Computers & Graphics (2017)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/results-in-physics

Metric-based Curve Clustering and Feature Extraction in Flow Visualization

Lieyu Shi[a], Guoning Chen[a]

[a]*Department of Computer Science, University of Houston, Houston, TX, USA, 77204*

ARTICLE INFO

Article history:

ABSTRACT

1. Supplemental Materials

1.1. Clustering Analysis

Clustering is a popular topic in unsupervised machine learning, and it was applied to flow visualization to extract and highlight information of flow data. A popular clustering technique is K-means which enjoys easy implementation and fast convergence. More importantly, it has a relatively low overhead with O(kMN) complexity. However, K-means has some restrictions, such as its strict demand of a hyper-spherical structure in the high-dimension space, same variance in observation functions, etc., as discussed in [1]. Improved K-means has been proposed, including K-means++ [2], IODATA [3] and kernel K-means [4], which achieves better clustering results at higher computational overhead.

Another prevailing method is the agglomerative hierarchical clustering (AHC) [5] which achieves a bottom-up clustering. Compared to K-means, AHC can provide better clustering quality because of fewer restrictions of data distribution in high-dimensional space. Nevertheless, it has much higher computation complexity than K-means. In general, its complexity is O($NM^2 \log M$), making it too slow for the processing of large datasets. Also, final clustering results are sensitive to user-defined threshold.

Spectral clustering (SC) makes use of the spectrum of similarity matrix to perform dimensionality reduction before clustering, which may achieve rather good results in streamline detection and clustering [6, 7, 8, 9]. Normalized cut is one spectral clustering widely used in image processing [10] and stream-tube clustering and analysis [11]. However, SC is even more computationally demanding than AHC due to the similarity matrix computation and SVD decomposition. SC also suffers from a demanding memory requirement with O(lM^2) (l is constant number of matrices) complexity.

Therefore, considering the performance and memory overhead of the conventional AHC and SC techniques, we mainly applied K-means clustering to assess the performance of various distance/similarity metrics on our flow datasets. We would leave implementations of our metrics with improved hierarchical clustering in future work.

1.2. Metric Design

1.2.1. Spatial-based Metrics

Spatial-based metrics are only concerning the spatial distance between two feature vectors. These examples include

1. Euclidean distance $\mathbf{L}_{Euc}$: The most intuitive spatial distance is the Euclidean distance. It's easy to implement and guarantees unconditional convergence especially for the K-means clustering. However, in high-dimensional space, Euclidean distance would not accurately demonstrate the same distribution as in lower-dimension.
2. Principal component analysis (PCA) distance metric $\mathbf{L}_{PCA}$: The PCA-based metric $\mathbf{L}_{PCA}$ is taken from [12], in which high-dimensional feature vectors of curves are reduced to lower-dimensional ones via PCA, followed by an Euclidean-based clustering. The benefit of this method is that it automatically determines the dimensionality in the lower-dimension space according to standard deviation accumulation.
3. Fraction distance metric $\mathbf{L}_{Frac}$ * [1]: Fraction distance metric in high-dimension space was firstly proposed in [13]. Due to the different distribution characteristics of data in the high-dimensional space than in the lower dimensional space, fraction distance metric enables to increase the contrast between points with the closest and furthest distance,

[1] * denotes the new metrics proposed in this work for flow visualization

respectively. We are the first to introduce this fraction distance metric into the flow visualization community to perform spatial clustering.

Specifically, given high-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ of same dimension $d = 3N$, their fraction distance is defined as

$$\mathbf{L}_{Frac}(\mathbf{x},\mathbf{y}) = (\sum_{i}^{d} \| \mathbf{x}_i - \mathbf{y}_i \|^p)^{\frac{1}{p}} \tag{1}$$

where $0 < p < 1$ is a constant.

Note that fraction distance does not satisfy triangle inequality. Since smaller p would increase the contrast between the closest and furthest points, in our experiment we set $p = 0.5$.

4. Normalized dot distance metric $\mathbf{L}_{Dot}$* : $\mathbf{L}_{Dot}$ is a generalized distance metric from the $3D$ Euclidean space to the high-dimensional space, which measures the pseudo intersection-angle between two feature vectors. It is defined as

$$\mathbf{L}_{Dot}(\mathbf{x},\mathbf{y}) = \arccos \frac{\widetilde{\mathbf{x}} \cdot \widetilde{\mathbf{y}}}{N} \tag{2}$$

$$\begin{aligned} \widetilde{\mathbf{x}} &= \{\widetilde{\mathbf{x}}_i\}_{i=1}^{d} \\ \widetilde{\mathbf{y}} &= \{\widetilde{\mathbf{y}}_i\}_{i=1}^{d} \end{aligned} \tag{3}$$

$$\begin{aligned} \widetilde{\mathbf{x}}_i &= \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|} \\ \widetilde{\mathbf{y}}_i &= \frac{\mathbf{y}_i}{\|\mathbf{x}_i\|} \end{aligned} \tag{4}$$

Notice that $\mathbf{L}_{Dot}$ is actually estimating the average intersection angle between each pair of normalized vertices along two curves $\mathbf{x}$ and $\mathbf{y}$, and it is spatially sensitive.

1.2.2. Statistics-based Metrics

The basic intuition of our design of the statistics-based metrics comes from the fact that each curve can be regarded as a Gaussian distribution of either single-variate or multi-variate from the law of large numbers as in [14]. Then, we are able to use the Bhattacharyya metric [15] to measure relative closeness of two Gaussian distributions. The benefit of this metric group is that it doesn't need pair-wise comparison, hence no need to elongate each curves to be exactly the same size. Despite performance improved by matrix computation, statistic metrics will work best with large enough N (N is number of vertices of curves).

To be specific, the discrete curvature is defined as

$$\tau_i = \arccos \frac{\mathbf{x}_i \cdot \mathbf{x}_{i+1}}{\| \mathbf{x}_i \| \cdot \| \mathbf{x}_{i+1} \|} \tag{5}$$

1. Bhattacharyya metric for curvature sequence $\mathbf{L}_{BMCS}$*:
   Each curve has a discrete curvature sequence computed by piece-wise angles, which theoretically forms a Guassian distribution for independent random curvatures along a curve. Then $\mathbf{L}_{BMCS}$ is represented as

$$\begin{aligned} \mathbf{L}_{BMCS}(\mathbf{x},\mathbf{y}) = & \frac{1}{4} \ln \left(\frac{1}{4} \left(\frac{\sigma_p^2}{\sigma_q^2} + \frac{\sigma_q^2}{\sigma_p^2} + 2 \right) \right) \\ & + \frac{1}{4} \left(\frac{(\mu_p - \mu_q)^2}{\sigma_p^2 + \sigma_q^2} \right) \end{aligned} \tag{6}$$

   where p, q are curvature sequences, respectively, for two streamlines $\mathbf{x}$ and $\mathbf{y}$. μ_p, μ_q are their means, and σ_p, σ_q are their standard deviations.

2. Bhattachryya metric for to-fixed-direction angles $\mathbf{L}_{BMTA}$*
   Different from the piece-wise curvature as in Eq. (5), $\mathbf{L}_{BMTA}$ measures the Bhattacharyya metric for the angle sequences of line segments to a fixed direction. This metric is more robust than $\mathbf{L}_{BMCS}$ because the latter is very sensitive to initial line segment direction. The two angle sequences are compared using the same format as in Eq. (6).

3. Bhattacharyya metric for normalized line direction $\mathbf{L}_{BMNLD}$*
   Similar to $\mathbf{L}_{Dot}$ in Eq. (4), we can obtain a normalized direction sequences $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{y}}$ for two curves $\mathbf{x}$ and $\mathbf{y}$. We could take $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{y}}$ as $3D$ independent random variables, which enables a multivariate Bhattachryya distance on them as below

$$\begin{aligned} \mathbf{L}_{BMNLD}(\mathbf{x},\mathbf{y}) = & \frac{1}{8} (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) \\ & + \frac{1}{2} \left(\frac{\det \Sigma}{\sqrt{\det \Sigma_1 \det \Sigma_2}} \right) \end{aligned} \tag{7}$$

   where μ_1 and μ_2 is respectively mean vector of $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{y}}$, and Σ_1 and Σ_2 is covariance matrix.

4. Bhattachryya metric for unnormalized line direction $\mathbf{L}_{BMULD}$*
   Different to $\mathbf{L}_{BMNLD}$, $\mathbf{L}_{BMULD}$ didn't normalized the direction vector of line segments, which is due to the fact that longer line segments with same direction should increase disimilarity result as ground truth. The computation of $\mathbf{L}_{BMULD}$ is exactly the same as $\mathbf{L}_{BMNLD}$ in Eq. (7), with line segment direction not normalized to convey more length-related information.

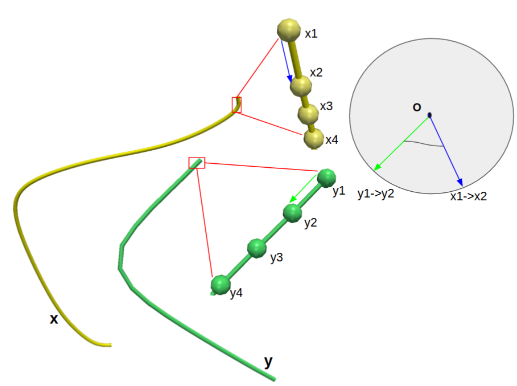

Fig. 1. Illustration for $\mathbf{L}_{GPW}$. Two streamlines x and y composed of line segments by vertices. We can measure the intersection angle between two piece-wise line segments, as α is angle between two piece-wise line segments [x1,x2] and [y1,y2]

1.3. Metric Analysis

Performance study We also conducted a simple performance study on the K-means clustering combined with our metrics us-

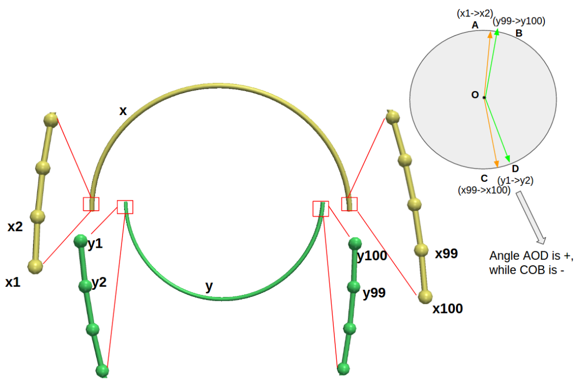

Fig. 2. Illustration for $\mathbf{L}_{SAPW}$. Two streamlines x and y composed of line segments by vertices which are similar in shape. We can measure the intersection angle between two piece-wise line segments, as $\angle AOD$ is angle between [x1,x2] and [y1,y2], and $\angle COB$ is angle between [v99,v100] and [s99,s100]. They have opposite sign as to Formula (signed) so will cancel out. Their distance under $\mathbf{L}_{SAPW}$ would be 0.

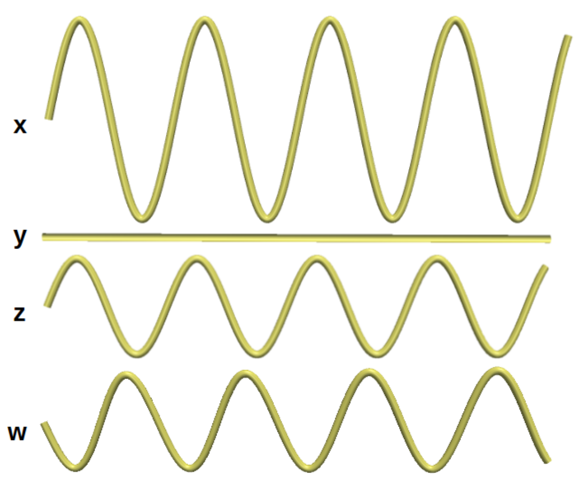

Fig. 3. Four streamlines x,y,z and w are measured to be zero distance under $\mathbf{L}_{SAPW}$, which means they're regarded as same shape. Then $\mathbf{L}_{SAPW}$ can eliminate these repeated and unimportant streamlines.

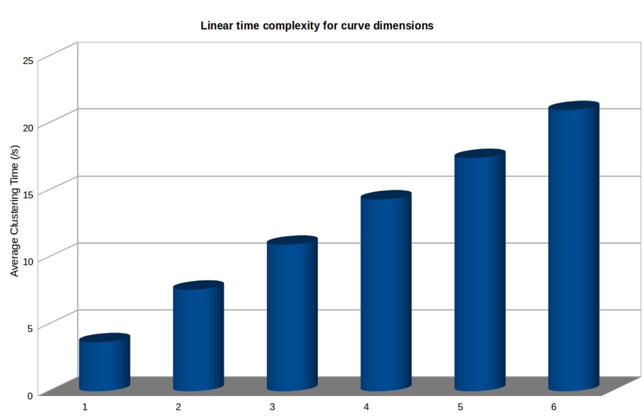

Fig. 4. Time spent for each iteration of the K-means clustering with our metrics on different dimensions of input curves of the flow behind cylinder with $\mathbf{L}_{GPW}$ metric. The preset cluster number is 120 and the maximal iterations are 20. We observe a linear increase of time w.r.t the curve dimensions.

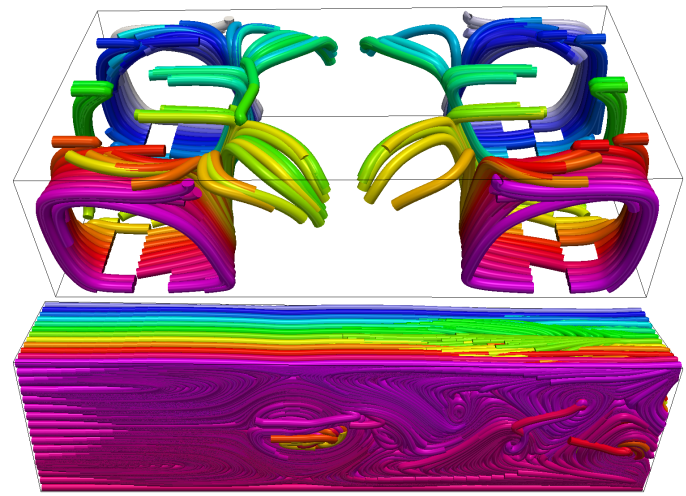

Fig. 5. Streamlines in flow datasets colormapping with increasing index Top first: sparse Bernard streamlines (256 streamlines). Top second: for flow behind cylinder (9266 streamlines).

ing the flow behind cylinder data (bottom in Figure 5). In particular, we chose $\mathbf{L}_{GPW}$ as an example for this study (other metrics have similar performance). The number of the input streamlines is 9266, and dimension of curves is increasing from 300 to 1800. Figure 4 shows the times spent on one iteration of the clustering of these increasing curve dimensions of streamlines. From this plot, we see that the time spent on the clustering increases approximately linearly w.r.t to the curve dimensions of input streamlines. .

Property study

Basic properties for different metrics designed in Section 3 of main contents are listed in Table 1. There're three important mathematical characteristics for a metric as below

- Homogeneity: It measures whether a metric is scaling-free or not, which is to testify $\mathbf{L}(a\mathbf{x}, \mathbf{y}) \equiv |a|\mathbf{L}(\mathbf{x}, \mathbf{y})$ satisfied or not, where a is a constant value.
- Tri-Inequality: It determines whether $\mathbf{L}_(\mathbf{x}, \mathbf{z}) \leq \mathbf{L}(\mathbf{x}, \mathbf{y}) + \mathbf{L}(\mathbf{y}, \mathbf{z})$ is satisfied or not.
- Definiteness: It tests whether the derivation is satisfied $\mathbf{L}(\mathbf{x}, \mathbf{y}) \equiv 0 \iff \mathbf{x} \equiv \mathbf{y}$ or not. It measures whether only two totally identical curves can have zero distance in metric or not.

Homogeneity measures scaling-free property for a defined metric. If a metric is scaling-free for curves inside the domain, e.g., $\mathbf{L}(a\mathbf{x}, \mathbf{y}) \equiv \mathbf{L}(\mathbf{x}, \mathbf{y})$, then this metric should violate homogeneity. Homogeneity is essentially spatially related, so most of spatial metrics, i.e., $\mathbf{L}_{Euc}$ and $\mathbf{L}_{Frac}$ can conform by homogeneity, while $\mathbf{L}_{Dot}$ didn't because it has a normalization process during metric computation.

Definiteness measures spatial uniqueness for two curves under a metric. If it's satisfied, topological compactness along with tri-inequality can be applied in the metric space so that mature mathematical concepts would be applied directly here.

However, since we study for curve similarity, definiteness could be violated easily because two separate curves should have zero distance in dissimilarity metric.

The most arguable property of metric goes to tri-inequality. For the tri-inequality, all spatial and geometric metrics satisfy this property except $\mathbf{L}_{Frac}$. Spatial metrics and geometric metrics are based on either Euclidean distance or intersection angle, but $\mathbf{L}_{Frac}$ exponentially increases contrast which violates the tri-inequality. Besides, all statistic-related metrics do not obey the tri-inequality, as also mentioned in [15] about the Bhattacharyya metric.

Proof of tri-inequality for geometric metrics would be very straightforward. Since they're defined on average of piece-wise parallelism or intersection angle, we simplify the proof to be the tri-inequality of single segment parallelism, i.e., $\arccos(\mathbf{x}_i, \mathbf{y}_i)$, which is definitely obeying tri-inequality. Hence, $\mathbf{L}_{GPW}$ satisfies triangle inequality. The same goes for $\mathbf{L}_{SAPW}$, where better situation is that left side of inequality eliminates piece-wise angle, and worse situation is right side eliminates piece-wise intersection angle but still at most the inequality equals to right. However, it's not easy to judge $\mathbf{L}_{GPWD}$ because it has an amplifier of standard deviation. We estimate it with doubt to obey triangle-inequality by deductions on extreme conditions.

Table 1 does provide a pragmatic guideline for metric selection based on the property that needs to be emphasized in a specific application setting. For example, if streamlines with strong tortuosity (i.e. with certain helical behavior) are of interest without considering different types of rotation, then the $\mathbf{L}_{SAPW}$ is preferred as an ideal metric.

1.4. *Metric-based Experiment*

1.4.1. *Bernard Dataset*

Figure 8 demonstrates the metric-based clustering and extraction result on simplified Bernard. We chosen k-means intial controids as random curves in the spatial domain, and preset cluster number to be 20. The final clusters are often smaller than 20 due to some centroid having no curves assigned.

We first tested our metrics on the Bernard data. As shown in Figure 8, $\mathbf{L}_{PCA}$ (Row 2), $\mathbf{L}_{Frac}$ (Row 3), $\mathbf{L}_{GPW}$ (Row 5) and $\mathbf{L}_{SAPW}$ (Row 6) are able to approximately preserve the two vortex structures. Among these four metrics, $\mathbf{L}_{GPW}$ (Row 5) and $\mathbf{L}_{SAPW}$ (Row 6) could best preserve the vortex structures using the closest streamlines (middle column), while $\mathbf{L}_{PCA}$ (Row 2) and $\mathbf{L}_{Frac}$ (Row 3) are better in terms of furthest streamlines (right column). Entropy of different metrics is shown in Figure 6. Among all metrics, the results produced with $\mathbf{L}_{PCA}$ ranks the highest, followed by $\mathbf{L}_{Frac}$ and $\mathbf{L}_{GPW}$. As we indicated in Table 1, statistic-based metrics may not work well for small-size curves, and their corresponding entropy values also show that most of them are zeros. This is mostly because small samples cannot provide sufficient variance of an effective and accurate statistics-based distance computation. Therefore, we can conclude based on this experiment that if the input set of curves is sparse enough, we should consider spatial-based metrics over the statistics-based metrics. Also, our geometric metric $\mathbf{L}_{GPW}$ and spatial metric $\mathbf{L}_{Frac}$ has relatively close entropy value to $\mathbf{L}_{PCA}$, while $\mathbf{L}_{GPW}$ preserves better vortex-similar structures in closest trajectories.

1.4.2. *PBF Two-Half-Merging*

The second PBF simulation data we experiment with is the two-half-merging scenario. $300K$ trajectories with 250 points for each trajectory are used. The clustering results and their corresponding representative trajectories are shown in Figure 10 . From the results, we see that in general our $\mathbf{L}_{Frac}$ (row 3) is able to provide the same or similar representation of the general particle trends as $\mathbf{L}_{PCA}$ (row 2). In contrast, our geometric metrics (rows 4, 5 and 6) focus more on the details and tortuous trajectory extraction. For example, the representative trajectories that are furthest away from the centroids using $\mathbf{L}_{GPW}$ (row 4) and $\mathbf{L}_{GPWD}$ (row 6) have lots of small-size winding trajectories, while the trajectories closest to the centroids of the same metrics focus on longer trajectories. Our geometric metrics identify these short and tortuous trajectories as potential features, which again to some extend demonstrate that these metrics favor the tortusity of the trajectories.

The entropy values of the clustering results with different metrics for this dataset is shown in Figure 2 of main content, and our two geometric metrics (highlighted in red) are ranked top two, which supports our observation above.

Also we notice that our $\mathbf{L}_{Frac}$ performs as well as $\mathbf{L}_{PCA}$, as already seen in previous datasets. The preference of entropy on geometric metrics is that they are able to extract small but tortuous trajectories as features inside particle-based datasets, as rows 4 and 6 in Figure 10, especially trajectories furthest away from centroids have much more tortuous behaviors than the other metrics.

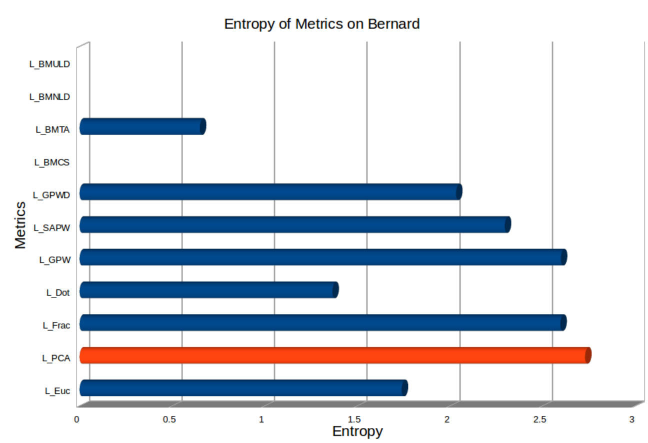

Fig. 6. Entropy values for metrics on Bernard datasets, and $\mathbf{L}_{PCA}$ is ranked the highest (highlighted by red).

References

[1] Yadav, J, Sharma, M. A review of k-means algorithm. International Journal of Engineering Trends and Technology (IJETT) 2013;4.

[2] K-means++: the advantages of careful seeding. SODA '07; 2007. ISBN 978-0-898716-24-5.

[3] Sampath, S, Ramya, B. Rough isodata algorithm. Int J Fuzzy Syst Appl 2013;3(4):1–14.

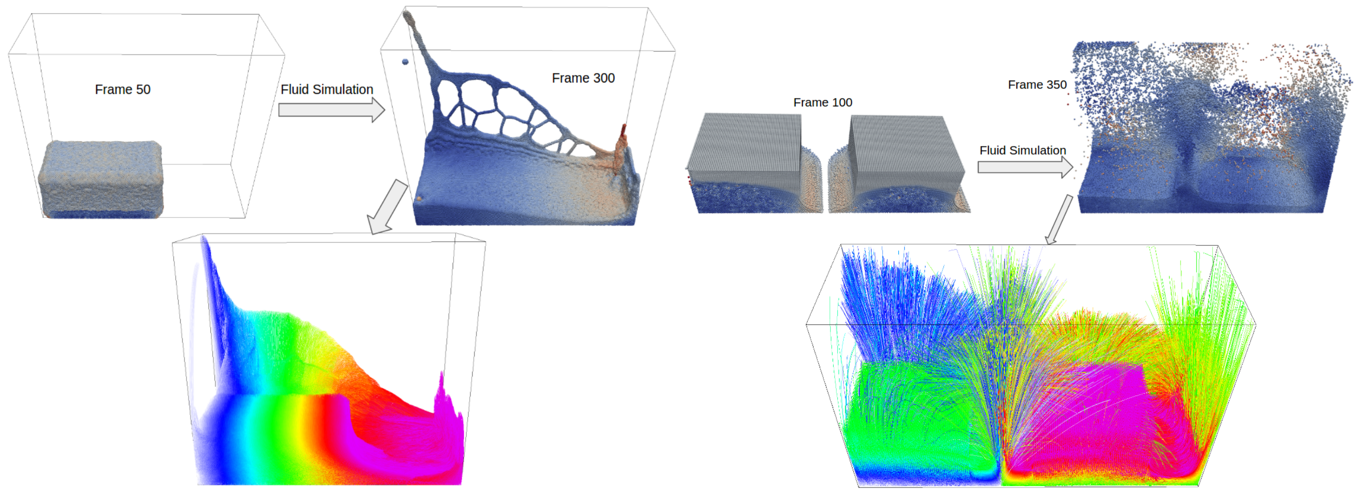

Fig. 7. Particle trajectories from position-based fluid simulations. Left: dam-breaking ($128K$ trajectories within frame 50 and 300) with signed-distance boundary handling. Right: two-half-merging ($300K$ trajectories within frame 100 and 350) with boundary-particle handling [16] and free-slip condition [17]. Both time step sizes are $0.016s$ as required in [18].

[4] Filippone, M, Camastra, F, Masulli, F, Rovetta, S. A survey of kernel and spectral methods for clustering. Pattern Recognition 2008;41(1):176 – 190.

[5] Blei, DM. Hierarchical clustering. `http://www.cs.princeton.edu/courses/archive/spr08/cos424/slides/clustering-2.pdf`; 2008. [Online; accessed 18-April-2017].

[6] Hadjighasem, A, Karrasch, D, Teramoto, H, Haller, G. Spectral-clustering approach to lagrangian vortex detection. Phys Rev E 2016;93:063107. URL: `https://link.aps.org/doi/10.1103/PhysRevE.93.063107`. doi:10.1103/PhysRevE.93.063107.

[7] Chennubhotla, SC. Spectral methods for multi-scale feature extraction and data clustering. Ph.D. thesis; Toronto, Ont., Canada, Canada; 2004. AAINQ91672.

[8] Zare, H, Shooshtari, P, Gupta, A, Brinkman, RR. Data reduction for spectral clustering to analyze high throughput flow cytometry. BMC Bioinformatics 2010;11.

[9] Giforos, J, Longmire, E, Marusic, I, Hutchins, N, Papanikolopoulos, N. Spectral clustering for identifying coherent eddy structure in turbulent boundary layers. In: International Conference on Intelligent Symstems and Computing: Theory and Applications. 2006,.

[10] Shi, J, Malik, J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 2000;22(8):888–905.

[11] Oeltze, S, Lehmann, DJ, Theisel, H, Preim, B. Evaluation of streamline clustering techniques for blood flow data. 2012.

[12] Ferstl, F, Brger, K, Westermann, R. Streamline variability plots for characterizing the uncertainty in vector field ensembles. IEEE Transactions on Visualization and Computer Graphics 2016;22:767–776.

[13] On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. ICDT '01; 2001. ISBN 3-540-41456-8.

[14] Dinov, ID, Christou, N, Gould, R. Law of large numbers: the theory, applications and technology-based education. J Stat Educ 2009;17:1–19.

[15] Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of the Calcutta Mathematical Society 1943;35:99–109.

[16] Akinci, N, Ihmsen, M, Akinci, G, Solenthaler, B, Teschner, M. Versatile rigid-fluid coupling for incompressible sph. ACM Trans Graph 2012;31(4):62:1–62:8. URL: `http://doi.acm.org/10.1145/2185520.2185558`. doi:10.1145/2185520.2185558.

[17] Kang, N, Sagong, D. Incompressible sph using the divergence-free condition. Computer Graphics Forum 2014;33(7):219–228. URL: `http://dx.doi.org/10.1111/cgf.12490`. doi:10.1111/cgf.12490.

[18] Macklin, M, Müller, M. Position based fluids. ACM Transactions on Graphics (TOG)-SIGGRAPH 2013 2013;32.

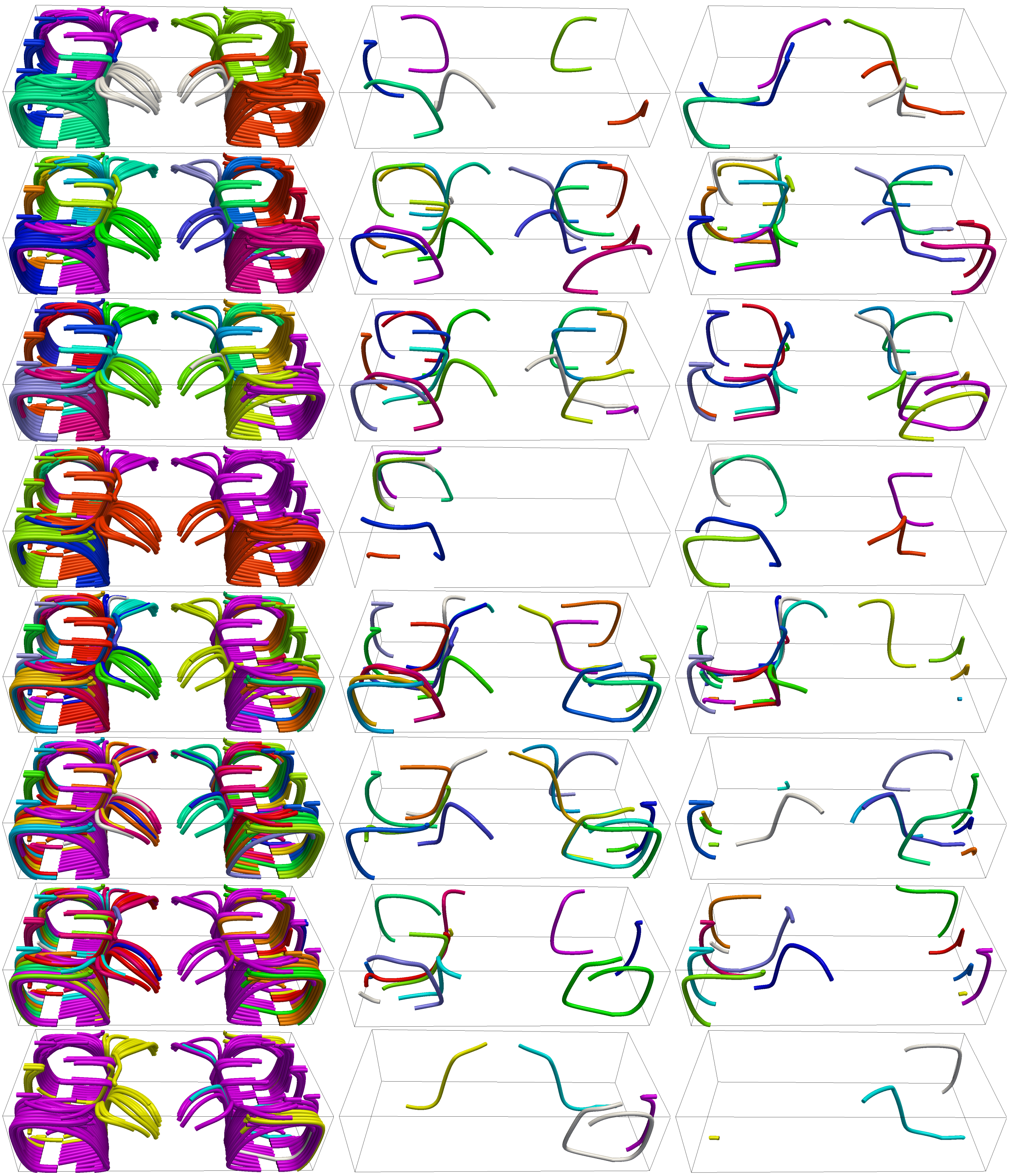

Fig. 8. Metric evaluation on simplified Bernard with the K-means clustering. Several metrics cannot extract more than 2 clusters, so we leave them out. From top to bottom, results with L_{Euc}, L_{PCA}, L_{Frac}, L_{Dot}, L_{GPW}, L_{SAPW}, L_{GPWD} and L_{BMTA} are shown, respectively. From left to right of each row, grouping streamlines, closest streamlines to centroid, and farthest streamlines to centroid are visualized, respectively. Streamlines are color coded based on which clusters they belong to. From visual comparisons, we see that L_{PCA} (Row 2), L_{Frac} (Row 3), L_{GPW} (Row 5) and L_{SAPW} (Row 6) work better than the other metrics.

Table 1. Brief reviews on properties of our distance metrics defined in Section 3 Metric Designing

Group	Metric	Homogeneity	Tri-Inequality	Definiteness	Remarks	Common characteristics
Spatial	$\mathbf{L}\_Euc$	Yes	Yes	Yes	Strictly hyper-spherical structures in higher-dimension spaces.	Translation and rotation sensitive. Scaling-sensitive. Not similarity in shape. Tells how close two curves are.
	$\mathbf{L}_{PCA}$	Yes	Yes	No	More then O(d). SVD for larger-scale is super expensive.	
	$\mathbf{L}_{Frac}*$	Yes	No	Yes	Easy for overflow. Parameter-sensitive. Increase contrast in higher-dimension space.	
	$\mathbf{L}_{Dot}*$	No	Yes	No	Weak in spatial measurement.	
Geometric	$\mathbf{L}_{GPW}*$	No	Yes	No		Translation-invariant and spatial-independent. Scaling-free. Only dissimilarity in shape. Tells how parallel two curves are.
	$\mathbf{L}_{SAPW}*$	No	Yes	No	Rotation-invariant. Eliminate cosine-similar curve to straight line.	
	$\mathbf{L}_{GPWD}*$	No	Yes	No	Somehow rotation-invariant. Can group curves with same curvature distribution.	
Statistic	$\mathbf{L}_{BMCS}*$	No	No	No	Sensitive to first line segment direction. Can't distinct curves with similar curvature trends. Curvature-oriented	Translation-invariant and spatial-independent. Scaling-free and order-free. Only measures distribution trends. Not working well especially for small-size curves faster in matrix operation
	$\mathbf{L}_{BMTA}*$	No	No	No	Can't distinct curves which winds an axis at same angles. Intersection-angle based	
	$\mathbf{L}_{BMNLD}*$	No	No	No	Only concerning directions.	
	$\mathbf{L}_{BMULD}*$	No	No	No	Convey length information compared to $\mathbf{L}_{BMND}$	

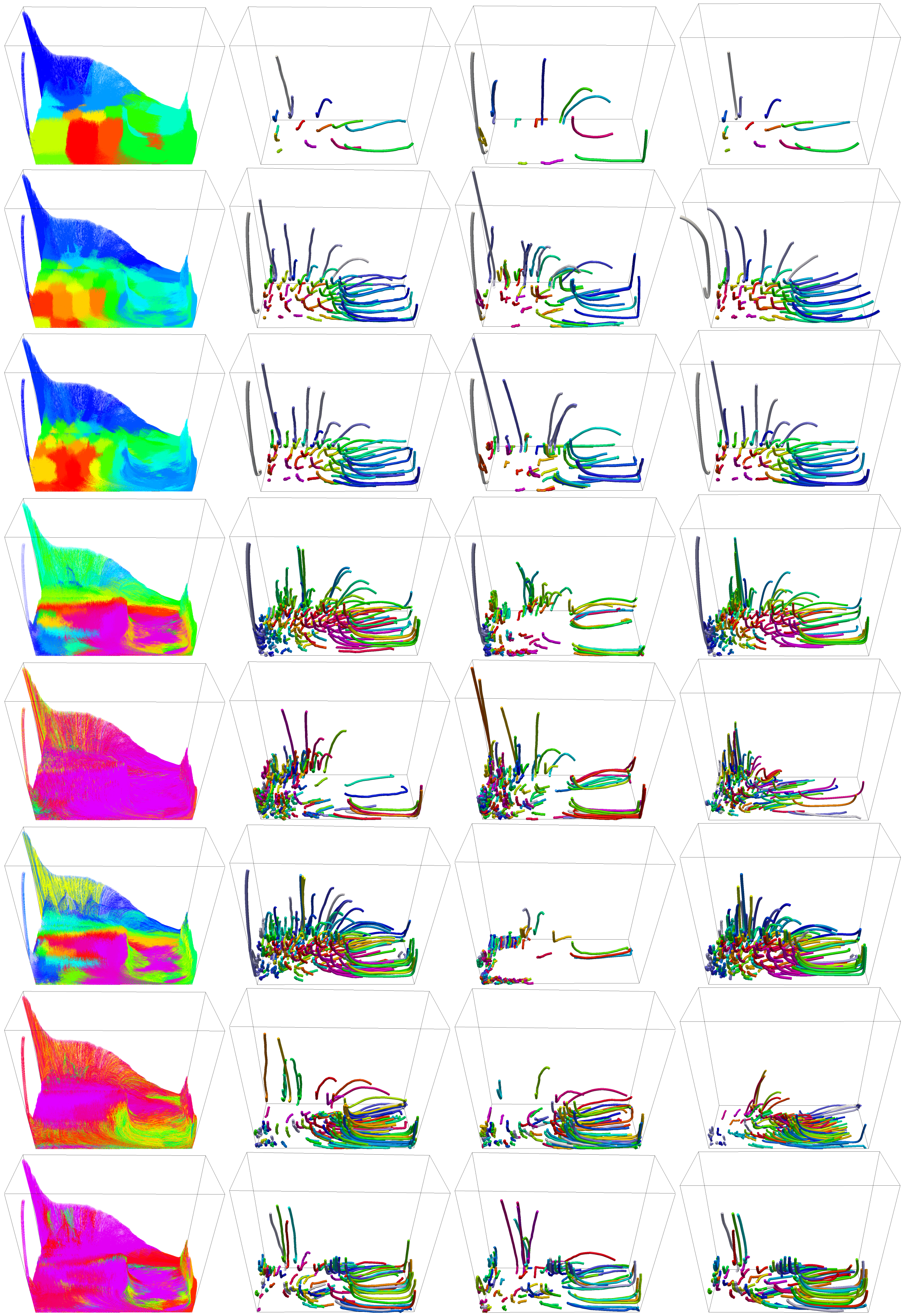

Fig. 9. Metric evaluation on the dam-breaking data with the K-means clustering. Several metrics cannot extract more than 10 clusters, so we leave them out. From top to bottom, results with L_{Euc}, L_{PCA}, L_{Frac}, L_{GPW}, L_{SAPW}, L_{GPWD}, L_{BMTA} and L_{BMNLD} are shown, respectively. Left to right of each row, grouping trajectory, trajectories closest, furthest away from centroid and centroid of each cluster are shown, respectively. In all three representative trajectory identification, L_{GPW} (row 4) works the best with the latter having more details. The other metrics have more or less omitted trajectories of importance.

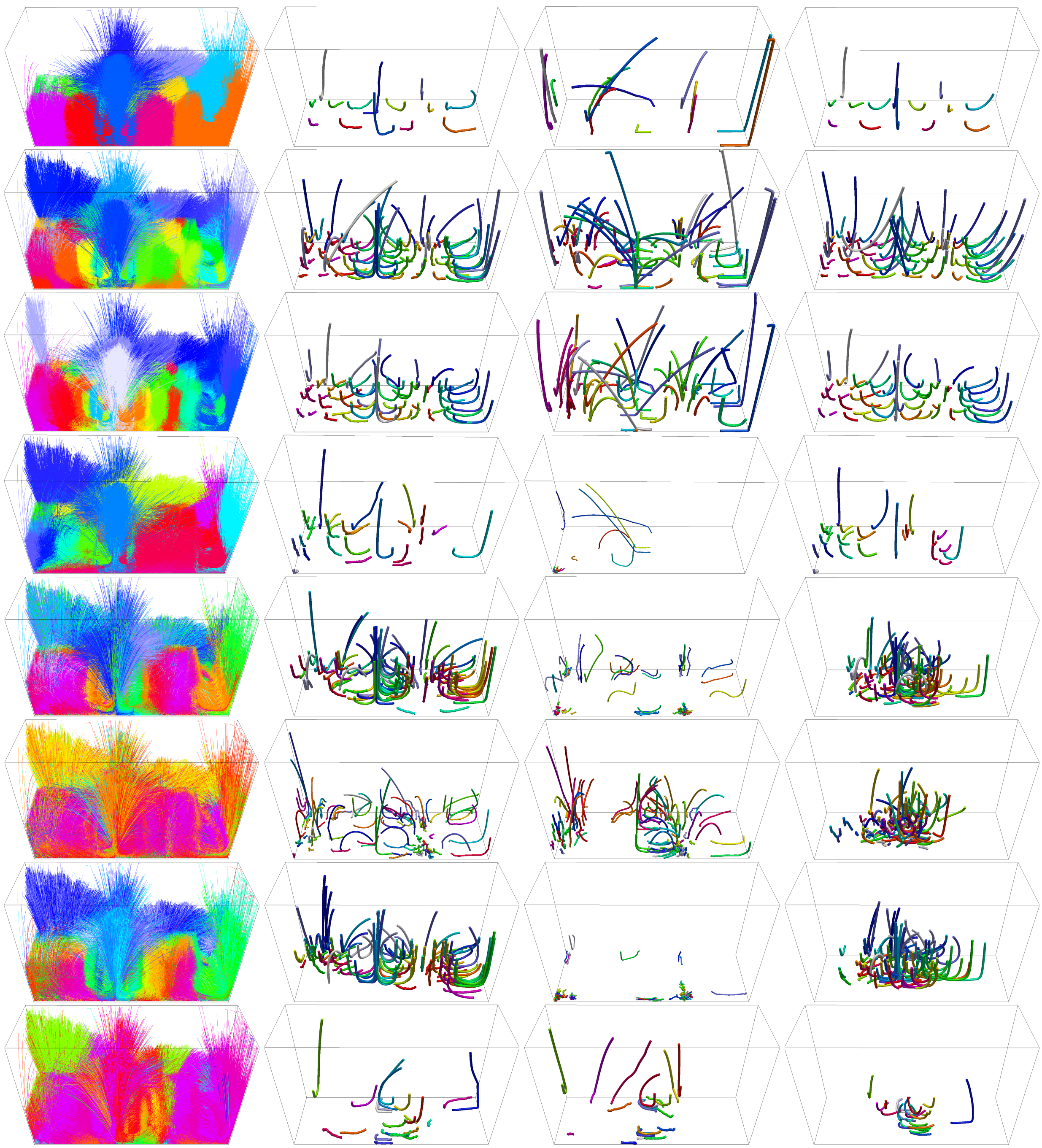

Fig. 10. Metric evaluation on the two-half-merging data with the K-means. Several metrics cannot extract more than 10 clusters, so we leave them out. From top to bottom, the results with L_{Euc}, L_{PCA}, L_{Frac}, L_{Dot}, L_{GPW}, L_{SAPW}, L_{GPWD} and L_{BMTA} are shown, respectively. Left to right of each row, grouping trajectories, trajectories closest to centroids, trajectories furthest away from centroids and centroid of each cluster are shown, respectively. L_{PCA} (row 2) and L_{Frac} (row 3) work the best which enables the visualization representing the approximate structure of trajectories, while our geometric metrics (row 3, 4 and 5) tend to extract small-size tortuous trajectories due to their more emphasis on geometric shape.