ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Tensor Decomposition of Dense SIFT Descriptors in Object Recognition

Tan Vo[1] and Dat Tran[1] and Wanli Ma[1]

1- Faculty of Education, Science, Technology and Mathematics
University of Canberra, Australia

Abstract. In machine vision, Scale-invariant feature transform (SIFT) and its variants have been widely used in image classification task. However, the high dimensionality nature of SIFT features, usually in the order of multiple thousands per image, would require careful consideration in place to achieve accurate and timely categorization of objects within images. This paper explores the possibility of processing SIFT features as tensors and uses tensor decomposition techniques on high-order SIFT tensors for dimensionality reduction. The method focuses on both accuracy and efficiency aspects and the validation result with the Caltech 101 dataset confirms the improvement with notable margins.

1 Introduction

In recent years, the role that image classification plays in visual data mining is getting more and more important. This is even more specific in the case of large scale visual recognition, where substantial amounts of images are required to be processed and categorized by machine in a both effective and accurate manner. The performance of image classification relies heavily on the process that extracting image features. Among the options in the category, scale-invariant feature transform (SIFT) has been considered an effective algorithm for producing these features [1], which consists of local characteristic across images. Introduced by Lowe et al [2], SIFT has the ability of deducing SIFT keypoints as well as producing SIFT descriptors for those keypoints. Due to the illumination and scale invariance nature of SIFT descriptors [1, 2], object recognition with this algorithm is considered more robust against affine distortions such as rotation, scale and position as well as lighting distortions of objects in images.

In general, a bag-of-words (BoW) of these SIFT descriptors is constructed at training stage. A learning algorithm (normally a linear support vector machine) will utilize the frequency histogram of an image's SIFT descriptors calculated by the BoW to classify the content of the image. The method proposed by Ke et al, PCA-SIFT [3] has been known as a way to produce more compact and robust features out of SIFT based on principal component analysis. This method, together with the biometric work by Liu et al. [4] in iris recognition have inspired us to propose our approach, in which dense SIFT descriptors are considered and processed as tensor (multi-dimensional scalars) . In particular, canonical polyadic decomposition (CP decomposition, or CANDECOMP) [5] was used as a mean to dimensionality reducing these SIFT tensors as well as producing more compact and distinct features used for image classification.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

2 Proposed Method

2.1 Dense SIFT keypoints generation

A dense SIFT operation starts with segmenting a gray-scale image into smaller segments, or patches, of size 8×8 pixels as demonstrated in Figure 1 (a). Each of these patches is further divided into 2×2 smaller segments as shown in Figure 1 (b). For each of these sixteen (4×4) segments, which represent the neighborhoods around the feature point (center of the patch), the image gradients were calculated. Since there are eight directions (N, NE, E, SE, S, SW, W, NW), a smoothed weighted histogram of eight bins is created based on the gradient value. Figure 1 (c) visualizes the values of a histogram built for one cell, with the length of each arrow corresponding to the magnitude of that histogram entry.

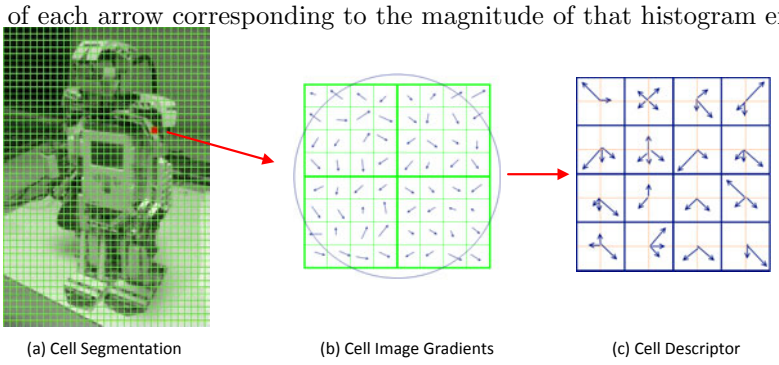


Fig. 1: Process of producing a SIFT descriptor.

Depending on its size, an image will be divided into D patches, of which has a three dimensional array of size $4 \times 4 \times 8$. If we organized these patches in an order of left to right, top to bottom, we will obtain a four dimensional array, or a fourth-order tensor of size $D \times 4 \times 4 \times 8$. Since the dimension of D is arranged in a logical order, one could assume that if he decomposes other dimensions along this dimension, the end result would still be a valid representation of the original tensor.

2.2 CP Tensor Decomposition

The process of decomposing a tensor involves factorizing it into a sum of component rank-one tensors. In this situation, given a fourth-order tensor $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4}$ and a positive integer tensor-rank R, this process is denoted as:

$$\boldsymbol{X} \approx [\![\boldsymbol{\lambda}; \boldsymbol{A}^{(1)}, \boldsymbol{A}^{(2)}, \boldsymbol{A}^{(3)}, \boldsymbol{A}^{(4)}]\!] = \sum_{r=1}^{R} \lambda_r \, \boldsymbol{a}_r^{(1)} \circ \boldsymbol{a}_r^{(2)} \circ \boldsymbol{a}_r^{(3)} \circ \boldsymbol{a}_r^{(4)} \qquad (1)$$

where operator "$\circ$" is the vector outer product, with $\boldsymbol{a}_r^{(n)} \in \mathbb{R}^{I_n}$. Factor matrices $\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(4)}$ are combinations of those rank-one components (i.e.,

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

$\boldsymbol{A}^{(n)} = \begin{bmatrix} \boldsymbol{a}_1^{(n)} & \boldsymbol{a}_2^{(n)} & \cdots & \boldsymbol{a}_R^{(n)} \end{bmatrix}$). Core vector $\boldsymbol{\lambda} \in \mathbb{R}^R$ is used to normalize the columns of factor matrices to length one. Equation 1 can be rewritten as:

$$\boldsymbol{X}^{(n)} \approx \boldsymbol{A}^{(n)} \operatorname{diag}(\boldsymbol{\lambda})(\boldsymbol{A}^{(4)} \odot \ldots \odot \boldsymbol{A}^{(n+1)} \odot \boldsymbol{A}^{(n-1)} \odot \ldots \odot \boldsymbol{A}^{(1)})^\top, \quad (2)$$

where $n = 1, \ldots, 4$, the operator "$\odot$" is Khatri-Rao product of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ [5]. The decomposition process of tensor $\boldsymbol{X}$ is to identify a composition $\boldsymbol{X}'$ such that it satisfy the condition:

$$\min_{\hat{\boldsymbol{A}}^{(n)}} \left\| \boldsymbol{X}^{(n)} - \hat{\boldsymbol{A}}^{(n)}(\boldsymbol{A}^{(4)} \odot \ldots \odot \boldsymbol{A}^{(n+1)} \odot \boldsymbol{A}^{(n-1)} \odot \ldots \odot \boldsymbol{A}^{(1)})^\top \right\| \quad (3)$$

where $\hat{\boldsymbol{A}}^{(n)} = \boldsymbol{A}^{(n)}.\operatorname{diag}(\boldsymbol{\lambda})$, or the column vector normalization of $\boldsymbol{A}^{(n)}$. The optimal solution is then given by:

$$\hat{\boldsymbol{A}}^{(n)} = \boldsymbol{X}^{(n)} \left[(\boldsymbol{A}^{(4)} \odot \ldots \odot \boldsymbol{A}^{(n+1)} \odot \boldsymbol{A}^{(n-1)} \odot \ldots \odot \boldsymbol{A}^{(1)})^\top \right]^\dagger, \quad (4)$$

where $(\boldsymbol{A} \odot \boldsymbol{B})^\dagger$ indicates the pseudoinverse [5] of $\boldsymbol{A} \odot \boldsymbol{B}$. It is a convenience that the pseudoinverse of a Khatri-Rao product has the following property:

$$(\boldsymbol{A} \odot \boldsymbol{B})^\dagger = (\boldsymbol{A}^\top \boldsymbol{A} * \boldsymbol{B}^\top \boldsymbol{B})^\dagger (\boldsymbol{A} \odot \boldsymbol{B})^\top, \quad (5)$$

where the matrix operator "$*$" represents the Hadamard element-wise matrix product [5]. With that, the construction of the alternating least squares (ALS) method [6] is possible. ALS computes an estimate of the best R-ranked CP model of a tensor $\boldsymbol{X}$. It iteratively calculates and normalizes $\boldsymbol{A}^{(n)}$ based on the remaining $\boldsymbol{A}^{(i)}$ ($i \neq n$ and $i = N, \ldots, 1$), hence the name alternating. The outcome of a R-ranked CP decomposition on a fourth-order tensor ($D \times 4 \times 4 \times 8$) comprises of a core tensor $\boldsymbol{\lambda}$ and four component tensors $\boldsymbol{A}^{(1)}$, $\boldsymbol{A}^{(2)}$, $\boldsymbol{A}^{(3)}$ and $\boldsymbol{A}^{(4)}$ of size $D \times R$, $4 \times R$, $4 \times R$ and $8 \times R$, respectively.

2.3 Construction of Features

Traditionally, a SIFT descriptor is vectorized into a 128 length vector before being used to build a visual BoW. This applies for the D descriptors produced of a particular image. At later stage, to represent a given set of SIFT descriptors of an image, histograms of corresponding entries of those descriptors in the BoW dictionary are generated and used as features representing that image.

In this paper, we propose to only further use the first tensor $\boldsymbol{A}^{(1)}$ of $R \times D$ dimension from the outcome of the CP decomposition. This implies that if R has a value of 2, we will achieve a reduction in the order of 64 times ($128 \div 2$) in dimensionality.

With the reduction in data dimension, the most apparent gain is the effective processing time of the image descriptors. In particular, it is the time of building the BoW vocabulary. But in our opinion, an even more valuable gain in our

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

proposed method is the improved accuracy in object classification task. We suspect that the CP decomposition process has produce more robust and concise representation of more convoluted dense SIFT descriptors. A set of object classification experiments has been performed to confirm the validity of these gains in using the proposed method against the conventional method.

3 Experiment

The Caltech 101 image dataset [7] was used for this image classification experiment. Five classes were chosen from the dataset: Faces, Faces-easy, Leopards, Motorbikes and Airplanes [7]. For each class, 160 images were selected for generating SIFT descriptors and two scenarios were defined: (i) Scenario One, or conventional method, in which full SIFT descriptors are used as the input and, (ii) Scenario Two, or our proposed method, where the tensors $\boldsymbol{A}^{(1)}$ constructed from CP-decomposition are used as input.

3.1 Vocabulary building

For scenario One, a BoW vocabulary is created while in the other scenario requires multiple sets of vocabulary for different R-rank decomposition. K-means clustering equipped with accelerated Elkan optimization [8] was used to construct the vocabulary of the BoWs. The number of cluster centers (dictionary size) is 25. The clustering time and the upper bound for the Elkan algorithm were recorded to compare the clustering process (Table 1). Table 1 also indicates that the process of building visual vocabulary in scenario Two is more superior to the one in the conventional scenario. Aside from the total value, the much lower Elkan upper bounds achieved in scenario Two indicates that a much more efficient convergence during the clustering process.

Scenarios	Total time	Total distance	Upper bound
One	29.00 seconds	6.89×10^7	5.08×10^{10}
Two (R = 4)	16.91 seconds	2.46×10^7	1.75×10^2
Two (R = 3)	13.96 seconds	1.30×10^7	6.59×10^1
Two (R = 2)	9.58 seconds	8.07×10^6	8.09×10^0
Two (R = 1)	8.04 seconds	5.05×10^6	9.59×10^{-2}

Table 1: K-means clustering performance.

3.2 Classification performance

For the purpose of verifying the effectiveness of the decomposed image feature, the same set of 160 images will be used. For each set of SIFT descriptors from a

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

test image, kd-tree nearest neighbor is used to compute the histograms of corresponding entries from the constructed BoW dictionaries in the previous step. In a similar fashion, scenario One just uses the unprocessed SIFT descriptors while the other makes use of the features produced from CP-decomposing the same set of descriptors, with R-rank values of 1, 2, 3 and 4. The BoW's vocabulary size dictates the histogram is a vector of 50 bins.

A one-versus-all multiclass SVM configuration was used for this experiment. For that, the label for an input vector $\boldsymbol{x}$) is the $\arg\max_x f(\boldsymbol{x})$, where $f(\boldsymbol{x})$ is the value function of a linear SVM for each class:

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + \boldsymbol{b}, \tag{6}$$

where $\boldsymbol{w}$ is the weight vector and $\boldsymbol{b}$ is the bias. Prior being fed into SVM, feature vector X is normalized with Z-score.

The performance of SVM is validated with 10-Folds cross-validation. The precision and recall for each class were recorded as well as the overall precision for each scenario. Table 2 summaries and compares the results obtained. The result achieved in scenario Two is obtained with a R-rank value of 1. Table 2 shows the superior in classification performance of the proposed method over the conventional with approximately a five percent gain in overall precision. The precisions and recall rates of individual class also follow the same trend.

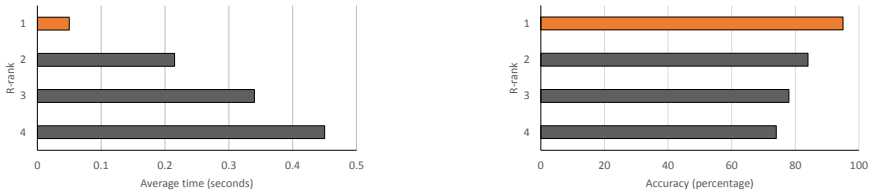
Scenario One	Overall	Class 1	Class 2	Class 3	Class 4	Class 5
Precision	0.90	0.85	0.82	0.99	0.93	0.92
Recall	-	0.79	0.89	0.97	0.88	0.97
Scenario Two	Overall	Class 1	Class 2	Class 3	Class 4	Class 5
Precision	0.95	0.99	0.93	0.98	0.93	0.93
Recall	-	0.97	0.97	0.99	0.93	0.89

Table 2: Classification performance (R-rank = 1 with scenario Two).

3.3 Identifying the optimal value of R-rank parameter

The effects of the R-rank parameter were also recorded to identify the most suited value for R, i.e the value that best compromises the precision rate and the decomposition efficiency. We ran the same experiment with four values of R: 1, 2 3 and 4 and recorded the overall precision as well as the average time it took to decompose a set of SIFT descriptors in seconds. Figure 2 visualizes the result. From that, it is clear that the value of 1 is best suited for R (One-rank decomposition). One assumption can be made to explain the drop in precision as R increases beyond 1 is that, by CP-decomposing an ordered set of SIFT descriptors with a value of $R > 1$, the decomposed components may no longer follow the original order that the SIFT descriptors follow.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.



(a) Average decomposing time (b) Overall precision

Fig. 2: Effects of R-rank parameter on CP-decompostion

4 Conclusion

In this paper, the effectiveness of using CP-decomposition on existing dense SIFT descriptors has been proven to be very positive. Based on a simple observation, we have proposed effective and elegant approach that provides gains for a conventional existing method in image recognition in both accuracy and efficiency categories. This also hints a lot more potentials in using tensor methods, which has been quite successful in computer vision.

References

[1] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2):43–72, November 2005.

[2] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.

[3] Yan Ke and R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–506–II–513 Vol.2, 2004.

[4] Xiaomin Liu and Peihua Li. Tensor decomposition of sift descriptors for person identification. In *Biometrics (ICB), 2012 5th IAPR International Conference on*, pages 265–270, 2012.

[5] T. Kolda and B. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[6] J.Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, April 2007.

[8] Charles Elkan. Using the Triangle Inequality to Accelerate K-Means. 2003.