Product Images Classification with Multiple Features Combination

Shi-jie Jia[1] Xiang-wei Kong[2] Hai-yan Fu[2] Guang Jin[2]

[1]Faculty of Electronic Information & Electrical Engineering, Dalian University of Technology, Dalian 116023, China
Email: jsj@djtu.edu.cn
[2]College of Electrical & Information, Dalian Jiaotong University, Dalian 116028, China

Abstract

Visual-based automatic product image classification is a great need and challenge work for e-commerce field. Previous work tested few number product categories with one or two descriptors. For the task of product classification over large number categories, we employed kernel-based SVM classifier combining multiple features, including one global descriptor and three complimentary local descriptors. Furthermore, we investigate four ways to combine discriminative features for SVM classifier. Experiments on the product image dataset (PI 100) showed the performance improved significantly by features fusion.

Keywords: Product image classification, Multiple Features, SVM, PI 100

1. Introduction

We investigate the problem of visual-based automatic product image classification for e-commerce, which has potential use in product tagging[1] and tagging-based image searching [2]. Just as the saying goes, "One picture is worth a thousand words", product tagging with human is not only labor-intensive but also hard to be accurate and complete [11]. Consequently, it is a great need of e-commerce development to achieve visual-based automatic product image classification.

In this paper, we aim for categorizing images automatically by the product types (i.e., such as piano or guitar) or by some subtle distinctions (such as pointynesss or non- pointyness of toes). It is one specific application of content-based image classification, most methods of which mainly use supervised learning methods, combining underlying feature-based modeling with intermediate semantic analysis to achieve classification results [3-7]. Although our tasks are often with tamer pictures than natural images (where the object dominates the image and there is little background clutter), it is still a challenging problem for the large number of categories and intra-class variations. Ref. [1] explored the feasibility of tagging products through supervised image classification. They achieved accuracies between 66% and 98% on 2- and 3-classes classification tasks. Ref. [11] proposed a fast supervised image classifier which is based on class-specific descriptor and Image-to-Class distance and achieved 84% for 30 product classes. However, such a small number they tested was far from real application. In fact, for a large number of classes, single descriptor cannot be optimal in all situations to alleviate the effect of intra class variations. Moreover, the importance of different types of features varies with from task to task. It may require a proper combination of relevant

features to yield a good representation. Bosch etc [3] made state-of-the-art performance on Caltech101 object classification test with combining linearly two local descriptors (PHOW and PHOG) and spatial pyramid kernels. As an extension of the pioneer work, we employ one global descriptor (GIST [4]) with three complimentary local descriptors (shape descriptor PHOG [5], appearance descriptor PHOW[6], texture descriptor P-LBP[7]) as the image representation set. Furthermore, we investigate the ways to combine discriminative features with diverse type of kernels for Support Vector Machine (SVM) classifier. Kernels define (possibly nonlinear) similarities between data points and allow abstracting learning algorithms from data, and different kernels generate different structures in the embedding space. We intend to construct a number of kernel matrices corresponding to each specific type of image feature. Consequently, combining features is equivalent to fusing kernel matrices. We tested four methods to combine the kernels for boosting the performance (see Part 3). The algorithms are tested on publicly available product datasets PI 100 and exhibit good performance.

The rest of our paper is organized as follows. Section 2 describes the image representation with multiple features (GIST, PHOG, PHOW, PLBP), while Section 3 gives details of SVM Classifier with the multiple kernels combination. Experiment setup and typical results are described in Section 4. The final part concludes with suggestions for future research.

2. Image Representation

None of the feature descriptors have the same discriminative power for all classes. For example, shape may be a good feature to distinguish between cars and airplanes but it is not good to distinguish between horses and zebras [3]. It is better to adaptively combine a set of diverse and complementary features (such as the global and local, appearance, shape and texture) to discriminate each class from all other classes.

2.1. GIST (Global descriptor)

Humans can recognize the gist of a novel image in a single glance as they use global scene factors before analyzing the image in detail. The Global descriptor GIST [4] consists of a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that represent the dominant spatial structure of an image. These dimensions may be reliably estimated using spectral and coarsely localized information. To compute the color GIST description the image is segmented by a 4 by 4 grid for which orientation histograms are extracted. Our implementation takes as input a square image of fixed size and produces a vector of dimension 960.

2.2. PHOW (Local appearance descriptor)

The descriptor consists of visual words computed on a dense grid, which will be referred to as appearance. It works by partitioning the images into increasingly fine sub-regions and computing histograms of local features found inside each sub-region. The local features are extracted with dense sampling and represented with SIFT descriptor. To cope with empty patches, we zero all SIFT descriptors with L2 norm below a threshold (200). The K-means clustering is performed over 10 training images per category selected at random; we thereafter created a 200-elements vocabulary. Therefore, an image is represented as a Pyramid Histogram Of Words (PHOW) descriptor. In our experiments, the sampling interval was set to 8 pixels, each 16×16 pixel block formed a 128-

dimensional SIFT feature vector. The optimal setting of pyramid level L is 3, which followed Ref. [6]. The PHOW is normalized to sum to unity taking into account all the pyramid levels.

2.3. PHOG (Local shape descriptor)

Histogram of Orientated Gradients (HOG) [8] is a useful shape descriptor which is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid. HOG has taken the spatial distribution of the image into account. However, it is ignored that the combination at different spatial scales has a space effect on the performance of retrieval and classification. In view of this, Bosch etc [5] proposed the PHOG descriptor which is captured by titling the image into regions at multiple resolutions and consists of a histogram of orientation gradients over each image sub-region at each resolution level. Followed Ref [10], the number of bins K was set to 40, and the pyramid levels was set to 3 (in fact, the performance was found not to be very sensitive to the value of K). The PHOG is normalized to sum to unity, which ensures that the images with more edges are not weighted more strongly than others [10].

2.4. PLBP (Local texture descriptor)

Local Binary Pattern (LBP) is a simple yet very efficient texture operator which labels the pixels of an image by thresholding the neighborhood of each pixel with the value of the center pixel and considers the result as a binary number [7]. The original LBP operator forms labels for the image pixels by thresholding the 3 ×3 neighborhood of each pixel with the center value and the values of the pixels in the thresholded neighborhood are multiplied by the binomial weights given to the corresponding pixels. Finally, the values of the eight pixels are summed to obtain the LBP number for this neighborhood. The 256-bin LBP histogram computed over a region is used for texture description. LBP is invariant against any monotonic gray scale transformation and has characters of computational simplicity. However, the original LBP descriptor hasn't taken the spatial distribution into account. To overcome the challenge, we borrowed the idea of pyramid representation from PHOW [6] and PHOG [5] to build the pyramid spatial LBP (PLBP). In view of the trade-off between the effective and compute efficiency, the pyramid level was set to 2 in our experiments.

3. SVM classifier with multiple kernels

For a given test image the learned classifier has to decide which class the image belongs to. Support vector machines (SVMs) are considered a good candidate from many learning choices (decision trees, neural networks etc.) because of its high generation performance without the need of a-prior knowledge. In support vector machines (SVMs), the data representation is implicitly chosen through the so-called kernel $K(x_i; x_j)$, which implicitly maps examples x to a feature space given by a feature map $\Phi(x)$. via K (xi; xj) $= \Phi(x_i)*\Phi(x_j)$. This kernel defines the similarity between two examples xi and xj. Through the 'kernel trick', classfiers can be learnt and applied without explicitly computing $\Phi(x)$. The only requirement is that K must be a positive definite (PD) function, which guarantees the existence of a feature map.

The choice of the kernel largely influences the performance of the algorithm, and it is imperative to choose a suitable kernel for a given learning task. The single kernel types we tested are as follows:

(1) Linear kernel

$$k(h,h') = h^T h'; \quad (1)$$

(2) Gaussian kernel

$$k(h,h') = \exp(-\| h-h'\|^2 / 2\delta^2) \quad (\text{For } \delta > 0); \quad (2)$$

(3) Chi-square kernel

$$k(h,h') = \exp(-\gamma\chi^2(h,h')/d) \quad (3)$$

$$\chi^2(h,h') = \sum_k -\frac{(h_k - h_k')^2}{h_k + h_k'}$$

(3) Intersection kernel

$$k(h,h') = \sum_k \min(h_k, h_k') \quad (4)$$

Where h_k, h_k' are two bins of feature distribution function (i.e. histogram) h and h', the band-width d is set to the mean chi-square distance between all pairs of training samples. In total 4 features are used, each with pyramid levels except global descriptor GIST. Therein, nine feature kernels are obtained with four kernel types.

Four modes are implemented to combine the kernels for boosting the performance.

(1) Global kernel selection (GKS). For each specific task, all the classes have the same weights, the best (global) kernel is automatically identified by the fivefold cross-validation.

(2) Global kernel linear combination (GKLC) .The combining kernel is set as follows.

$$k(h,h') = \sum_{fl} d_{fl} K(h_{fl}, h_{fl}') \quad (5)$$

Where d_{fl} is the weight of the specific kernel corresponding to the feature f and its pyramid level l. In this case, all the classes have the same weights for a particular kernel. This means that we only need learn L (the number of kernels) parameters regardless of the (large) number of categories. Nevertheless, hand-tuning kernel parameters can be difficult. Multiple Kernel Learning [9] (MKL) seeks to address this issue by learning the linear combination kernel from training data. We use the open MKL software (SimpleMKL Toolbox, http:// asi.insa-rouen.fr/enseignants/ ~arako to m/code/mklindex.html) to learn an optimal combination of kernels, each of which captures a different feature channel.

(3) Class-specific kernel selection (CSKS). In this mode, the 'best kernel' varies from class to class. Instead of searching a global best kernel, we select specific optimum kernel for each specific class by the fivefold cross validation..

(4) Class-specific kernel linear combination (CSKL).

$$k^c(h,h') = \sum_{fl} d_{fl}^c K(h_{fl}^c, h'^c_{fl}) \quad (6)$$

Different categories will require the different combination of descriptors to distinguish them. Instead of learning weights common across all classes, the weights are learnt for each class separately to optimize classification performance for that class. It is necessary to learn N (the number of categories) times parameter values than the global combination mode. We also use SimpleMKL Toolbox to get the optimum kernel combination for each specific class. For above all, we resolve multiclass decisions by using one-versus-one classifiers.

4. Experiments

4.1. Image data set

We tested our algorithms on Microsoft research's product image categorization data set (PI 100), which was collected from the MSN shopping web site http:// shopping. msn. com /). PI 100 contains ten thousands low resolution (100×100) images in 100 categories, each image contains a single object or one dominant object in relatively stable forms, just as most product images appear on the Web. Table 1 shows some sample images from PI 100.The experiments were performed on an Intel Pentium CPU 2.66GHz computer running Windows XP and MATLAB7.1 with 1GB RAM. The LibSVM implementation and the SimpleMKL Toolbox are used to train the classifier. In this experiment, we split the total images of each category into two sets: 10% as the query set, 15%, 30% as the training set, respectively. All experiments are repeated ten times with different randomly selected training and testing images, and the average of per-class recognition rate is recorded for each run.

4.2. Results & Discussion

1) The performance of each descriptor with single kernel.

Firstly, we tested the SVM classifier with each feature at each pyramid levels. The kernel types we computed include linear kernel, RBF kernel, intersection kernel and chi-square kernel. A summary of test results is listed in Tab.2.

As shown in Table 2, we arrived at the following review.(i) compared to the histogram intersection kernel and chi-square kernel, linear kernel and RBF kernel performed much worse, which suggested that the two kernels are not fitted for the histogram-based image feature representation. This also indicates that the performance of histogram representation is sensitive to choice of kernel metrics and relies heavily on the classifier. That is, the superior performance of chi-square kernels comes from the specific nature of the histogram representation. (ii)The overall accuracies grow as the pyramid level increase. In general, the spatial matching plays an important role in boosting product classification performance. However, it is not exactly for each specific class. For the categories with much inner-variation (e.g. the Cellphones), high pyramid level matching could degrade the performance [3] . (iii) For the overall performance, the SIFT based appearance descriptor PHOW perform better slightly than three other descriptors (yet not exactly for each specific image).

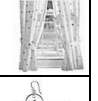
helmet				
Crib				
Curtain				
erring				
flower				
jacket				

Table 1: some samples of product image set (PI 100).

2) Combination of multiple features

We tested four combination methods as mentioned in part Ⅲ, in which chi-square kernel[1] is selected for its comparative superior performance. A summary of test results is listed in Tab.3.

[1] we also tested with combination of several different kernel types, but the performance of combination is inferior slightly to that of the best one.

		Linear	RBF	Intersec-tion	Chi-square
PHOG	L0	36	32.4	41.4	47
	L1	60.6	49.8	69	71.6
	L2	69	53.2	71	77.4
PHOW	L0	39.6	35.2	44.4	45.8
	L1	61.8	53	70.8	72.4
	L2	71	64.4	76.2	78.1
PLBP	L0	41.4	34.4	60.2	62.4
	L1	52.4	46.2	72.6	74.4
GIST		71.4	64.4	75.8	76.5

Table 2: Performance (%) of SVM classifier with single kernel (for RBF kernel: g=0.07,C=1000).

	GKS	GKLC	CSKS	CSKL
PHOG	77.4	78.8	77.9	78.4
PHOW	78.1	79	79.1	80.8
PLBP	74.4	75.9	76.2	76.8
GIST	76.5	77.4	78.2	78.6
Com-biantion	78.1	80.2	79.1	83.7

Table 3: Result of four combining methods(%).

Table 3 illustrates the class-specific kernel methods boost the performance apparently by comparison with the two global ways, while the performance of kernel selection is slightly inferior to that of the kernel linear combination. The advantage of learning class-specific feature-weights is that classes have the freedom to adapt if there is more or less intra-class spatial variation. The disadvantage is that the solutions need much more computation work.

Compared with Ref.[11] which adopted single descriptor and obtained 84% accuracy on just 30 product classes, our algorithm achieved similar performance on much more categories through features combination .

5. Conclusion

The first contribution of the paper is the image representation combining both local and global as well as appearance, shape and texture based features, thereby a versatile product representation is derived. Moreover, a series of experiments have shown the properties of each single kernel type and their combining approach which yielded competitive performance in the product image classification. As no single feature is sufficient for handling diverse intra variation among broad categories, it is a long way to design more discriminative robust visual features and effective ways of fusing various complementary informative kernels. On the other hand, it is a potential effective paradigm to turn some pre-processing choices (such as the codebook size and the spatial kernel set) into kernel parameters [10]. And last but not least, for resolving product classifications over a large number of categories, there is of necessity to construct the framework matching human visual system and embed rich a-prior knowledge through various ways.

Acknowledgement

This work was supported by the National Science Foundation of Major Funded Project (No. 70890083).

References

[1] Tomasik, B., P. Thiha, and D. Turnbull, Tagging products using image classification, in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 2009, ACM: Boston, MA, USA.

[2] Lin, X., et al. Visual search engine for product images.Proc. SPIE, Vol. 6820, 68200M (2008);
[3] Bosch, A., A. Zisserman, and X. Munoz, Image classification using ROIs and multiple kernel learning. International Journal of Computer Vision, 2008.
[4] Oliva, A. and A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. International Journal of Computer Vision, 2001. 42(3): p. 145-175.
[5] Bosch, A., A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. in CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval. 2007: ACM Press.
[6] Lazebnik, S., C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. 2006.
[7] Ojala, T., M. Pietikainen, and T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence. IEEE Transactions on, 2002. 24(7): p. 971-987.
[8] Dalal, N. and B. Triggs. Histograms of Oriented Gradients for Human Detection. in CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1. 2005: IEEE Computer Society.
[9] Rakotomamonjy, A., et al., SimpleMKL. Journal of Machine Learning Research, 2008. 9: p. 2491-2521.
[10] Gehler, P.V. and S. Nowozin. Let the kernel figure it out; Principled learning of pre-processing for kernel classifiers. in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. 2009.
[11] Jia Shi-jie, Kong Xiang-wei, Jin Gua. Automatic fast classification of product images with class-specific descriptor. Journal of electronic in China,2010, in press.