Downloaded from SAE International by Jiajin Wu, Sunday, January 02, 2022

2021-01-0425 Published 06 Apr 2021

Hybridizing Unsupervised Clustering Methods for In-Cylinder Vortex Motion Analysis under Different Swirl Ratio Conditions

Fengnian Zhao, Mengqi Liu, Weihan Fan, Jiajin Wu, Junxiang Zhang, and David Hung UM-SJTU JI - Shanghai Jiao Tong University

Citation: Zhao, F., Liu, M., Fan, W., Wu, J. et al., "Hybridizing Unsupervised Clustering Methods for In-Cylinder Vortex Motion Analysis under Different Swirl Ratio Conditions," SAE Technical Paper 2021-01-0425, 2021, doi:10.4271/2021-01-0425.

Abstract

Large-scale vortex motion could enhance the fuel-air mixing and the combustion stability inside a direct-injection engine. For in-cylinder vortex motion analysis, detection of the vortex features is usually a challenging task because of large cyclic variations of vortex structure, number, and locations. In previous study, K-means clustering has been successfully applied to detect the vortex features and quantify the cyclic variations. However, K-means algorithm is somewhat limited in analyzing the complex flow with multiple vortex zones and vortex merging. Therefore, we propose a hybrid clustering method which blends in two clustering algorithms, namely, Gaussian mixture model (GMM) clustering and K-means clustering, to optimize the analysis of vortex classification, vortex zone detection, and cyclic variation quantification under different flow conditions. In this study, in-cylinder flow fields with varying degrees of swirling behavior were recorded by high-speed particle image velocimetry (PIV) under high and low swirl ratio conditions. Owing to the probability density contour feature of GMM, the clustering results show that the hybrid clustering method improves the accuracy of vortex classification and vortex zone detection, especially for low swirl cases where there exist large cluster overlaps. Additionally, this hybrid clustering method is able to retain the quantification of cyclic variations by blending K-means clustering to quantify the distance. In summary, the transient features and cyclic variations of the in-cylinder vortex motion under different swirl ratio conditions can be accurately revealed by this hybrid clustering method.

Introduction

In-cylinder flow motion and its cyclic variations affect the fuel-air mixing and combustion stability inside a gasoline direct-injection (GDI) engine. Strong swirling flow could induce large-scale vortex motion and reduce spatiotemporal variations of the in-cylinder flow structures such that the fuel-air mixing and combustion stability could be effectively enhanced [1, 2, 3]. With these benefits of in-cylinder vortex motion, it is necessary to acquire a complete understanding of the in-cylinder swirling flow phenomena.

To accurately capture the in-cylinder swirling flow behavior, temporally-resolved flow field measurements can be collected over massive engine cycles using high-speed particle image velocimetry (PIV), which allow researchers to study the ensemble swirling flow of multiple cycles and analyze the cycle-to-cycle variations by statistical and decomposition methods [4, 5, 6, 7, 8]. However, temporal averaging approach does not provide accurate flow information if the flow is highly transient with strong cyclic variations. The enormous flow field datasets could easily make the statistical and decomposition analysis computationally expensive [9,10]. Recently, advancement of machine learning in big-data analysis allows more researchers to adopt data-driven algorithms for analyzing large engine datasets [11, 12, 13, 14, 15]. For instance, previous work of Zhao et al. [11] utilized the K-means clustering, which is an unsupervised clustering method, to detect the time-resolved vortex patterns and cyclic variations. It was found that K-means clustering could cluster the vortex centers into different zones automatically and quantify their cyclic variations. K-means algorithm was proven to outperform temporal averaging analysis by minimizing the issue of vortex misidentification.

Owing to its relatively simple and efficient implementation, K-means clustering method has been widely used in many aspects for pattern recognitions, classifications, and data predictions among others [16, 17, 18]. The idea of K-means clustering is to generate K centroids for K clusters. These centroids should be located at the farthest distance between each other. In previous study [11], K-means clustering is accurate and efficient since the in-cylinder flow is under high swirl ratio condition where the vortex regions are distinct with small number of clusters. However, the clustering accuracy of K-means algorithm could reduce if the cluster number and/or the overlaps of clusters increase [19, 20]. For typical in-cylinder flow during compression stroke, the intake valves are mostly closed which lead to more stable flow fields with repeatable vortex structure. On the contrary, during intake stoke where the intake valves are mostly open, the flow fields are much more complicated where multiple vortices and vortex merging exist at different durations. Moreover, under

Downloaded from SAE International by Jiajin Wu, Sunday, January 02, 2022

different swirl conditions, the bulk flow structure and vortex motion could be totally different. High swirl flows induce highly regulated flow structures while low swirl flows exhibit much more random motions with merging vortex interactions. Additionally, the cyclic variation of vortex features could be high even under the same swirl ratio condition. Therefore, the capability of K-means clustering on complicated in-cylinder flow fields must be further improved.

To overcome these limitations of K-means clustering, Gaussian mixture model (GMM) clustering, which is a soft clustering method, is introduced to hybridize the K-Means clustering method previously demonstrated [11]. Unlike K-means clustering which partitions the observations based on similarity distances, GMM clustering uses a distribution-based method. The basic idea of distribution-based clustering is that the data generated from the same distribution would be assigned to the same cluster if the entire dataset is represented by a parametric distribution or by a mixture of these distributions [21, 22]. Specifically, for GMM, the data obeying the same independent Gaussian distribution is considered to belong to the same cluster. Thus, GMM clustering provides a rigorous framework to access the parameter such as probability distribution in the clustering process. With the use of probability distributions, GMM clustering is well adopted for pattern identification and image classification when the original dataset is with larger complexity [23, 24]. Therefore, GMM is more realistic to give the probability of belongings comparing with K-means clustering. For in-cylinder vortex motion analysis in this work, it appears that the vortex center locations extracted from multiple engine cycles follow closely the Gaussian distributions. Thus, GMM clustering can effectively improve the shortcomings of K-means clustering. However, K-means can still be used to determine the centroid and quantify the distance of the data points within each cluster which cannot be done by GMM. Therefore, a hybrid clustering scheme combining GMM and K-means is implemented to improve the overall clustering performance and accuracy regarding complicated in-cylinder flow data.

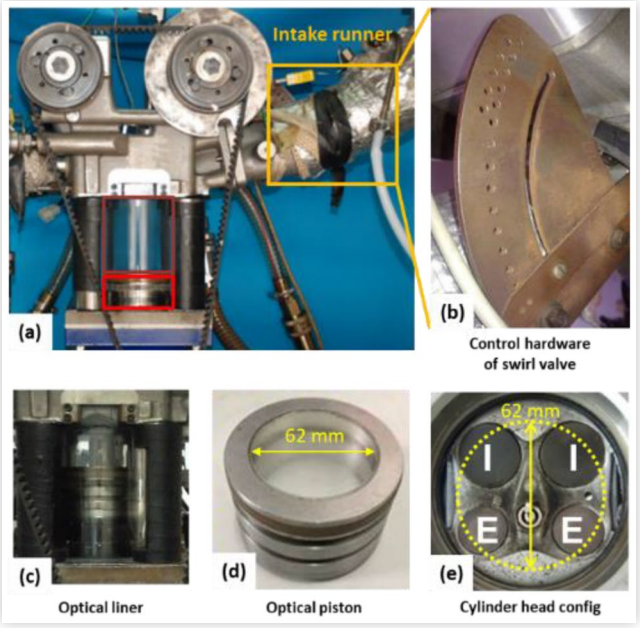
In summary, this research focuses on improving the clustering algorithm for in-cylinder flow analysis under different swirl conditions by revealing the vortex motion features with a hybrid clustering algorithm. In this paper, two datasets of in-cylinder flow fields under high and low swirl ratio conditions were measured by high-speed PIV from early intake stroke to late compression stroke at every 2 crank angle degrees (CAD) for 100 consecutive cycles. A hybrid method blending the GMM clustering and K-means clustering is implemented to detect the in-cylinder vortex characteristics by vortex classification, vortex zone detection, and cyclic variations quantification. In the sections below, the experimental setup is introduced first, followed by the detailed explanations of the hybrid clustering method. In the results and discussion section, the datasets are divided into three parts: 1) a dataset description for both high-swirl and low-swirl flow fields; then for clustering analysis (vortex classification, vortex zone detection and cyclic variation quantification) according to the complexity of flow fields, 2) high swirl flow data analysis; and 3) low swirl flow data analysis. In the end, conclusions are drawn according to the clustering results to highlight the benefits of this hybrid method.

Experimental Setup

The PIV measurements were carried out in a four-stroke single cylinder optical GDI engine with four valves. Figure 1 shows the main components of the optical engine including swirl control valve, optical piston, optical liner, and cylinder head configuration.

The swirl control valve was installed in one of the two intake ports. Figure 1b shows the control hardware of the swirl valve. There are 20 preset positions to control the swirl valve from opening to closing for swirl ratio adjustments. The intake port without the swirl control valve is the "primary intake port" and the other one is the "secondary intake port". High swirl ratio can be achieved by closing the secondary intake port completely. When the secondary intake port is open, intake air enters from both intake ports and generates a lower swirl ratio. The swirl ratio of this engine was measured on a cylinder head flow bench [8]. In this work, flow data of high swirl ratio of 4.41 and low swirl ratio of 2.29 was selected for analysis. Figure 1c shows the quartz liner, which provides the optical access for laser sheet to illuminate a specific plane for PIV measurements. The optical piston in Figure 1d allows the high-speed camera to focus on the specific location to record the flow fields along the swirl plane. The engine cylinder head configuration with two intake valves and exhaust valves can be visualized through the optical piston. The engine speed was motored by the AVL AC dynamometer at 800 revolutions per minute (RPM). Since the laser sheet could be blocked by the piston assembly, the range of crank angle recorded by PIV is from –300 CAD after top dead center (ATDC) to –60 CAD ATDC, which covered the time period from early intake stroke to late compression stroke. Table 1 shows the key parameters of the engine and operating parameters.

FIGURE 1 (a) Optical engine; (b) control hardware of swirl valve; (c) optical liner; (d) optical piston; (e) cylinder head configuration [11].



© SAE International.

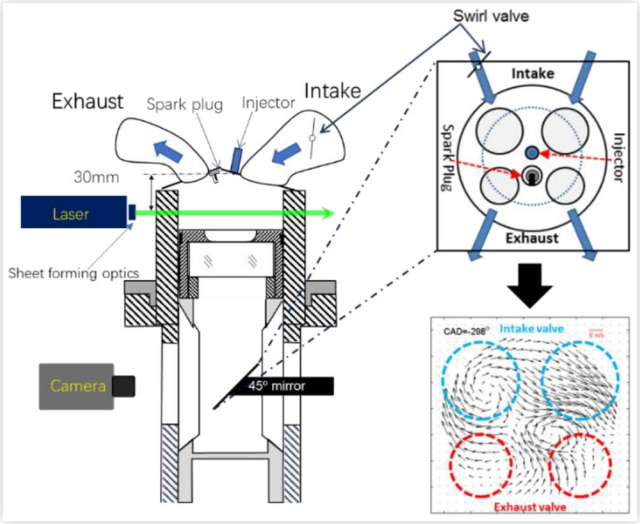
Downloaded from SAE International by Jiajin Wu, Sunday, January 02, 2022

TABLE 1 Engine and operating parameters.

Parameter	Value
Engine speed	800 RPM
Bore	86 mm
Stroke	94.6 mm
Clearance volume	49.96 cm^3
Displacement	549.51 cm^3
Compression ratio	12:1
Manifold absolute pressure	100 kPa
Intake swirl ratios	4.41 & 2.29
Number of valves	4
Intake valve open (IVO)	-370 CAD ATDC
Intake valve close (IVC)	-123 CAD ATDC

© SAE International.

FIGURE 2 Setup of PIV measurements with view of swirl plane.



© SAE International.

Figure 2 depicts the schematic of the PIV setup and a sample flow field data. A high-speed Nd:YLF laser (Litron LDY303HE) with 527 nm wavelength was used as the light source to illuminate the seeding particles. The laser sheet illuminated the plane 30 mm below the injector tip such that the flow field data was measured along a selected swirl plane which is parallel with the piston top. Through a 45-degree mirror, the high-speed camera (Phantom V7.3) was used to take particle images inside the cylinder. The particle images were processed into velocity fields using the commercial code of LaVision Davis 8.4. The circled regions in the velocity field represent the valve configurations where the diameter of the optical insert is 62 mm. Table 2 summarizes the detailed parameters of PIV measurements.

Analysis Approach

To improve the clustering accuracy of vortex motions and quantify cyclic variation, a hybrid clustering method combining GMM and K-means is implemented on two datasets with high & low swirl ratio conditions. The in-cylinder flow characteristics include vortex classification, vortex zone detection, and cyclic variation quantification. The source code for K-means and GMM clustering is adopted from the MATLAB Statistics and Machine Learning Toolbox [25]. The following sections describe the basic formulations of the clustering schemes.

TABLE 2 PIV measurement parameters.

Parameter	Value
Particle image size	3 ~ 4 pixels / particle
Interrogation size and overlap	32 × 32 pixels and 50%
Particle image density / interrogation window	10 ~ 15 particles
Laser sheet thickness	1 mm
Seeding particle diameter	1 μm
Time interval between two pulses	20 μs
Frame resolution	1 pair of frames / 2 CAD

© SAE International.

K-means Clustering

Before applying K-mean clustering method, the vortex center dataset is first determined by vortex center index and vorticity maps [9]. In this work, the K-means algorithm for the vortex motion analysis is the same as in a previous study [11]. K-means clustering first initializes the K centroids. Then, each data point is assigned to its closest centroid which leads to a re-calculation of a new centroid. The distance between the data point and centroid is defined as the sum of squared Euclidean distance (SSE), which is the objective function to be minimized:

$$\mathrm{SSE} = \sum_{i=1}^{n} \left\| p_i - c_j \right\|^2 \tag{1}$$

where data points are $P = \{p_1, p_2, \ldots, p_n\}$, c_j is the closest centroid of the data point p_i, $\|p_i - c_j\|$ is the Euclidean distance. After the new centroid position is determined, the data point assignment and centroid re-calculation will continue for a predefined number of iterations or until there is no change in the centroid position. In particular, determining the K value for K-means clustering is usually a tricky task since the K value determines how many clusters should be classified. In this work, the K value selection for vortex center clustering is based upon gap criterion [26]. The detailed explanation of the K value selection procedure can be referred to a previous work [11].

K-means clustering is a distance-based method. With the centroid calculated by K-means algorithm, the average distance in previous work [11] can be used to quantitatively reflect the cyclic variations of the vortex motion inside the engine. The average distance is defined as follows:

$$\text{Average distance}(j) = \frac{\sum_{p_i \in S_j} \left\| p_i - c_j \right\|}{n_j} \tag{2}$$

Downloaded from SAE International by Jiajin Wu, Sunday, January 02, 2022

Within a cluster j, $\|p_i - c_j\|$ is the Euclidean distance. The data set $P = \{p_1, p_2, p_3, \ldots, p_n\}$ is partitioned into K clusters $S = \{S_1, S_2, S_3, \ldots, S_K\}$ where $S_m \bigcap S_n = \emptyset$, for $1 \le m \ne n \le K$. p_i is the data point in cluster S_j and c_j is the centroid. A larger average distance represents higher cyclic variations of vortex locations.

GMM Clustering

GMM is the expansion of single Gaussian model, which applies the superposition of several independent Gaussian sub-models with certain parameters to describe the observed distribution. The probability density function of a Gaussian distribution can be represented as follows:

$$N\left(X|\mu,\Sigma\right) = \frac{1}{\left(2\pi\right)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} exp\left(-\frac{\left(X-\mu\right)^T \Sigma^{-1}\left(X-\mu\right)}{2}\right) \quad (3)$$

where $\boldsymbol{X}$ is the vector of D dimensions, μ and Σ are the mean and covariance matrix, respectively. GMM combines several Gaussian sub-models such that it fits the dataset by optimizing the linear combination of the sub-model results:

$$P\left(X\right) = \sum_{i=1}^{K} \alpha_i N\left(X|\mu_i,\Sigma_i\right) \quad (4)$$

where $\boldsymbol{X}$ is the data point vector, N is the Gaussian distribution, K is the number of clusters, μ_i is the mean matrix of the cluster, Σ_i is the covariance matrix, and α_i is the weight to be learnt by GMM algorithm. The sum of α_i over all K is equal to 1. The parameters to be optimized of a GMM is $\{\alpha_i, \mu_i, \Sigma_i\}$ with $1 \le i \le K$. The main idea of fitting the GMM to the dataset is the expectation-maximization (EM) algorithm, which consists of two steps: E-step and M-step. For an observation $X_j \in \boldsymbol{X}\{X_1, X_2, \ldots, X_N\}$, the E-step calculates the posteriori probability with initialed parameters $\{\alpha_k, \mu_k, \Sigma_k\}$ as follows:

$$\gamma_{jk} = \frac{\alpha_k N\left(X_j|\mu_k,\Sigma_k\right)}{\sum_{k=1}^{K} \alpha_i N\left(X_j|\mu_k,\Sigma_k\right)} \quad (5)$$

γ_{jk} is the posteriori probability of the kth Gaussian mixture component given the observation X_j. Afterwards, the M-step is conducted to update the new set of $\{\alpha_k, \mu_k, \Sigma_k\}$ by:

$$\mu_k = \frac{\sum_{j=1}^{N} \gamma_{jk} X_j}{\sum_{j=1}^{N} \gamma_{jk}} \quad (6)$$

$$\Sigma_k = \frac{\sum_{j=1}^{N} \gamma_{jk}\left(X_j - \mu_k\right)^2}{\sum_{j=1}^{N} \gamma_{jk}} \quad (7)$$

$$\alpha_k = \frac{\sum_{j=1}^{N} \gamma_{jk}}{N} \quad (8)$$

where $k = 1, 2, \ldots, K$. These two steps are iteratively computed until the convergence is reached with the following likelihood function:

$$L\left(\mu_k,\Sigma_k|X\right) = \sum_{i=1}^{N} \log\left[\sum_{k=1}^{K} \alpha_k N\left(X|\mu_i,\Sigma_i\right)\right]. \quad (9)$$

Unlike the cluster number selection of K-means algorithm, the Bayesian Information Criterion (BIC), which outperforms other criteria for GMMs in wide application domains [27], is proposed for selecting the number of clusters.

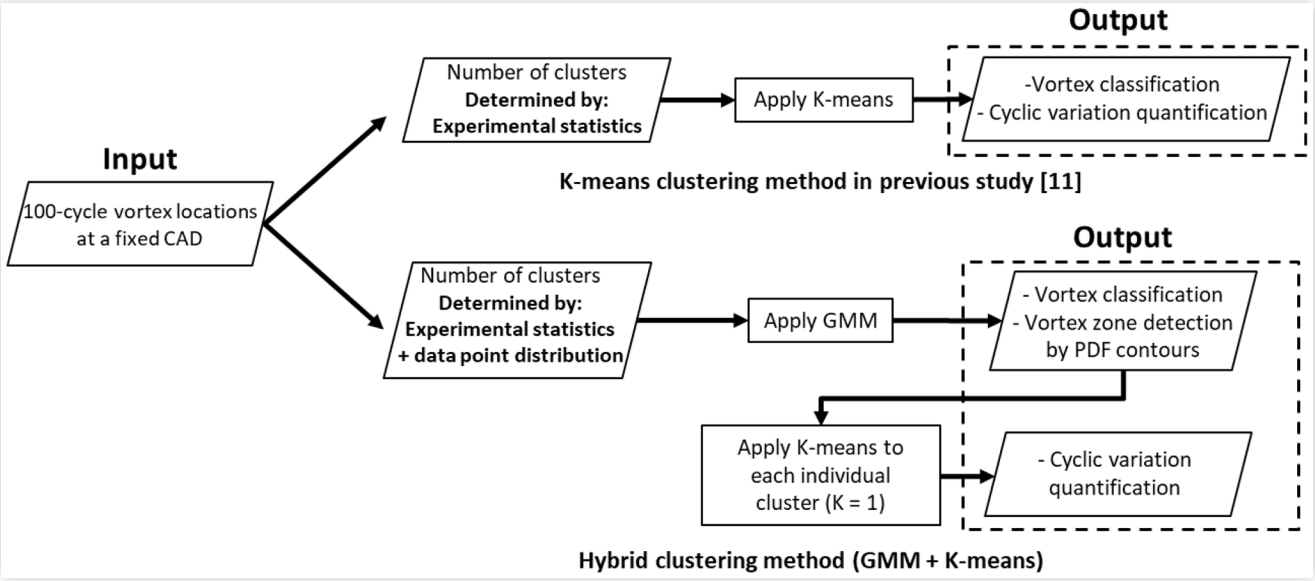
Hybridization of K-means and GMM

As discussed in previous section, while the K-means is a hard clustering method which calculates the centroids and SSE to quantify the vortex variation, GMM is a soft clustering which uses distribution-based method as clustering criterion. Therefore, the capabilities of each clustering method are combined into this hybrid clustering scheme. Figure 3 shows the flow charts of the K-means method in previous study [11] and the hybrid clustering method in this study. Basically, in the hybrid clustering approach, GMM is first utilized to assign probability to each data point and generate the probability density function (PDF) contour of the clusters for improved vortex zone detection particularly in the presence of multiple vortices and vortex merging. Then, K-means is applied for vortex variation quantifications after the determination of vortex clusters by GMM. The outputs of these two methods are highlighted in the figure.

Figure 4 illustrates the clustering results including input and output data of these two methods. The input data is extracted from the low-swirl condition. The large vortex variations make distinct differences in the clustering results of these two methods. K-means method assigns the vortex locations into one cluster and calculates a single centroid to represent the vortex location. It cannot detect the vortex zones automatically. As for the hybrid method, it clearly divides the vortex locations into two clusters and detects the vortex zones using PDF contour lines. For flow fields with large vortex variations, using vortex zone with the probability density to represent the vortex location is more accurate than using a single centroid point. It is evident that the distribution-based clustering of this hybrid method makes the vortex classification more precise and provides more meaningful clustering results of vortex zone detection than the K-means method.

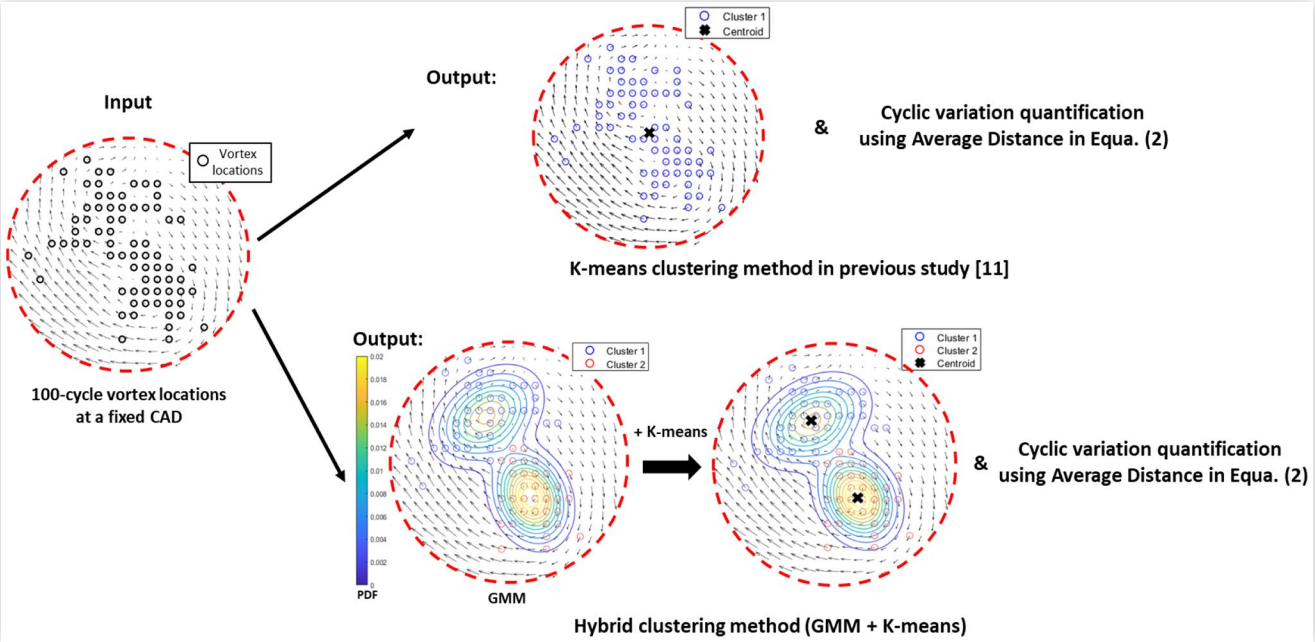
Downloaded from SAE International by Jiajin Wu, Sunday, January 02, 2022

FIGURE 3 Flow charts of the K-means clustering in previous study [11] and hybrid clustering method.



© SAE International.

FIGURE 4 Illustration of clustering results including input and output data of the K-means clustering in previous study [11] and hybrid clustering method.



© SAE International.

Results and Discussion

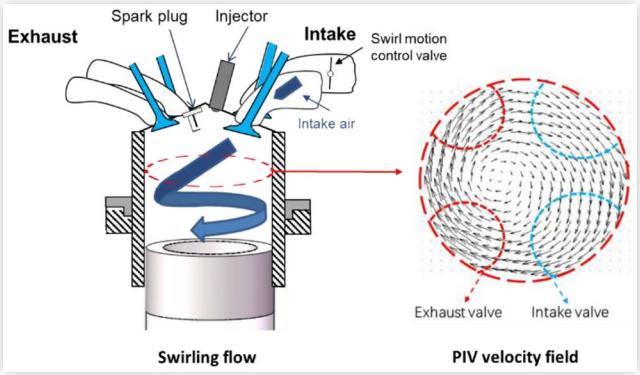
Flow data Description

Swirl motion with large-scale rotational flow behavior can be considered as vortex structure which significantly affects engine working process especially for engine idle condition with weak air charge motion. The breakdown of vortex structures at the end of the compression stroke could enhance the fuel spray mixing and increase the turbulence intensity [28]. Therefore, it is important to understand the in-cylinder flow vortex characteristics. Figure 5 shows a strong 3D swirl motion illustration inside an engine cylinder and the velocity field along a swirl measurement plane.

In this work, two swirl ratio values of 4.41 and 2.29 were generated. Figure 6 shows the ensemble flow fields and vortex motion locations of both high swirl (4.41) and low swirl (2.29) conditions from early intake to late compression. The vortex locations of 100 engine cycles at various CAD are indicated with blue and red symbols, which represent the vortex in clockwise and count-clockwise directions, respectively. It can be seen from Figure 6 that significant differences of swirl behavior happen between high swirl and low swirl conditions.

Downloaded from SAE International by Jiajin Wu, Sunday, January 02, 2022

FIGURE 5 Swirling flow inside engine and related PIV velocity field.

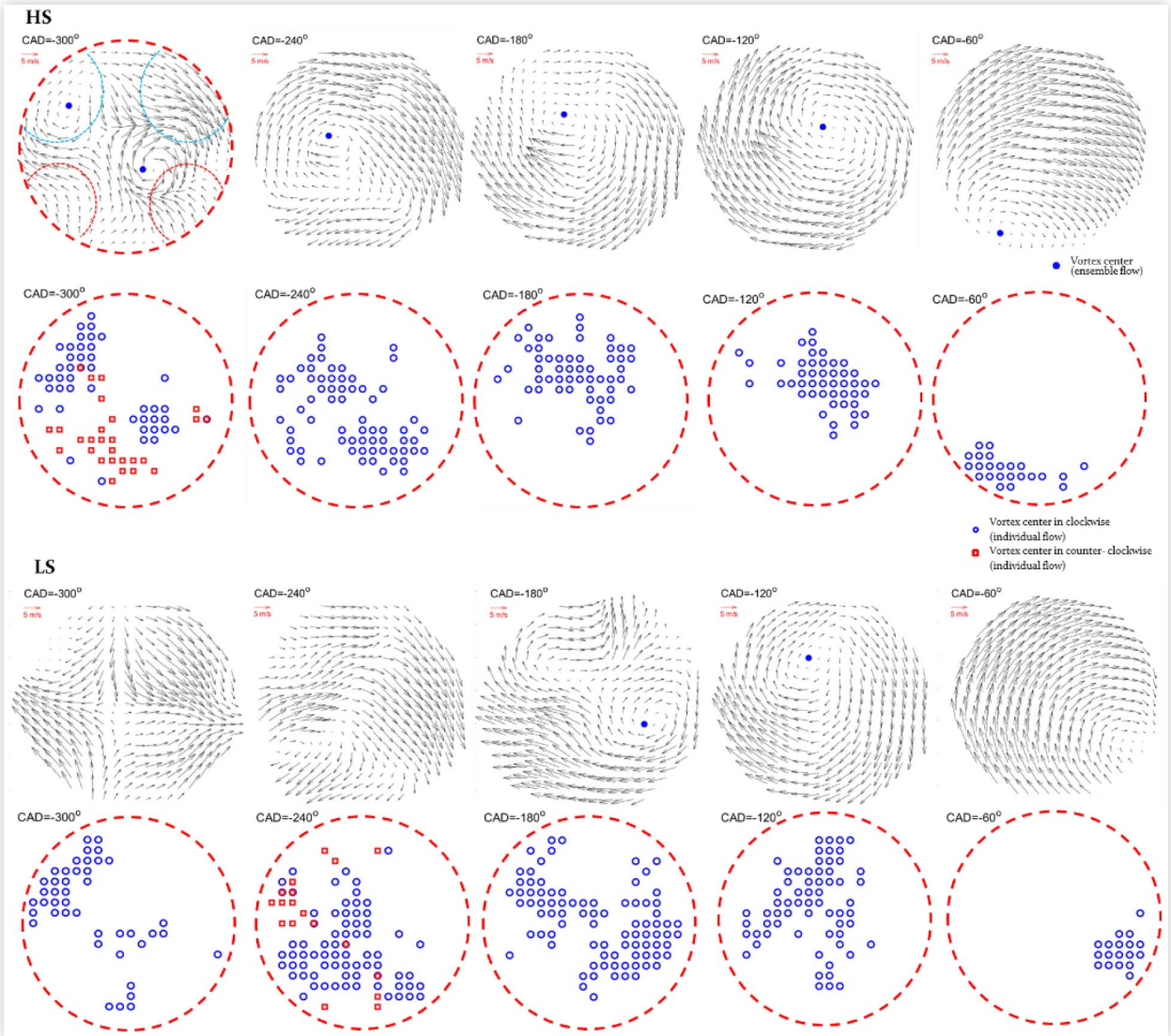


© SAE International.

Under high swirl condition, multiple vortex structures from ensemble flow fields can be clearly found from early intake stroke to late compression stroke. However, under low swirl condition, the vortex motion is much weaker. The dominant vortex motion is established until late compression under low swirl condition while the vortex motion becomes stable during intake stoke for high swirl case. However, the detailed vortex locations in individual cycles as plotted in Figure 6 are very different with the locations in ensemble flow fields. For example, under low swirl case, ensemble flow fields at –300 CAD and –240 CAD indicate no vortex motion exists in the flow fields but extensive vortex centers are detected from individual cycles. The vortex center distributions indicate that analyzing vortex characteristics only by ensemble flow is sometimes limited. Further analysis needs to be conducted.

With the vortex center locations detected from 100 consecutive engine cycles, K-means clustering method was utilized to automatically classify the vortex locations and provide more insights on the dataset features to mitigate the limitation of the ensemble flow analysis. Under high swirl condition, the vortex centers of individual cycles locate at multiple regions during intake stroke. Transitioning from intake to compression stroke, the multiple regions of vortex centers gradually become into a single dominant region. It is because the intake valve gradually closes such that the flow fields become stable without strong intake air. Although vortex locations during intake stroke are with relatively larger

FIGURE 6 Ensemble flow and vortex motion locations of both high swirl (4.41) and low swirl (2.29) conditions.



© SAE International.

Downloaded from SAE International by Jiajin Wu, Sunday, January 02, 2022

cyclic variations, these vortex center regions are still distinct. Therefore, in previous study [11], K-means algorithm was applied to detect vortex motions for a high-swirl flow case with a pre-determined cluster number from the experimental statistics. The reason for choosing K-means clustering is its simple and effective implementation and it is accurate enough for high-swirl flow data comprises distinct clusters. However, for low swirl flow fields, the vortex distributions are much more random. During early intake stroke at –300 CAD and –240 CAD, vortex centers are extensively detected in individual cycles but the vortex center regions are not as separated as those under high swirl condition. At –120 CAD, although the ensemble flow shows a similar bulk flow pattern with one dominant vortex structure as high swirl case, the vortex centers are still randomly distributed. Comparing with the flow fields under high swirl condition, the vortex features under low swirl condition are more complex with large cyclic variations. Therefore, under low swirl condition, K-means cannot be fully utilized to analyze the complex flow since the flow field contains multiple vortex zones and vortex merging (cluster overlapping).

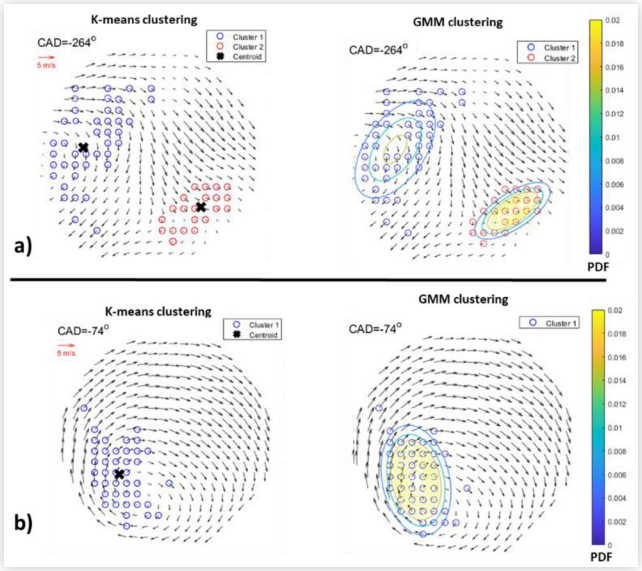
Figure 7 shows the vortex center distributions under high and low swirl conditions. Upon a further check of the vortex center distributions as depicted in Figure 7, each data cluster exhibits a single-peak shape (unimodal) in the middle of the cluster, which is similar to the Gaussian distribution. In the figure, the vortex center regions following Gaussian distributions are circled. One can notice that for the low swirl case at –88 CAD, the data distribution still shows two distinct vortex regions although they are merged together. Thus, GMM clustering can supplement the K-means clustering on extracting the vortex features under low swirl cases with multiple vortex zones and vortex merging.

High Swirl Flow Fields Analysis

Vortex Classification & Vortex Zone Detection Highly regulated vortex motion under high swirl condition makes the vortex zone(s) from intake to compression strokes clear and distinct. Figure 8a illustrates that K-means clustering could accurately classify the vortex center locations into different groups during intake at –264 CAD and during compression at –74 CAD. Figure 8b shows the GMM clustering results at the same CADs. It is clear that both K-means and GMM clustering can accurately classify the vortex locations. The differences are that K-means calculates the specific centroid positions while GMM generates PDF contours of the clusters. For the high swirl case with stable vortex structures, K-means clustering is sufficient for vortex location classification. However, K-means clustering has limitations on vortex zone detection. It is because K-means is a hard clustering method, that is, each data point can only be partitioned into one class. It cannot assign the probability for each data point which describes how similar each point is to each cluster, as GMM clustering does. Therefore, GMM can generate the PDF contours to detect the vortex zones instead of using the discrete distributions.

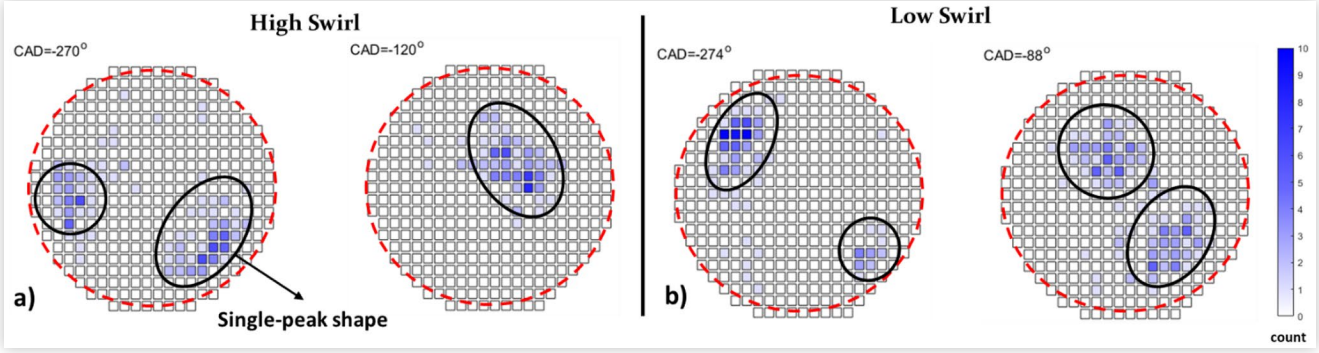
In the previous study using K-means clustering, the vortex zone of each cluster is determined manually to include the majority of the data points within each cluster. In such way, the vortex zone could be larger than its true size because of the presence of outliers. As the cyclic variations becomes larger, the vortex zones are overlapping more such that the vortex zone could be more difficult to detect. GMM can be used to resolve these issues. Specifically, each data point within the same cluster is assigned a probability by GMM

FIGURE 8 K-means and GMM clustering results at a) –264 and b) –74 CADs of high swirl ratio.



© SAE International.

FIGURE 7 Sample vortex center distributions of both a) high and b) low swirl conditions.



© SAE International.

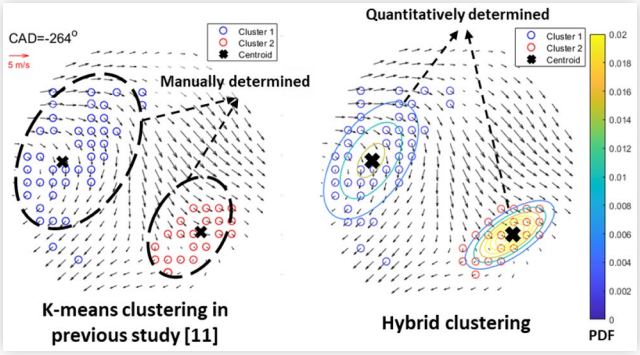
Downloaded from SAE International by Jiajin Wu, Sunday, January 02, 2022

such that the PDF contour lines can be generated. Although the cluster overlapping does not happen under high swirl condition, the PDF contour could also be helpful to identify outliers and detect vortex zone(s). Figure 9 shows the comparison between K-means clustering and hybrid clustering on detecting the vortex zones at –264 CAD. For K-means clustering results, the vortex zones are determined manually to include the majority of the observations in a cluster (Figure 9). However, such manual determination of vortex zone is only partially intuitive and qualitative without rigorous standards. Instead, with hybridizing GMM clustering, the contour lines partition several zones with quantitative probability density, which is more informative on vortex zone detection and prediction.

Cyclic Variation Quantification After vortex classification and vortex zone detection, next step of applying K-means clustering makes the hybrid clustering method be capable for vortex variations quantification. Figure 10 shows the average distance curve to represent the cyclic variations from intake to compression. The distance curve shows that the cyclic variations of counter-clockwise vortex are stronger than clockwise vortex since the intake system is designed to induce a strong clockwise swirl motion. From the crank angle-based perspective, the distance curve shows an increasing trend of cyclic variations during intake stroke. At early intake stage, the intake air strongly affects the vortex structures resulting in larger vortex variations. As the intake air gradually forms a dominant vortex motion, it can be found that the distance curve after bottom dead center (BDC) exhibits a deceasing trend. Since the intake valves gradually close during late compression stroke, the intake air becomes weaker and the piston motion is the only dominant source impacting the vortex motion such that the cyclic variations are further reduced.

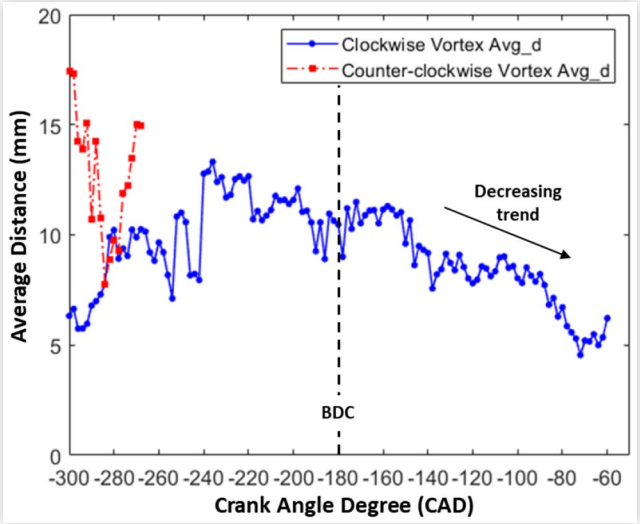
In summary, the hybrid clustering method integrates the competencies of both GMM and K-means clustering techniques. For high swirl flow with distinct vortex structures, although K-means clustering is accurate and efficient in vortex classification, the vortex zone cannot be detected automatically by K-means alone. To overcome this shortcoming, the hybrid method utilizes quantitative probability density contours of GMM to detect the vortex zones.

FIGURE 9 Comparison between K-means clustering and hybrid clustering.



© SAE International.

FIGURE 10 Average distance curve under high swirl condition.



© SAE International.

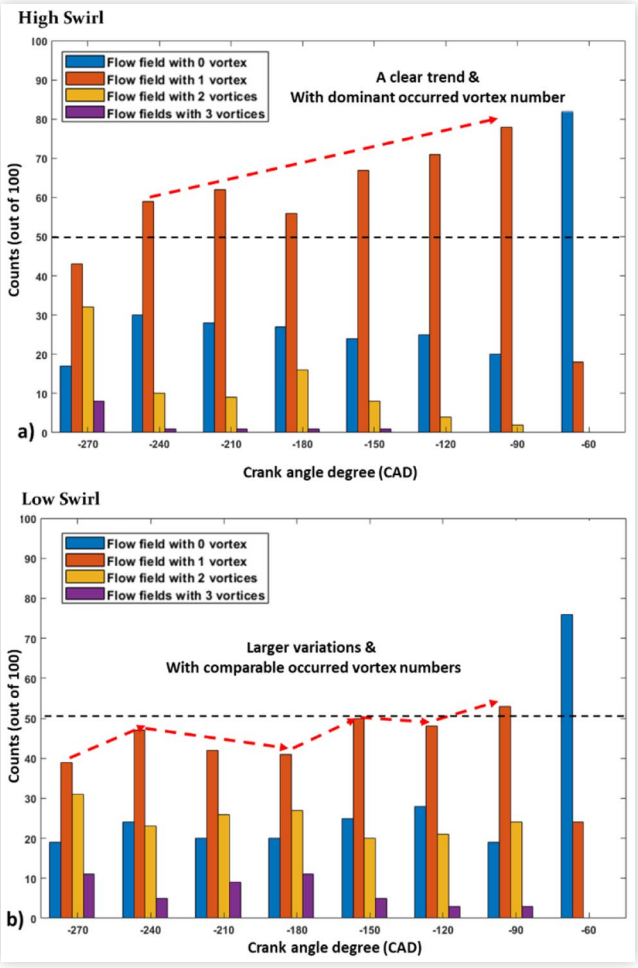
Low Swirl Flow Fields Analysis

Vortex Classification & Vortex Zone Detection. Unlike high swirl flow data, the vortex features under low swirl condition exhibit much larger variations. The clustering results could be inaccurate using K-means method in previous study [11]. Figure 11 shows the bar diagram which accounts for the number of vortex present in individual cycle at several CADs from intake to compression for both high and low swirl conditions. At a fixed CAD, the experimental statistics of vortex number per cycle do not have dominant values as high swirl data, especially for intake stroke. The red arrows show the trends of the most occurred vortex number. The 50% lines (dashed line) are also drawn in Figure 11 for both high swirl and low swirl conditions to compare the counts of the most occurred vortex number. For high swirl condition, the counts of the most occurred vortex number are over 50 from intake to the compression and show a clear increasing trend. It indicates that stable vortex structure is established from intake stroke to late compression under high swirl condition. Unlike the high swirl case, the counts of the most occurred vortex number exceed 50 until –90 CAD for low swirl case. It can be found that multiple comparable occurred vortex numbers exist before late compression stroke for low swirl case. Therefore, with the cluster number selection procedure based on the experimental statistics, the classification results for low swirl case may not be accurate because of the improper selection.

Figure 12 illustrates an example of the K-means clustering result and vortex center distribution at –150 CAD. There is no dominant vortex number at –150 CAD. By vortex number statistics, the cluster number is determined to be one. However, from the vortex center distributions in Figure 12b, it is clear that there are two distinct regions where a clear vortex structure exists frequently. If the cluster number of one is used, the centroid will be forced to locate at an inaccurate position (Figure 12a) where vortex structure exists occasionally during these 100 engine cycles. Therefore, cluster number selection

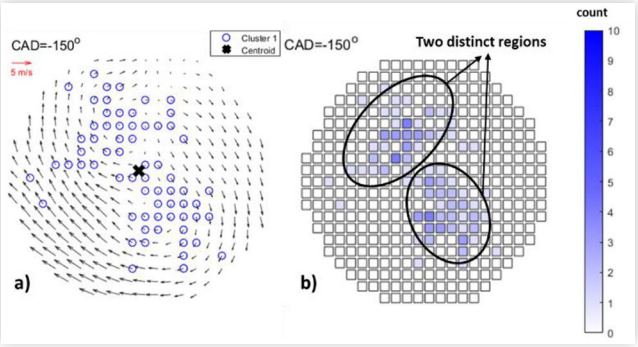
Downloaded from SAE International by Jiajin Wu, Sunday, January 02, 2022

FIGURE 11 Percentage diagram of the vortex number based on individual flow fields under a) high and b) low swirl conditions.



© SAE International.

FIGURE 12 a) K-means clustering result and b) vortex center distribution at –150 CAD.

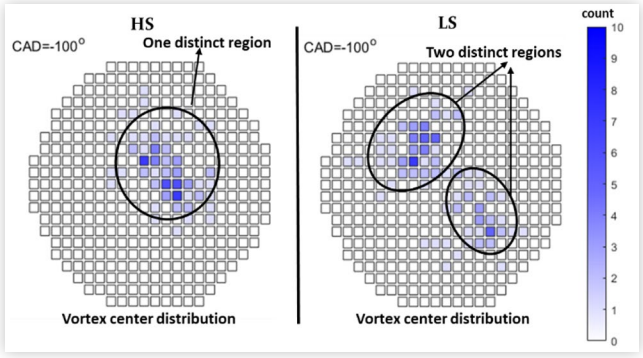


© SAE International.

relying on vortex number statistics with large variations is not always reliable to generate a reasonable clustering result.

Additionally, although there is a dominant vortex number during late compression stroke, the clustering results could still be inaccurate for low swirl case. Figure 13 shows two vortex center distributions at late compression of –100 CAD under high and low swirl conditions. While the most

FIGURE 13 Vortex center distributions at –100 CAD under low swirl condition.



© SAE International.

frequently occurred vortex numbers are both one, the vortex center distributions are obviously different. The vortex distribution of high swirl data indicates that there is only one distinct vortex motion region frequently. For low swirl condition, there are also two vortex regions as the case at –150 CAD. The vortex classification could become inaccurate if using cluster number of one. The reason can be revealed by checking the vortex locations of individual cycles at –100 CAD under low swirl condition. Figure 14 shows several randomly selected flow fields at –100 CAD with indicated vortex centers. It is true that most cycles only show one dominant vortex motion. However, the vortex center locations of different cycles vary a lot and they do not gather together to form a dominant vortex cluster. Consequently, clustering with cluster number of one fails to accurately classify the vortex locations. With incorrect vortex classification results, the zone detection of the vortex cluster also becomes inaccurate.

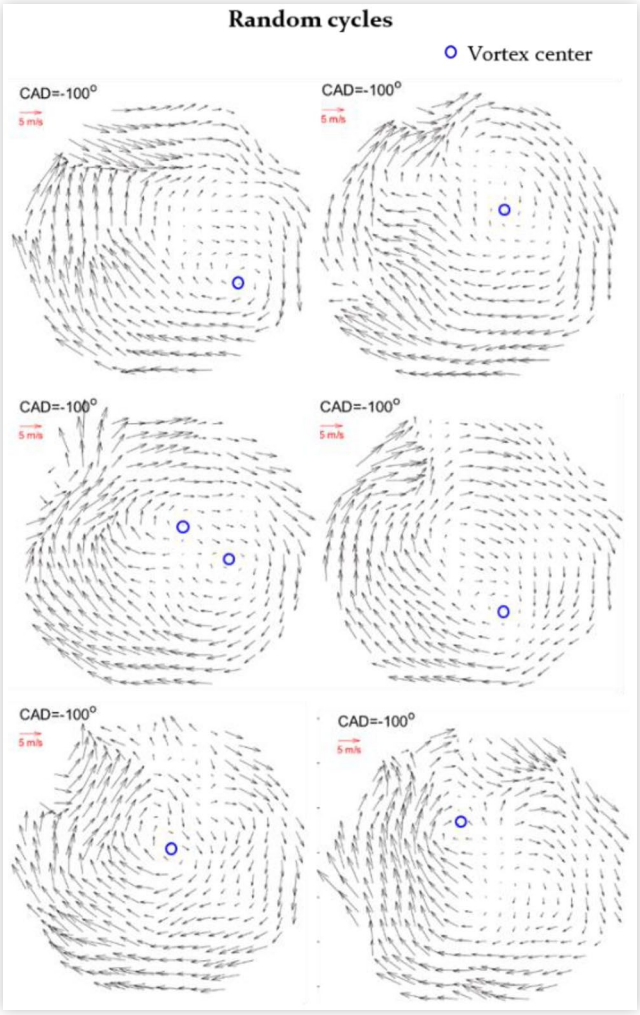
For low swirl case with such large cyclic variations as shown in Figures 12 and 13, predetermined values of the vortex number might be inaccurate due to large randomness in the experimental statistics. Therefore, for flow fields with large variations of vortex locations, it is better to concentrate on the vortex center distributions. As mentioned in the previous section of flow data description, the vortex distributions of low swirl case are similar to the Gaussian distribution or mixture of several Gaussian distributions, which allows GMM clustering to accurately classify the vortex and detect zones for low swirl flow case.

Figure 15 shows the hybrid clustering and K-means clustering results at –150 and –100 CAD under low swirl condition. Comparing with K-means clustering results, hybrid clustering accurately detects these two vortex zones shown in the vortex distributions of Figures 12 and 13. This hybrid method calculates the probability of each observation such that the cluster overlapping issue could be alleviated. From Figure 15, the data points in the border of clusters as well as the outliers are assigned with very low probabilities. With a proper probability threshold, the data points with low probabilities (outliers) can be eliminated to generate more accurate clustering results without the loss of major vortex features.

Cyclic Variation Quantification Figure 16 shows the average distance curve from –300 CAD to –60 CAD. As expected, the cyclic variations for low swirl case are larger

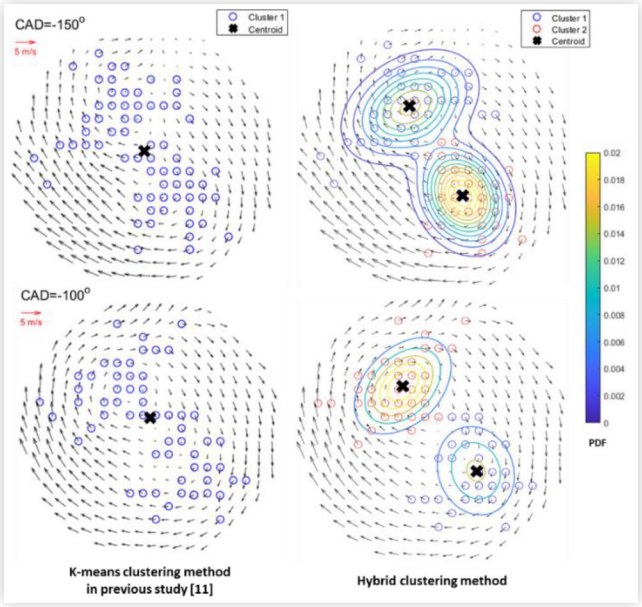
Downloaded from SAE International by Jiajin Wu, Sunday, January 02, 2022

FIGURE 14 Flow fields at –100 CAD of random cycles under low swirl condition.



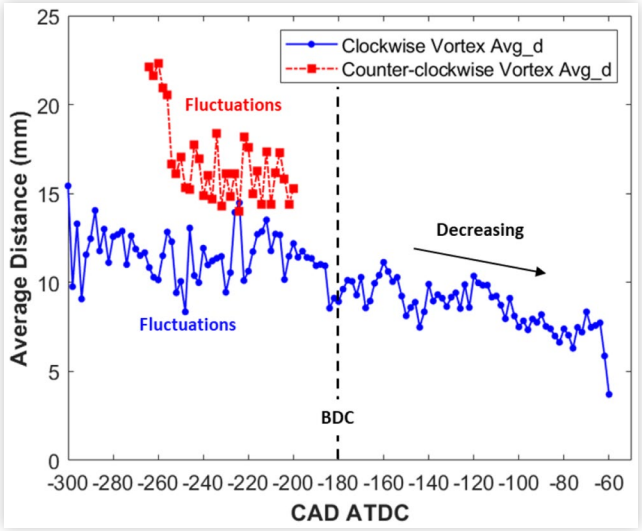
© SAE International.

FIGURE 15 K-means and hybrid clustering results at –150 and –100 CADs under low swirl condition.



© SAE International.

FIGURE 16 Average distance curve under low swirl condition.



© SAE International.

comparing to the results under high swirl condition (Figure 10), except for the CADs near BDC. During intake stroke of low swirl case, the vortex motion is not strong nor is it highly regulated as high swirl case such that the vortex movements would be easily affected. As a consequence, the cluster number varies easily from CAD to CAD and the average distance will be strongly affected. Therefore, during intake stroke, the cyclic variation quantification becomes more difficult since the distance curves fluctuate a lot. Near BDC, the cyclic variations under high swirl and low swirl conditions are comparable. It is because that the reverse motion of the piston is the main source which enlarges the variations for both swirl ratio conditions. After BDC, the flow fields exhibit stable flow structures. As the intake valves close, the curve shows a decreasing trend of cyclic variations of the vortex motion. As the piston moves upwards, the vortex gradually moves out of the measurement area and the cyclic variations reduce again.

To summarize the clustering results of low swirl condition, the hybrid clustering method outperforms the K-means method with a higher clustering accuracy due to the PDF contour capability of GMM. It is because the clustering errors of K-means clustering increase since the experimental statistics under low swirl condition are highly random, which causes more cyclic variations of the predetermined number of the clusters. Additionally, the hybrid clustering assigns probability to each data point which is also helpful to detect outliers and overcome the cluster overlapping issue. For cyclic variation analysis, K-means clustering is effective to quantify the vortex motion variations using the average distance defined in Equation 2. Therefore, hybridizing GMM and K-means clustering not only improves the vortex classification and vortex zone detection accuracy, but also quantifies the cyclic variations of vortex motion at different crank angle degrees.

Downloaded from SAE International by Jiajin Wu, Sunday, January 02, 2022

Summary

In this work, a hybrid clustering method combining GMM and K-means clustering is implemented for in-cylinder flow field characterization in a GDI engine under both high swirl and low swirl flow conditions. This hybrid clustering method has a sequential structure. First, GMM clustering is applied to classify vortex and detect vortex zone using PDF contour lines. Then, K-means clustering is used to quantify the cyclic variations with both centroid location and average distance based on GMM results. These two methods supplement each other in improving the accuracy of vortex feature extraction including vortex classification, vortex zone detection, and cyclic variation quantification.

For high swirl flow with distinct vortex structures, although K-means clustering is accurate and efficient in vortex classification, the vortex zone detection is still an intractable issue. To overcome this shortcoming, the hybrid method utilizes quantitative probability density contours of GMM to detect the vortex zones. For the low swirl case with higher vortex motion variations, both the vortex zone detection and vortex classification accuracy are improved using this hybrid method. In all, this hybrid clustering method blends in the competencies of both GMM and K-means, which can sufficiently extract the vortex features including vortex classification, vortex zone detection, and cyclic variations for both swirl flow conditions. With this hybrid clustering method, the temporal evolution and transient features of the in-cylinder vortex motion under different swirl ratio conditions can be better revealed than using the K-Means clustering method alone.

References

1. Mittal, M., Hung, D., Zhu, G., and Schock, H., "High-speed flow and Combustion Visualization to Study the Effects Of Charge Motion Control on Fuel Spray Development and Combustion Inside a Direct-injection Spark-ignition Engine," *SAE Int. J. Engines* 4:1469–1480, 2011, https://doi.org/10.4271/2011-01-1213.
2. Porpatham, E., Ramesh, A., and Nagalingam, B., "Effect of Swirl on the Performance and Combustion of a Biogas Fuelled Spark Ignition Engine," *Energy Convers. Manage,* 76:463–471, 2013, doi:10.1016/j.enconman.2013.07.071.
3. Wang, T., Li, W., Jia, M., Liu, D. et al., "Large-eddy Simulation of in-cylinder flow in a DISI Engine with Charge Motion Control Valve: Proper Orthogonal Decomposition Analysis and Cyclic Variation," *Appl. Therm. Eng.* 75:561–574, 2015, doi:10.1016/j.applthermaleng.2014.10.081.
4. Reuss, D., "Cyclic Variability of Large-scale Turbulent Structures in Directed and Undirected IC Engine flows," SAE Technical Paper 2000-01-0246, 2000, https://doi.org/10.4271/2000-01-0246.
5. Druault, P., Guibert, P., and Alizon, F., "Use of Proper Orthogonal Decomposition for Time Interpolation from PIV data. Application to the Cycle-to-cycle Variation Analysis of in-cylinder Engine Flows," *Exp. Fluids* 39:1009–1023, 2005, doi:10.1007/s00348-005-0035-3.
6. Chen, H., Reuss, D., and Sick, V., "On the Use and Interpretation of Proper Orthogonal Decomposition of in-cylinder Engine Flows," *Measurement Science and Technology* 23(8):085302, 2012, doi:10.1088/0957-0233/23/8/085302.
7. Chen, H., Xu, M., and Hung, D., "Analyzing in-cylinder Flow Evolution and Variations in a Spark-ignition Direct-injection Engine Using Phase-invariant Proper Orthogonal Decomposition Technique," SAE Technical Paper 2014-01-1174, 2014, https://doi.org/10.4271/2014-01-1174.
8. Zhuang, H. and Hung, D., "Characterization of the Effect of Intake Air Swirl Motion on Time-resolved in-cylinder Flow Field Using Quadruple Proper Orthogonal Decomposition," *Energy Convers. Manage* 108:366–376, 2016, doi:10.1016/j.enconman.2015.10.080.
9. Zhao, F., Ge, P., Zhuang, H., and Hung, D., "Analysis of Crank Angle-resolved Vortex Characteristics Under High Swirl Condition in a Spark-ignition Direct-injection Engine," *J. Eng. Gas Turbines Power* 140(9):092807, 2018, doi:10.1115/1.4039082.
10. Liu, Z., Teh, K., Ge, P., Zhao, F. et al., "Tumble Vortex Characterization by Complex Moments," SAE Technical Paper 2018-01-0207, 2018, https://doi.org/10.4271/2018-01-0207.
11. Zhao, F., Hung, D., and Wu, S., "K-means Clustering-driven Detection of Time-resolved Vortex Patterns and Cyclic Variations Inside a Direct Injection Engine," *Appl. Therm. Eng.* 180(115810), 2020, doi:10.1016/j.applthermaleng.2020.115810.
12. Zhao, F., Ruan, Z., Yue, Z., Hung, D. et al., "Time-sequenced Flow Field Prediction in an Optical Spark-ignition Direct-injection Engine Using Bidirectional Recurrent Neural Network (bi-RNN) with Long Short-term Memory," *Appl. Therm. Eng* 173(115253), 2020, doi:10.1016/j.applthermaleng.2020.115253.
13. Hanuschkin, A., Schober, S., Bode, J., Schorr, J. et al., "Machine Learning-based Analysis of in-cylinder Flow Fields to Predict Combustion Engine Performance," *Int. J. Engine Res*, 2019, doi:10.1177/1468087419833269.
14. Kodavasal, J., Moiz, A., Ameen, M., and Som, S., "Using Machine Learning to Analyze Factors Determining Cycle-to-cycle Variation in a Spark-ignited Gasoline Engine," *Journal of Energy Resources and Technology* 140(10):102204, 2018, doi:10.1115/1.4040062.
15. Cay, Y., "Prediction of a Gasoline Engine Performance with Artificial Neural Network," *Fuel* 111:324–331, 2013, doi:10.1016/j.fuel.2012.12.040.
16. Yu, W., Zhao, F., Yang, W., and Xu, H., "Integrated Analysis of CFD Simulation Data with K-means Clustering Algorithm for Soot Formation Under Varied Combustion Conditions," *Appl. Therm. Eng.* 153:299–305, 2019, doi:10.1016/j.applthermaleng.2019.03.011.
17. Satre-Meloy, A., Diakonova, M., and Grünewald, P., "Cluster Analysis and Prediction of Residential Peak Demand Profiles Using Occupant Activity Data," *Appl. Energy* 260(114246), 2020, doi:10.1016/j.apenergy.2019.114246.
18. Teichgraeber, H. and Brandt, A., "Clustering Methods to find Representative Periods for the Optimization of Energy Systems: An Initial Framework and Comparison", *Appl.*

Downloaded from SAE International by Jiajin Wu, Sunday, January 02, 2022

Energy 239:1283–1293, 2019, doi:10.1016/j.apenergy.2019.02.012.

19. Fränti, P. and Sieranoja, S., "K-means Properties on Six Clustering Benchmark Datasets," *Applied Intell* 48(12):4743–4759, 2018, doi:10.1007/s10489-018-1238-7.
20. Fränti, P. and Sieranoja, S., "How Much Can k-means be Improved by Using Better Initialization and Repeats?" *Pattern Recogn.* 93:95–112, 2019, doi:10.1016/j.patcog.2019.04.014.
21. Preheim, S., Perrotta, A., Martin-Platero, A., Gupta, A. et al., "Distribution-based Clustering: Using Ecology to Refine the Operational Taxonomic Unit," *Appl Environ Microbiol* 79:6593–6603, 2013, doi:10.1128/AEM.00342-13.
22. Xu, D. and Tian, Y., "A Comprehensive Survey of Clustering Algorithms," *Annals of Data Science* 2:165–193, 2015, doi:10.1007/s40745-015-0040-1.
23. Lu, Y., Tian, Z., Peng, P., Niu, J. et al., "GMM Clustering for Heating Load Patterns in-depth Identification and Prediction Model Accuracy Improvement of District Heating System," *Energy & Buildings* 190:49–60, 2019, doi:10.1016/j.enbuild.2019.02.014.
24. Hou, X., Zhang, T., Xiong, G., Lu, Z. et al., "A Novel Steganalysis Framework of Heterogeneous Images Based on GMM Clustering," *Signal Processing: Image Communication* 29:385–399, 2014, doi:10.1016/j.image.2014.01.006.
25. MATLAB and Statistics and Machine Learning Toolbox Release 2020a, The MathWorks, Inc., Natick, Massachusetts, United States.
26. Tibshirani, R., Walther, G., and Hastie, T., "Estimating the Number of Clusters in a Data Set Via the Gap Statistic," *J. Roy. Stat. Soc. B* 63(2):411–423, 2001, doi:10.1111/1467-9868.00293.
27. Steele, R. and Raftery, A., "Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models". Technical Report 559, University of Washington, Dept. of Statistics, 2009.
28. Lumley, J., *Engines: An Introduction* (Cambridge University Press, 1999).

Definitions/Abbreviations

GMM - Gaussian mixture model
PIV - Particle image velocimetry
DI - Direct injection
GDI - Gasoline Direct injection
CAD - Crank Angle Degree
ATDC - After Top Dead Center
SSE - Sum of squared Euclidean distance
EM - Expectation-maximization
BIC - Bayesian Information Criterion
PDF - Probability density function
BDC - Bottom dead center

© 2021 SAE International. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of SAE International.

Positions and opinions advanced in this work are those of the author(s) and not necessarily those of SAE International. Responsibility for the content of the work lies solely with the author(s).

ISSN 0148-7191