Robustifying Descriptor Instability using Fisher Vectors

Ivo Everts, Jan C. van Gemert, Thomas Mensink, Theo Gevers, *Member, IEEE*

***Abstract*—Many computer vision applications including image classification, matching and retrieval use global image representations such as the Fisher Vector to encode a set of local image patches. To describe these patches, many local descriptors have been designed to be robust against lighting changes and noise. However, local image descriptors are unstable when the underlying image signal is low. Such low-signal patches are sensitive to small image perturbations which might come e.g. from camera noise or lighting effects. In this paper we first quantify the relation between the signal strength of a patch and the instability of that patch, and second we extend the standard Fisher Vector framework to explicitly take the descriptor instabilities into account. In comparison to common approaches to dealing with descriptor instabilities our results show that modeling local descriptor instability is beneficial for object matching, image retrieval and classification.**

***Index Terms*—Feature Extraction, Image Representation, Object Recognition**

I. Introduction

COMPUTER VISION tasks such as (object) recognition, image matching and retrieval typically depend on local image descriptors. Many robust image descriptors have been designed [24] or optimized [2] to deal with small changes in image geometry and photometry. For such robust descriptors, ideally, small geometric and photometric changes in the image recording conditions correspond to a negligible change in the image descriptor.

We focus on the popular descriptor family of gradient orientation histograms such as SIFT [14], HOG [5], SURF [1], or color SIFTS [24]. With a strong gradient signal, the SIFT descriptor is indeed robust to small perturbations in the image. However, if the gradient signal is weak, small changes in the image signal could result in huge variations in the local descriptor after ℓ_2-normalization. Such change in the descriptor after a small variation is what we denote as descriptor (in)stability. We show that there exists a parametric relation between descriptor stability and signal strength, which also applies to robust image descriptors and cannot be resolved by noise filtering.

To illustrate the problem, in Figure 1 we show for a few image patches the influence of adding small amounts of zero mean additive Gaussian noise to the image. Stable patches containing strong gradient signals are robust to the additive noise, and remain close to the original patch in descriptor space. In contrast, image patches with weak gradient signals make large shifts in the descriptor space after being distorted. This could severely influence the encoding scheme used to transform the local descriptors into an image representation.

There are two common approaches that address descriptor instabilities, either explicitly or implicitly. First, unstable patches may be identified based on a threshold on the gradient strength and subsequently mapped to a NULL descriptor [15], [25]. However, this may potentially lead to severe performance loss, whereas the optimal threshold is highly dataset dependent otherwise. Second, the instabilities may be taken for granted in the image representation and instead assumed to be modeled by a classifier as many variations are observed in a training set [27].

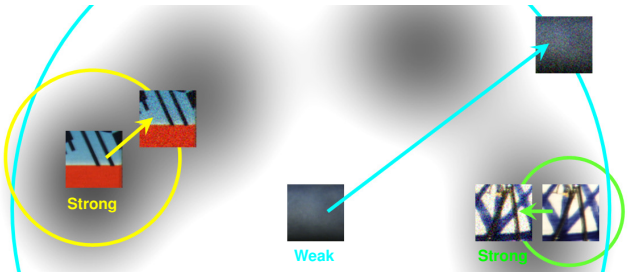


Fig. 1. Different image patches exhibit different behavior in SIFT descriptor space when subject to small image perturbations. Image perturbations for low-signal patches cause large jumps in descriptor space. We aim to model the descriptor instability (colored circles) based on the signal strength.

In contrast to fully relying on a classifier, thresholding on the gradient signal or changing the descriptor itself [8], we model descriptor instability in the Fisher Vector (FV) framework. The FV encodes local descriptors into a global image representation [18], [20] which can be used for classification, retrieval or matching. We choose the FV framework for two reasons, (i) it has proven to be one of the most powerful encoding schemes for image classification and retrieval [3], [11], and (ii) it offers a principled way to model descriptor instabilities in the underlying graphical model. The FV is based on the Fisher Kernel [9], and it consists of characterizing a set of local image descriptors by its deviation, measured by the gradient with respect to the log-likelihood, of a generative Gaussian mixture model (GMM). The GMM corresponds to a probabilistic version of the visual dictionary as used in the bag-of-visual-words approach [4], [13], [23]. We will show in this paper that descriptor instability modeling with FVs substantially improves recognition performance in comparison to signal thresholding for matching and classification tasks.

The rest of the paper is organized as follows. Next, we relate the signal strength to the instability of the descriptors. In Section III, we introduce our modification of the FV framework to incorporate descriptor instability, which we use in Section IV for image matching, retrieval and classification. Finally, we summarize our contributions in Section V.

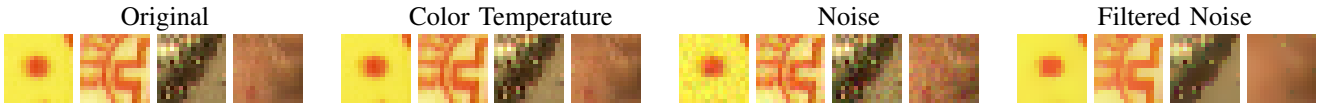


Fig. 2. Example patches from ALOI. The original patch and its near-copies due to: changed illumination color, additive Gaussian noise with $\sigma^2_{noise} = 10^{-3}$, and the noisy patch after applying the noise-reduction filter. See Figure 7 for examples of full ALOI images.

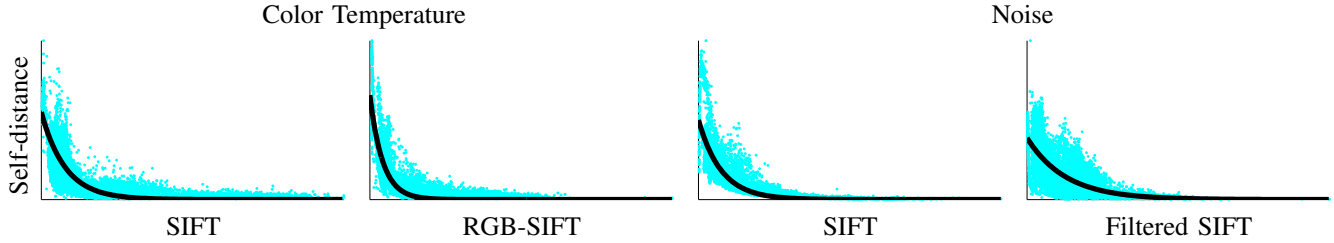


Fig. 3. Self-distance scattered against signal strength (x-axis) for color temperature change and additive i.i.d. Gaussian noise. The solid line is a least-squared fit of an exponential function. SIFT is computed from densely sampled 24x24 image patches from ALOI.

II. Quantifying the Relation between Signal Strength and Descriptor Instability

The stability of a descriptor is related to the signal strength of an image patch, as illustrated in Figure 1. Here, we aim to quantify the relation such that descriptor instability can be measured and interpreted as observational variance.

The signal strength of an image patch I is measured by the ℓ_2-norm of the gradient magnitudes $||\nabla I||_2$, since we describe image patches by the gradient-based SIFT descriptor. For gauging stability in descriptor space we use a near-copy of the same image patch, which is created by either (1) a stochastic change by adding a small amount of Gaussian noise, or by (2) a photometric change by a re-recording of the image patch under a slightly different light color. For each near-copy we extract its SIFT descriptor and compute the distance to the original descriptor. Ideally, these self-distances are close to zero since the underlying image content remains unaltered.

For the stochastic variant (1) to obtain a near-copy we use a small amount of i.i.d. zero-mean additive Gaussian noise, which is the standard model of amplifier noise [8], [16]. For these near-copies we also evaluate the impact of applying a noise reduction technique prior to descriptor extraction by an edge preserving anisotropic diffusion filter [17]. For the photometric variant (2), the near-copies are obtained by two recordings in the ALOI set [7] (see section IV) of the same object under a nearly imperceptible different illumination color temperature (2975^oK vs 3075^oK) where cameras are white balanced at 3075^oK. For the difference in illumination color, we also evaluated RGB-SIFT, which is invariant to changes in the illumination color [24]. Example patches are depicted in Figure 2.

In Figure 3 we show the relation between signal strength and descriptor instability for 10K randomly sampled image patches from ALOI. As illustrated, there is a strong relationship between signal strength and image descriptor instability. Strong signal patches are stable, i.e., close to the near-copies in descriptor space. Low-signal patches, however, are unstable as illustrated by large self-distances. Moreover, any attempt to remove the differences between near-copies by either photometric invariance (RGB-SIFT) or noise reduction (Filtered SIFT) does not diminish the instability. The experiment has also been repeated with filtered original patches to verify that the filtering routine is not influencing the observed relation. This results in essentially the same graphs (results not shown).

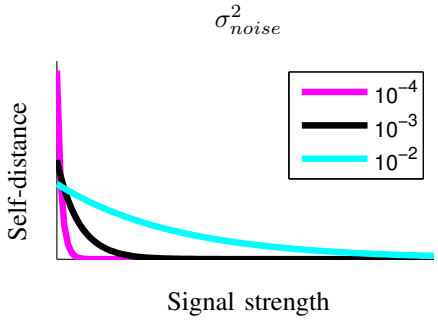


Fig. 4. Instability curves resulting from different levels of noise.

We propose to use the signal strength to model descriptor instability and incorporate the descriptor instability in the Fisher Kernel model. We use signal strength to model the descriptor instability $C(\cdot)$ as the average descriptor distance to itself, i.e., the variance of the descriptor distance. The relation between signal strength $x = ||\nabla I||_2$ and instability $C(x)$ can be described with an exponential function,

$$C(x) = \alpha e^{\beta x}, \tag{1}$$

where we use least-squares to fit α and β, as illustrated by the black line in Figure 3.

Note that a larger difference between near-copies will influence the self-distances. A higher noise level or larger color temperature difference will also increase the self-distance of stronger signal patches. To take this into account, we consider several noise levels where the variance of the applied Gaussian noise σ^2_{noise} is a parameter to be optimized on a hold out set. Figure 4 illustrates the effect of different noise levels on the instability curve.

The advantage of using the relation in Eq. (1) is that merely computing the image gradient norm allows us to estimate the descriptor instability in terms of its variance as a single scalar value. A scalar variance suffices as there is no reason to assume

a priori that the variance is not uniformly distributed over SIFT dimensions.

In this paper, we consider the inferred variance in addition to the descriptor itself as image measurement. The Fisher Vector representation is next reformulated such that descriptor instability is incorporated as measurement variance. This offers a principled approach to dealing with descriptor instability, as opposed to thresholding on the gradient signal or fully relying on a classifier.

III. Fisher Vectors from Unstable Descriptors

The Fisher Vector (FV) approach for image classification [18], [20] models a visual word vocabulary by a Gaussian mixture model (GMM) and characterizes a set of local image descriptors X by their gradient w.r.t. the parameters θ of the GMM under a log-likelihood model, i.e. $\text{FV} \equiv F_\theta \nabla_\theta \log p(X)$, where F_θ is the Fisher Information Matrix. In this paper, we also model the data using an GMM. However, the descriptor instabilities are incorporated while learning the parameters of the GMM. We analyze this by relating the responsibilities of the GMM components to the signal strength of the associated patches. As Fisher Vectors we extract gradients w.r.t. a different objective function to encode the noisy descriptors into the image representation.

Following [18], [20] we assume that our GMM has diagonal covariance and is defined as:

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_k p(\boldsymbol{x}|k)p(k) = \sum_k \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k^2)\ w_k, \qquad (2)$$

where $\boldsymbol{x}$ is an arbitrary point in the d dimensional descriptor space $\mathbb{R}^d$, k denotes a mixture component, $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k^2)$ is the multi-variate Gaussian distribution of component k with mean $\boldsymbol{\mu}_k$ and variances $\boldsymbol{\sigma}_k^2$, and w_k is the mixing weight (with constraints $\forall k : w_k \geq 0$ and $\sum_k w_k = 1$). When a signal threshold t_{signal} is used, $\boldsymbol{x}$ is mapped to a NULL descriptor (i.e. containing only zero elements) if the gradient strength of the underlying image patch falls below the threshold:

$$\boldsymbol{x} = \begin{cases} g(I) & \text{if } t_{signal} < ||\nabla I||_2 \\ \mathbf{0} & \text{otherwise.} \end{cases} \qquad (3)$$

Here, I is the image patch and $g(\cdot)$ denotes the descriptor extraction algorithm (i.e. SIFT). Note that the proposed method does not rely on such a threshold. The set of parameters to be estimated is $\theta = \{w_k, \mu_k, \sigma_k\}_{k=1}^K$, for a K component mixture.

The parameters of the GMM $\boldsymbol{\theta}$ in the FV framework are usually learned on a set $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ of local descriptors using the EM algorithm to maximize the log-likelihood $\sum_j \log p(\boldsymbol{x}_j)$. To gain insight in the influence of the descriptor instability on the FV encoding, we use again the patches from the ALOI dataset used in Section II, and train a GMM with $k = 16$ components. In Figure 5 (*top-row*), we show the following. We compute the most likely component k^* for the original patch j: $k^* = \arg\max_k q_{jk}$, where $q_{jk} \propto p(\boldsymbol{x}_j|k)p(k)$ is the posterior of component k for patch j. We show the difference $d_{jj'} = q_{jk^*} - q_{j'k^*}$ between the posteriors of the original patch j and its near-copy j', for component k^*. We relate this difference to the signal strength, similar as in Figure 3. We observe that for all descriptors there is a clear relation between the signal strength and the difference in posterior. Especially for patches with a low signal strength there are substantial changes in the posterior values.

We now incorporate the descriptor instabilities derived from signal strength for learning the parameters $\boldsymbol{\theta}$ of the GMM, to better model the uncertainties of the descriptors. We follow the EM approach of [26] to learn a GMM from noisy observations, which we coin N-EM for clarity in the rest of this paper. Their method is summarized as follows. Using all descriptors, together with their (diagonal) covariance matrices $\{\boldsymbol{C}_1,\ldots,\boldsymbol{C}_n\}$, we can define a variable kernel density estimator as:

$$f(\boldsymbol{x}) = \frac{1}{n}\sum_j f(\boldsymbol{x}|j) = \frac{1}{n}\sum_j \mathcal{N}(\boldsymbol{x};\boldsymbol{x}_j,\boldsymbol{C}_j). \qquad (4)$$

This kernel density estimator represents a non-parametric distribution over the descriptor space.

The learning problem is now expressed as the minimization of the Kullback-Leibler divergence between the kernel estimator and the unknown mixture, i.e.
$\theta^* = \arg\min_\theta D_{\text{KL}}\left[f(\boldsymbol{x})||p(\boldsymbol{x};\boldsymbol{\theta})\right]$, which yields the following function to maximize:

$$L = \sum_j \int_x f(\boldsymbol{x}|j)\log p(\boldsymbol{x};\boldsymbol{\theta})dx. \qquad (5)$$

Instead of directly maximizing L, an EM approach is considered to maximize a lower bound of L, which reads:

$$F = \sum_j\sum_k q_{jk}\left[\int_x f(\boldsymbol{x}|j)\log p(\boldsymbol{x}|k)dx + \log p(k) - \log q_{jk}\right], \qquad (6)$$

where q_{jk} is the posterior $p(k|\boldsymbol{x}_j)$ between a descriptor $\boldsymbol{x}_j$ and component k, also known as the responsibility. The integral in Eq. (6) is analytically solved as:

$$\int_x f(\boldsymbol{x}|j)\log p(\boldsymbol{x}|k)dx = \log\mathcal{N}(\boldsymbol{x}_j;\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k^2) - \frac{1}{2}\langle\boldsymbol{\sigma}_k^{-2},\boldsymbol{C}_j\rangle, \qquad (7)$$

where $\langle\cdot,\cdot\rangle$ denotes the dot-product between two vectors.

The N-EM update equations.
Iteratively maximizing F results in update equations which are very similar to the EM algorithm for noise-free data. First, in the expectation step the responsibilities are computed as follows:

$$q_{jk} = \frac{\mathcal{N}(\boldsymbol{x}_j;\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k^2)\ w_k\ \exp\left(-\frac{1}{2}\langle\boldsymbol{\sigma}_k^{-2},\boldsymbol{C}_j\rangle\right)}{\sum_{k'}\mathcal{N}(\boldsymbol{x}_j;\boldsymbol{\mu}_{k'},\boldsymbol{\sigma}_{k'}^2)\ w_{k'}\ \exp\left(-\frac{1}{2}\langle\boldsymbol{\sigma}_{k'}^{-2},\boldsymbol{C}_j\rangle\right)}. \qquad (8)$$

Second, in the maximization step the mixture parameters are updated as follows:

$$w_k = \frac{1}{n}\sum_j q_{jk} \qquad (9)$$

$$\boldsymbol{\mu}_k = \frac{1}{n\,w_k}\sum_j q_{jk}\,\boldsymbol{x}_j \qquad (10)$$

$$\boldsymbol{\sigma}_k^2 = \frac{1}{n\,w_k}\sum_j q_{jk}\left((\boldsymbol{x}_j - \boldsymbol{\mu}_k)^2 + \boldsymbol{C}_j\right), \qquad (11)$$

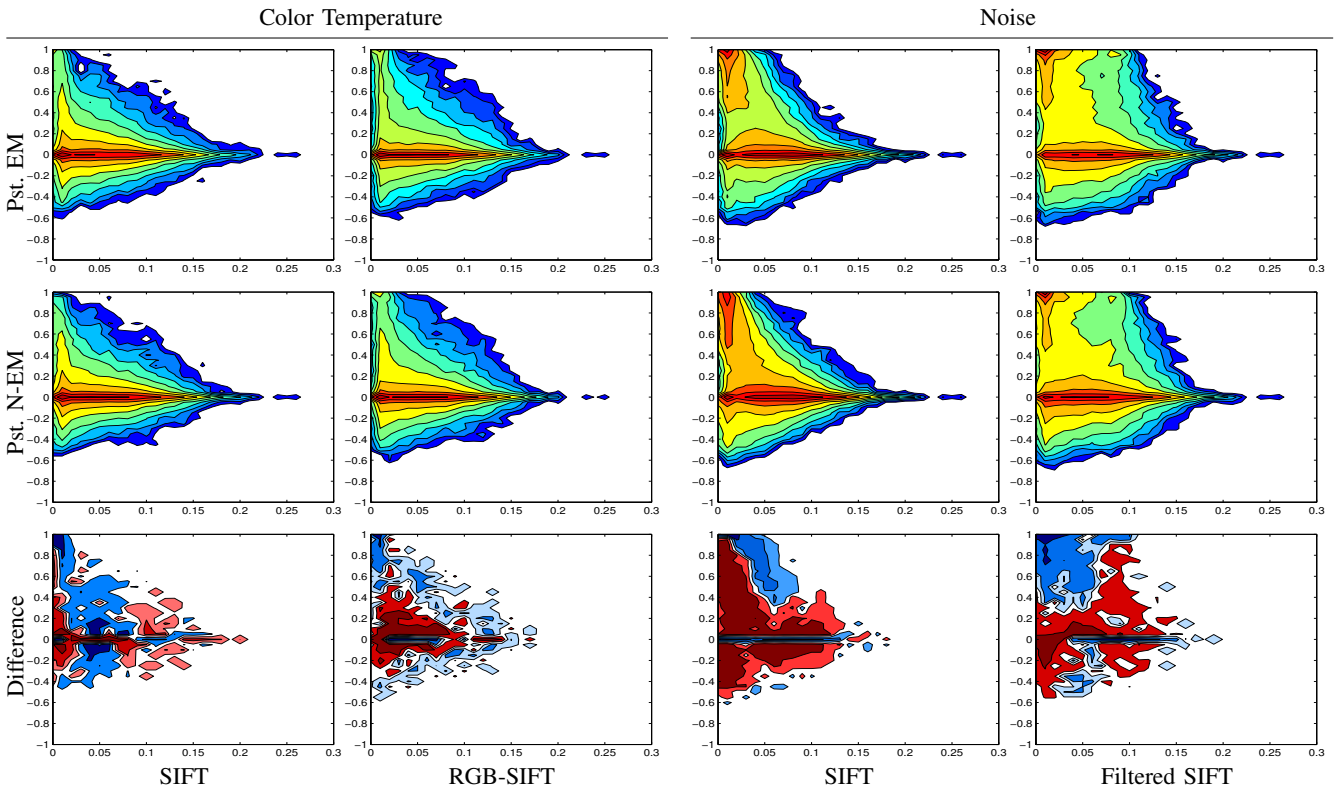


Fig. 5. Illustration of the behavior of the posterior distribution of the mixture components as a relation of the signal strength (x-axis). We show heat maps of patch frequencies for EM (*top*), N-EM (*middle*) and for the difference between the two methods (*bottom*). In the top and middle row, the red color indicates a higher density, in the bottom row red indicates that N-EM has higher values than EM. Considering the bimodal distributions appearing in the 'noise' panels, we note that the descriptor of a low signal patch may be so unstable that the corresponding noisy descriptor is located distant enough for completely removing the responsibility of the initial most likely GMM component, which explains the top mode of the distribution. The bottom mode (elongated around a difference in posterior probability of 0) indicates that descriptors may be stable irrespective of the associated signal strength.

where the exponentiation of a vector should be understood as a term-by-term operation.

Once more we use the patches from the ALOI dataset, to show the difference in the max-posterior in the GMM trained using N-EM in Figure 5 (*middle-row*). In this case q_{jk} is defined as in Eq. (8). The plot illustrates that indeed the posterior values of patch j and its near-copy j' are slightly more stable (i.e. similar to each other). This is also shown when plotting the difference between the distributions when using EM and N-EM in Figure 5 (*bottom-row*), where the blue region indicates higher mass for EM, and red for N-EM. Generally, more EM than N-EM mass is observed for larger differences in the posterior. This means that the probability of the most likely GMM component of a patch is less affected by distortion (illumination or noise) if N-EM is used (and thus the instability of the original descriptor is modeled). N-EM yields less variable assignments and appears more stable under distortions of the image.

To intuitively illustrate the N-EM algorithm and the effect of minimizing the proposed Kullback-Leibler divergence, we also compare, in Figure 6, a $k = 5$ GMM learned with EM (*left*) and with N-EM (*right*), together with synthetic 2 dimensional noisy data (*middle*). The plot illustrates that the mixture components also try to model the uncertainties of the data, e.g. the variance of the mixture components becomes higher in areas where the data has a high variance c.f. the blue and red components.

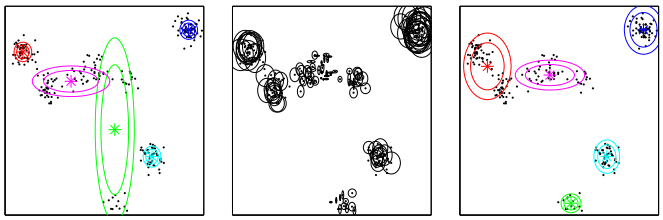


Fig. 6. Illustration of the EM algorithm using noisy observations on synthetic data. We show the GMM, with $K = 5$, learned using standard EM (*left*), the observation noise (*middle*), and the GMM after learning with N-EM for noisy observations (*right*). Both models start from the same initialization.

A. Fisher Vector Gradients

We next detail on how we modify the Fisher Vector model to incorporate the descriptor instabilities. Instead of using the gradients w.r.t. the log-likelihood, $\text{FV} \equiv \nabla_\theta \log p(X)$, we propose to use the gradient of the observations of an image w.r.t. the lower bound F in Eq. (6). In this model it is assumed that the responsibilities q_{jk} are given by Eq. (8), and that they are fixed, i.e. they do not yield a gradient signal w.r.t. θ (see also [21]).

We follow [12] and use $w_k = \frac{\exp \alpha_k}{\sum_{k'} \exp \alpha_{k'}}$, to obtain the

following gradients with respect to $\{\alpha_m, \mu_m, \sigma_m\}_{m=1}^K$:

$$\nabla_{\alpha_m} F = \sum_j \sum_k q_{jk}(\llbracket k = m \rrbracket - w_m), \quad (12)$$

$$\nabla_{\boldsymbol{\mu}_m} F = \sum_j q_{jm} \frac{(\boldsymbol{x}_j - \boldsymbol{\mu}_m)}{\boldsymbol{\sigma}_m^2}, \quad (13)$$

$$\nabla_{\boldsymbol{\sigma}_m} F = \sum_j q_{jm} \left(\frac{(\boldsymbol{x}_j - \boldsymbol{\mu}_m)^2}{\boldsymbol{\sigma}_m^3} - \frac{1}{\boldsymbol{\sigma}_m} + \frac{\boldsymbol{C}_j}{\boldsymbol{\sigma}_m^3} \right), \quad (14)$$

where $\llbracket z \rrbracket$ denotes the Iverson brackets which is 1 if z is true and zero otherwise, and where the division between vectors or the exponentiation of a vector should be understood as term-by-term operations.

Intuitively, these gradient vectors include the descriptor instability both when computing the posterior q_{jk} according to Eq. (8), and in the gradients with respect to the variances.

Comparison to Standard FVs.
To compare our model to the normal FVs used in [20], lets assume that the kernel density function $f(x|j) = \delta(x, x_j)$ is a Dirac delta function. In that case the integral of Eq. (7) is trivially solved as $\int_x f(x|j) \log p(x|k) dx = \log p(x_j|k)$, and the lower bound, Eq. (6), reads:

$$F_\delta = \sum_j \sum_k q_{jk} \left[\log p(x_j|k) + \log p(k) - \log q_{jk} \right]. \quad (15)$$

When assuming, as above, that the responsibilities q_{jk} are fixed, it is easy to show that the gradients of Eq. (15) w.r.t. $\{\alpha_m, \mu_m, \sigma_m\}$ yield the normal FV equations, see e.g. Eq. (9)-(11) in [18]. Note that the responsibilities are now defined as usual as $q_{jk} \propto p(\boldsymbol{x}_j|k)p(k)$.

Conforming to [20], as the final image representation we use $\text{FV} = [\nabla_{\boldsymbol{\mu}_k} F \ \nabla_{\boldsymbol{\sigma}_k} F]_{k=1}^K$, and we apply power-normalization $z \leftarrow \text{sign}(z)|z|^{-1/2}$, followed by ℓ_2 normalization $z \leftarrow \frac{1}{||z||_2} z$. When multiple spatial pyramid levels are used, each pyramid cell is individually normalized.

Note that, in this paper, we consider the scalar variance $C_j = C(x)\mathbf{I}$ from Eq. (1) which results from its definition in terms of descriptor instability. This is however not a restriction of the model.

IV. Experiments

We compare our models to the standard FV framework for three different computer vision tasks: object matching, image retrieval and object category recognition. We use a classifier (SVM) for object category recognition whereas the other tasks are performed by directly matching the FVs (i.e. nearest neighbor classification and distance-based ranking). As we model SIFT instability directly in the FV representation, it is expected that matching-based approaches will especially benefit from the proposed method. Opposed to this, we expect learning-based approaches to already exhibit robustness due to the observed variations in the training data. The basic setup is the same for all tasks.

A. Experimental Setup

Patch extraction proceeds by sampling 24x24 patches on a dense grid every 4 pixels. Images are processed on 5 scales by iterative down-sampling with a factor of $\sqrt{0.5}$. The signal strength for a patch is measured by the ℓ_2-norm of its image gradient. SIFT descriptors [14] are extracted using VLFEAT [25] and we apply PCA to reduce the 128 dimensions of SIFT to 64, as commonly done in the FV framework [22].

GMM parameters are estimated from a set of 1M randomly sampled descriptors. The same initialization of the standard EM is used for our N-EM approach, taking into account the descriptor instabilities. For all experiments we estimate the parameters of the GMM and the instability curve on a separate dataset. We study the effect of instability modeling for different values of k, the number of GMM components. Instability curves are modeled separately per scale with additive Gaussian noise in $\{10^{-4}, 10^{-3}, 10^{-2}\}$. The gradient signal threshold is considered in $\{0.0025, 0.005, 0.01\}$ where 0.005 is the default setting of the SIFT implementation.

The reported performance measures are averaged over 3 runs using different seeds for training the GMM. We have observed standard deviations ranging from 0.2 to 0.5 percentage points. Based on unpaired t-tests, the best improvements on all datasets were found to be significantly different from the baseline at a standard 5% confidence level.

B. Matching Task: ALOI Object Matching

The Amsterdam Library of Object Images (ALOI) image set [7] contains 1000 objects with systematic variations in viewing angle, illumination angle, and illumination color. We use this set because it allows systematic evaluation of changing a single appearance variable. We focus on lighting arrangement change since this has proven to be difficult for SIFT [24]. We match the canonical image, i.e., with all lamps turned on, with a paired random illumination arrangement in a set of all other objects. Images are cropped such that no background is visible. Performance is measured by the percentage of correct closest object images using the Euclidean distance to the canonical image. See Figure 7 for some examples of ALOI's illumination conditions.

Observing the results in figure 9, it appears that instability modeling has a considerable effect. That is, the already near-perfect baseline of 97.3% is improved to 98.9% for k=64, whereas performance improves with 2 percentage points for lower values such as k=16. Furthermore, SIFT mapping by signal thresholding may lead to incidental improvements (i.e. for $k = 16$, $t_{signal} = 0.005$), but performance degrades in general. Thus, despite the unstable behavior associated to low-signal descriptors, discriminative information may be lost when they are all mapped to the same NULL descriptor.

As object matching in the ALOI dataset is a simple problem, and the appearance changes are controlled and expected to be advantageous to our method, we conduct a more challenging image retrieval experiment in which the image content and recording conditions are more complex.

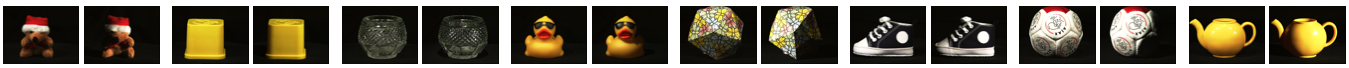


Fig. 7. Example images from ALOI under varying illumination conditions. The left image is the canonical image, i.e., with all lamps turned on. For every canonical image, we have randomly picked a single match from a different illumination arrangement for creating pairs in the dataset. The dark background is ignored in the experiments.

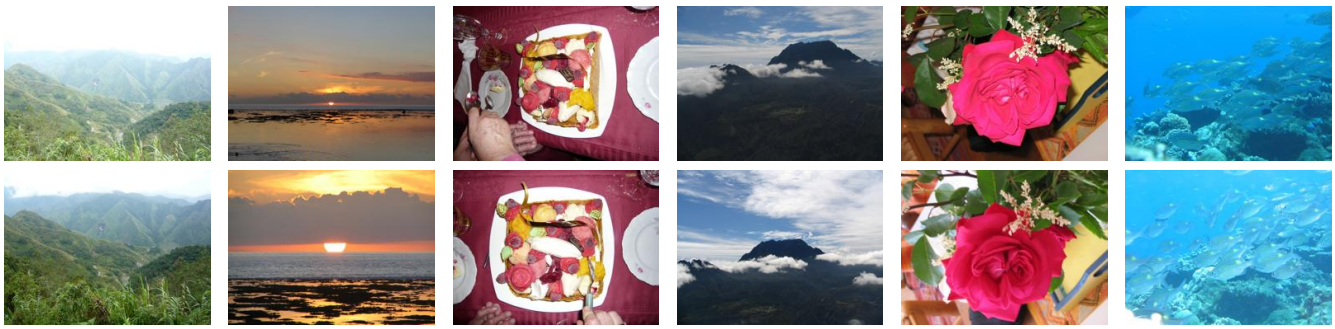


Fig. 8. Example images from INRIA Holidays used in the image retrieval experiment. Matching image sets consist of 2-4 images.

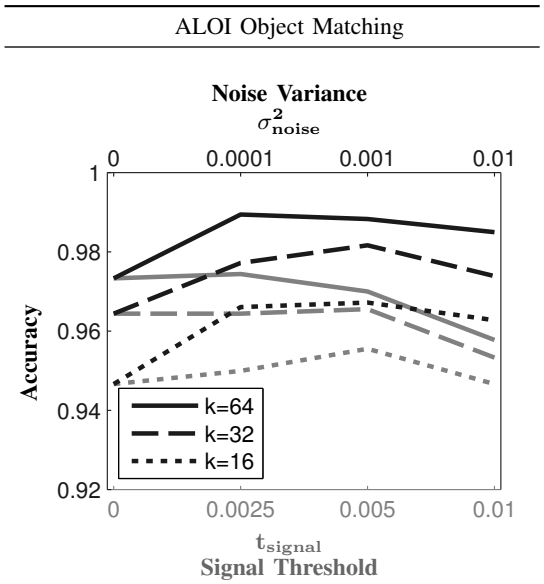


Fig. 9. Object matching results. This task is performed based on direct FV matching. The overall results generally improve due to instability modeling, of which several settings for the corresponding noise levels (variance) are plotted in black. Matching performance of several settings for the signal threshold are plotted in gray.

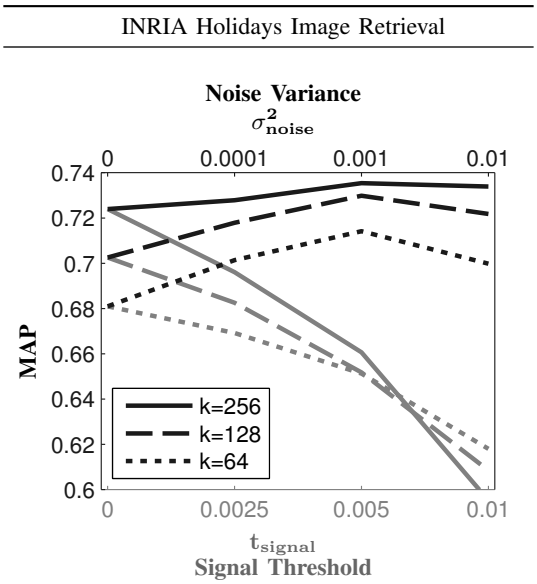


Fig. 10. Image retrieval performance. The effect of instability modeling is similar to ALOI results (black plot). Special treatment of unstable descriptors by thresholding on the gradient signal may lead to drastic performance loss (gray plot).

C. Matching Task: INRIA Holidays Image Retrieval

For the image retrieval task, we use the INRIA Holidays dataset [10], which consists of approximately 1500 images, see Figure 8. For each of the 500 queries, the remaining images are ranked and average precision (AP) is computed. The final performance is measured as the mean AP over all queries (MAP).

The results for various k values in Figure 10 show results similar to ALOI as performance increases along with k. Our baseline result of 72.5 MAP for $k = 256$ compares favorably to the 0.70 of the Fisher Vector approach in [19]. Instability modeling leads to substantial performance gains, where the improvement increases as k decreases. The reason for this lies in the fact that compensating for the incidental location of descriptors with respect to the GMM clusters has most effect when the GMM is sparse: the responsibilities become even more stable as observations are 'spread' through the descriptor space by considering their associated instabilities as measurements of variance (which is also illustrated in Figure 5).

It is interesting to observe the dramatic performance degradation on the Holidays dataset when a threshold on the gradient signal is used for dealing with unstable descriptors. Also here, we conclude that discriminative information is ignored by mapping unstable SIFT descriptors to a NULL

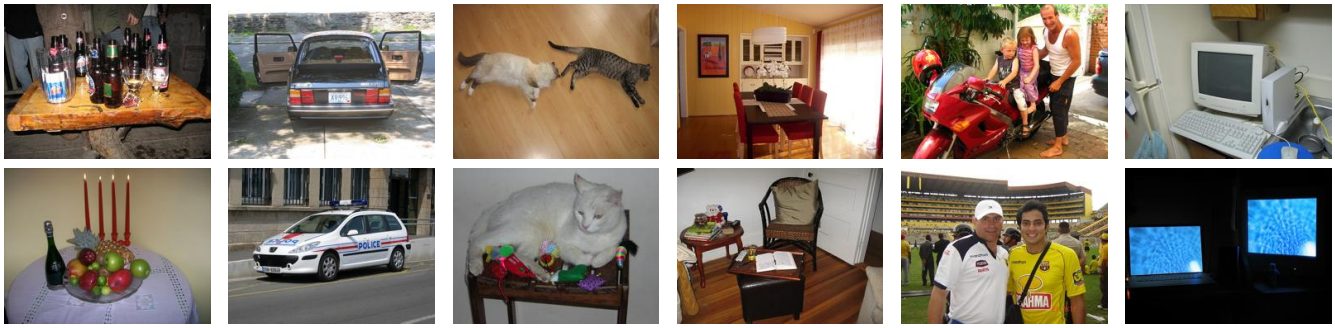
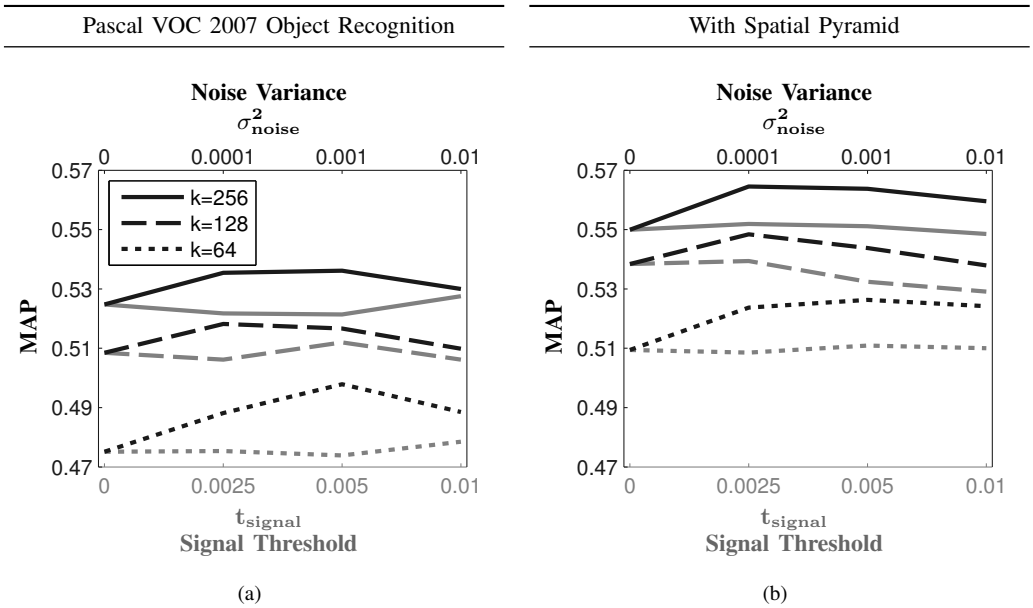


Fig. 11. Example images from Pascal VOC 2007. This dataset exihibits much larger intra-class variation than the ALOI and Holidays datasets, and consists of train and test sets.



(a) (b)

Fig. 12. VOC 2007 validation scores in MAP for varying noise levels (black) and signal thresholds (gray). Without (Figure 12(a)) and with (Figure 12(b)) spatial pyramid.

descriptor. The effect is more pronounced here than for the ALOI object matching task because the retrieval task is harder and thus suffers more from a cut down of discriminative information.

The matching and retrieval experiments showed that modeling the instability of SIFT descriptors effectively maintains discriminative power of low-signal patches in the FV representation. Another approach to (implicitly) dealing with descriptor instability is to model the variations in unstable image content by training a classifier on many examples.

D. Classification Task: Pascal VOC 2007 Object Category Recognition

Object recognition is evaluated on the Pascal VOC 2007 visual recognition benchmark [6]. This is a well known set for image categorization and consists of 20 categories such as *Aeroplane, Bottle, Cat, Dog*, etc. (see Figure 11), with train, validation and test sets of 2501, 2510, and 5011 images respectively. We train a linear SVM on the two variants of the FV and evaluate the effect of a 3x1 spatial pyramid [13]. Performance is evaluated by mean average precision (MAP), the area under the precision-recall curve.

The results in Figure 12(a) show the effect of varying the noise levels and signal thresholds on the Pascal VOC validation set. As with the other datasets, a high noise variance of $\sigma^2_{noise} = 0.01$ for instability modeling decreases performance in comparison to lower values. High descriptor uncertainties may render the responsibilities ambiguous. Opposed to this, the representations benefit substantially from instability modeling, especially those that are based on a sparser GMM ($k = 64$). However, the performance differences between FVs with and without instability modeling appear somewhat less pronounced than for the ALOI and Holidays datasets. This is because the SVM effectively exploits the variations in unstable image content by observing the training examples. Furthermore, in contrast with the substantial performance drops resulting from descriptor NULL-ling in the retrieval task, we observe that a signal threshold may slightly improve classification results. However, this highly depends on the settings for k and the threshold t_{signal}, and also varies across datasets. Opposed to this, instability modeling consistently improves over the baseline on all datasets and k, where a noise variance σ^2_{noise} of 10^{-2} or 10^{-3} has to be chosen.

Figure 12(b) shows the same effect as Figure 12(a), but with the use of a spatial pyramid level in the representation, which is commonly used to boost the performance. Here, the improvements also hold and even become more pronounced for the often used setting of k=256 [22].

Based on the observations made on the validation set, we conclude to not perform descriptor mapping based on signal thresholding, and to adopt a noise level of $\sigma^2_{noise} = 10^{-3}$ for instability modeling. These settings are applied on the Pascal VOC 2007 test set, for which results are reported in Table I. The results show improvements for every FV component and combinations thereof.

In summary, it is always beneficial to incorporate descriptor instability in the FV as long as the noise variance for instability modeling is not too high (i.e. $\leq 10^3$).

TABLE I
RECOGNITION PERFORMANCE ON PASCAL VOC 2007 TEST SET, USING SPATIAL PYRAMIDS AND k=256 GMM COMPONENTS. INCLUDED ARE THE RESULTS FOR 0^{th} ORDER (w), 1^{st} ORDER (μ) AND 2^{nd} ORDER (σ) STATISTICS, AND COMBINATIONS THEREOF.

	w	μ	σ	$\mu + \sigma$	$w + \mu + \sigma$
Standard FVs	39.0%	55.4%	56.7%	58.9%	58.9%
Proposed FVs	40.0%	56.2%	57.1%	60.5%	60.7%

1) The Effect of Variance Estimation: We next present a number of recognition experiments on the Pascal VOC 2007 validation set in which variations of the proposed method are considered. First, instead of estimating the variance by instability modeling, we consider assigning the same variance to all descriptors. Second, we apply instability modeling either during GMM learning or FV coding in order to determine where it has most effect. The results are presented in Table II.

TABLE II
VARIATIONS OF THE PROPOSED METHOD (ON THE PASCAL VOC 2007 VALIDATION SET USING $k = 256$ AND SPATIAL PYRAMIDS). INSTEAD OF ESTIMATING THE VARIANCE PER DESCRIPTOR, A FIXED VARIANCE CAN BE USED FOR ALL DESCRIPTORS. VARIANCE ESTIMATION BY INSTABILITY MODELING CAN BE PERFORMED EITHER DURING GMM LEARNING (N-EM) OR FV CODING (N-FV), OR BOTH. USING A FIXED VARIANCE OF 0 CONSTITUTES THE BASELINE, WHEREAS THE PROPOSED METHOD IS N-EM+N-FV.

Fixed Variance				Estimated Variance		
0	10^{-5}	10^{-4}	10^{-3}	N-EM	N-FV	N-EM+N-FV
55.0%	54.75%	54.79%	54.71%	55.3%	56.0%	56.4%

Using the same non-zero variance for all descriptors has a marginal negative effect, because stable descriptors may be spread out too much whereas the opposite holds for unstable descriptors. Note that a fixed variance of 0 constitutes the standard FV. Furthermore, the table shows that per-descriptor variance estimation by instability modeling has most effect in the FV coding step, as compared to GMM learning. This illustrates that the enhanced stability of the GMM (as depicted in Figure 5) not necessarily implies very substantial performance improvements. More gain is obtained from the FV coding step because this directly affects the image representation.

2) Run-time Comparison: The proposed method is computationally somewhat more expensive than standard FVs. Computing the responsibilities from noisy observations requires an extra dot product (in the log domain) in Eq. (8). Furthermore, the instabilities propagate to the computation of second order statistics in Eq. (11) for GMM training and Eq. (14) for FV coding, involving, for all observations, an extra element-wise summation in both Eq. (11) and Eq. (14), and an extra division in Eq. (14). The proposed method is 8.9% slower as compared to NULL-ing the low signal patches ($t_{signal} = 0.01$), which is determined by computing the total runtime of extracting all descriptors from the VOC2007 train set.

V. CONCLUSION

In this paper we make the observation that local image descriptors extracted from low-signal image patches are unstable in feature space. We fit an exponential relation between signal strength and descriptor instability and exploit the estimated instability as measurement variance in a novel Fisher Vector feature encoding scheme. The proposed framework allow to model the descriptor instability in a principled way, as opposed to employing a threshold on the gradient signal. In effect, the discriminative information of these unstable descriptors is better preserved. The results show improvements for image classification, retrieval and matching. The proposed method can be especially beneficial in settings where classification is performed by direct descriptor matching.

REFERENCES

[1] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. SURF: Speeded up robust features. *CVIU*, 110:346–359, 2008.
[2] M. Brown, G. Hua, and S. Winder. Discriminative learning of local image descriptors. *PAMI*, 2011.
[3] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.
[4] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Int. Workshop on Stat. Learning in Computer Vision*, 2004.
[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
[6] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop, 2007.
[7] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The amsterdam library of object images. *IJCV*, 2005.
[8] T. Gevers and H. Stokman. Robust histogram construction from color invariants for object recognition. *PAMI*, 26(1), 2004.
[9] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, 1999.
[10] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.
[11] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Trans. PAMI*, 2012. to appear.
[12] J. Krapac, J. Verbeek, and F. Jurie. Modeling spatial layout with fisher vectors for image categorization. In *ICCV*, 2011.
[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.
[15] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 2005.
[16] J. Ohta. *Smart CMOS Image Sensors and Applications*. CRC PressINC, 2008.
[17] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *PAMI*, 1990.

[18] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.
[19] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2011.
[20] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, 2010.
[21] R. Salakhutdinov, S. Roweis, and Z. Ghahramani. Optimization with em and expectation-conjugate-gradient. In *ICML*, pages 672–679, 2003.
[22] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 2013.
[23] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.
[24] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 2010.
[25] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.
[26] N. Vlassis and J. Verbeek. Gaussian mixture learning from noisy data. Technical Report IAS-UVA-04-01, Universiteit van Amsterdam, 2004.
[27] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 2006.