Discriminative Feature Fusion for Image Classification

Basura Fernando[1], Elisa Fromont[2], Damien Muselet[2] and Marc Sebban[2]
[1]K.U.Leuven, ESAT-PSI, Leuven, Belgium
[2]CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France
Université de Saint-Étienne, Jean-Monnet, F-42000, Saint-Étienne, France
basura.fernando@esat.kuleuven.be, {elisa.fromont,damien.muselet,marc.sebban}@univ-st-etienne.fr

Abstract

Bag-of-words-based image classification approaches mostly rely on low level local shape features. However, it has been shown that combining multiple cues such as color, texture, or shape is a challenging and promising task which can improve the classification accuracy. Most of the state-of-the-art feature fusion methods usually aim to weight the cues without considering their statistical dependence in the application at hand. In this paper, we present a new logistic regression-based fusion method, called **LRFF***, which takes advantage of the different cues without being tied to any of them. We also design a new marginalized kernel by making use of the output of the regression model. We show that such kernels, surprisingly ignored so far by the computer vision community, are particularly well suited to achieve image classification tasks. We compare our approach with existing methods that combine color and shape on three datasets. The proposed learning-based feature fusion process clearly outperforms the state-of-the art fusion methods for image classification.*

1. Introduction

During the past few years, bag-of-words approaches have allowed significant advances in image classification [6]. Most of these methods use the well-known SIFT descriptor [20]. Therefore, they are mostly based on local shape information, although it has been shown that color information can also be an efficient cue in some image classification tasks [27, 31]. However, the way to efficiently combine multiple cues is still an open problem because the relevance of each individual cue (color, shape, texture, etc.) is highly dependent on the images to classify [12, 25]. For instance, to discriminate soccer players from two teams, color information is crucial. On the other hand, the shape is essential to separate bananas from yellow apples, while we usually need both cues to discriminate most of the flowers.

Instead of using all the cues and all the visual words for all the classes (as done in [4, 12, 23, 28]), we claim that it would be more relevant to adaptively select (weight) a set of diverse and complementary visual words (color, shape or texture visual words) in order to better discriminate each class from the others. This prevents us from using confusing visual words while keeping only the most relevant ones for a given classification task. To achieve this task, we propose in this paper to first create a visual dictionary for each cue. Then, we use a Logistic Regression (LR) method [8] to deduce from the multiple dictionaries the most class-specific discriminative visual words. Finally, we take advantage of the LR outputs (not only the conditional probabilities but also some geometrical information w.r.t. the learned hyperplanes) to design a new efficient marginalized kernel [14, 16].

The rest of this paper is organized as follows: first, we present the related work and contributions in Section 2. Section 3 is devoted to the notations and definitions. Then, we present our fusion method in Section 4 and we introduce our new marginalized kernel which will be used to learn a SVM classifier. The experimental results are presented in Section 5 and a conclusion and promising lines of research in Section 6.

2. Related Work and Contributions

There exist several ways to combine multiple cues in an image classification task. When cues are combined at the **pixel level**, each single dimension of the descriptor (extracted from a local region) represents a mixture of the information coming from each cue. Such a combination requires the use of multiple cue descriptors, *e.g.* spatio-colorimetric descriptors such as the color-SIFT [1, 4, 5, 27]. These methods lead to spatio-colorimetric visual words whose each single dimension represents both color and shape information. Since an independent weighting for each cue is impossible, these models are recommended in applications where all the cues are necessary to discriminate

the classes. Moreover, because they mix multiple information, they usually require a large number of visual words to represent an object class [5, 27].

When the combination of the cues is applied at the **local region level**, different descriptors are extracted from each local region, each one representing only one cue. Then, they are concatenated into a local feature descriptor using a given weighting scheme. The resulting representation is then used to learn a single visual dictionary [9, 28]. Compared with the approaches that combine color and shape at the pixel level, these methods allow to weight each cue. But again, this level of combination leads to spatio-colorimetric visual words which may reduce the final classification accuracy by introducing confusing information when only one cue (or a subset of the cues in general) is relevant to learn a concept. For example, Khan *et al.* [25] propose to combine color and shape at the local region level by simulating the human visual attention. They characterize each image with histograms of shape visual words (one histogram per concept) in which the frequency of each visual word is weighted by its (color) probability to belong to the considered concept. Likewise, Elsayad *et al.* [10] and Chen *et al.* [7] propose to weight the contribution of each shape visual words in the histogram by using a probability derived from color information. The drawback of these approaches is that the resulting representation is a shape-based histogram, *i.e.* the primary visual cue is assumed to be the shape.

The last combination strategy consists in merging all the cues at the **global image level**. In this case, multiple dictionaries are created, one for each cue, and the global description of the image informs us about the cues present in the image without binding them neither at the pixel level nor at the local region level. Nilsback *et al.* [23] apply this kind of approach to classify flowers. They use a multiple kernel learning (MKL)-based feature fusion [12] where each kernel deals with one specific cue. Note that such a description informs us about the shapes and colors present in the images but does not provide any information neither about the color of each shape nor about the shape of each colored region.

In the computer vision community, the **global image level** fusion is usually known as late fusion. When it is applied without any weighting scheme, it is known as standard late fusion (**SLF**). Even though there are some differences between **pixel level** and **local region level** fusion methods, both of them are referred to as early fusion because the resulting visual words contain mixed information.

All the previous approaches share a common feature: somehow, they combine all the cues without neither (i) taking into account their dependence nor (ii) selecting the most relevant visual words for the classification task at hand. To overcome these drawbacks, we propose in this paper a new method which combines multiple cue information (in our experiments color and shape) by implicitly weighting the visual words belonging to different dictionaries according to their relevance for the considered classification task. These dictionaries can have been generated both at a **pixel level** and at a **local region level** that allows us to take advantage of the associated combination methods, and they are concatenated at a global level which enables us to benefit from the **global** combination method. Technically, our contribution is two-fold:

First, we propose to train a L1-logistic regression (LR) model for each class (versus all the others) that allows us to deduce in a sparse way the most discriminative visual words from multiple dictionaries. The main advantage of a LR model is that it does not assume that the unlabeled data follow the same class distribution as the labeled training examples. Note that regression models have already been used for different purposes in computer vision [32, 33]. And more recently, some interest has been shown in group sparsity [19, 33] in regression models.

Second, we make use of these class-specific LR models to design a new performing marginalized kernel [16] which takes into account not only the probability for two images to belong to the same class, but also an image-to-class similarity measuring in a way the margin between the images and the learned hyperplanes. Surprisingly, unlike Fischer kernels which have been widely used in image classification [17, 24], marginalized kernels have been almost ignored by the computer vision community (except, *e.g.*, [2] and [18]). In this paper, we show experimental evidences that our kernel is very effective and by using both non-linear (the conditional probabilities given by the discriminative LR model) and linear (the distance to the hyper-plane) information in the calculation of the similarity, allows us to improve the classification accuracy.

3. Notations and Definitions

Let us consider we have a training set $S = \{(I_i, y_i)\}_{i=1\ldots m}$ of m labeled images, where each image I_i belongs to some image space $\mathcal{I}$ and each label y_i is in the set $\mathcal{Y} = \{1, \ldots, N\}$. For each image, we consider that a set of key-points is extracted, each of them being mapped in n different descriptor spaces. For each of the n descriptor spaces, we assume that a visual word dictionary is learned of size $d_1, \ldots, d_n$ respectively. Usually, a bag-of-words model consists in assigning to each key-point the closest visual word in the considered descriptor space. In our case, each extracted key-point is assigned to n visual words, one for each dictionary. Applying this principle for all the key-points extracted from a given image I, it is then possible to represent I in the form of a set of n normalized feature vectors $\mathbf{x^i} = (x^i_1, \ldots x^i_{d_i}), i = 1 \ldots n$, where $\sum_j x^i_j = 1$ and where each component x^i_j represents the normalized occurrence frequency of the j^{th} visual word of the i^{th} dictionary

in the image. Finally, the n vectors are merged to obtain a single feature vector $\mathbf{x}$ of size $d = \sum_i d_i$ which will constitute the input data of our algorithm. Our objective is to fuse multiple cues utilizing the most relevant visual words w.r.t. the image classification task at hand and to use these discriminative words to classify images using a new marginalized kernel. We suggest in this paper to learn the relative importance of each visual word by means of a regularized logistic regression model [8]. Logistic regression assumes, for a binary classification task (*i.e.* $y \in \{-1, 1\}$), that the following relation holds:

$$log(\frac{p(y=1|\mathbf{x};\alpha,\boldsymbol{\beta})}{p(y=-1|\mathbf{x};\alpha,\boldsymbol{\beta})}) = \alpha + \sum_{j=1}^{d} \beta_j x_j, \quad (1)$$

where α and $\boldsymbol{\beta}$ are the parameters to learn. From Eq.(1), we deduce that

$$p(y=1|\mathbf{x};\alpha,\boldsymbol{\beta}) = \frac{1}{1+exp(-[\alpha+\sum_{j=1}^{d}\beta_j x_j])}. \quad (2)$$

Note that $p(y=-1|\mathbf{x};\alpha,\boldsymbol{\beta}) = 1 - p(y=1|\mathbf{x};\alpha,\boldsymbol{\beta})$. Roughly speaking, logistic regression consists in modeling the posterior probability of the class membership using a linear function. α is known as the bias parameter and $\boldsymbol{\beta}$ as the weight vector of the function. By considering that $\alpha = \beta_0$ and $x_0 = 1$, Eq.(2) can be rewritten as $p(y = 1|\mathbf{x};\boldsymbol{\beta}) = 1/(1 + exp - (\boldsymbol{\beta}^T\mathbf{x}))$. The optimal parameters $\hat{\boldsymbol{\beta}}$ of the logistic regression model are usually obtained by optimizing the conditional log-likelihood $\mathcal{L}$ [22], such that:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= argmax_{\boldsymbol{\beta}} log \prod_i p(y_i|\mathbf{x}_i;\boldsymbol{\beta}) \\ &= argmin_{\boldsymbol{\beta}} \sum_i log(1+exp(-y_i\boldsymbol{\beta}^T\mathbf{x}_i)). \quad (3)\end{aligned}$$

4. Feature Fusion Algorithm

4.1. Regularized logistic regression-based method

Most of the feature fusion approaches apply a single weight for all the visual words of the same cue and whatever the classification problem [4, 12, 23, 28], whereas the importance of the visual words can vary within a cue and between the binary classification tasks. In the following, we propose an alternative by learning the relative importance of each visual word derived from a dictionary *for each* underlying binary classification task. We use a logistic regression model in order to assess the ability of each visual word to separate a class from all the others. However, the dimensionality of the resulting vector $\mathbf{x}$ is the cumulative sum of the different sizes of the visual dictionaries. Consequently, the optimization problem described by Eq.(3) can lead to an overfitting phenomenon if the log-likelihood is optimized without any regularization. To prevent overfitting, a L2-regularization term is usually used that boils down to restricting large value components [13]. In our case, we not only aim to avoid overfitting (due to the large number of features) but also to control the number of visual words involved in the classification. To do this, we suggest in this paper to resort to a L1-regularization which creates sparse answers. Therefore, the objective function of Eq.(3) can be rewritten as follows:

$$\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}}(\textstyle\sum_i log(1+exp(-y_i\boldsymbol{\beta}^T\mathbf{x}_i)) + \lambda||\boldsymbol{\beta}||_1), \quad (4)$$

where $\lambda > 0$ is the regularization parameter. Unlike the L2-regularization which restricts large values, the L1-regularization term penalizes all factors equally. Note that the non-differentiability of the L1-norm in Eq.(4) can be efficiently treated by interior-point methods.

4.2. Logistic Regression Marginalized Kernel

Once the parameters $\boldsymbol{\beta}$ have been learned for each class y (versus the others), hereafter denoted by $\boldsymbol{\beta_y}$, we suggest in our approach to take advantage of the output of the regression model to design a new efficient kernel. As reported in [25], a standard approach in image classification consists in using the intersection kernel that usually allows us to obtain good results with SVMs [30]. This kernel is defined as follows: $K(\mathbf{x},\mathbf{x}') = \sum_{i=1}^{n} min(x_i, x_i')$. We claim that this kernel is not well suited to our logistic regression context. In the following, we rather make use of the information provided by our model to design a new marginalized kernel [14, 16], which is defined in its original form as follows:

$$K(\mathbf{x},\mathbf{x}') = \sum_y \sum_{y'} P(y|\mathbf{x})P(y'|\mathbf{x}')K_{\mathbf{z}}(\mathbf{z},\mathbf{z}'), \quad (5)$$

where $y, y' \in Y$ and $K_{\mathbf{z}}(\mathbf{z},\mathbf{z}')$ is a joint kernel over labeled examples $\mathbf{z} = (\mathbf{x}, y)$. In our specific framework of logistic regression, we define a new marginalized kernel from the following joint kernel:

$$K_{\mathbf{z}}(\mathbf{z},\mathbf{z}') = S(y,y') \times \boldsymbol{\beta_y}^T\mathbf{x} \times \boldsymbol{\beta_{y'}}^T\mathbf{x}', \quad (6)$$

where $S(y,y')$ is the similarity between two classes y and y' ($0 \leq S(y,y') \leq 1$). $K_{\mathbf{z}}(\mathbf{z},\mathbf{z}')$ takes into account not only the similarity between the classes of $\mathbf{x}$ and $\mathbf{x}'$ but also the image-to-class similarities in the form of $\boldsymbol{\beta_y}^T\mathbf{x}$. The image-to-class distances are already efficiently used in [3, 26] in nearest-neighbor based image classification approaches. Roughly speaking, this means that $K_{\mathbf{z}}(\mathbf{z},\mathbf{z}')$ will return a high similarity for two images located on the same side (*i.e.* leading to a positive product $\boldsymbol{\beta_y}^T\mathbf{x} \times \boldsymbol{\beta_{y'}}^T\mathbf{x}'$) of the hyperplanes associated to classes y and y' that are

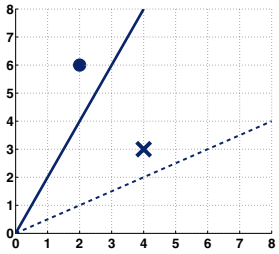

Figure 1. Graphical explanation of our joint kernel. The two labeled images $\mathbf{z} = (\mathbf{x}, y)$ and $\mathbf{z}' = (\mathbf{x'}, y')$ (represented by the circle and the cross) will be considered as similar if (i) the two classes (solid and dashed lines) are similar (according to $S(y, y')$) and (ii) the images are located on the same side and almost at the same distance from their corresponding separators.

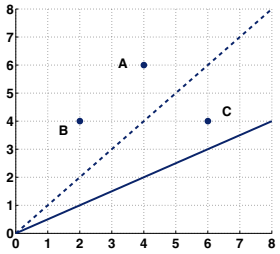

Figure 2. Graphical explanation of our marginalized kernel. While the similarities $K(A, B)$ and $K(B, C)$ are the same with an intersection kernel, the location of the points w.r.t. to the hyperplanes allows our marginalized kernel to state that $K(A, B) > K(C, B)$.

similar (*i.e.* $S(y, y')$ is close to 1). The intuition behind $K_{\mathbf{z}}(\mathbf{z}, \mathbf{z}')$ is graphically described in Figure 1. According to our joint kernel, the two images (represented by the circle and the cross) will be considered as similar if (i) the two classes (solid and dashed lines) are similar and (ii) the images are located on the same side and more or less at the same distance from the separators. Note that the distance between the two classes is not represented graphically.

For a general kernel function to be valid, it needs to be positive semi-definite (PSD). Since the class of PSD kernels are closed under addition and multiplication, a marginalized kernel is PSD as long as the joint kernel $K_z(z, z')$ is PSD itself. According to the Mercer's theorem, any valid kernel function admits a representation as a simple inner product between suitably defined feature vectors, *i.e.*, $K_z(Z, Z') = \phi_Z^T \phi_{Z'}$, where the feature vectors come from some fixed mapping $Z \rightarrow \phi_Z$. In our case, a simple way to satisfy the PSD constraint consists in setting $S(y, y') = 1$ when $y = y'$ and 0 otherwise[1]. Therefore, in this case, $\phi_Z = \boldsymbol{\beta}_{\boldsymbol{y}}^{\boldsymbol{T}} \mathbf{x}$ and $K_z(Z, Z') = \phi_Z^T \phi_{Z'}$; hence $K_z(Z, Z')$ is PSD. Plugging $K_z(Z, Z')$ in Eq. (5) boils down to only considering the cases where $y = y'$ leading to the following marginalized kernel for multi-class classification problems, called **LRMK**.

$$K(\mathbf{x}, \mathbf{x}') = \sum_{y \in Y} p(y|\mathbf{x}, \boldsymbol{\beta}_{\boldsymbol{y}}) \times p(y|\mathbf{x}', \boldsymbol{\beta}_{\boldsymbol{y}}) \times \boldsymbol{\beta}_{\boldsymbol{y}}^{\boldsymbol{T}} \mathbf{x} \times \boldsymbol{\beta}_{\boldsymbol{y}}^{\boldsymbol{T}} \mathbf{x}'. \tag{7}$$

Beyond the fact that considering that $S(y, y') = 1$ if $y = y'$ (and 0 otherwise) allows us to design a valid joint kernel, it also enables us to make use of our kernel without any information about the label of $\mathbf{x}$ and $\mathbf{x}'$, that is crucial in an image recognition task. Moreover, note that our new kernel considers both non-linear (the conditional probabilities given by the discriminative logistic regression model) and linear (the distance to the hyperplane) information in the calculation of the similarity. More precisely, it returns a weighted sum of joint similarities over all the of classes. The weights are the learned conditional probabilities from the logistic models, and give a kind of confidence in the output of the logistic regression. A graphical explanation of the interest of $K(\mathbf{x}, \mathbf{x}')$ is presented in Figure 2. Let us assume we have three examples A, B, C and two learned logistic models. If we compute the standard histogram intersection kernel $K(\mathbf{x}, \mathbf{x}')$, we deduce that $K(A, B) = K(B, C)$. So in this sense, A is as similar to B as it is to C. Using our new marginalized kernel leads to $K(A, B) > K(B, C)$, because A and B (i) are always on the same side of the hyperplanes and (ii) are almost at the same distance from the separators.

4.3. Logistic Regression-based Feature Fusion

Data: A set $S = \{(\mathbf{x_k}, y_k)\}_{k=1...m}$ of labeled images mapped in a d-dimensional descriptor space, where $y_k \in \{1, \ldots, N\}$
Result: A final classifier $H(\mathbf{x})$
for *Each class* $i = 1, \ldots, N$ **do**
  1. Create a training set $S' = \{(\mathbf{x_k}, y'_k)\}_{k=1...m}$ from S s.t. $y'_k = +1$ if $y_k = i$ and $y'_k = -1$ otherwise;
  2. Learn a LR model $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ from S' solving equation (4);
3. Use the marginalized kernel $K(\mathbf{x}, \mathbf{x}')$ of Eq.(7) to learn a SVM classifier H;
Return the final classifier H;

Algorithm 1: Pseudo-code of **LRFF**.

The pseudo-code of our algorithm, called **LRFF** (for **L**ogistic **R**egression-based **F**eature **F**usion), is presented in Algorithm 1. **LRFF** takes as input a training set of labeled images $S = \{(\mathbf{x_k}, y_k)\}_{k=1...m}$. Then, for each binary problem represented by its corresponding set of images S' for a

[1] We are aware that this manner reduces the potential expressive power of our joint kernel. But note that $K_z(z, z')$, as described in Eq. (6), opens the door to interesting perspectives in metric learning.

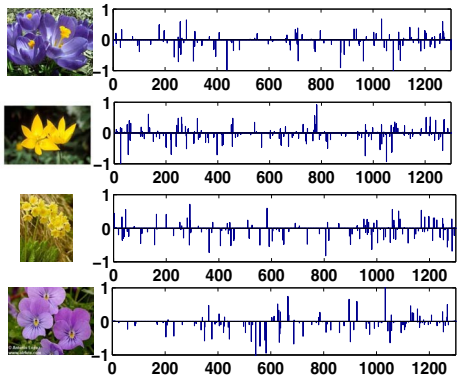

Figure 3. Learned weights for some flower categories in the *Flower-17* dataset. The X axis represents the visual words where the numbers 1 to 1000 represent shape words and 1001 to 1300 color words.

given class i (step 1), a L1-regularized logistic regression model is learned by solving equation (4). The output of this model is the conditional probability for an image to belong either to the class y_i or to any other possible class (step 2). The regularization term ensures that the model is sparse in terms of the considered visual words. A high value β implies that the corresponding visual word contributes a lot to the positive class. On the other hand, a high negative value means that the considered visual word contributes a lot to the negative class. Any other visual word with a very small weight does not contribute significantly to discriminate positive from negative examples. The conditional probabilities are then exploited in our marginalized kernel which is used to train a SVM classifier H (step 3).

4.4. Sparse and class-specific visual words

Figure 3 shows the weights learned for some categories of the *Flower-17*[2] dataset. First, we notice that, due to the use of the L1-norm, few visual-words are selected for each class. Second, the fact that the weight distributions are different between the considered categories validates our intuition that the contribution of each word has to be assessed with respect to the given classification task.

5. Experiments

To assess the relevance of our fusion method, we carry out in this section a series of experiments on three datasets and compare **LRFF** with some state-of-the-art approaches. After having presented the method we use to create the bag-of-visual words (BoW), we give some details about the datasets and the experimental setup. Finally, we present the results that show the effectiveness of our approach.

[2]http://www.robots.ox.ac.uk/∼vgg/data/flowers/

5.1. BoW models: feature detection, feature description and codeword generation

In our experiments, we use a standard BoW creation technique in which the key-points are extracted using multiple key-point detectors (Harris-Laplace key-point detector and dense sampling). The visual dictionaries are constructed using the K-means clustering algorithm [21]. Feature vectors (bag-of-word histograms) are built for each cue separately by using the corresponding visual dictionary. In this series of experiments, we use two visual cues: the shape and the color. Shape information is extracted from SIFT descriptors (which are the most used in the literature), and color information is extracted from Hue-histograms [28] (written "Hue.") and Color Name (CN) descriptors [29]. For each dataset, we also use the best known color-shape descriptors: HSV-SIFT [27], C-SIFT [5] or Opponent-SIFT [27].

5.2. Experimental setup

We first present baseline results which consist in classifying the data using single cue descriptors (shape or color): *SIFT* for shape, *Hue.* and *CN* for color. We compare our proposed method, **LRFF**, with the following three methods: (i) HSV-SIFT, C-SIFT or Opponent-SIFT which are **pixel level** fusion approaches (ii) the standard late fusion method (**SLF**) which concatenates the individual dictionaries in a **global level**, and (iii) the Color Attention (**CA**) mechanism presented in [25] which is also a **local region** based approach. We evaluate all these methods using three datasets : the *PASCAL-VOC-2007* [11], the *Soccer*[3] and the *Flower-17* datasets. All the reported results (except for **LRFF**) are obtained using a SVM [30] learned from the feature vectors of the different methods with the same cost parameter $(C = 1)$ and using an intersection kernel. For **LRFF**, we use the marginalized kernel of Eq.(7). Note that we use the mean average precision (*M. AP.* in the results) as evaluation criterion for the *PASCAL-VOC-2007* dataset as it is the measure commonly used in the literature while we use the standard accuracy rate (*Score* in the results) for the two other datasets. Finally, we analyze in Section 5.6, the relevance of the new kernel compared to a multiple kernel (**MKL**) method and a standard intersection kernel.

5.3. Results on the PASCAL-VOC-2007 dataset (shape dominant)

This dataset contains 20 object classes, 5,011 training images, 4,952 test images and is known to be shape dominant. The experimental results are shown in Table 1 where the column *Dictionary* indicates the number of visual words used in the experiments and column *Descriptors* indicates the type of descriptors used for fusion.

[3]http://lear.inrialpes.fr/people/vandeweijer/soccer/soccer_data.tar

We use *C-SIFT* as the **pixel level** fusion method for this dataset because it has been shown to obtain the best results [27].

Despite the fact that this dataset is shape dominant, we suggest to add step by step color information to assess the ability of the fusion methods to improve the baseline results. We can note (with descriptors *SIFT+Hue.*) that our fusion method **LRFF** is able to more efficiently take advantage of the color information to improve the results (50.19 versus 49.15 and 43.18 for **CA** and **SLF** respectively). This is confirmed when adding the CN descriptor. The same remark can be made when we merge all the dictionaries (*SIFT+Hue.+CN+C-SIFT*). **LRFF** significantly outperforms SLF (53.81 versus 47.04)[4]. To obtain state-of-the-art results using our method on PASCAL, it is necessary to include spatial information and more advanced encoding methods [6], but this is not the aim of this experiment.

Method	*Descriptors*	*Dictionary*	*M. AP.*
Shape Cue	*SIFT*	1000	43.13
Color Cue	*Hue.*	400	21.44
Color Cue	*CN*	300	21.44
Pixel Level	*C-SIFT*	4000	46.81
SLF	*SIFT+Hue.*	1000+400	43.18
CA	*SIFT+Hue.*	1000+400	49.15
LRFF	*SIFT+Hue.*	1000+400	**50.19**
SLF	*SIFT+Hue.+CN*	1000+400+300	44.84
CA	*SIFT+Hue.+CN*	1000+400+300	50.25
LRFF	*SIFT+Hue.+CN*	1000+400+300	**51.99**
SLF	*SIFT+Hue.+CN+C-SIFT*	1000+400+300+4000	47.04
LRFF	*SIFT+Hue.+CN+C-SIFT*	1000+400+300+4000	**53.81**

Table 1. Results on the PASCAL-VOC-2007 dataset.

5.4. Results on the Soccer dataset (color dominant)

This dataset contains 280 images from 7 football teams, 175 images are used as training examples and 105 are kept in a test set. For comparisons, we have selected *HSV-SIFT* and color-opponent-SIFT (*OPP.SIFT*) as the **pixel level** fusion method as proposed in [29, 25]. The results obtained for this dataset, which is color dominant, are shown in Table 2.

Method	*Descriptors*	*Dictionary*	*Score*
Shape Cue	*SIFT*	400	0.49
Color Cue	*Hue.*	300	0.68
Color Cue	*CN*	300	0.71
Pixel Level	*HSV-SIFT*	1000	0.72
Pixel Level	*OPP.SIFT*	1000	0.84
SLF	*SIFT+Hue.*	400+300	0.79
CA	*SIFT+Hue.*	400+300	0.82
LRFF	*SIFT+Hue.*	400+300	**0.86**
SLF	*SIFT+Hue.+CN*	400+300+300	0.88
CA	*SIFT+Hue.+CN*	400+300+300	0.92
LRFF	*SIFT+Hue.+CN*	400+300+300	**0.94**
SLF	*SIFT+Hue.+CN+OPP.SIFT*	400+300+300+1000	0.91
LRFF	*SIFT+Hue.+CN+OPP.SIFT*	400+300+300+1000	**0.96**

Table 2. Results on the Soccer dataset.

[4]It is not recommanded to use color-shape descriptors (**Pixel Level** fusion methods) such as *C-SIFT* with the **CA** method [15].

As expected, the shape descriptors are less relevant than the color ones to classify soccer images. For example, the single Color Name descriptor with a very small dictionary of 300 words gives a rather good (0.71) classification rate, while the SIFT descriptor provides a poor performance (0.49). Moreover, we can note that once again, **LRFF** outperforms all the other methods. Indeed, it efficiently makes use of the shape information to improve the results of the color-based methods. On the other hand, it is worth noting that the pixel-level method *HSV-SIFT*, gives the worst results among all the fusion approaches for this dataset (72%) even with a large dictionary of 1,000 words.

Our method is able to take advantage of color dominant dimensions and to fuse useful shape information leading to very good results (0.86 using *Hue.+SIFT* descriptors and of 0.94 using *CN+Hue.+SIFT*). When *SIFT* is combined with *CN+Hue.+OPP.SIFT*, **LRFF** achieves a superior classification score of 0.96.

5.5. Results on the Flower dataset (both color and shape dominant)

This dataset contains 17 categories of flowers, 1,020 training examples and 340 test images which require both shape and color information to be correctly classified. The results obtained for this dataset are shown in Table 3.

Whatever the descriptors, **LRFF** improves the results of **SLF** and **CA** (by 4 and 1 points respectively using *SIFT+Hue.* and by 5 and 3 points respectively using *SIFT+CN+Hue.*). These results provide an experimental evidence that our fusion method is very useful to combine different sources of information by selecting the most relevant color, shape and color-shape visual words for a given classification problem. To reinforce this claim on this specific dataset which requires to use both shape and color information, we also compare our method with **SLF** using the state-of-the art Color-opponent-SIFT (*OPP.SIFT*) descriptor [27]. Once again the results suggest that the proposed **LRFF** method has the ability to take advantage of all levels of feature fusion where **LRFF** reports a superior classification score of **0.93** on this challenging dataset.

Method	*Descriptors*	*Dictionary*	*Score*
Shape Cue	*SIFT*	1000	0.60
Color Cue	*Hue.*	300	0.48
Color Cue	*CN*	300	0.62
Pixel Level	*OPP. SIFT*	2000	0.80
Pixel Level	*HSV-SIFT*	2000	0.78
SLF	*SIFT+Hue.*	1000+300	0.81
CA	*SIFT+Hue.*	1000+300	0.84
LRFF	*SIFT+Hue.*	1000+300	**0.85**
SLF	*SIFT+Hue.+CN*	1000 + 300 + 300	0.86
CA	*SIFT+Hue.+CN*	1000 + 300 + 300	0.88
LRFF	*SIFT+Hue.+CN*	1000 + 300 + 300	**0.91**
SLF	*SIFT+Hue.+CN+OPP. SIFT*	1000 + 300 +300 + 2000	0.87
LRFF	*SIFT+Hue+CN+OPP. SIFT*	1000 + 300 +300 + 2000	**0.93**

Table 3. Results on the Flower dataset.

5.6. Evaluation of the new kernel on classification

One of the interesting contributions of this paper is the proposed new logistic regression marginalized kernel (**LRMK**). We claim that this kind of kernels is extremely effective and relevant in computer vision. To evaluate the performance of **LRMK**, we compare our final classification results (for a given set of descriptors) with those obtained using: **(i)** the Logistic Regression (**LR**) outputs ($y^* = argmax_{y_i} P(y_i|x)$), **(ii)** a standard late fusion and a multiclass SVM with an intersection kernel (**MSIK**), **(iii)** the sub-kernel $K(\mathbf{x}, \mathbf{x}') = \sum_{y \in Y} p(y|\mathbf{x}, \boldsymbol{\beta_y}) \times p(y|\mathbf{x}', \boldsymbol{\beta_y})$ (**CPK**) corresponding to the first part of Eq. (7). This allows us to estimate the discriminative contribution of the conditional probabilities; **(iv)** the sub-kernel $K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\beta}_{\boldsymbol{y}}^{\boldsymbol{T}} \mathbf{x} \times \boldsymbol{\beta}_{\boldsymbol{y}}^{\boldsymbol{T}} \mathbf{x}'$ (**ICSK**) corresponding to the second part of Eq. (7). This allows us to estimate the discriminative contribution of the image-to-class similarities; **(v)** A multiple kernel learning (**MKL**) method [12, 23] where $K(x, x') = \sum_f \alpha_f K_f(x, x')$, $\alpha_f > 0$, $\sum_f \alpha_f = 1$ and α_f is the weight of the f^{th} cue[5] and $K_f(x, x')$ is the corresponding kernel for the f^{th} cue (here we use intersection kernels again).

From the results of Table 4, we can make the following remarks: first, our marginalized kernel outperforms all the other kernels that confirms that exploiting the information provided by the logistic regression is a good way to improve the discriminative power of the visual words; second, the results confirm that the best behavior is obtained by combining in our kernel not only the conditional probabilities but also the image-to-class similarities. Taken alone, each of these parts leads to poorer performances; finally, our kernel outperforms a multiple kernel method.

Soccer		
Method	Vector Size	Accuracy
LR	700	72
MSIK	700	79
CPK	700	77
ICSK	700	84
MKL	700	80
LRMK	700	**86**
Flower		
LR	1600	78
MSIK	1600	86
CPK	1600	88
ICSK	1600	90
MKL	1600	88
LRMK	1600	**91**

Table 4. Comparison between different kernels.

6. Conclusion and Future Work

In this paper we presented a new approach to fuse multiple cues by adaptively weighting a set of diverse and complementary visual words for a given class using a sparse logistic regression model and a new marginalized kernel. This kernel takes into account not only the learned conditional probabilities of the LR model, but also the image-to-class similarity to define the similarity between 2 images. We compared our method (called **LRFF**) with other feature fusion approaches on three datasets and showed that it outperforms the state-of-the-art methods.

So far, we did not take advantage of the similarity $S(y, y')$ between two classes y and y' by stating that $S(y, y') = 1$ only if $y = y'$. We plan to use metric learning approaches to capture background knowledge from the training data in order to assess the proximity between two different classes. For example, in the *Soccer* dataset, the visual categories $AC - Milan$ and PSV seem more similar than $AC - Milan$ and $Chelsea$, so their similarity score should be higher. Learning automatically this type of prior information could reduce the confusions at the classification step.

Acknowledgement This work is supported by the Pascal 2 Network of Excellence.

[5] We cross-validated all the possible combinations for the two datasets and deduced that the best weights for the soccer dataset are 0.6 and 0.4 for the color and shape respectively, and 0.35 and 0.65 for the flower dataset.

References

[1] A. Abdel-Hakim and A. Farag. Csift: A sift descriptor with color invariant characteristics. In *CVPR*, volume 2, pages 1978–1983, 2006.

[2] E. Aldea, J. Atif, and I. Bloch. Image classification using marginalized kernels for graphs. In *GbRPR'07*, pages 103–113, Berlin, Heidelberg, 2007. Springer-Verlag.

[3] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, pages 1 –8, june 2008.

[4] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via plsa. In *ECCV*, pages 517–530, 2006.

[5] G. Burghouts and J.-M. Geusebroek. Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, 113 (1):48–62, 2009.

[6] 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.

[7] X. Chen, X. Hu, and X. Shen. Spatial weighting for bag-of-visual-words and its application in content-based image retrieval. In *PAKDD '09*, pages 867–874. Springer-Verlag, Berlin, Heidelberg, 2009.

[8] M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, adaboost and bregman distances. *Machine Learning*, 48:253–285, September 2002.

[9] A. Dahl and H. Aanaes. Effective image database search via dimensionality reduction. In *CVPR Workshop*, pages 1–6, 2008.

[10] I. Elsayad, J. Martinet, T. Urruty, and C. Djeraba. A new spatial weighting scheme for bag-of-visual-words. In *CBMI*, pages 1 –6, 2010.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[12] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, pages 221 –228, 29 2009-oct. 2 2009.

[13] A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 42:80–86, February 2000.

[14] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *ICML*, pages 321–328. AAAI Press, 2003.

[15] F. S. Khan, J. van de Weijer, and M. Vanrell. Modulating shape features by color attention for object recognition. *International Journal of Computer Vision*, 2011, to appear.

[16] T. K. Koji Tsuda and K. Asai. Marginalized kernels for biological sequences. *Bioinformatics*, 1:1–8, 2002.

[17] J. Krapac, J. Verbeek, and F. Jurie. Modeling spatial layout with fisher vectors for image categorization. In *CVPR*, Barcelona, Spain, Nov 2011.

[18] J. T. Kwok and P.-M. Cheung. Marginalized multi-instance kernels. In *IJCAI*, pages 901–906, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[19] E. P. X. Li-Jia Li, Hao Su and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, Vancouver, Canada, December 2010.

[20] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150–1157 vol.2. IEEE Computer Society, August 1999.

[21] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman, editors, *Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[22] T. P. Minka. A comparison of numerical optimizers for logistic regression. Technical report, Machine Learning and Perception Group Microsoft Research (Cambridge, UK), 2003.

[23] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics Image Processing*, pages 722 –729, 2008.

[24] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, pages 1 –8, june 2007.

[25] F. Shahbaz Khan, J. van de Weijer, and M. Vanrell. Top-down color attention for object recognition. In *ICCV*, pages 979 –986, 2009.

[26] T. Tuytelaars, M. Fritz, C. Saenko, and T. Darrell. The nbnn kernel. In *ICCV*, 2011.

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

[28] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *ECCV*, volume 3952, pages 334–348. Springer, 2006.

[29] J. van de Weijer and C. Schmid. Applying color names to image description. In *ICIP*, pages 493–496, 2007.

[30] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer New York Inc., New York, NY, USA, 1995.

[31] D. Vigo, F. Khan, J. van de Weijer, and T. Gevers. The impact of color on bag-of-words based object recognition. In *ICPR*, pages 1549 –1553, 2010.

[32] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 31:210–227, 2009.

[33] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metaxas. Automatic image annotation using group sparsity. In *CVPR*, pages 3312 –3319, 2010.