LETTER

Statistical Property Guided Feature Extraction for Volume Data

Li WANG[†a)], **Xiaoan TANG**[†], ***Nonmembers***, **Junda ZHANG**[†], ***Student Member***, ***and*** **Dongdong GUAN**[†], ***Nonmember***

SUMMARY Feature visualization is of great significances in volume visualization, and feature extraction has been becoming extremely popular in feature visualization. While precise definition of features is usually absent which makes the extraction difficult. This paper employs probability density function (PDF) as statistical property, and proposes a statistical property guided approach to extract features for volume data. Basing on feature matching, it combines simple liner iterative cluster (SLIC) with Gaussian mixture model (GMM), and could do extraction without accurate feature definition. Further, GMM is paired with a normality test to reduce time cost and storage requirement. We demonstrate its applicability and superiority by successfully applying it on homogeneous and non-homogeneous features.

key words: *feature extraction, probability density function (PDF), statistical property, simple liner iterative clustering (SLIC), Gaussian Mixture Model (GMM)*

1. Introduction

Feature visualization is of great significances in volume visualization because direct visualization for the whole dataset may cost massive hardware resource and result in heavy shelter when displaying, as the size of dataset has increased to TB and beyond. With the development of feature visualization, feature extraction has been becoming an extremely popular issue for research for it could filter out irrelevant data and reduce visualization mapping [1]. Further, people would gain a better scientific insight of the dataset by feature visualization.

There have been plenty of woks for feature extraction, and majority of them are definition-dependent methods, as they generally assume that the features are predefined and extraction of features is deterministic. Such as Gu [2] proposed a C^2-continuous framework to extract high-quality topological structure. Gyulassy [3] characterized the range of features and extracted them from a Morse-Smale complex. The contour tree [4] and Reeb graph [5] are also used to define the features of interest, etc.

However, these methods are limited within their specific datasets and features, if we get a new kind of dataset, we have to define the features newly, and this is unacceptable in practice. What is worse, if the descriptions of the features are fuzzy or the definitions could not be got by analytical equations, extraction process would be difficult. Unfortunately, most datasets to visualize belong to these two cases as precise definition of features is usually absent.

According to this, definition-independent methods have been becoming promising. Chaudhuri [6] proposed an integral distribution based method. Lee [7] used integral distributions with discrete wavelet transformation to analyze the local statistical properties. To evaluate the performance of query driven visualization, Gosink [8] utilized PDF effectively. Wang [9] introduced an importance-driven method for volume visualization using information theory basing on distributions. A prominent technology is that Xie [10] proposes a fast uncertainty-driven refinement method, it combines simple liner iterative clustering (SLIC) with multi-resolution technology and refinement approach, and do well for kinds of volumes and features without definition. While multi-resolution technology is time and hardware resource consuming, what is worse, extraction deteriorates when facing with some non-homogeneous features such as topology structure or texture, and these two disadvantages constrain its application. Overall, our paper aims at overcome the problems of [10] and focus on definition-independent method.

In this work, probability density function (PDF) is employed as statistical property, and a statistical property guided approach for feature extraction is proposed, our main contributions are:

- It could do extraction without precise feature definition.
- It is applicable to kinds of features including homogeneous and non-homogeneous features.
- A normality test is applied to PDF estimation which reduces the time cost and storage requirement.

2. Proposed Method

2.1 Overview

The primary contents of our method are three parts: Firstly the dataset is segmented by SLIC and series of supervoxels generate, simultaneously we preselect a small region from the volume as reference feature ($f_{reference}$); Secondly, PDF is employed as statistical property, then PDFs of all supervoxels and $f_{reference}$ are estimated by Gaussian mixture model

Manuscript received August 24, 2017.
Manuscript publicized October 13, 2017.
[†]The authors are with School of Electronic Science and Engineering, National University of Defense Technology, Changsha, 410073, China.
a) E-mail: wangli08a@126.com (Corresponding author)
DOI: 10.1587/transinf.2017EDL8188

Copyright © 2018 The Institute of Electronics, Information and Communication Engineers

(GMM); Finally, PDF based distance between each supervoxel and $f_{reference}$ is calculated, and a matching threshold is applied, if the distance is lower than the threshold, that is the supervoxel matches the $f_{reference}$ and all these matched supervoxels compose the feature extracted.

2.2 SLIC Based Segmentation

SLIC [11] is originally proposed for 2D image segmentation, and it has been extended to 3D volume. Comparing with other cluster-based segmentation methods, SLIC is superior for its better segmentation quality and higher execution efficiency [10].

Iterative cluster for SLIC is similar to K-means: (a) each voxel is assigned to a supervoxel with the smallest distance; (b) the cluster center is updated after every assignment; (c) a threshold is predefined, if the difference between every two adjacent iterations is lower than the threshold, the algorithm is terminated, otherwise, the algorithm runs to its next iteration.

The primary disparities with K-means are: (a) search space of SLIC for each voxel is constrained within a window around each cluster center instead of the whole volume; (b) distance metric in cluster combines intensity and spatial proximity which controls the compactness of cluster.

It is to be noted that by localizing the search within a window in cluster, computational complexity of SLIC is extremely reduced comparing with K-means. The complexity of K-means is $O(kN)$ and SLIC is $O(N)$ [10] for each iteration. When the dataset is with a huge dimension, obviously SLIC would have a much faster performance.

2.3 GMM Based PDF Estimation and Normality Test

In this paper, PDF is employed as statistical property, and GMM is introduced to estimate PDFs of these supervoxels. Comparing with typical kernel density estimation (KED), GMM has lower storage requirement than KED, and could better represent PDFs of distribution-free datasets [12], [13].

GMM represents the PDF of a dataset as

$$f(x) = \sum_{k=1}^{K} \omega_k N(x : \mu_k, \Sigma_k) \tag{1}$$

Where K stands for the number of Gaussian kernels mixed, ω_k is the mixing weight and μ_k, Σ_k are the mean vector and covariance matrix of Gaussian kernel respectively, generally they are fixed by Expectation-Maximum (EM) [14].

However, storage requirement of GMM is still large as the volume dataset is usually with a big size, and EM has large time complexity because EM gets a better estimation iteratively. To settle these weaknesses, a strategy called normality test [15] is introduced, if a supervoxel satisfies the test, a single Gaussian model (SGM) is enough to estimate its PDF, and otherwise, GMM is employed. By this strategy, when most supervoxels satisfies the test, no doubt the total time and storage cost will largely decrease.

2.4 PDF Based Feature Matching and Extraction

To match all supervoxels with the $f_{reference}$, PDF based distance between them is calculated, the smaller the distance is, the more similar they are, and the supervoxel is more likely to be the feature. Simultaneously a matching threshold is applied, if distance is lower than the threshold, it means a successful matching and the supervoxel would be part of the feature, otherwise, it would be regarded as the background. All these supervoxels successfully matched compose the extracted feature.

Here the Bhattacharyya-based distance metric [16] is utilized as it is generally fast and leads to good results. It can be expressed as

$$\Psi(p, p') = \sum_{i=1}^{n} \sum_{j=1}^{m} w_i w'_j \psi(p_i, p'_j) \tag{2}$$

Where p and p' are GMMs, n and m are number of mixture components of p and p' respectively. $\psi(\cdot)$ is distance between two Gaussian kernels:

$$\psi(p, p') = \frac{1}{8}(u - u')^T \left(\frac{\Sigma + \Sigma'}{2}\right)^{-1} (u - u') + \frac{1}{2} \ln \left[\frac{|\Sigma + \Sigma'|}{2\sqrt{|\Sigma||\Sigma'|}}\right] \tag{3}$$

Here u, u' and Σ, Σ' are mean and covariance of the Gaussian kernel p and p' respectively.

Obviously $\psi(\cdot)$ is symmetric, positive, and is zero when p and p' are equal, which is well in corresponding with intuitive judgment.

3. Experiments

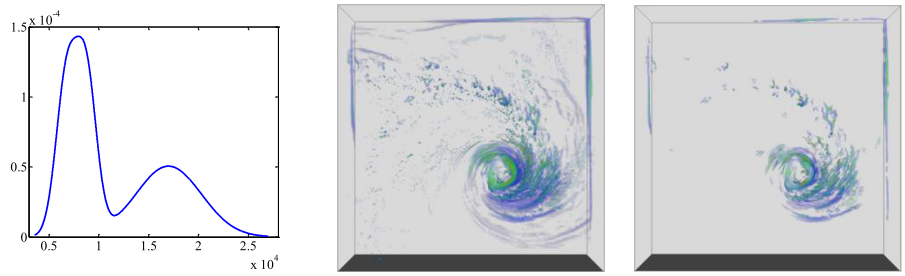
The experiments are performed on a PC equipped with Intel Core i5-6500 CPU@3.20 GHz and 8.00 GB DDR, software platforms are Visual Studio 2010 and MATLAB 2012a.

To demonstrate the extraction results, two datasets are used: (a) Hurricane Isabel data [17]. It is a very typical dataset with 13 scalars and is used for the IEEE Visualization 2004 Contest. The resolution is $500 \times 500 \times 100$, we select the wind field as the feature to be extracted, and it is non-homogeneous; (b) Blunt Fin Data [18]. The resolution is $256 \times 128 \times 64$, body of combustion is taken as the feature, and it is homogeneous.

3.1 Qualitative Evaluations

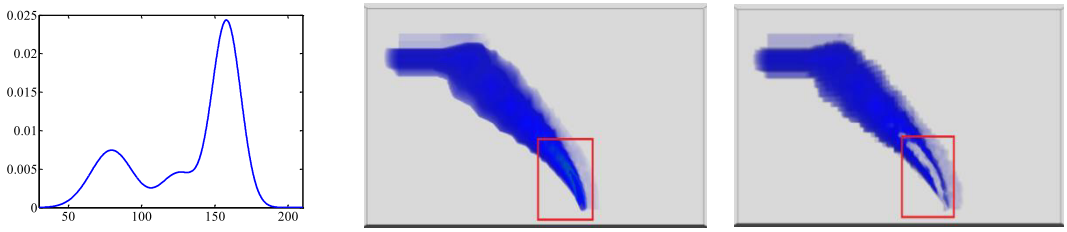
Two datasets are pre-partitioned for SLIC with a size of $5 \times 5 \times 5$ and $3 \times 3 \times 3$ respectively, and matching threshold is 0.45 and 0.20 respectively. Extracted features are exhibited by volume rending technology [19].

Figures 1 and 2 show qualitative evaluations. In Fig. 1, Fig. 1 (c) just identifies the core of the wind, but the smaller



(a) GMM of $f_{reference}$ (b) Proposed method (c) Method in [10]

Fig. 1 Extraction results of Hurricane Isabel Data.



(a) GMM of $f_{reference}$ (b) Proposed method (c) Method in [10]

Fig. 2 Extraction results of Blunt Fin Data.

band of the vortex is mostly missing or identified with low confidence. However, Fig. 1 (b) extracts not only the vortex core but also the smaller band, so the whole wind field is better extracted with higher accuracy. Comparing Fig. 1 (b) with Fig. 1 (c), we could find that the smaller band is also valuable as they help to represent the distribution of the wind field, so we can gain a better understanding of the dataset. In Fig. 2, Fig. 2 (b) is with much subtler details and it is much sharper than Fig. 2 (c) which is with vague and jagged outlines. What is more, Fig. 2 (c) leads to under-extraction because it loses important part where represents the combustion burner (as marked with red boxes). To sum up, extractions of our method are more complete.

3.2 Quantitative Evaluations

Essentially feature extraction is segmentation as the volume is divided into two parts: feature and background. To quantitatively evaluate the extraction performance, uniformity of region (*UR*) and disparity of regions (*DIR*) are introduced [20], [21]. *UR* is to measure the uniformity within-class and *DIR* is to measure disparity among-classes, they would be as large as possible for a perfect segmentation. Particularly, distance in *UR* and *DIR* is calculated by Eq. (2) to improve the criteria.

Table 1 and Table 2 show the quantitative evaluations. As depicted, for two datasets, *UR* of our method largely increases by 50.151% and 24.716% respectively, *DIR* largely increases by 34.545% and 33.032% respectively, and it reveals that feature extracted by our method is much more different and identifiable from the background.

Moreover, time cost of our method largely decreases by 56.777% and 35.066% respectively, which means our method is with much higher execution efficiency.

Overall, qualitative and quantitative results demonstrate that extraction performance of our method is much more excellent.

Table 1 Quantitative evaluations for Hurricane data.

	Method in [10]	Proposed method	Increase ratio
UR	0.662	0.994	50.151 %
DIR	1.320	1.776	34.545 %
Time cost (/*s*)	3062.336	1323.626	-56.777 %

Table 2 Quantitative evaluations for Blunt Fin data.

	Method in [10]	Proposed method	Increase ratio
UR	0.793	0.989	24.716 %
DIR	10.384	13.814	33.032 %
Time cost (/*s*)	597.474	387.965	-35.066 %

4. Conclusion

This paper employs PDF as statistical property, and proposes a statistical property guided method to extract features. Basing on feature matching, it combines SLIC with GMM, and could do extraction without precise feature definition. With the normality test introduced, the time cost and storage requirement reduce. Experiments illustrate its applicability to non-homogeneous and homogeneous features, and superiority with better performance.

Acknowledgements

Hurricane Isabel data produced by the Weather Research and Forecast (WRF) model, courtesy of NACR and the U.S. National Science Foundation (NSF).

References

[1] T.-Y. Lee, X. Tong, H.-W. Shen, P.C. Wong, S. Hagos, and L.R. Leung, "Feature tracking and visualization of the Madden-Julian Oscillation in climate simulation," IEEE Computer Graphics and Applications, vol.33, no.4, pp.29–37, 2013.

[2] W. Gu, M.-E. Fang, and L. Ma, "High-quality topological structure extraction of volumetric data on C2-continuous framework," Computer Aided Geometric Design, Elsevier, vol.35-36, pp.215–224, 2015.

[3] A. Gyulassy, N. Kotava, M. Kim, C.D. Hansen, H. Hagen, and V. Pascucci, "Direct Feature Visualization Using Morse-Smale Complexes," IEEE Trans. Vis. Comput. Graphics, vol.18, no.9, pp.1549–1562, 2012.

[4] H. Carr, J. Snoeyink, and U. Axen, "Computing contour trees in all dimensions," Computational Geometry, vol.24, no.2, pp.75–94, 2003.

[5] G. Weber, P.-T. Bremer, M. Day, J. Bell, and V. Pascucci, "Feature tracking using reeb graphs," In Topological Methods in Data Analysis and Visualization, Springer, pp.241–253, 2011.

[6] A. Chaudhuri, T.H. Wei, T.Y. Lee, H.W. Shen, and T. Peterka, "Efficient range distribution query for visualizing scientific data," In Pacific Visualization Symposium (Pacific Vis), 2014 IEEE Pacific, IEEE, pp.201–208, 2014.

[7] T.-Y. Lee and H.-W. Shen, "Efficient local statistical analysis via integral histograms with discrete wavelet transforms," IEEE Trans. Vis. Comput. Graphics, vol.19, no.12, pp.2693–2702, 2013.

[8] L.J. Gosink, C. Garth, J.C. Anderson, E.W. Bethel, and K.I. Joy, "An application of multivariate statistical analysis for query-driven visualization," IEEE Trans. Vis. Comput. Graphics, vol.17, no.3, pp.264–275, 2011.

[9] C. Wang, H. Yu, and K.-L. Ma, "Importance-driven time-varying data visualization," IEEE Trans. Vis. Comput. Graphics, vol.14, no.6, pp.1547–1554, 2008.

[10] J. Xie, F. Sauer, and K.-L. Ma, "Fast uncertainty-driven large-scale volume feature extraction on desktop pcs," 2015 IEEE 5th Symposium on Large Data Analysis and Visualization (LDAV), pp.17–24, 2015.

[11] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.11, pp.2274–2282, 2012.

[12] S. Dutta and H.-W. Shen, "Distribution driven extraction and tracking of features for time-varying data analysis," IEEE Trans. Vis. Comput. Graphics, vol.22, no.1, pp.837–846, 2016.

[13] Y. Wang, W. Chen, J. Zhang, T. Dong, G. Shan, and X. Chi, "Efficient volume exploration using the Gaussian mixture model," IEEE Trans. Vis. Comput. Graphics, vol.17, no.11, pp.1560–1573, 2011.

[14] J. Bilmes, "A gentle tutorial of the em algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," International Computer Science Institute, vol.4, no.510, p.126, 1998.

[15] R.B. D'agostino, A. Belanger, and R.B.D. Jr, "A suggestion for using powerful and informative tests of normality," The American Statistician, vol.44, no.4, pp.316–321, 1990.

[16] G. Sfikas, C. Constantinopoulos, A. Likas, and N.P. Galatsanos, "An analytic distance metric for Gaussian mixture models with application in image retrieval," Artificial Neural Networks: Formal Models and Their Applications-ICANN 2005, pp.835–840, Springer, 2005.

[17] IEEE Visualization 2004 Contest, http://vis.computer.org/vis2004contest/.

[18] The Volume Library, http://www9.informatik.uni-erlangen.de/External/vollib/.

[19] GitHub, https://github.com/csuzhangxc/vvv/tree/master/viewer.

[20] S. Zhang, J.-W. Dong, and L.-H. She, "The Methodology of Evaluating Segmentation Algorithms on Medical Image," Journal of Image and Graphics, vol.14, no.9, pp.1872–1880, 2009.

[21] F.-Y. Xie, Digital Image Processing and Application, Publishing House of Electronics Industry, Beijing, 2014.