Compact Global Descriptor for Visual Search using Sparse Coding

Jun-Woo Lee, Jong-Wook Kang, Jae-Hyup Jeong and Dong-Seok Jeong

Electronic Engineering, Inha University, Incheon, 402-751 Korea

Summary
This paper studies compact global descriptor for visual search using sparse coding. BoF(Bag of Feature)[1] is widely used in visual search application. But BoF has some problem of large memory usage and update inefficiency of new reference image. To overcome this problem, we propose compact global descriptor which needs small amount of memory. For attain better retrieval accuracy, we divide SIFT[2] features into two part according to its statistical property. And sparse coding[3] is applied for aggregation of local descriptors instead of k-means clustering that is used in existing global descriptor algorithm. The evaluation shows that our approach perform better than existing Test Model

Key words:
Compact Descriptor, Aggregation of local descriptor, Sparse coding

1. Introduction

Recent development of mobile devices with high performance such as smart phones eventually will enable image-based retrieval techniques to substitute for text-based retrieval techniques. Subsequently, demand for visual search systems via mobile devices has been increasing. According to these trends, the international standardization organization, MPEG-7, is working on the new standardization called Compact Descriptor for Visual Search (CDVS)[4]. The main purpose of the CDVS is to achieve a descriptor for visual search via mobile devices that is compact, scalable, and robust. The requirements of CDVS are shown in Table 1.

For the image retrieval in large-scale database, VLAD[5] compute the vector difference between each feature descriptor and the nearest visual word, which is called a word residual and the sum of word residual surrounding each visual word. The aggregated word residual for all k visual words are concatenated together to form an image signature. For a memory efficient representation, principal component analysis (PCA) and product quantization (PQ) are subsequently applied to the word residual vector. REVV[6] also compute the vector difference but aggregation method is average instead of summation. A power law is applied to the word residual. For efficient dimension reduction, linear discriminant analysis (LDA) is applied.

Table 1: Requirements of CDVS

Requirement	Description
Sufficiency	Descriptors shall be self-contained, in the sense that no other data are necessary for matching
Format independence	Descriptors shall be independent of the image format
Robustness	High matching accuracy shall be achieved at least for images of textured rigid objects, landmarks, and printed documents. The matching accuracy shall be robust to changes in vantage point, camera parameters, lighting conditions, as well as in the presence of partial occlusions
Compactness	Shall minimize lengths/size of image descriptors
Scalability	1. Shall allow adaptation of descriptor lengths to support the required performance level and database size. 2. Shall enable design of web-scale visual search applications and databases.
Extraction complexity	Shall allow descriptor extraction with low complexity (in terms of memory and computation)
Matching complexity	1.Shall allow matching of descriptors with low complexity (in terms of memory and computation). 2.If decoding of descriptors is required for matching, such decoding shall also be possible with low complexity.
Localization	1.Shall support visual search algorithms that identify and localize matching regions of the query image and the database image 2.Shall support visual search algorithms that provide an estimate of a geometric transformation between matching regions of the query image and the database image

In this paper, we develop a compact global descriptor for visual search using sparse coding. The proposed global descriptor use sparse coding coefficient as a weighted average while VLAD[5] and REVV[6] use a k-means clustering.

The paper is organized as follows: In section 2, we describe the proposed compact global descriptor. Experimental results in section 3. Section 4 concludes with a discussion

Manuscript received December 5, 2012
Manuscript revised December 20, 2012

2. Compact global descriptor

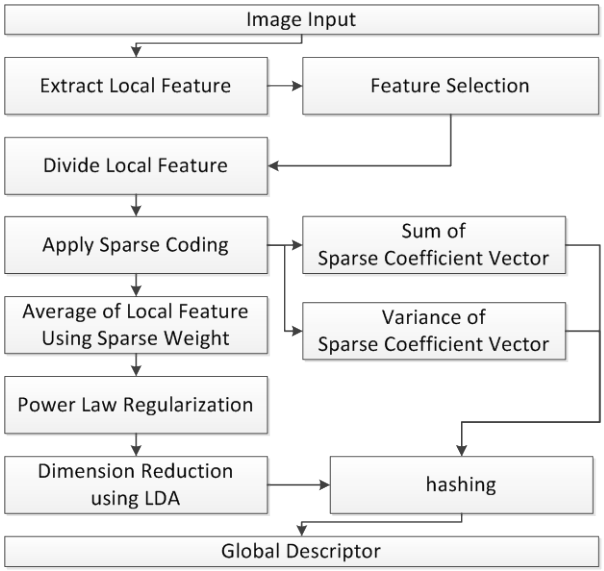

Fig. 1 flowchart of proposed global descriptor

The flowchart of proposed compact global descriptor is in Fig. 1. Sparse Coding is applied for the proposed compact global descriptor for visual search using the aggregation of local descriptors. When image input, SIFT[2] features are extracted. Using the feature selection method[7], local descriptors are sorted in order of importance. The local descriptors are separated into two parts using their statistical properties. As shown in Fig. 2, average of large number of SIFT descriptor have a symmetry statistical property.

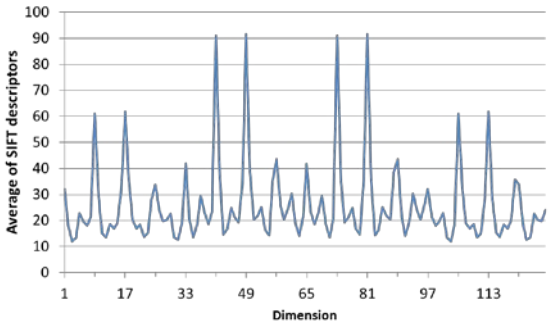

Fig. 2 Average of 7 million SIFT descriptor

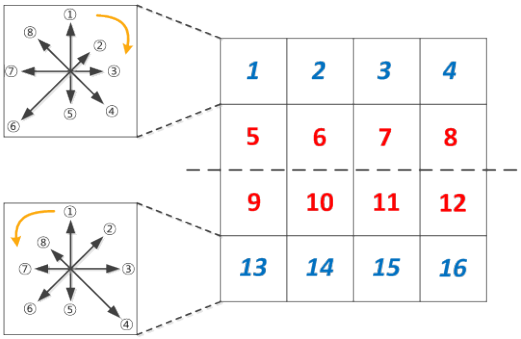

Fig. 3 Statistical property of SIFT feature

Fig. 3 represent the two part of SIFT feature. Afterward, each SIFT features are then represented in the form of sparse coefficient vectors using sparse coding.

VLAD[5] and REVV[6] use a k-means algorithm for aggregation of local descriptors. But k-means algorithm is hard decision process. Some features which are located in boundary between two Voronoi cells are unstable. These features affect performance of retrieval. So we apply a sparse coding instead of k-means clustering

Let $\mathbf{X}$ be a set of SIFT descriptor in a m-dimensional feature space, i.e. $\mathbf{X}=[\mathbf{x}_1\ \mathbf{x}_2\ \dots\ \mathbf{x}_n]\in\mathbb{R}^{m\times n}$, $\mathbf{D}$ be a set of k basis dictionary, i.e. $\mathbf{D}=[\mathbf{d}_1\ \mathbf{d}_2\ \dots\ \mathbf{d}_k]\in\mathbb{R}^{m\times k}$ and $\boldsymbol{\alpha}$ be a set of sparse coefficient about each n signal, i.e. $\boldsymbol{\alpha}=[\boldsymbol{\alpha}_1\ \boldsymbol{\alpha}_2\ \dots\ \boldsymbol{\alpha}_n]\in\mathbb{R}^{k\times n}$. Sparse coding is represented as

$$\min_{\boldsymbol{\alpha},\mathbf{D}}\sum_{j=1}^{n}\left(\left\|\mathbf{x}_j-\sum_{i=1}^{k}\alpha_{ij}\mathbf{d}_i\right\|^2+\lambda\sum_{i=1}^{k}|\alpha_{ij}|\right). \qquad (2.1)$$

Using Eq (2.1), the SIFT feature $\mathbf{X}$ can be represent to sparse coefficient $\boldsymbol{\alpha}$. The sum and variance of the 250 local descriptors, newly represented in the form of sparse coefficient vectors, are calculated and binarized based on their median values. The example of sum and variance for sparse coding is shown in Fig. 4 and Table 2.

Next, the weighted average of the sparse coefficient vectors of local descriptors, each linked to a basis dictionary, is calculated using sparse coding. Let $\boldsymbol{\varepsilon}_i$ be a set of average vector related with each basis dictionary, i.e. $\boldsymbol{\varepsilon}_i=[\varepsilon_{i1}\ \varepsilon_{i2}\ \dots\ \varepsilon_{i64}]\ (i=1,2,\dots,k)$, $\mathbf{X}=[\mathbf{x}_1\ \mathbf{x}_2\ \dots\ \mathbf{x}_n]\in\mathbb{R}^{64\times n}$ be a feature vector of input image, the $\boldsymbol{\varepsilon}_i$ can be represented as

$$\varepsilon_i = \frac{\sum_{j=1}^{n}(x_j^T \times \alpha_{ij})}{\sum_{j=1}^{n}\alpha_{ij}} . \quad (2.2)$$

The power law is applied to the previously calculated weighted average in order to reduce the influence of peaky components that are difficult to match. Afterward, LDA (Linear Discriminant Analysis) is used to reduce the dimensions of average vector and, subsequently, the average vectors with reduce dimensions are binarized according to their signs.

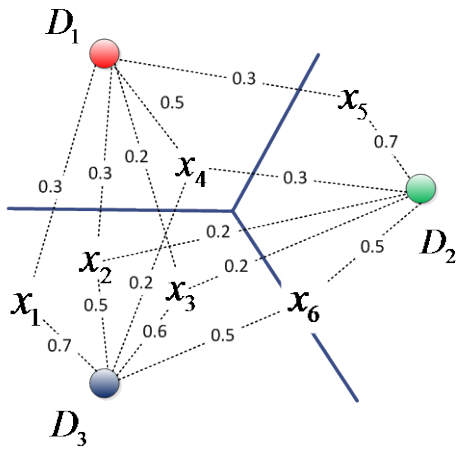

Fig. 4 Example of sparse coding

Table 2 : Example of sum and variance according to sparse coding of Fig.2

	x_1	x_2	x_3	x_4	x_5	x_6	S	V
D_1	0.3	0.3	0.2	0.5	0.3	0	1.6	0.012
D_2	0	0.2	0.2	0.3	0.7	0.5	1.9	0.047
D_3	0.7	0.5	0.6	0.2	0	0.5	2.4	0.035

3. Experimental results

A retrieval test is performed to evaluate the performance of the proposed method. The CDVS standard database that consists of graphics, paintings, video frames, landmarks and common objects is used in addition to the 1 million distractor databases that are gathered via internet. The CDVS standard database is shown in Table 3.

The performance is measured by mAP(mean Average Precision). Let $P(r)$ be a precision at a given cut-off rank r for a single query, i.e.

Table 3 : The CDVS standard database

Experiment	Category	Annotations
1	Mixed text + graphics	a) 3000 (m : matching pair), 30000 (nm:non-matching pair), localization data b) with reduced resolution(max side<=640 pixels). 300(m) 3000(nm) (for one of the images in the pair), localization data c) same as (b) with JPEG compression factor 20, 3000(m) 3000(nm)(for one of the images in the pair), localization data
2	Paintings	400 (m), 4000 (nm)
3	Frames from video clips	400 (m), 4000 (nm)
4	Landmarks / buildings	ZuBud: 575 (m), 5750 (nm) Stanford: 2000 (m), 20000 (nm) ETRI: 500 (m), 5000 (nm) Peking: 960 (m), 9918 (nm) TI: 715 (m), 7199 (nm) SudParis: 396 (m), 3960 (nm) Huawei: 200 (m), 2000 (nm)
5	Common objects	7650 (m), 76500 (nm)

$$P(r) = \frac{(the\ number\ of\ relevant\ images\ of\ rank\ r\ or\ less)}{r} . \quad (3.1)$$

Then, average precision is represented as

$$AveP = \frac{1}{R}\sum_{r=1}^{N} P(r)rel(r) , \quad (3.2)$$

where N is the number of retrieved images, R is the number of relevant images, and $rel(r) = 1$ if image at rank r is relevant, 0 otherwise. Average precision for a set of queries is defined as follows:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q} AveP(q) . \quad (3.3)$$

where Q is the number of queries.

Retrieval results are shown in Table 4. The maximum mAP performance improvement rates were 3.03% with average mAP rates of 1.54% in the retrieval test.

Table 4 : Comparison of mAP between Test Model and Proposed global descriptor

Experiment number	Query descriptor lengths	Test Model	**Proposed Global Descriptor**
1a	512	73.66	75.42
	1K	78.18	80.32
	2k	82.11	84.42
	4k	83.71	84.92
	8k	83.92	85.29
	16k	84.28	85.81
1b	512	73.65	75.54
	1K	78.39	79.86
	2k	82.27	84.28
	4k	83.79	85.02
	8k	84.22	85.38
	16k	84.52	85.91
1c	512	69.43	71.25
	1K	75.41	76.04
	2k	79.62	80.08
	4k	81.21	82.72
	8k	81.59	82.93
	16k	81.94	83.28
2	512	76.06	77.85
	1K	81.73	82.74
	2k	83.53	84.02
	4k	83.68	84.92
	8k	83.82	85.27
	16k	83.82	85.27
3	512	84.07	86.24
	1K	90.54	91.84
	2k	91.84	92.4
	4k	91.15	92.43
	8k	91.55	92.91
	16k	91.33	93.05
4	512	52.58	54.92
	1K	53.31	55.51
	2k	54.59	55.92
	4k	55.6	56.89
	8k	56.29	57.12
	16k	56.38	57.96
5	512	56.21	58.01
	1K	59.58	60.84
	2k	63.66	65.17
	4k	64.57	66.83
	8k	64.92	67.93
	16k	65.16	68.19

4. Conclusion

This paper describes compact global descriptor for visual search using sparse coding. The shortcoming of existing local descriptor aggregation method that perform hard decision complement by using sparse coding that have soft decision property. Furthermore we divide the SIFT feature according to its statistical property. Consequently, the robustness of the descriptor becomes better than existing Test Model. Therefore, it follows that the proposed method is well suited for mobile visual search based on limited descriptor lengths.

Acknowledgment

This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Research & Development Program 2012.

References

[1] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," Proc. IEEE International Conference on Computer Vision, vol. 2, pp. 1470-1477, Oct., 2003.
[2] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, pp. 91-110, 2004.
[3] B. Olshausen and D. Field, "Sparse Coding with ans Overcomplete Basis Set: A Strategy Employed by V1?," Vision Research, vol. 37, pp. 3311-3325, 1997.
[4] "Call for Proposals for Compact Descriptors for Visual Search," ISO/IEC JTC1/SC29/WG11, vol. N12201, Jul., 2011.
[5] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3304-3311, Jun., 2010.
[6] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod, "Residual enhanced visual vector as a compact signature for mobile visual search," Signal Processing, Jun., 2012.
[7] J.W. Lee, J.H. Jeong, J.W. Kang, S.I. Na, and D.S. Jeong, "Image feature point selection method using nearest neighbor distance ratio matching," Journal of the Institute of Electronic Engineers of Korea, vol. 49, Dec., 2012.

Jun-Woo Lee received the B.S. degree in electronic engineering and the M.S. degree in information engineering from Inha University, Korea, in 2006 and 2008, respectively, and then the Ph.D. candidate in information engineering from Inha University, Korea, from 2008 to now. His research interest includes visual search, MPEG-7 and Image & Video Signature, pattern recognition.

Jong-Wook Kang received the B.S. degree in electronic engineering and the M.S. degree in information engineering from Inha University, Korea, in 2010 and 2012, respectively, and then the Ph.D. candidate in information engineering from Inha University, Korea, from 2012 to now. His research interest includes visual search, pattern recognition, SLAM

Jae-Hyup Jeong received the B.S. degree in electronic engineering and the M.S. degree in information engineering from Inha University, Korea, in 2009 and 2011, respectively, and then the Ph.D. candidate in information engineering from Inha University, Korea, from 2011 to now. His research interest includes visual search, pattern recognition, computer vision

Dong-Seok Jeong received the B.S. degree from Seoul National Univ. in 1977. And he worked as research engineer in Agency for Defense Development. He received the M.S. degree and Ph.D. degree from Virginia Tech, USA, in 1985 and 1988, respectively. Currently he is an associate professor of the Department of Electronic Engineering at Inha University and president of Inha Technical College. He is a senior member of IEEE. His research interests include image and video processing, image and video signature and forensic watermarking.