. Vegetation is bright and both sky and water are dark in the NIR channel of a scene.

For evaluation, we followed the same protocol as [2]: we randomly selected 11 images per class for testing (99 total) and trained the classifier using the remaining images. We repeated this process 10 times, and we report mean and standard deviation of the classification accuracy.

Visible Benchmark Datasets. We also consider the following two categorisation datasets.

The **MIT scene-8 classification dataset (MIT)** [13] is composed of similar scene categories as the EPFL dataset. It contains 2688 images of 8 outdoor scene categories: coast, mountain, forest, open country, street, inside city, tall buildings and highways. Following [13], we randomly select 100 images per class for training, while the remaining images are used for testing. We repeat this process 5 times, and report mean and standard deviation of the average accuracy.

We also test our observations on the popular **PASCAL VOC Challenge 2007 dataset (PASCAL)** [4] to show that similar conclusions can be deduced for object categorisation. It contains 9963 images in total and 20 object classes. In our experiments, we use the provided split into a training set and a testing set, and we compute mean average precision (MAP) over the classes. This allows us to fairly compare our results with the state-of-the art.

3 Categorisation Method

For our study, we use the Fisher Vector (FV) as an image representation. The FV [14] is an extension of the popular Bag-of-Features (BOF) representation [3], which considers higher order statistics instead of describing images as histograms of visual word occurrences [8, 14].

The principle is the following. First, a set of low level features is extracted from each image. Patches are extracted according to a regular grid (dense detector) and a local descriptor is computed for each patch. We apply a principle component analysis (PCA) projection to the descriptors. Projected descriptors are used to build a visual codebook, which describes the descriptor space as a Gaussian mixture model (GMM), with N Gaussians. The GMM distribution can be written as:

$$u_\lambda(x) = \sum_{i=1}^{N} w_i \mathcal{N}(x|\mu_i, \Sigma_i). \tag{1}$$

This visual codebook is used to transform each image into a global signature (FV). To compute the FV for one image, we consider its set of low level features $X = \{x_t, t = 1 \ldots T\}$. The FV $\mathcal{G}_\lambda^X$ characterises the sample X by its deviation from the distribution u_λ (with parameters $\lambda = \{\mu_i, \Sigma_i, i = 1..N\}$):

$$\mathcal{G}_\lambda^X = L_\lambda G_\lambda^X, \tag{2}$$

where G_λ^X is the gradient of the log-likelihood with respect to λ:

$$G_\lambda^X = \frac{1}{T} \nabla_\lambda \log u_\lambda(X). \tag{3}$$

L_λ is the Cholesky decomposition of the inverse of the Fisher information matrix F_λ of u_λ, *i.e.* $F_\lambda^{-1} = L'_\lambda L_\lambda$, where by definition:

$$F_\lambda = E_{x \sim u_\lambda} \left[\nabla_\lambda \log u_\lambda(x) \nabla_\lambda \log u_\lambda(x)' \right]. \tag{4}$$

More details on the FV computation and its theoretical foundations can be found in [8, 14]. As suggested in [15], we further apply square-root and L2-normalisation to the FV based image signature.

We varied the number of Gaussians in the visual codebook, from 64 to 2048 and observed very little difference between classification scores, for both SIFT and colour descriptors. We set 128 Gaussians for all our experiments on NIR. Similarly, for the PASCAL VOC dataset we varied this number and obtained no more improvement with more than 256 Gaussians. Consequently we used that number for the two visible datasets.

During training, the signatures from all training images are used to train a classifier. This classifier is then applied to each testing image signature. We use linear SVMs with a hinge loss as classifier and stochastic Gradient Descent (SGD) algorithm [1] to optimise it in the primal formulation.

Image channels and feature vectors. We use two types of features. First, we consider the popular *SIFT* [11] descriptor. We use the classical 4-by-4 bins decomposition of the patch, with 8 orientation histograms computed for each bin. The concatenation of these histograms leads to a 128 dimensional descriptor.

The second type of features looks directly at the intensity values in each image channel [15]. For a given channel, we consider a 4-by-4 bin decomposition of the patch, and each bin is described by the mean and standard deviation of the intensity of pixels in that bin. For a given channel, a 32 dimensional vector is produced. The final feature is obtained by concatenation of the vectors of all relevant channels. We refer to this feature as *COL*.

For visible-only datasets, images are described by 3 colour channels: r, g and b. For the NIR dataset, the original images contain 4 different channels, r, g, and b for the visible part,

	r	g	b	n	$p1$	$p2$	$p3$	$p4$	l
$SIFT$	$SIFT_r$	$SIFT_g$	$SIFT_b$	$SIFT_n$	$SIFT_{p1}$	$SIFT_{p2}$	$SIFT_{p3}$	$SIFT_{p4}$	$SIFT_l$
COL	COL_r	COL_g	COL_b	COL_n	COL_{p1}	COL_{p2}	COL_{p3}	COL_{p4}	-

Table 1: Summary of basic features considered for combination.

and a near-infrared (NIR) channel denoted by n. We additionally consider the luminance channel denoted by l, which can be computed on the visible part (r,g,b).

As in [2], we consider an alternative way of dividing images into channels. The initial 4D (r,g,b,n) vector is decorrelated and a PCA projection is computed. Afterwards, each pixel can be described in this new 4 dimensional space, as $p1$, $p2$, $p3$, $p4$. This new colour space can replace the conventional channels. This projection takes place so that the first component, *i.e.* $p1$, explains the maximum amount of variance of the data. In other words, the first component captures the largest amount of information in the decomposition. [2] showed that it is composed of a positive and almost equal contribution of all 4 channels.

The SIFT and COL descriptors can be applied and combined in different ways on these channels, hence we obtain a large set of possible features for building the image signatures. Table 1 summarises the basic features we consider for further combinations. We use the following notations. $SIFT_{i,j,k}$ denotes the concatenation of the $SIFT_i$, $SIFT_j$, and $SIFT_k$ features, computed on the channels i, j, and k, respectively. Similarly, $COL_{i,j,k}$ concatenates the COL_i, COL_j, and COL_k descriptors into a 96 dimensional descriptor. For 2 and 4 channels (such as $SIFT_{l,n}$ and $COL_{r,g,b,n}$) similar rules are applied.

4 Experimental Study

In this section we describe the conducted experiments. First, we show the positive influence of the PCA applied to both local descriptors. Then, we analyse the behaviour of the SIFT and colour descriptors independently. Finally, in order to determine the best strategy, different combinations of these descriptors are discussed.

4.1 The Influence of PCA on Local Features

In the previously described pipeline, a PCA projection is usually applied to reduce the dimensionality of the local descriptors. Here, we study the influence of this step and compare three configurations: i) no PCA is applied and the original descriptors are directly used to build the codebook, ii) PCA is used in order to reduce the dimensionality to D=64 (as in [15]), and iii) PCA is applied and descriptors are projected into the PC space, however, the full dimensionality is kept. For the latter case, referred to as Full PCA, the projected feature vectors have the same dimensionality as the original descriptors. The results of these three options are compared for “visible-only” descriptors (*i.e.* $SIFT_l$ and $COL_{r,g,b}$ that use only RGB channels). Results are reported in the first three sub-tables of Table 2, for the three datasets, namely, PASCAL, MIT, and EPFL. To study the PCA effect when the NIR channel is available, we also report the results on $SIFT_{p1,p2,p3,p4}$, used in [2], and $COL_{r,g,b,n}$, as a direct extension of $COL_{r,g,b}$ (see the last sub-table of Table 2 for the results).

First, we confirm that this PCA step has a crucial role and is needed in the pipeline (Table 2). If we compare Full PCA to the original descriptors, for visible only features, this difference is of a few percent for $SIFT_l$ (7.8% for PASCAL, 2.2% for MIT, and 4.3% for

	Descriptor	$SIFT_l$	$COLr,g,b$
PASCAL	Orig. desc.	51.6	37.6
	PCA with 64D	59.4	48.2
	Full PCA	59.4	48.2
MIT	Orig. desc.	90.1 ±(0.6)	75.8 ±(1.1)
	PCA with 64D	92.1 ±(0.2)	86.9 ±(0.3)
	Full PCA	92.2 ±(0.2)	86.9 ± (0.4)
EPFL	Orig. desc.	79.1 ± (4.7)	71.6 ± (2.6)
	PCA with 64D	83.4 ± (2.6)	80.2 ± (4.1)
	Full PCA	83.4 ± (2.8)	81.7 ± (2.8)
	Descriptor	$SIFT_{p1,p2,p3,p4}$	$COLr,g,b,n$
EPFL	Orig. desc.	83.8 ± (3.4)	71.1 ± (3.9)
	PCA with 64D	85.1 ± (3.4)	81.7 ± (1.8)
	Full PCA	85.9 ± (4.0)	82.2 ± (1.6)

Table 2: Mean average precision (MAP) for PASCAL, and class accuracy (mean ± std) for different PCA configurations for MIT and EPFL.

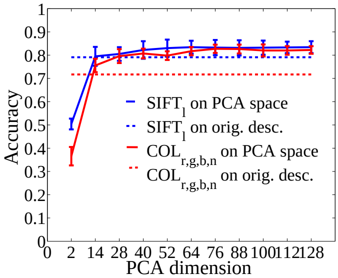


Figure 2: EPFL dataset: varying PCA dimensions for $SIFT_i$ and $COL_{r,g,b,n}$

EPFL). For $COL_{r,g,b}$ differences are larger at about 10%. Similar results are achieved when the NIR channel is available (few percent for $SIFT_{p1,p2,p3,p4}$ and 11.1% for $COL_{r,g,b,n}$).

We now look at PCA as a dimensionality reduction method. Table 2 reports results for a reduction to 64 dimensions. Also, we compare the final accuracy for a full range of dimensions in PCA projection, for both RGB+NIR descriptors (Figure 2). Results show that in most of the cases, by increasing the PCA dimensions no accuracy loss is observed, although it makes computations more expensive. It is interesting to note that no over-fitting is observed. Therefore, in the rest of the paper we choose to use the full resolution descriptors on the PCA space. Nevertheless, the stability observed over the set of dimensions indicates that PCA can also be used to reduce the feature dimension and fasten the computations for larger scale experiments.

4.2 Local Descriptors Study

We begin our study by looking at individual descriptors. As mentioned earlier, we considered two complementary features. The SIFT descriptor encodes gradient orientation, and consequently captures both the image contour, and the texture contained in a patch (Section 4.2.1). As colour information is also characteristic for scenes, and as an additional channel (NIR) is available, we look at the influence of the simple colour descriptor we described (Section 4.2.2). We will show evidence of their complementarity in Section 4.3.

4.2.1 SIFT Features

In this section, we investigate different ways of computing SIFT descriptors on RGB+NIR images. The standard SIFT feature is $SIFT_l$, which considers the visible-only luminance channel. We consider $SIFT_n$, as the NIR channel is also a mixture of red, green, and blue sensor responses, but in the invisible domain. We also consider $SIFT_{p1}$ as, by construction, the channel $p1$ incorporates a large amount of information from r,g,b and n. Finally, we combine the first two descriptors, $SIFT_i$ and $SIFT_n$ by concatenating them, as they contain non-overlapping information. Classification accuracies are shown in Table 3.

First, we observe that all these descriptors perform quite similarly. As expected, $SIFT_{p1}$ has a rather similar performance with $SIFT_l$, as it is achromatic and contains a positive and almost equal contribution of all 4 original channels. More surprising, $SIFT_n$ performs

	Descriptor	$SIFT_l$	$SIFT_{l,n}$	$SIFT_n$	$SIFT_{p1}$	$SIFT_{p1,p2,p3,p4}$	$SIFT_{r,g,b,n}$
EPFL	Accuracy	83.4 ± (2.6)	84.3 ± (3.3)	83.7 ± (2.4)	83.0 ± (2.3)	85.9 ± (4.0)	82.7 ± (3.1)

Table 3: Accuracy (mean ± std) on EPFL for SIFT features extracted on different channels.

	Descriptor	$COL_{r,g,b}$	$COL_{r,g,b,n}$	$COL_{p1,p2,p3,p4}$	$COL_{p1,p2,p3}$
EPFL	Accuracy	81.7 ± (2.8)	82.2 ± (1.6)	83.0 ± (2.2)	83.2 ± (2.4)

Table 4: Accuracy (mean ± std) on EPFL, colour features on different channel combinations.

comparably, while SIFT computed on any single visible colour channel performs worse (*e.g.* $SIFT_g$ gets $82.0\% \pm (2.8)$). Nevertheless the loss is not too important, showing that some texture information can be kept in any of these channels. This is further confirmed, on the 2 other datasets (For MIT for instance, $SIFT_l$ is $92.2 \pm (0.2)$, while $SIFT_r$ gives $91.9 \pm (0.3)$)

Then, considering $SIFT_{l,n}$ the concatenation of $SIFT_l$ and $SIFT_n$ gives an accuracy of 84.3%, showing a small improvement over both individual results and confirming that they indeed contain complementary information.

Instead of using a dedicated colour descriptor, as we will do in the next section, the colour information can be included in SIFT, by concatenating SIFT descriptors per channel over the full 4D space. Results for $SIFT_{r,g,b,n}$ and $SIFT_{p1,p2,p3,p4}$ are also reported in Table 3. $SIFT_{r,g,b,n}$ performs equally, and not better than luminance SIFT, although it uses additional NIR information. For the two visible datasets, $SIFT_{r,g,b}$'s accuracy even drops compared to $SIFT_l$ (MAP=54.1% on PASCAL and ACC= 86.9% on MIT). $SIFT_{p1,p2,p3,p4}$, the multi-spectral SIFT descriptor proposed by [2], stands out with 85.9% accuracy. By using a more adapted colour space, it makes better use of both the texture and the colour information.

We could have compared these with other versions of the colour SIFT descriptor, as proposed by [18], but [2] showed that the $SIFT_{p1,p2,p3,p4}$ performs better than those features.

4.2.2 Colour Features

Another feature relevant to classification is colour, which looks at the relation between r, g, b, and n channel values. In this part, we compare 4 different colour descriptors. $COL_{r,g,b}$ is the standard one and considers only visible information. Its direct extension to the NIR domain is $COL_{r,g,b,n}$. We also consider RGB+NIR information in the alternative colour PCA space, $COL_{p1,p2,p3,p4}$. Finally, as the last channel of this descriptor contains less information and a significant amount of noise, we also look at the $COL_{p1,p2,p3}$ descriptor. Table 4 compares the classification accuracy of the aforementioned descriptors.

The first observation is the good performances obtained by these colour descriptors, which encode only simple statistics. For EPFL, $COL_{r,g,b,n}$ is for instance only 1.5% below $SIFT_n$. Altough not comparable with SIFT, we observed the fair performance, considering the simplicity of this descriptor, of $COL_{r,g,b}$ on visible datasets too (48.2% on PASCAL and 86.9% on MIT).

Table 4 also confirms that incorporating NIR information increases the classification accuracy. The visible baseline $COL_{r,g,b}$ is slightly below $COL_{r,g,b,n}$, and is clearly outperformed by $COL_{p1,p2,p3,p4}$. This shows that the PCA projection in the colour space, which de-correlates the channels r, g, b, and n, further improves the results of the COL descriptor.

Finally, we can also observe that removing $p4$, which contains the least amount of information, does not change the classification accuracy, while slightly reducing the cost (smaller feature dimension).

Classifier	Descriptor1	Descriptor2	Fusion type	Accuracy
FV + SVM, with NIR	$COL_{r,g,b}$	$SIFT_{l,n}$	EF	84.3 ± (1.7)
			LF	**87.4** ± **(2.7)**
		$SIFT_n$	EF	84.1 ± (2.6)
			LF	86.2 ± (2.0)
	$COL_{r,g,b,n}$	$SIFT_{l,n}$	EF	84.4 ± (2.8)
			LF	85.5 ± (2.1)
		$SIFT_n$	EF	84.1 ± (2.9)
			LF	86.5 ± (2.4)
	$COL_{p1,p2,p3}$	$SIFT_{l,n}$	EF	83.3 ± (3.4)
			LF	86.7 ± (2.7)
		$SIFT_n$	EF	82.9 ± (2.7)
			LF	**87.5** ± **(2.4)**
	$COL_{p1,p2,p3,p4}$	$SIFT_{l,n}$	EF	85.8 ± (2.6)
			LF	86.5 ± (2.8)
		$SIFT_n$	EF	83.2 ± (3.0)
			LF	**87.9** ± **(2.2)**
FV + SVM, no NIR	$COL_{r,g,b}$	$SIFT_l$	LF	84.5 ± (2.3)
Brown and Süsstrunk [2]	–	–	–	72.0 ± (2.9)

Table 5: Accuracy (mean ± std) for different fusions on the EPFL dataset

4.3 Fusion of SIFT and colour information

Now that we have studied SIFT and colour descriptors independently, we are interested in strategies that combine texture and colour information. We already considered the $SIFT_{r,g,b,n}$ and the more successful $SIFT_{p1,p2,p3,p4}$, which encode both of them using the SIFT descriptors on multiple channels. In the following, we show that we can go beyond the accuracies obtained by these features by combining the previously studied SIFT and COL features.

We consider two types of combination. *Early fusion* is done at the descriptor level *i.e.* we concatenate both extracted features for each patch, and then apply a full PCA projection on the concatenated descriptors. *Late fusion* combines the features at the latest stage by averaging the classifier outputs obtained for both descriptors. In this study we apply, late fusion with equal weights. We skipped the optimisation of these weights on a validation set considering the already small size of our training set[1].

For SIFT, we kept $SIFT_{l,n}$ and $SIFT_n$ descriptors as they obtain best performances in the first part of our study. For the colour feature, we decided to test all features considered in 4.2.2 to see how complementary they are compared to the above mentioned two SIFT features. Table 5 shows the performance of early fusion (EF) and late fusion (LF) strategies on all 8 possible SIFT/COL pairs. We can draw several observations from this table.

First, it is clear that late fusion always outperforms early fusion, when SIFT and colour are combined. The difference is often significant (3% in average). Similar results are also observed for the two visible-only datasets. For the MIT and PASCAL datasets, early fusion of $SIFT_l$ and $COL_{r,g,b}$ gets 90.7% (± 0.4) and 54.9% while late fusion obtains 92.4% (± 0.6) and 61%, respectively. This could be explained by the very different nature of the combined descriptors. In that case, building specific classifiers for each descriptor, and doing a combination at the decision level is better suited.

Second, we can see that all the considered LF combinations, for which at least one descriptor contains NIR cues, outperform the "visible-only" baseline, *i.e.* the results obtained by the late fusion of $SIFT_l$ with $COL_{r,g,b}$. The most dramatic gain is observed for the classes country and mountain (from 68% to 75% for country and 86% to 97% for mountain). This again underlines the usefulness of the NIR information for the categorisation task.

[1] We did not study more complex combinations, simple strategies were shown to yield competitive results [7].

Finally, we observe the benefit obtained by the proposed strategy: the fusion of independently trained SIFT and colour descriptor-based signatures resulted in 87.9% accuracy, and outperforms the multi-spectral SIFT ($SIFT_{p1,p2,p3,p4}$) with only 85.9%. The best results are obtained with $SIFT_n + COL_{p1,p2,p3,p4}$, but $SIFT_n + COL_{p1,p2,p3}$ and $SIFT_{l,n} + COL_{r,g,b}$ lead to very similar performances. For visible only datasets, the same observation holds. While for PASCAL (resp. MIT), $SIFT_{r,g,b}$ obtained 54.1% (resp. 86.9%), the late fusion of $SIFT_l$ and $COL_{r,g,b}$ gets 61.0% (resp. 92.4%). This shows that for both standard and RGB+NIR datasets, colour can better improve SIFT based classifiers when used as a specific descriptor on its own and combined with SIFT at a late fusion stage.

Comparison with previous work. To improve our results, since the geometry is expected to be consistent across scene categories, we have combined the FV representation with the spatial pyramid (SP) [9], as suggested in [15]. For the visible datasets, we applied SP on $SIFT_l$ and $COL_{r,g,b}$ independently and combined them with late fusion. Results[2] were indeed further improved on MIT, from 92.4% to 93.6% and on PASCAL, from 61.0% to 63.7%.

However, when we applied SP on EPFL, on $COL_{p1,p2,p3}$ and $SIFT_n$, their late fusion gave 87.5% that is similar to its non-SP counterpart (87.5%). The absence of improvement can be justified by the limited size of our training set.

All reported results significantly outperform the 72% (± 2.9) on the EPFL dataset reported in [2]. This big difference is justified by the two different categorisation methods. While [2] uses a nearest neighbour based classifier directly at the local descriptor level, we train a linear SVM classifier on FV representations that aggregate local descriptors. Using the same local descriptor as [2], *i.e.* $SIFT_{p1,p2,p3,p4}$ reduced to 128D, leads already to a much higher accuracy (85.1 (± 3.4)). Nevertheless, our experiments confirm that their idea of decorrelating the four colour channels allows to improve the categorisation accuracy. Based on this colour space, our best strategy obtains a final accuracy of 87.9%.

For the MIT and PASCAL datasets, we obtain similar or better results than state-of-the art. On the MIT dataset, Oliva *et al.* [13] report a classification accuracy of 83.7% with GIST features. Our best result 92.4% was obtained with $SIFT_l + COL_{r,g,b}$ (LF) and SP. On the PASCAL dataset we obtain 63.7%, which is comparable to the 64.0% of [22].

5 Conclusion

In this paper, we have presented a thorough study of scene categorisation in the case where NIR information that can be captured with a normal digital camera is available in addition to visible light. This study is based on the FV representation, a generic and powerful categorisation framework. As a conclusion specific to that method, we have shown the usefulness of applying a PCA projection to local descriptors.

Through the study of two generic local descriptors, we have obtained NIR specific conclusions. First, we showed that NIR is useful information that combined with visible cues can improve recognition. In particular, the SIFT descriptor in the NIR domain performed similar to the standard luminance-based SIFT. The two latter are outperformed by the concatenation of SIFT descriptors computed on each channel of an alternative well-designed colour space ($p1, p2, p3, p4$) obtained through de-correlation of the colour channels, and NIR.

Second, we also investigated the best way how to include the 4D colour information in our categorisation method. Instead of encoding colour through a multi-channel SIFT, we

[2] As SP increases the image signature dimensionality, we reduced the features to 64D for these last experiments.

proposed to use a dedicated colour descriptor, that encodes local statistics in each channel. This simple descriptor performed almost as well as SIFT features, in particular in the alternative PCA colour space instead of the conventional r, g, b, and n space. The late fusion of FV signatures computed on this best colour descriptor ($COL_{p1,p2,p3,p4}$) and on the NIR SIFT descriptor ($SIFT_n$) was shown to be the best categorisation strategy and outperforms multi-channel SIFT descriptors.

This observation generalised to visible datasets, showing that colour information is better used when considering colour in a specific descriptor. A classifier trained on colour-descriptor based signatures, combined as late fusion with a complementary SIFT based classifier, outperforms the multi-channel SIFT descriptor ($SIFT_{r,g,b}$).

Acknowledgement This work was supported by the Swiss National Science Foundation under grant number 200021-124796/1 and Xerox foundation. We would like to thank F. Perronnin and J. Sánchez for providing the Fisher Vector implementation.

References

[1] L. Bottou. SGD. http://leon.bottou.org/ projects/sgd/.

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

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV SLCV Workshop*, 2004.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes challenge 2007 (VOC2007) results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[5] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. *CVPR*, 2005.

[6] C. Fredembach and S. Süsstrunk. Colouring the near-infrared. In *IS&T/SID CIC16*, 2008.

[7] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009.

[8] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, 1999.

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[10] S. Z. Li, R. Chu, S. Liao, and L. Zhang. Illumination invariant face recognition using near-infrared images. *IEEE TPAMI*, 29:627–639, 2007.

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[12] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42:145–175, 2001.

[13] A. Oliva and A. Torralba. http://people.csail.mit.edu/torralba/code/spatialenvelope/.

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

[15] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

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

[17] L. Schaul, C. Fredembach, and S. Süsstrunk. Color image dehazing using the near-infrared. In *ICIP*, 2009.

[18] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE TPAMI*, 32(9):1582–1596, 2010.

[19] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *IJCV*, 72:133–157, 2007.

[20] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[21] L. Zhang, B. Wu, and R. Nevatia. Pedestrian detection in infrared images based on local shape features. In *CVPR*, 2007.

[22] X. Zhou, K. Yu, T. Zhang, and T. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010.