Neurocomputing 167 (2015) 209–217

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Higher-level feature combination via multiple kernel learning for image classification

Wei Luo*, Jian Yang, Wei Xu, Jun Li, Jian Zhang

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, PR China

ARTICLE INFO

Article history:
Received 14 October 2014
Received in revised form
5 April 2015
Accepted 28 April 2015
Communicated by Xiaofei He.
Available online 16 May 2015

Keywords:
Feature coding
Image classification
Feature combination
Multiple kernel learning

ABSTRACT

Feature combination is an effective way for image classification. Most of the work in this line mainly considers feature combination based on different low-level image descriptors, while ignoring the complementary property of different higher-level image features derived from the same type of low-level descriptor. In this paper, we explore the complementary property of different image features generated from one single type of low-level descriptor for image classification. Specifically, we propose a soft salient coding (SSaC) method, which overcomes the information suppression problem in the original salient coding (SaC) method. We analyse the physical meaning of the SSaC feature and the other two types of image features in the framework of Spatial Pyramid Matching (SPM), and propose using multiple kernel learning (MKL) to combine these features for classification tasks. Experiments on three image databases (Caltech-101, UIUC 8-Sports and 15-Scenes) not only verify the effectiveness of the proposed MKL combination method, but also reveal that collaboration is more important than selection for classification when limited types of image features are employed.

1. Introduction

Image classification has gained much attention in recent years. Due to large variations, e.g. illumination, occlusion, intraclass variation and so on, posed in images, it becomes a nontrivial task to be solved. Recent years, with the spatial pyramid matching (SPM) [1] becoming the *de facto* standard for image classification, much work has been contributed to this topic [2–7]. SPM partitions an image into increasingly fine subregions and computes statistical features inside each subregion. Specifically, it includes four steps to generate an image feature: (1) Densely extracting descriptors from an image, e.g. SIFT. (2) Encoding descriptors using a supervised/unsupervised learned dictionary. (3) Pooling and concatenating the codes in each subregion. (4) Training linear/nonlinear SVMs for classification. The step (2) is the most essential stage for its critical role to transform low-level descriptors to higher-level image codes. Different encoding strategies employed in this stage will directly affect the following pooling features, and thus result in different image features.

Essentially, many kinds of descriptors can be employed in the SPM framework to generate correspondingly different image features. Therefore, a straightforward strategy to improve the classification performance is combining different image features to form a more power one. Intuitively, hard combination like concatenation may degenerate the efficiency and performance, since different features may have different scales and hard combination may cause high feature redundancy. Recent studies on multiple kernel learning (MKL) [8] have revealed that combining different features through kernels can effectively improve the classification performance, and the combining coefficients can be adaptively determined to reflect the importance of different features for different classes. Especially, ℓ_1 norm is employed to constrain sparse combinations [9,8]. The combination methods vary from linear to nonlinear, and from the same type of kernel to different types of kernels [8–11].

In the current MKL studies, feature combination mainly focuses on features derived from different types of low-level descriptors. For image classification tasks, different descriptors can capture different properties of images and preserve different degrees of discriminative power and invariance, such as PHOG [12] captures shape information while SIFT [13] captures appearance information. Therefore, combining these features through MKL can make the final feature incorporates more information for classification than only using one single type of feature, and consequently resulting in higher performance.

In this paper, we propose using MKL to combine different image features derived from the same type of low-level descriptor for image classification (see Fig. 1). Our motivation is based on the encoding and the pooling stages in SPM, where different encoding and pooling strategies can be seemed to capture different properties of an image. For example, the hard voting with average pooling can be considered as computing the frequency of each visual codeword in an image, while the salient coding with max pooling can be seemed to reflect

* Corresponding author.
E-mail address: cswluo@gmail.com (W. Luo).

http://dx.doi.org/10.1016/j.neucom.2015.04.075
0925-2312/© 2015 Elsevier B.V. All rights reserved.

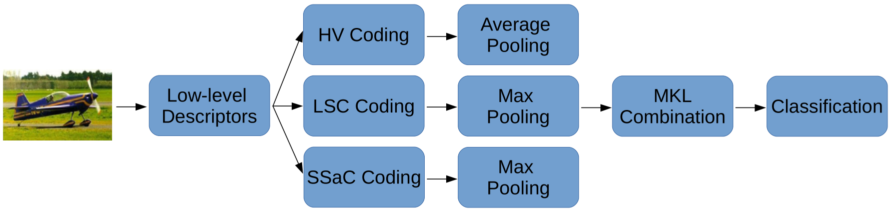

Fig. 1. Schematic diagram of the proposed MKL feature combination for classification.

the salient degree of each visual codeword in an image. Based on these observations, we analyse the physical meaning of each component in image features, and empirically demonstrate that these different image features derived from different encoding and pooling strategies can be compatibly combined together through MKL to improve the image classification performance. We consider three representative encoding methods and two common pooling strategies in our work, namely hard voting (HV) [1], localized soft-assignment coding (LSC) [2], salient coding (SaC) [14] and correspondingly average and max pooling. SaC is essentially a hard voting scheme, thus much of information contained in an image may be suppressed when max pooling is employed to pool features. To this end, we propose a soft SaC (SSaC) to alleviate this situation and empirically validate its effectiveness in this paper. To combine these features, we utilize MKL to adaptively learn the combining coefficients, and further analyse the performance of the MKL combined features with different regularizers. Specifically, we use ℓ_1 and ℓ_2 norm to regularize MKL in this paper, respectively. Our experiments reveal that the performance of MKL with different regularizers is sensitive to the number of training data when the types of image features are limited. And in this case, the performance of MKL with ℓ_2 regularizer always outperforms it with ℓ_1 regularizer, and also better than the performances of all its individual coding methods. This signifies that different image features capture different properties of an image and they can complement each other and collaborate to achieve a better performance. Further, it indicates that collaboration is more important than selection when limited types of image features are used. Specifically, we make the following three concrete contributions:

1. We propose SSaC method to alleviate the suppression of information problem in the original SaC method, thus much more information of an image can be exploited for classification. We further empirically compare the performance of SSaC with SaC and a group-code size based SaC (GSaC) to verify its effectiveness.
2. We analyse the meaning of codes generated through different encoding and pooling methods, and study what properties of an image are reflected from these different codes. Essentially, different encoding and pooling methods will reflect different properties of an image.
3. We combine different image features using kernel methods for classification. Specifically, we employ MKL to adaptively combine image features for different classes. We further analyse the influence of the performance when MKL is regularized with ℓ_1 and ℓ_2 norm, respectively. Experiments on three image datasets are implemented to verify the effectiveness of the proposed method.

The reminder of this paper is organized as follows: Related encoding, pooling and MKL methods are introduced in Section 2. Then in Section 3 we present the proposed softened salient coding. The properties of different image features derived from different encoding and pooling methods are then detailed in Section 4, followed by the combination strategy through MKL. In Section 5, we first evaluate the performance of SSaC and then present the experimental results in three image datasets and analyse the performance in detail. Finally, we conclude this paper in Section 6.

2. Related work

In the framework of SPM [1], a large amount of work contributes to the encoding step. For an input **x** and a given dictionary **D**, the hard-voting (HV) originally employed in SPM assigns 1 to the basis which is the nearest neighbor of **x**. The authors in [5] relaxed this constraint to assign each basis a value based on a Gaussian-shape kernel. However, this strategy increases the computational cost. To alleviate this problem, the localized soft-assignment coding (LSC) [2] was proposed to encode **x** by only considering its k-nearest bases in **D**. To make the codes preserve reconstruction ability, the authors in [3] leveraged the sparse coding (ScSPM) technique to encode **x**. Locality-constrained linear coding (LLC) [4] moves forward by further considering the local smoothness of codes. The authors in [15] extended LLC to considering the global smoothness of codes. Although these reconstruction based methods work well, the physical meaning of the code is not straightforward. In contrast, salient coding (SaC) [14] was proposed encoding **x** like HV but with the code reflecting its salient degree to **x**. All the aforementioned coding methods encode inputs independently, which means they encode one input per time. Recently, the authors in [16,17] proposed to encode a group of inputs per time by exploiting the spatial structure information. We here mainly focus on the single coding method, because we mainly want to study what properties of an image captured by different coding methods and whether they can be complemented each other to obtain a better performance while not developing a new coding method. To this end, we select HV, LSC and SaC as three representative coding methods in this study, because they each correspond to assignment-based, locality-constrained and salient-based coding methods, respectively.

In the framework of MKL, originally proposed in [18], different kernels are used to construct corresponding kernel maps for different feature descriptors. These different descriptors capture different properties of an image, and the weights for corresponding kernels can thus be adaptively learned through MKL for a specific task. In order to learn an appropriate kernel combination, various regularizers have been introduced for MKL, e.g. ℓ_1 norm [8] and ℓ_p norm ($p>1$) [19]. The objective function of MKL is usually formulated as that of SVMs [8–10,20], thus the corresponding optimization procedure usually involves a step of gradient descent to update the combining coefficients and a step of optimizing the SVM parameters. In [10], the authors proposed LPBoost approach to learn individual parameter sets $\{\alpha_m, b_m\}$, the Lagrangian multipliers and bias, for each SVM, thus each individual SVM can be trained to yield maximal generalization. In [8], the authors advocated all SVMs sharing parameters, but put the ℓ_1 norm regularization on the combination coefficient in the objective function, which would discover a minimal set of invariances and thus prevent overfitting if many base kernels are used. While the aforementioned MKL is constrained to linear combination of the same type of kernel, the authors in [9] extended it to linear and nonlinear combination of different types of kernels, which is especially useful for the feature combination problem faced