Efficient Supervised Dimensionality Reduction for Image Categorization

Rachid Benmokhtar, Jonathan Delhumeau, Philippe-Henri Gosselin

EFFICIENT SUPERVISED DIMENSIONALITY REDUCTION FOR IMAGE CATEGORIZATION

Rachid Benmokhtar, Jonathan Delhumeau and Philippe-Henri Gosselin

INRIA Rennes, France

ABSTRACT

This paper addresses the problem of large scale image representation for object recognition and classification. Our work deals with the problem of optimizing the classification accuracy and the dimensionality of the image representation. We propose to iteratively select sets of projections from an external dataset, using Bagging and feature selection thanks to SVM normals. Features are selected using weights of SVM normals in orthogonalized sets of projections. The Bagging strategy is employed to improve the results and provide more stable selection. The overall algorithm linearly scales with the size of features, and thus is able to process the large state-of-the-art image representation. Given Spatial Fisher Vectors as input, our method consistently improves the classification accuracy for smaller vector dimensionality, as demonstrated by our results on the popular and challenging PASCAL VOC 2007 benchmark.

Index Terms— Image representation, dimensionality reduction, spatial layout, Fisher vectors, PASCAL VOC dataset

1. INTRODUCTION AND RELATED WORK

With the increasing availability of high quality capture devices and smartphones, the amount of images available in digital format has dramatically increased in the last decade. In parallel, major advances in the field of image analysis and classification have arguably reached the age of maturity. Several commercial applications such as *Google Goggles* now offers automatic interpretation of image content. More specifically, we consider the problem of high-accuracy image classification, where the goal is to automatically assign some textual labels to images. In this context, the state-of-the-art approaches rely on very high-dimensional vectors, up to millions of components per image, which raises both a computational and a storage problem when dealing with large image databases. This paper aims at proposing a dimensionality reduction method to improve the trade-off between the compactness of the representation and its recognition performance.

One of the most popular method for image classification and retrieval is a combination of the so-called Bag-of-Words (BoW) image representation and a strong nonlinear classifier such as a support vectors machines (SVM). BoW characterizes an image by extracting numerous local features, and by subsequently quantizing them into “visual words”. The normalized histogram of visual words serve as the image representation. The performance of the BoW model is reported to be better with large quantizer codebooks [1], at the cost of high complexity: First, quantizing the local features to their nearest visual word is computationally expensive, as it’s complexity scales as the product of the number of visual words, the number of regions and the local feature dimensionality. Second, the learning complexity of nonlinear SVMs ranges from $O(N^2)$ to $O(N^3)$, where N is the number of training images, which becomes impractical for large datasets. In contrast, a linear SVM offers a training cost in $O(N)$, at the cost of inferior performances.

Recently, some accurate encoding techniques suited to linear classifiers have been proposed, such as the Fisher Vector (FV) [2]. FV characterizes an image with the gradient vector of the parameters associated with a pre-defined generative probability model, for instance a Gaussian mixture model in [2]. This method has been evaluated and compared to many other techniques by Chatfield et al. [1] and has shown its superiority for image classification. To take into account the rough geometry of scene, the Spatial Pyramid Matching (SPM) proposed by Lazebnick et al. [3] divides the image into block regions and concatenates all the histograms to form a vector descriptor. Recently, Krapac et al. [4] have proposed an alternative encoding of spatial layout information. Based on the FV principle, the spatial location of the image regions assigned to visual words is included in the probabilistic model, leading to results similar to FV with SPM but with a shorter representation.

In this paper, we will mainly focus on more scalable approaches in the spirit of recent works on compact images representations. Dimensionality reduction can be grouped in various ways: (1) supervised or unsupervised (2) linear or nonlinear. For instance, principal components analysis (PCA) and linear discriminant analysis (LDA) are regarded as the most fundamental tools for extracting features from input data, depending on the availability of the class label. Furthermore, kernel functions can be used to extend these linear techniques to non-linear problems.

In literature, many supervised techniques have been proposed, in particular Recursive Support Vector Machines (RSVM) [5], Margin Maximizing Discriminant Analysis (MMDA) [6] and SVM-based Dimensionality Reduction (SVMDRC) [7]. They all use SVM which is ascribed to the rigorously theoretical basis and strong practical capability. RSVM presents two steps: 1) determine the discriminant direction for separating different classes and 2) generate a new sample set by projecting the samples into a subspace that is orthogonal to the direction calculated. MMDA and SVMDRC project the inputs onto a subspace spanned by a series of normal vectors of orthogonal maximum margin hyperplanes of SVM. The reduced representation is used to feed an Artificial Neural Network in [6], K-Nearest Neighbors in [5], or kernalized SVM in [7]. However, these methods cannot efficiently scale with state-of-the-art image features.

In contrast to those, we propose a scalable supervised feature extraction method, where the main idea is to iteratively select a set of linear projections per category. The paper is organized as follows: Section 2 presents our proposed model including projection candidates, feature selection using Bagging and SVM normals. Section 3 reports the experimental results conducted on the PASCAL VOC 2007 dataset, and compared them to the state-of-the-art methods discussed in Section 1.

2. PROPOSED METHOD

In this section, we present a novel method for dimensionality reduction of image features in the context of binary image classification. Thus, we consider a labelled set of images $(\mathbf{X}, \mathbf{y}) = ((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2) ... (\mathbf{x}_n, \mathbf{y}_n))$, where $y_i \in \{-1, 1\}$ is the label of image $\mathbf{x}_i \in \mathbb{R}^p$.

Our aim is to find a linear mapping $\phi(\mathbf{x})$ that leads to a subspace $\mathbb{R}^d \subset \mathbb{R}^p$ where images are better classified:

$$\phi(\mathbf{x}) = \mathbf{P}^\top \mathbf{x} \tag{1}$$

with $\mathbf{P} \in \mathbb{R}^{p \times d}$. In order to evaluate image classification, we consider hyperplane classifiers in the space induced by the mapping ϕ:

$$f(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + b \tag{2}$$

with $\mathbf{w} \in \mathbb{R}^d$ the normal of the separating hyperplane, and $b \in \mathbb{R}$. It should be noted that we consider linear rather than non-linear mapping techniques and classifiers because of the typically high dimensionality of image features, usually between 10k and 100k, and in some cases more than 300k.

The general problem we address in this paper can be expressed as:

$$\mathrm{argmin}_{\mathbf{P} \in \mathcal{P}} \, \mathrm{argmin}_{\mathbf{w} \in \mathcal{W}(\mathbf{P}^\top \mathbf{X}, \mathbf{y})} \, R(\mathbf{P}^\top \mathbf{X}, \mathbf{y}, \mathbf{w}) \tag{3}$$

where $\mathcal{P}$ is the set of projection candidates, $\mathcal{W}(\mathbf{P}^\top \mathbf{X}, \mathbf{y})$ is the set of satisfactory normals considering classification problem $(\mathbf{P}^\top \mathbf{X}, \mathbf{y})$, and $R(\mathbf{P}^\top \mathbf{X}, \mathbf{y}, \mathbf{w})$ is the classification risk. The last two parameters are related to the choice of the classification trainer. In this paper, we use Support Vector Machines (SVM), whose performance in content-based image categorisation is widely recognised [1].

In order to find a good solution of the problem of Eq. (3) considering the constraints of image classification, we propose a new method which components are described in the following sections.

2.1. Set of Projection Candidates

Solving the problem of Eq. (3) for any projection sets, i.e. $\mathcal{P} = \mathbb{R}^{p \times d}$, can be a very complex task because of the very large dimensionality p of image features. Consequently, we propose to reduce the set of projection candidates to a set $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_s)$ of image features $\mathbf{e}_i \in \mathbb{R}^p$ computed on a random selection of images.

Furthermore, we orthogonalize these features in order to ease the selection in the following step. We perform this procedure using a QR algorithm: $\mathbf{E} = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{R}^{p \times s'}$ and is orthogonal, $\mathbf{R} \in \mathbb{R}^{s' \times s}$ and $s' \leq s$ since $p > s$, i.e. the dimension of features p is larger than the size s of training sets. The vectors $\mathbf{q}_i$ of matrix $\mathbf{Q}$ are then considered as projection candidates.

2.2. Projection selection with linear SVM

In order to select a relevant subset of projection from a set of orthogonalized projection candidates $\mathbf{Q}$, we propose to use a technique based on the values of the normal of a separating hyperplane.

We first train a SVM using the training set $(\mathbf{Q}^\top \mathbf{X}, \mathbf{y})$ in the space induced by a linear mapping with $\mathbf{Q}$. The resulting decision function is :

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{Q}^\top \mathbf{x} = \sum_{i=1}^{s'} w_i \mathbf{q}_i^\top \mathbf{x} \tag{4}$$

Then, we select the vectors $\mathbf{q}_i$ of matrix $\mathbf{Q}$ for which the corresponding weights $|w_i|$ are the highest, and store this projectors into a matrix $\mathbf{Q}^\star$. This is a usual feature selection technique whose relevance is shown for ℓ_2 norm [8]. Let us note that we first focused on ℓ_2 norm SVM trainers because of the large size of visual features. However, norms like ℓ_1 that lead to more sparse normals could be more effective, as long as computational complexity is not an issue.

2.3. Stable selection with Bagging

In order to compute more stable projection selection, we propose to follow a Bagging strategy, also known as “bootstrap aggregation” [9]. This strategy can improve the performance of various individual models due to the reduced variance of the bagged model.

First we randomly draw several subsets $\mathbf{Q}^c = \mathbf{Q}(I^c)$ where I^c is the subset of indices. Then we train a SVM using the training set for the space induced by each subset $\mathbf{Q}^c$. As a result, we obtain a set of hyperplane normals $\mathbf{w}^c$, each one being expressed in the space induced by $\mathbf{Q}^c$. Note that each projector $\mathbf{q}_i$ may only appears in subset $\mathbf{Q}^c$ of indices $C_i = \{c|i \in I^c\}$. Finally, for each projector $\mathbf{q}_i$ we combine values from all normals $\mathbf{w}^c$ that were train in a subspace $\mathbf{Q}^c$ that contains $\mathbf{q}_i$:

$$w_i = \frac{1}{|C_i|} \sum_{c \in C_i} w_j^c \quad \text{with } j = I^c(i) \tag{5}$$

The selection of projection set $\mathbf{Q}^\star$ is the same as before: we select the projectors $\mathbf{q}_i$ whith the largest $|w_i|$. In the following experiments for image classification, we observed that using 5 subsets $\mathbf{Q}^c$, each $2/3$ the size of $\mathbf{Q}$, is appropriate for stable selections.

2.4. Iterative selection of projection sets

The previous components we presented do not necessarily lead to an improvement in image classification. Moreover, the size of the random sets $\mathbf{E}$ of image features is limited by computational complexity, mainly because of the QR decomposition. In order to handle these problems, we propose to perform an iterative selection of projection sets.

We first select a projection set $\mathbf{Q}_1^\star$ from a random set $\mathbf{E}_1$ of image features using the techniques presented in the previous section. Then, we proceed iteratively, and select for each round t a projection set $\mathbf{Q}_t^\star$ from a random set $\mathbf{E}_t$ of image features. In the case where the combination of all projections $\bigcup_t \mathbf{Q}_t^\star$ leads to better classification performance on a validation set, we proceed to the next iteration. Otherwise, another random set of images features is drawn, until improvement is observed. It should be noted note that in our experiments, all random sets of image features leaded to an increase of performance.

This iterative selection algorithm can be compared to a basic Boosting algorithm. As a result, it is likely that further improvement could be achieved using more advanced Boosting techniques.

3. EXPERIMENTATIONS

In this section, we evaluate the effectiveness of the proposed method on the challenging *PASCAL VOC 2007* image classification dataset. This dataset contains about 10,000 images split into *train*, *validation* and *test* sets, of 20 object categories (*cf.* Fig. 1). A 1-versus-all SVM classifier is trained for each category and result is evaluated using the official protocol of PASCAL VOC, in terms of average precision (AP). Overall performance is measured as mean average precision (mAP).

All our experiments are based on the same type of feature: SIFT descriptors computed on a dense grid. More precisely, we adopt the same parameters as used in [1, 2, 10], i.e. a spatial stride of 3 pixels with multiple resolutions. SIFT descriptor dimensionality is reduced from 128 to 80, as it is done in [2]. Spatial Fisher Vectors (SFV) encoding is then computed as it offers good performance with linear classifiers. An additional reason for this choice is that SPM-based representations suffer scalability issues in real-world applications with the Spatial Coordinate Coding strategy of SFV avoids (More details about the SFV encoding can be found in [4]). We follow with *power-law* and ℓ_2 normalizations of the SFV. The projection candidates are randomly drawn from Flickr[1].

Fig. 1. Images from VOC 2007 database.

Figure 2 shows the mAP performances for different reduced feature size. Each curve presents the performance for a specific number of selected projections per iteration. For instance, the blue curve selects 500 projections per iteration. For all curves, and for a specific iteration, projections are always selected from the same set of 15k images randomly drawn from Flickr. In these experiments, performance increases with the number of iterations and the number of selected projections. We can also observe that for a fixed number of selected projections, better performance is achieved when more projections are selected per iteration. For instance, a single iteration selecting 8k projections is more effective than two iterations selecting 4k projections. We have also evaluated the improvement gained from orthogonalization of projection candidate sets, and observe a gain of around 5% of mAP.

Table 1 compares our method with state-of-the-art. With similar parameters, we notice that to reach a mAP of 60% or more, the feature size of the image representation have to be larger than $40k$ for Spatial Fisher Vectors (SFV*) and larger

[1] http://www.flickr.com

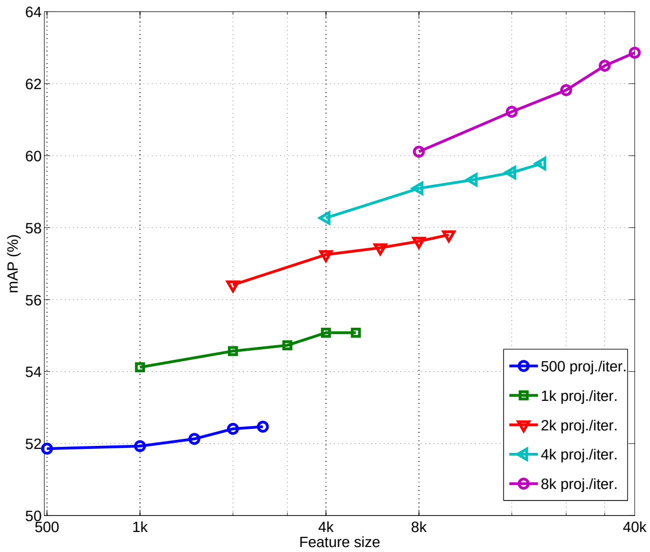

Fig. 2. mAP performances for different reduced feature sizes.

than $320k$ for SPM-based techniques. In contrast, our method achieves a mAP higher than 60% for a feature size of $8k$, and the overall best results of 62.86% for a feature size of $40k$. To our knowledge, our method is the only one with 8 times less dimension than standard FV+SPM and which outperforms the best methods on PASCAL VOC 2007 dataset.

4. CONCLUSION

In this paper, we introduced an efficient method for supervised dimensionality reduction for image categorization. Our method iteratively selects sets of projections from an external dataset, using Bagging and feature selection thanks to SVM normals. The mapping of image to classify using these projections leads to a smaller representation of images while achieving good performances. Furthermore, experiments showed that better performance can be achieved when selecting large number of projections per iteration. On PASCAL VOC 2007, we reported state-of-the-art results only using SIFT features and linear classifiers. This makes our system scalable to large image datasets.

5. REFERENCES

[1] Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman, “The devil is in the details: An evaluation of recent feature encoding methods,” in *BMVC*, 2011, pp. 76.1–76.12.

[2] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, “Improving the fisher kernel for large-scale image classification,” in *ECCV*, 2010, pp. 143–156.

Methods	**Grids**	**Feature size**	**mAP** (%)
Std. FV [2]	-	≈ 41k	55.30
SFV [4]	SCC	≈ 42k	55.50
SV [1]	SPM	-	58.16
HE [10]	SPM	≈ 16k	58.34
SFV*	SCC	≈ 42k	60.43
FV [2]	SPM	≈ 329k	58.30
FV [1]	SPM	≈ 329k	61.69
HE+FV [10]	SPM	≈ 345k	62.78
Our method **based on** **reduced SFV***	SCC	**4k** **8k** **16k** **40k**	**58.27** **60.11** **61.22** **62.86**

Table 1. Comparison of the proposed method with the state-of-the-art image representation on PASCAL VOC 2007. SFV*: Improved implementation of SFV; SCC: Spatial Coordinate Coding; SPM: Spatial Pyramid Matching.

[3] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *CVPR*, 2006, pp. 2169–2178.

[4] Josip Krapac, Jakob Verbeek, and Frédéric Jurie, “Modeling spatial layout with Fisher vectors for image categorization,” in *ICCV*, Nov 2011.

[5] Q. Tao, D. Chu, and J. Wang, “Recursive support vector machines for dimensionality reduction,” *IEEE Transactions on Neural Networks*, pp. 189–193, 2008.

[6] András Kocsor, Kornél Kovács, and Csaba Szepesvári, “Margin maximizing discriminant analysis,” in *ECML*, 2004, pp. 227–238.

[7] Bo Yang, “SVM-induced dimensionality reduction and classification,” in *ICICTA*, 2009, pp. 275–278.

[8] Janez Brank, Marko Grobelnik, Nataa Milic’-Frayling, and Dunja Mladenic’, “Feature selection using linear support vector machines,” Tech. Rep., Microsoft Research, 2002.

[9] Leo Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[10] Mihir Jain, Rachid Benmokhtar, Hervé Jégou, and Patrick Gros, “Hamming embedding similarity-based image classification,” in *ICMR*, 2012, pp. 1–8.