DEEP-BASED FISHER VECTOR FOR MOBILE VISUAL SEARCH

*Chen Huang   Shengchuan Zhang   Xianming Lin   Xiangrong liu   Rongrong Ji**

School of Information Science and Engineering, Xiamen University, 361005, China
*Corresponding author: rrji@xmu.edu.cn

ABSTRACT

We tackle the problem of mobile visual search. Moving pictures experts group (MPEG) has completed a standard named compact descriptor for visual search (CDVS) to provide a standardized syntax in the context of image retrieval application. CDVS applies principal components analysis to reduce the dimension of local feature descriptor as the input of global descriptor pipeline, and utilizes traditional fisher vector as the local feature descriptor aggregation algorithm. However, the descriptor components of SIFT and Fisher Vector (FV) have highly non-Gaussian statistics, and applying a single PCA transform can in-fact hurt compression performance at high rates. We develop a net-based architecture combining neural networks with FV layer to obtain fisher vector. There are two advantages in our architecture comparing with CDVS global descriptor pipeline. One is that we employ "autoencoder" networks to reduce the dimensionality of data, the other is that we exploit a trainable system to learn parameters after the FV codebook obtained. The experiments demonstrate an obvious advantage of our proposed architecture in terms of CDVS retrieval task.

***Index Terms*—**CDVS, mobile visual search, Fisher Vector, autoencoder, fisher layer.

1. INTRODUCTION

The calculative ability, storage size and bandwidth delay bring a great challenge on the experience of mobile visual search. Over the course of the standardization process, remarkable improvements were achieved in reducing the size of image feature data in the feature extraction process [1]-[3]. The CDVS image retrieval pipelines consist of two blocks: (1) retrieving a subset of images from the database that are similar, and (2) using Geometric Consistency Checks (GCC) (e.g., based on RANSAC) for finding relevant database images with high precision. The GCC step is computationally complex and can only be performed on a small number of images (tens to hundreds). As a result, the first step of the pipeline is critical to achieve high recall.

In the first step, CDVS standard was the adoption of the 512 byte Residual Enhanced Visual Vector (REVV) global descriptor into the test model TM2.0 [16]. Combining global and local feature descriptors improved the performance at each descriptor length over prior approaches based solely on local feature descriptors [16]. REVV is similar to the VLAD descriptor [17], but it incorporates several improvements to close the performance gap between VLAD and BoW representations [18], [19]. Key enhancements include more effective residual aggregation, dimensionality reduction of residuals using LDA projection matrices, and weighted distance measures for matching. An enhancement titled Robust Visual Descriptor (RVD) [20] was introduced to incorporate soft assignment to quantized words in REVV.

Scalable compressed Fisher Vector (SCFV) is adopted into the test mode TM4.0[4], [23]-[26] after extensive experimentation. It incorporates several new ideas that REVV had introduced, such as learning correlation weights used in signature comparison. SCFV requires significantly less memory than REVV. SCFV allows for fast retrieving in large database with the Hamming distance and different length of global descriptors matching by storing a small set of header bits indicate which components are selected for each aggregated global feature.

We bring the following two innovations in our architecture to improve the SCFV pipeline: (1) we utilize "autoencoder" networks to perform dimensionality reduction instead of PCA. (2) we introduce fisher layer proposed in [15] to our architecture which is an end-to-end trainable system.

An overview of the rest of the paper is as follows: in section 2 we review the SCFV pipeline. Section 3 defines the "autoencoder" network model. Section 4 describe the fisher layer architecture used. Finally, in section 5 we present some quantitative results of our method compared with SCFV.

2. RELATED WORK

Figure 1 illustrates the SCFV pipeline which is built upon the baseline FV model of [5] and [6]-[9]. Firstly, it extracts 128- dimensional SIFT features from query image. Secondly, it reduces the dimensionality of SIFT features by using PCA

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

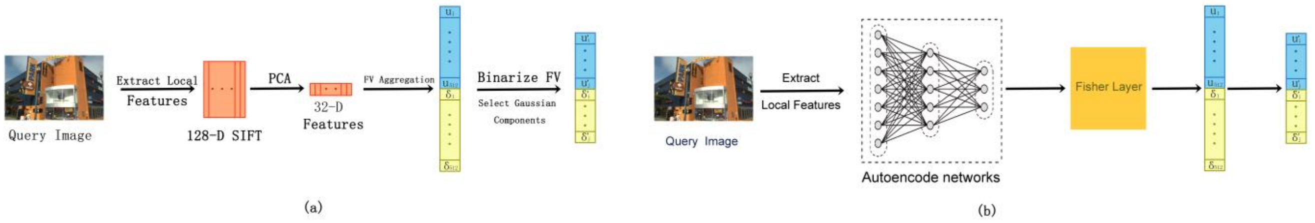

Figure 1 (a) shows SCFV pipeline in the CDVS standard. (b) shows architecture applies "autoencode" networks instead of PCA on dimensionality reduction of local features and Fisher layer to aggregate local features into fisher vector.

which finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. Thirdly, it applies a Gaussian Mixture Model (GMM) with 512 components to capture the distribution of up to 250 local feature descriptors. The gradient of the log-likelihood for an observed set of local feature descriptors with respect to the mean (for higher bitrates) and variances of the GMM are concatenated to form the FV representation [8], [10]. Finally, to compress the FV, SCFV uses one-bit scalar quantizer and selects a subset of Gaussian components (with an average size of 304, 384, 404, 1117 bytes for different specified bitrates) in the GMM based on the standard deviation of certain components of the fully populated fisher vector and retains only the information associated with the selected components [11].

But as shown in [12], applying a single PCA transform can in-fact hurt compression performance at high rates since the descriptor components of SIFT features and Fisher Vectors are known to have highly non-Gaussian statistics. And, the parameters estimation of GMM in traditional FV usually applies EM algorithm. But, EM algorithm has a significant defect that it is easy to trap in local optimum which will affect the performance of cluster result. We bring the following two innovations in our architecture to improve the performance of local feature descriptor aggregation pipeline. First, we utilize "autoencoder" networks which work much better than PCA as a tool to reduce the dimensionality of data [14] by applying the 'minimising contrastive divergence' (MCD rule) [13] to reduce the dimensionality of SIFT features. Second, inspired by [15], we introduce FV layer to our architecture. FV layer, however, has learnable parameters and can be trained jointly with neural networks.

As show in Section 5, we first apply MCD to learn parameters of "autoencode" networks. Then we insert them to a fisher layer and learn both parameters in an end-to-end manner, by standard back-propagation. Similar with SCFV, we exploit rate-scalable representations by selecting a subset of Gaussian components in the GMM based on the standard deviation of certain components of the fully populated fisher vector and retain the information associated with the selected components [11]. So our method also allow for fast retrieving in large database with the Hamming distance and different length of global descriptors matching. The right picture in figure 1 show the detail architecture.

Experiments demonstrate that the proposed architecture significantly outperforms non-learnt FV representations and improves over state-of-the-art compact image representations on standard CDVS image retrieval benchmarks.

3. AUTOENCODE NETWORKS

Hinton and Salakhutdinov [14] show the binary Continuous Restricted Boltzmann Machine model which convert high dimensional vectors to low-dimensional codes perform significantly better than PCA for dimensional reduction. However, the binary RBM is not ideally suited to modelling continuous data. Inspired by that, we propose Continuous Restricted Boltzmann Machine (CRMB) to convert SIFT descriptor to 32 dimensions as the input of fisher layer encoding.

CRMB introduces a continuous stochastic unit by adding a zero-mean Gaussian noise to the input of a sampled sigmoid unit. Let S_j be the output of neuron j, with inputs from neurons with states $\{S_i\}$.

$$s_j = \varphi_j(\sum_i w_{ij}s_i + \sigma\bullet N_j(0,1)) \qquad (1)$$

While

$$\varphi_j(x_j) = \Theta_L + (\Theta_H - \Theta_L)\bullet\frac{1}{1+\exp(-a_j x_j)} \qquad (2)$$

Where $N_j(0,1)$ represents a Gaussian random variable with zero mean and unit variance. The constant σ and $N_j(0,1)$ thus constitute a noise input component $n_j = \sigma\bullet N_j(0,1)$ according to a probability distribution

$$p(n_j) = \frac{1}{\sigma\sqrt{2\pi}}\exp(\frac{-n_j^2}{2\sigma^2}) \qquad (3)$$

$\varphi_j(x)$ is a sigmoid function with asymptotes at Θ_L and Θ_H. Parameter a_j controls the slope of the sigmoid function.

Replacing the binary stochastic unit in RBM by this continuous form of stochastic unit leads to a continuous RBM, within which a_j is a 'noise-control' parameter, allowing a smooth transition from noise-free, deterministic behavior to binary-stochastic behavior.

We use two hidden units and decrease the dimensionality of hidden layers by a factor of 2 and the MCD training rules for CRBM to replace the relaxation search of Gibbs sampling. Equations (4) and (5) indicate the training rules for CRBM's weights $\{w_{ij}\}$ and 'noise-control' parameters $\{a_j\}$.

$$\Delta \hat{w} = \eta_w(< s_i s_j > - < \hat{s}_i \hat{s}_j >) \qquad (4)$$

$$\Delta \hat{a}_j = \frac{\eta_a}{a_j^2}(< s_j^2 > - < \hat{s}_j^2 >) \qquad (5)$$

Where S_j denotes the one-step sampled state of unit j and $<\cdot>$ refers to the mean over the training data.

4. FISHER LAYER

The parameters of traditional FV are fixed once codebook is constructed. We exploit fisher layer [15] after "autoencode" networks that being an end-to-end system to generate fisher vector. Fisher layer makes two simplification of the original FV:

(1) In GMM, it assumes all GMM components have equal weights.
(2) Simplify the k-th Gaussian distribution that can be written as Equation (6), which is similar to covariance matrices sharing the same determinants.

$$u_k(x) = \frac{1}{(2\pi)^{D/2}} \exp\{-\frac{1}{2}(X-\mu_k)^T \sum\nolimits_k^{-1} (X-\mu_k)\} \quad (6)$$

And for any local features X_{ij}, its fisher vector $\varphi(X_{ij}) = [g_{\mu_1}^T, ..., g_{\mu_K}^T, g_{\sigma_1}^T ..., g_{\sigma k}^T]^T \in \mathbb{R}^{2KD\times 1}$, where the simplified gradient of mean vectors $g_{\mu_k}^{X_{ij}}$ and the simplified gradient of variance vectors $g_{\sigma k}^{X_{ij}}$ can be written as Equations (7) and (8).

$$g_{\mu_k}^{X_{ij}} = \gamma_j(k)[w_k \odot (X_{ij} + b_k)] \qquad (7)$$

$$g_{\sigma k}^{X_{ij}} = \gamma_j(k)\frac{1}{\sqrt{2}}[(W_k \odot (X_{ij} + b_k))^2 - 1] \qquad (8)$$

TABLE 1
Recall@500 vs. different descriptor length over the graphic, objects and building datasets, combines with the distractor set FLICKR 1M.

method	Recall@500			
	Length	Graphic	Objects	Building
SCFV	304bytes	88.9%	85.8%	66.6%
	384bytes	91.9%	88.1%	68.2%
	404bytes	92.5%	90.2%	70.1%
	1117bytes	94.1%	92.5%	72.3%
Autoencoder+Fisher layer	304bytes	90.3%	86.4%	68.4%
	384bytes	92.2%	90.6%	71.2%
	404bytes	93.9%	92.3%	73.5%
	1117bytes	95.5%	94.2%	75.6%

The $\gamma_j(k)$ in Equations (7) and (8) is posterior probability which assumes all GMM components have equal weights as in Equations (9).

$$\gamma_j(k) = \frac{\exp\{-\frac{1}{2}(W_k \odot (X_{ij} + b_k))^T (W_k \odot (X_{ij} + b_k))\}}{\sum_{n=1}^{K} \exp\{-\frac{1}{2}(W_n \odot (X_{ij} + b_n))^T (W_n \odot (X_{ij} + b_n))\}} \quad (9)$$

While $W_k = 1/\sigma_k$ and $b_k = -\mu_k$, $\odot$ is an element-wise product operation. W_k and b_k are sets of learnable parameters for each GMM component k. Since the shared computation part $W_k \odot (X_{ij} + b_k)$ is obviously differentiable we can derive all parameters via back-propagation.

5. EXPERIMENT

We present large-scale retrieval experiments for CDVS data sets: *Buildings* (3499 queries), *Objects* (10200 queries), Graphics (1500 queries), Video Frames with 1 million distractor images from FLICKER. For instance retrieval, the GCC step is computationally complex and can only be performed on a small number of images. As a result, it is important for the relevant image to be present in the short list, so that the GCC step can find it. Hence, we present recall at typical operating points, R = 500 after the first step in the retrieval pipeline: matching of global descriptors. Given two images X and Y, we can calculate the Hamming distance-based similarity score [1] among global descriptors.

To train the fisher layer, we use a random set of 100,000 images in 1000 categories from ImageNet data set [21]. This training set does not have any overlap with the query and database data used in the retrieval experiments. We set the learning rate to 0.001 for the weight and bias parameters, momentum to 0.9, and run the training for a maximum 40 epochs with mini-batch size 5. The number of GMM

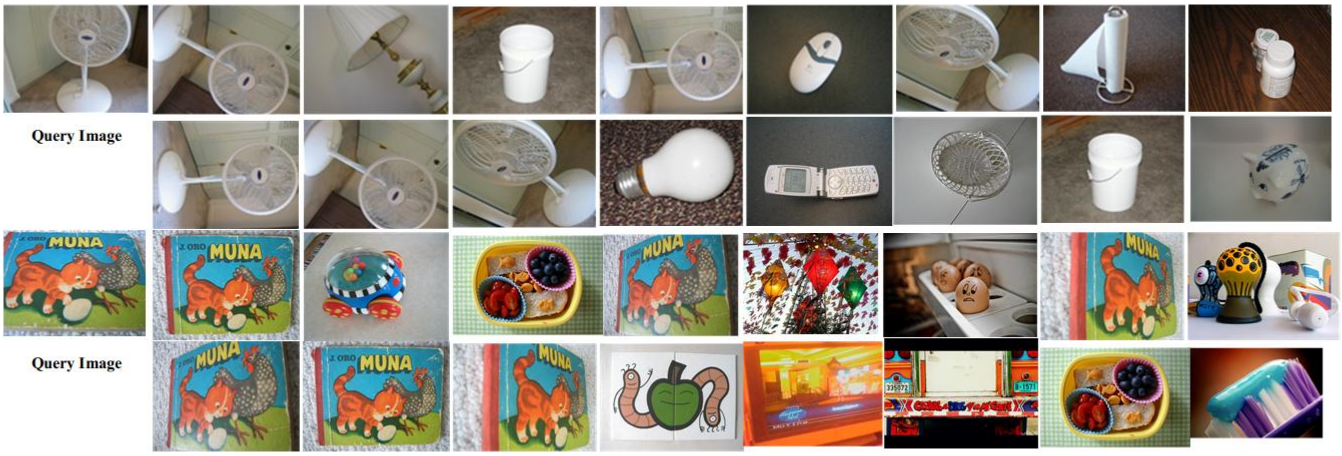

Figure 2. Queries from Objects datasets: retrieval using SCFV; retrieval using our architecture.

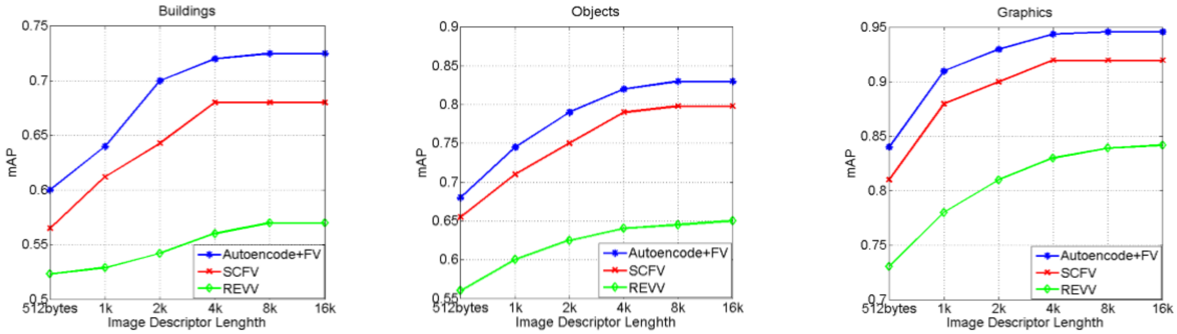

Figure 3. Results of Mean Average Precision of REVV, SCFV and our architecture over the CDVS evaluation framework

components is 512. Here we choose the muti-class sigmoid cross entropy loss. Best parameters for CRBM are chosen as described before.

Table 1 compares our architecture with SCFV at different descriptor length (304 bytes, 384 bytes, 404 bytes, 1117 bytes). It shows the retrieval results in terms of Recall@500 vs. different bit rates over three datasets. Comparison with SCFV, our method achieves competitive results on three datasets. For example, recall is increased from 72.3% at 1117 bytes to 75.6% on Buildings dataset and 85.8% at 304 bytes to 86.4% on Objects dataset.

Figure 2 shows two query results from Objects datasets by SCFV and our architecture. Figure 3 shows the mAP of REVV, SCFV, and our architecture over the CDVS evaluation framework. GCC includes a ratio test followed by a fast geometric model estimation algorithm [22]. We achieve mAP improvements of +4%, +3.25%, +2.85% on average for the Objects, Graphic and buildings datasets compared with SCFV on different image descriptor length. The results have demonstrated the effectiveness of fisher vector generated through "autoencoder" networks with fisher layer.

6. CONCLUSION

In this paper, we propose a novel architecture for the global descriptor pipeline in the CDVS standard. The architecture combines autoencode network with fisher layer to generate more effective fisher vector representation. After FV aggregation, We adopt the binary quantization method and deviation-based approach utilized in SCFV so it retain the benefit that different length global descriptor can calculate the Hamming distance-based similarity score. Experiments on CDVS evaluation framework show substantial improvements comparing with state-of-art results.

7. ACKNOWLEDGEMENT

This work is supported by the National Key Technology R&D Program (No. 2016YFB1001503), the Nature Science Foundation of China (No. 61422210, No. 61373076, No. 61402388, and No. 61572410)

8. REFERENCES

[1] Duan L, Chandrasekhar V, Chen J, et al. Overview of the MPEG-CDVS standard.[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2015, 25(1):179-194.

[2] Ji R, Duan L Y, Chen J, et al. Learning Compact Visual Descriptor for Low Bit Rate Mobile Landmark Search.[C]// IJCAI 2011, Proceedings of the, International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July. DBLP, 2011:2456-2463.

[3] Ji R, Duan L Y, Chen J, et al. Towards low bit rate mobile visual search with multiple-channel coding[C]// ACM International Conference on Multimedia. ACM, 2011:573-582.

[4] Lin J, Duan L Y, Huang Y, et al. Rate-adaptive Compact Fisher Codes for Mobile Visual Search[J]. Signal Processing Letters IEEE, 2014, 21(2):195-198.

[5] Lin J, Duan L Y, Huang T, et al. Robust fisher codes for large scale image retrieval[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013:1513-1517.

[6] Perronnin F, Dance C. Fisher Kernels on Visual Vocabularies for Image Categorization[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007:1-8.

[7] Perronnin F, Nchez J, Mensink T. Improving the fisher kernel for large-scale image classification[C]// Computer Vision - ECCV 2010, European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings. DBLP, 2010:143-156.

[8] Perronnin F, Liu Y, Sanchez J, et al. Large-scale image retrieval with compressed Fisher vectors[C]// CVPR. 2010:3384-3391.

[9] Sánchez J, Perronnin F, Mensink T, et al. Image Classification with the Fisher Vector: Theory and Practice[J]. International Journal of Computer Vision, 2013, 105(3):222-245.

[10] Jaakkola T S, Haussler D. Exploiting Generative Models in Discriminative Classifiers[J]. Advances in Neural Information Processing Systems, 1998, 11(11):487--493.

[11] Z. Wang et al., Response to CE2: Improved SCFV, document ISO/IEC JTC1/SC29/WG11/M33189, 2014.

[12] Chandrasekhar V, Chen D. Transform coding of image feature descriptors[C]// IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2009:725710-725710-9.

[13] Chen, Murray, A.F. Continuous restricted Boltzmann machine with an implementable training algorithm[J]. Vision, Image and Signal Processing, IEE Proceedings -, 2003, 150(3):153-158.

[14] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks.[J]. Science, 2006, 313(5786):504-7.

[15] Tang P, Wang X, Shi B, et al. Deep FisherNet for Object Classification[J]. 2016.

[16] Chen D, Chandrasekhar V, Takacs G, et al. Compact Descriptors for Visual Search: Improvements to the test model under consideration with a global descriptor[C]// Iso/iec Jtci/sc29/wg11 M. 2012.

[17] Jegou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation[J]. 2010, 238(6):3304-3311.

[18] Chen D, Tsai S, Chandrasekhar V, et al. Residual enhanced visual vector as a compact signature for mobile visual search[J]. Signal Processing, 2013, 93(8):2316-2327.

[19] Chen D, Tsai S, Chandrasekhar V, et al. Residual Enhanced Visual Vectors for on-device image matching[C]// 2011:850-854.

[20] Miroslaw B, Syed H, Stavros P, et al. Improvements to TM6.0 with a Robust Visual Descriptor – Proposal from University of Surrey and Visual Atoms[J]. 2013.

[21] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]// Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009:248-255.

[22] Lepsoy S, Francini G, Cordara G, et al. Statistical modelling of outliers for fast visual search.[C]// IEEE International Conference on Multimedia and Expo, ICME 2011, 11-15 July, 2011, Barcelona, Catalonia, Spain. DBLP, 2011:1-6.

[23] J. Lin et al., Peking University Response to CE1: Perfor mance Improvements of the Scalable Compressed Fisher Co des (SCFV), document ISO/IEC JTC1/SC29/WG11/M2806 1, Jan. 2013.

[24] J. Lin et al., Peking Scalable Low-Memory Global Desc riptor SCFV, document ISO/IEC JTC1/SC29/WG11/M2672 6, Oct. 2012.

[25] J. Lin et al., Peking University Response to CE2: The I mproved SCFV Global Descriptor, document ISO/IEC JTC1 /SC29/WG11/M32261, Jan. 2014.

[26] Z. Wang et al., Response to CE2: Improved SCFV, doc ument ISO/IEC JTC1/SC29/WG11/M33189, 2014.