F-divergence based Local Contrastive Descriptor for Image Classification

Sheng Guo[1,2] Weilin Huang[1,3] Chunjing Xu[4] Yu Qiao[1,3]
Shenzhen Key Lab of Comp. Vision and Patt. Recog.,Shenzhen Institutes of Advanced Technology, CAS, China[1]
University of Chinese Academy of Sciences,Beijing,China[2]
The Chinese University of Hong Kong, Hong Kong[3]
Huawei Technologies[4]
{*sheng.guo, wl.huang*}*@siat.ac.cn xuchunjing@huawei.com yu.qiao@siat.ac.cn*

***Abstract*— Recent studies showed that f-divergence based features have achieved great successes in speech recognition, synthesis and dialect classification. This paper proposes a novel local contrastive descriptor for image classification based on the f-divergence, referred as LCD. It extracts local image feature by computing the contrastive characteristic between image patches. Each patch is described as a discrete probability distribution of its properties (e.g. histogram of the intensity or gradient), and the contrast is measured by computing the f-divergence between different distributions. Then we build the bag-of-visual-words (BoVW) model based on the designed LCD and applied it for the task of image classification. We evaluated the proposed descriptor on the widely-used PASCAL VOC2007 benchmark dataset, and experimental results demonstrate that the LCD can work effectively and practically with reasonable accuracy achieved. In addition, we also showed that the LCD, encoding the local color information, can be used to compensate for the gradient-based features (e.g. SIFT) efficiently, with moderate improvements gained.**

***Index Terms*— f-divergence, sift, contrast characteristic, bags-of-visual-words.**

I. Introduction

Image feature extraction plays a key role in the task of content based image retrieval and analysis, such as object recognition and image classification. It has long been a crucial yet challenge issue for the computer vision research. A large amount of efforts have been devoted to developing a simple yet powerful local descriptor for image representation. Recently, a number of representative descriptors have been proposed, such as the scale invariant feature transform(SIFT) [1], local binary pattern(LBP) [2], [3], histogram of orientated gradient(HOG) [4] etc. They have been widely applied for various vision applications with great success such as image matching [1], object detection [4], [5], object recognition [6], [7], texture recognition [7], image retrieval [8].

Generally, local feature extraction includes two steps: detection of local interest points and the description of them [1]–[9]. Local interest points are often detected by computing the corners, briquettes, or entropy significant characteristics to extract the local feature information, e.g. textural structures, sharps or spatial distributions. Then, each detected point is described by a feature vector to represent the structure of local region or patch. Histogram based descriptor is often adopted to describe the occurrence of local patterns to obtain the spatial invariance and robustness. A comprehensive review and comparison of various local features is presented in [7], whose results demonstrate that the SIFT is the best choice in many tasks implemented.

The main objective of this paper is to develop a new descriptor to capture the local difference between image patches, which could effectively resemble the inherent structures of the image, while retaining computational simplicity. It has been proved that the f-divergence is the general form of various distance measures [10], [11], which enables it for efficiently measuring the difference between two distributions. It has already been widely applied in the communities of the statistics [10], [11], coding [12] and statistical learning [13], due to its mathematical elegance. One of its key advantages is the invariance against invertible transformations. Recently, Qiao et. al. has exploited f-divergence to construct effective representation of speech, and achieved promising performance in speech recognition, speech synthesis, computer aided speech evaluation and the classification of the dialect [14]–[17].

In this work, we propose a novel local contrastive descriptor by effectively leveraging the advantages of the f-divergence for encoding local color information. Based on the theories of Gestalt perceptual organization, relativity and structural is the basic characteristic of the visual perception. However, color relativity or contrast is often ignored in the design of the traditional local features. For example, SIFT, SURF and HOG are computed in the gradient space. In addition, applying f-divergence function for computing local distances between image patches is a simple yet efficient method to capture inherent structural context of the images. Three main contributions of the paper are list as bellow.

1) We introduce the$f$-divergence to the computer vision community by proposing a novel local feature for image classification, named as local contrastive descriptor(LCD). LCD delivers local image structure information by measuring the differences between different distributions of the neighboring

978-1-4799-4808-6 /14/$31.00 ©2014 IEEE

patches

2) With the advantages of the f-divergence, we leveraged the color local structure for image representation. We showed that this property can successfully compensate for traditional gradient based features, such as SIFT.

3) The proposed LCD was incorporated with BoVW model for the task of image classification. The efficiency of the LCD was evaluated on the PASCAL VOC 2007 benchmark dataset, and the experimental results indicates that LCD help to improve the performance.

The rest of the paper is organized as follows. Section 2 briefly introduces the formulations and properties of the f-divergence, along with its application to speech recognition. The details the LCD are presented in Section 3. In the Section 4, we show our experimental results on the PASCAL VOC 2007 dataset. Finally, we conclude the paper in Section 5.

II. RELATED WORK

In statistics and information theory, Csiszár f-divergence [10] (also known as Ali-Silvey distance [18]) measures the difference (dissimilarity) between two distributions. Formally, $f : (0, \infty) \rightarrow R$ is a real convex function and $f(1) = 0$, $p_i(x)$ and $p_j(x)$be density functions of two distributions defined on measurable $\Re$.Then

$$D_f(p_i, p_j) = \int_{\Re} p_j(x) f(\frac{p_i}{p_j}) dx \tag{1}$$

To sure the f-divergence between two identical distribution is zero, $D_f(p, p) = 0$, we have the constraint $f(1) = 0$.

It has been proved that the f-divergence has a number of remarkable properties. Csiszár et. al. [10], [19] proved the reflexivity of divergence. A number of important theorems were also been shown by Qiao et. al., who has prove that f-divergence could yield an invariant measure for speech recognition and successfully applied it for speech recognition [14]–[17]. Several important theorems closely related to our LCD descriptor are presented below:

Theorem The f-divergence between two distributions is invariant under differentiable and invertible transformation h

$$D_f(p_i, p_j) = D_f(q_i, q_j) \tag{2}$$

Qiao et al. adopt the advantages of the f-divergence to construct an invariant structural representation of the speech [14]. They also applied the structural feature for speech recognition [14]–[17] and speech synthesis [20]. Zisserman et.al. [21] used the SIFT feature with bag of words approach to the task of object matching. Structure and texture are two important properties of natural images. community that both structure and texture information can compensate strongly for each other. In this paper, we employed the KL-divergence as our f-divergence formulation. The proposed LCD is extracted from multiple channels of the image, and is incorporated with BoVW model for image representation.

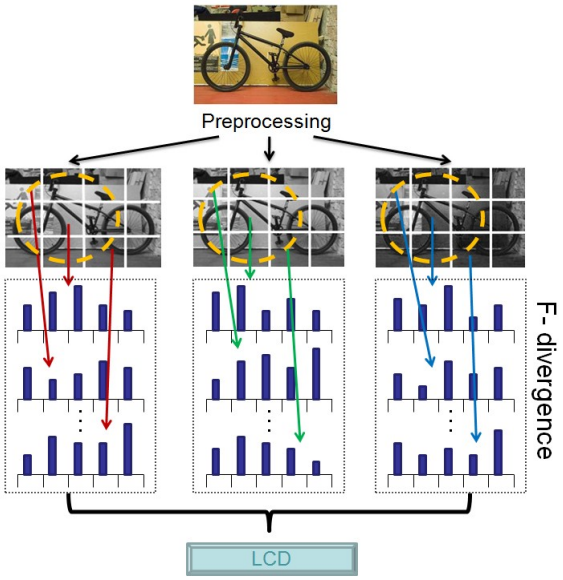

Fig. 1. Computation of the LCD from an image

III. LOCAL CONTRASTIVE DESCRIPTOR

In this section, we present the details of computing the LCD using the f-divergence in the form of the KL-divergence. From **Theorem** in the Section II, we can find that the f-divergence is invariant to transformation of the distributions between different probability spaces. Therefore, it is capable of robustly measuring inherent image structure against multiple variations, e.g. scales, deformations or outliers.

A. Intensive sampling

Good sampling methods can bring better experimental results, and intensive sampling is recognized as a way of sampling with good effect. We have an intensive sampling in this paper. As illustrated in Fig. 1, in order to derive a multiple structural representation of an image, we firstly have to generate some channels for image pre-processing (e.g. RGB or gray image). Each image channels is further divided into $n \times n$ patches. A probability density distribution is computed from each patch. Then a window with the size of $n_1 \times n_1 (n \times n)$ patches is sliding through the whole channel image with the step of one patch. In each window, we compute the f-divergence between the center patch and its adjacent neighboring patches using the defined probability distribution n each patch. Finally, three f-divergence vectors from the same window locations of different channels are concatenated together to generate the LCD vector, which is subsequently used in the BoW model to get the final representation of an image. More details for defining the probability density distribution and computing the f-divergence between patches are described below.

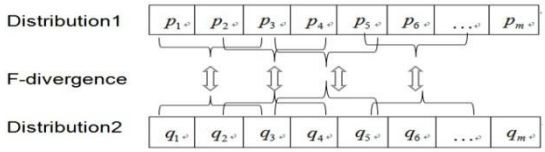

Fig. 2. Computation of f-divergence subspace

B. The probability density distribution

Definition of the probability space is crucial to generate a good LCD. An appropriate probability density distribution reflects the meaningful structure of the image. Histograms of the intensities, RGB color values and gradient orientations, are three common approaches to compute the discrete probability distribution within an image region. Three histograms could be used for generating the LCD. One of the key observations is that f-divergence with the color histogram best measures the underlying image structural information, which strongly compensates for the gradient-based features for image representation. As shown in Fig. 1, three channels are generated from an input image, and we define the distribution for each patch by computing a histogram of its intensity values with m bins, which is used for calculating the f-divergence in the following step.

C. Calculation of f-divergence

As mentioned above, there are various function forms for computing the f-divergence, such as the Bhattacharyya distance, the KL-divergence, the Total variation and the Hellinger distance. Different distance functions could measure different contrast structures between two distributions. We empirically choose the KL-divergence as our f-divergence function, which is defined as follow:

$$D_{KL}(P,Q) = \sum_{I} P(i) log_a \frac{P(i)}{Q(i)} \tag{3}$$

where P and Q are two probability distributions of the patches,corresponding to the histograms of the patches in our definition. In our LCD computation, the number of bins is a trade-off between the feature compactness and informativeness. The larger number of the bins means that more contrast information between patches is retained, leading to stronger discrimination. But this will result in a larger number of dimensions, which increases the computational demands and information redundancy. While a smaller number of the bins would get more information lost. We optimize the number of bins (m) between $[5, 30]$ in our setting

Furthermore, we introduce subspace computation for f-divergence in an effort to enhance the discrimination of the LCD. Details are shown in Fig. 2 and described as follow. We divide each distribution into a number of subspaces with a same number of bins $(m_1, m_1 < m)$. The subspaces are generated by moving a sliding window with size of m_1 from left to right sequentially. The m bins are decomposed into $(m - m_1 + 1)$subspaces, each of which includes m_1 bins. For two distributions$P = (p_1, p_2, ..., p_m)$ and$Q = (q_1, q_2, ..., q_m)$, we have:

$$a(j) = \sum_{i=j}^{j+m_1-1} P(i) log_a \frac{P(i)}{Q(i)} \tag{4}$$

where $j = 1, 2, ..., (m-m_1+1)$.By this way, the dimensions for the f-divergence between two patches are increased from 1 to $m - m_1 + 1 \times (n_1^2 - 1)$,$(n_1^2 - 1)$ is the number of the adjunct neighboring patches, n_2 is the number of the channels divided. We extract dense LCD features from an image, hence the total number of feature vectors is about $n \times n$.

D. BoVW Model

After extracting the LCD features from an image, we applied the bag-of-visual-words (BoVW) approach for image representation. In BoVW model, the information of text is represented by word frequencies with a pre-learned dictionary, while the spatial order is completely discarded. Then document classification is implemented by calculating these statistical frequencies. Each LCD feature is visual equivalent to a word for identification. Therefore, We naturally adopt a standard approach to represent the our descriptors with BoVW model [21], [22]. We first extract a set of LCD features from the training data set. Then a dictionary of visual words was created by the standard k-means clustering.

IV. EXPERIMENT

In this section, we evaluated the performance of the proposed LCD on the task of visual object categorization based on the PASCAL VOC 2007 standards [23]. On the well-known VOC2007 dataset [23], we conducted the first and most expansive set of experiments. This challenge is known as one of the most difficult image classification tasks because both in appearance, posture, and even with occlusions have due to significant a significant change and occlusion. There are 9,963 images with objects of 20 different categories. It consists of 5,011 training (train + validation sets) and 4,952 test images respectively. The performance was evaluated by the standard PASCAL protocol which computes average precision (AP) based on the precision/recall curve. We also reported the mean of AP (mAP) over 20 categories.

We first divided each image into 2,500 (50x50) patchs, and then densely extracted each LCD feature from an image region including 3x3 adjunct patches.The resulted LCD feature was with 432 dimensions (using subspace), which was empirically reduced to 216 by PCA [24]. Our method was implemented by modifying the VLFeat toolbox [25]. The dictionary was generated by running the K-means clustering on a subset of 819K LCD features extracted from the training images. Its size was set to 4,096 as most of previous work did. Some sets of experiments were conducted to evaluate

TABLE I

THE RESULT OF LCD (WITH SUBSPACE) AND SIFT FEATURE.

Category	SIFT	LCD	SIFT+LCD	Δ
aeroplane	0.6971	0.5415	0.7071	0.0100
bicyle	0.5470	0.2084	0.5552	0.0082
bird	0.3633	0.2983	0.4365	0.0732
boat	0.6277	0.4789	0.6502	0.0225
bottle	0.1655	0.0935	0.1676	0.0021
bus	0.5594	0.1274	0.5667	0.0073
car	0.7481	0.5058	0.7530	0.0049
cat	0.5171	0.1969	0.5183	0.0012
chair	0.4581	0.3131	0.4764	0.0183
cow	0.3062	0.1660	0.3281	0.0219
diningtable	0.4164	0.1680	0.4686	0.0522
dog	0.3405	0.1840	0.3680	0.0275
horse	0.7323	0.6127	0.7407	0.0084
motorbike	0.5632	0.3318	0.5869	0.0237
person	0.7976	0.7059	0.8117	0.0141
pottedplant	0.1694	0.1190	0.2168	0.0474
sheep	0.3219	0.2309	0.3855	0.0636
sofa	0.4119	0.1785	0.4242	0.0123
train	0.6951	0.4014	0.7214	0.0263
tvmonitor	0.4431	0.2193	0.4685	0.0254
mAP	**0.4941**	**0.3041**	**0.5176**	**0.0235**

the efficiency of the proposed LCD feature. We evaluated the performance of The multichannel LCD feature (with subspace) and the combined LCD feature with SIFT, as shown in Table 1.

Table 1 shows more detailed comparisons with SIFT on the 20 categories. We can observe that The multichannel LCD feature got very close performance to SIFT in some specific categories, such as person and pottedplant. By fusing with SIFT, the SIFT+LCD feature outperform the SIFT in all 20 categories with the overall increase of 2.3% in accuracy, while the improvements in some categories were significant,e.g.about 6% in the categories of bird, diningtable and sheep. These results clearly verify our observation that the LCD can be a good compensation feature to the SIFT for image representation.

V. CONCLUSIONS

In this section, a novel feature descriptor (LCD) was proposed for detecting the inherent structural information of the image. We introduced f-divergence based descriptor to the computer vision community and applied it for computing the local relationships between image patches. The proposed LCD computes the color contrast between local patches and is capable of capturing informative color information for image representation. A key observation is that this color information can strongly compensate for the traditional gradient-based features, and performance on image classification task was moderately improved by fusing the SIFT and LCD in the BoVW model. Extensive experiments were conducted on the PASCAL VOC 2007 dataset to show the efficiency of the proposed LCD, which clearly verified our observations.

ACKNOWLEDGMENT

This work is partly supported by National Natural Science Foundation of China (91320101), Shenzhen Basic Research Program (JC201005270350A, JCYJ20120903092050890, JCYJ20120617114614438), 100 Talents Programme of Chinese Academy of Sciences, and Guangdong Innovative Research Team Program (No.201001D0104648280).

REFERENCES

[1] Lowe D G. Distinctive image features from scale-invariant keypoints. International journal of computer vision, 2004, 60(2): 91-110.

[2] Ojala T, Pietikăinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern recognition, 1996, 29(1): 51-59.

[3] Huang W, Yin H. A dissimilarity kernel with local features for robust facial recognition.Image Processing (ICIP), 2010 17th IEEE International Conference on. IEEE, 2010: 3785-3788.

[4] Dalal N, Triggs B. Histograms of oriented gradients for human detection.Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 1: 886-893.

[5] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010, 32(9): 1627-1645.

[6] Mikolajczyk K, Leibe B, Schiele B. Local features for object class recognition.Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. IEEE, 2005, 2: 1792-1799.

[7] Zhang J, Marszałek M, Lazebnik S, et al. Local features and kernels for classification of texture and object categories: A comprehensive study. International journal of computer vision, 2007, 73(2): 213-238.

[8] Jegou H, Schmid C, Harzallah H, et al. Accurate image search using the contextual dissimilarity measure. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010, 32(1): 2-11.

[9] Mikolajczyk K, Schmid C. A performance evaluation of local descriptors. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005, 27(10): 1615-1630.

[10] I.Csiszar, Information-type measures of difference of probability distri butions and indirect, Stud. Sci. Math. Hung., vol. 2, pp. 299-318, 1967.

[11] Liese F, Vajda I. On divergences and informations in statistics and information theory. Information Theory, IEEE Transactions on, 2006, 52(10): 4394-4412.

[12] R.E.Blahut,Principles and Practice of Information Theory. Reading, MA: Adison-Wesley, 1987.

[13] Nguyen X L, Wainwright M J, Jordan M I. On surrogate loss functions and f-divergences. The Annals of Statistics, 2009: 876-904.

[14] Qiao Y, Minematsu N. A Study on Invariance of-Divergence and Its Application to Speech Recognition. Signal Processing, IEEE Transactions on, 2010, 58(7): 3884-3890.

[15] Qiao Y, Minematsu N. f-divergence is a generalized invariant measure between distributions.INTERSPEECH. 2008: 1349-1352.

[16] Qiao Y, Suzuki M, Minematsu N. Affine invariant features and its application to speech recognition,Proc.Int.Conf.Acoustics,Speech and Signal Processing(ICASSP), pp.4629-4632,2009

[17] Qiao Y, Suzuki M, Minematsu N. A study of Hidden Structure Model and its application of labeling sequences, Proc.Int.Workshop on Automatic Speech Recognition and Understanding(ASRU2009), pp.118-123,2009.

[18] Ali S M, Silvey S D. A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society. Series B (Methodological), 1966: 131-142.

[19] I.Csiszar and P. Shields, Information Theory And Statistics: A Tutorial. Now Publishers Inc, 2004.

[20] Saito D, Qiao Y, Minematsu N, et al. Optimal event search using a structural cost function-improvement of structure to speech conversion.INTERSPEECH. 2009, 9: 2047-2050.

[21] Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification.Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2010: 270-279.

[22] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos.Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE, 2003: 1470-1477.

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

[24] Huang W, Yin H. On nonlinear dimensionality reduction for face recognition. Image and Vision Computing, 2012, 30(4): 355-366.

[25] A.Vedaldi and B.Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.