Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Flow Field Data Mining based on a Compact Stream Line Representation

Lars Gräning

2015

Preprint:

This is an accepted article published in SAE 2015 World Congress & Exhibition. The final authenticated version is available online at: https://doi.org/[DOI not available]

2015-01-XXXX

Flow Field Data Mining Based on a Compact Streamline Representation

Author, co-author (Do NOT enter this information. It will be pulled from participant tab in MyTechZone)

Affiliation (Do NOT enter this information. It will be pulled from participant tab in MyTechZone)

Copyright © 2015 SAE International

Abstract

In many engineering domains like aerospace, vehicle or engine design the analysis of flow fields, acquired from computational fluid dynamics (CFD) simulations can reveal important insights on the behavior of the simulated objects. However, the huge amount of flow data produced by each simulation complicates the data processing and limits the application of data mining and machine learning tools for the flow analysis. The paper introduces a compact streamline and feature based representation of three dimensional flow fields, in order to overcome this limitation. The compact representation defines the basis for a subsequent quantification of flow field similarities. In this paper, we exploit the similarity measure to solve typical data mining tasks like the comparison of individual flow field similarities, the retrieval of similar flow fields from a database and the identification of groups of similar flow fields, which provide the basis for a more efficient decision making in the engineering design process. Computational experiments based on realistic data from the simulation of the exterior flow around passenger car shapes reveal the advantages of the compact streamline representation for solving the data mining tasks with respect to computational efficiency and memory consumption.

Introduction

The computational simulation and modeling of fluid dynamics has become a corner stone for a huge variety of engineering applications, e.g., related to aerospace, climate modeling, vehicle or engine design. More and more physical experiments like wind tunnel test are replaced by computational fluid dynamics (CFD) simulations, since the efficiency and accuracy of the simulation techniques increases. However, high-fidelity simulations of the fluid flow remain computational expensive, what requires an efficient management and handling of the simulation data. Exploiting the information contained in simulation data at its best is expected to be the key for improving decision making in the engineering design process in the future. Modern data mining technologies provide the means to extract relevant information from the design and simulation data. The authors in [2, 11] addressed this and suggested the use of techniques like self-organizing maps (SOM) or analysis of variance (ANOVA) to gain insights into the relationship between the geometry of a design and multi-criteria performance measurements, which are estimated based on the results of CFD simulations. In their work the application of data mining techniques is still restricted to single design processes given a well-defined and low dimensional design representation. In [7] the shape mining framework has been introduced, suggesting the utilization of unstructured surface meshes as a unified design representation to facilitate the application of data mining techniques beyond individual design processes.

All those studies condense the information of the simulated flow into individual scalar performance numbers and neglect most of the details of the resulting flow data. In practice modern visualization techniques and tools are used to highlight particular physical flow phenomena of the entire flow data, in order to understand aerodynamic effects in the design domain, see [1, 6, 5, 13] for some examples. The extraction of new and useful information from the visual inspection is often a demanding and time-consuming process. Integrating the information over hundreds of flow field data sets by such a manual process will become unfeasible in any realistic and competing design process. Going from a manual to an automatic process for analyzing the flow field data, by means of applying computational tools from data mining and machine learning, can overcome this difficulty.

The computational analysis and comparison of the acquired flow field data can provide relevant domain information related to the behavior of the technical system under consideration, e.g., revealing important interactions between design and flow features as suggested in [14]. The authors extract specific flow features and utilize methods from information theory to investigate the interrelations between design changes, changes of the overall performance and the extracted flow features. For any flow field analysis it is essential that the differences between flow fields can be quantified, either based on the raw data or a set of descriptive features. Thus, comparing flow fields in general requires a proper similarity measure being defined, which can, e.g., be adopted for discovering flow field categories [3] or to compare simulated against ideal flow, as for the design of combustion engines [9].

Quantitative methods require that explicit considerations about the flow field representation and the similarity measure are being made. Xu et al. [18] have studied probabilistic similarity measures based on featured flow fields. The measures take feature data from the entire simulation domain into account. For practical applications, this requires big data storage and is linked to high computational costs for the similarity evaluation. As a successor of [8], this paper studies the evaluation of flow field similarities based on a compact streamline and flow feature based representation. The similarity measure defines the basis for flow field data mining applications like the retrieval of flow fields from databases or the grouping of similar flow fields, towards the identification of aerodynamic concepts. Finally, based on a realistic data set of flow field data from the simulation of the exterior flow around a passenger car, we study and discuss the

advantage of the compact streamline representation for the considered data mining applications, before concluding the paper.

Flow Field Representation

Flow fields, as a result of computational fluid dynamics simulations (CFD), define properties of the fluid or air flow in interaction with solid objects. Beside pressure, density and temperature, the velocity vector field defines the speed of the air or fluid flow at fixed positions in the Euclidean space. Considering steady state flow, the vector field is formally defined as a function that maps each point in the three dimensional Euclidean space to a velocity vector, which remains constant over time,

$$\boldsymbol{v}(\boldsymbol{p}): (x, y, z)^T \mapsto \left(v_x, v_y, v_z\right)^T.$$

Tools for computational flow simulation model the flow physics with the Navier Stokes equations. In order to numerically solve these partial differential equations the simulation area is subdivided into discrete cells. Especially for high fidelity simulations the cell sizes need to be kept small to achieve a high spatial resolution, resulting in a huge velocity vector field representing the detailed flow of the entire simulation domain.

The storage and processing of the entire flow data is often impractical. For an efficient processing a low dimensional representation of the raw data is favored. The transformation into a problem specific feature representation can be the means to derive an efficient representation, see [13] for an introduction.

Local Flow Features

The orientation and magnitude of the velocity vector does not always provide all relevant information about the flow at a fixed point p. Therefore, additional features need to be derived, wherein the calculation of the features are always problem and application dependent. Following Xu et al. [18], features representing higher order properties derived from the gradient of the velocity vector field have been studied in this paper. The local gradient in a 3D velocity vector field is defined by an asymmetric 3D tensor:

$$T(p, v) = \begin{pmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{pmatrix} = \begin{pmatrix} \frac{\partial v_x}{\partial x} & \frac{\partial v_x}{\partial y} & \frac{\partial v_x}{\partial z} \\ \frac{\partial v_y}{\partial x} & \frac{\partial v_y}{\partial y} & \frac{\partial v_y}{\partial z} \\ \frac{\partial v_z}{\partial x} & \frac{\partial v_z}{\partial y} & \frac{\partial v_z}{\partial z} \end{pmatrix} = S + A,$$

which can be decomposed into its symmetric and asymmetric components, S and A:

$$S = \begin{pmatrix} \frac{\partial v_x}{\partial x} & \frac{1}{2}\left(\frac{\partial v_y}{\partial x} + \frac{\partial v_x}{\partial y}\right) & \frac{1}{2}\left(\frac{\partial v_x}{\partial z} + \frac{\partial v_z}{\partial x}\right) \\ \frac{1}{2}\left(\frac{\partial v_y}{\partial x} + \frac{\partial v_x}{\partial y}\right) & \frac{\partial v_y}{\partial y} & \frac{1}{2}\left(\frac{\partial v_z}{\partial y} + \frac{\partial v_y}{\partial z}\right) \\ \frac{1}{2}\left(\frac{\partial v_x}{\partial z} + \frac{\partial v_z}{\partial x}\right) & \frac{1}{2}\left(\frac{\partial v_z}{\partial y} + \frac{\partial v_y}{\partial z}\right) & \frac{\partial v_z}{\partial z} \end{pmatrix},$$

$$S = \begin{pmatrix} \varepsilon_1 & \frac{1}{2}\theta_3 & \frac{1}{2}\theta_2 \\ \frac{1}{2}\theta_3 & \varepsilon_2 & \frac{1}{2}\theta_1 \\ \frac{1}{2}\theta_2 & \frac{1}{2}\theta_1 & \varepsilon_3 \end{pmatrix},$$

$$A = \begin{pmatrix} 0 & -\frac{1}{2}\left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right) & \frac{1}{2}\left(\frac{\partial v_x}{\partial z} - \frac{\partial v_z}{\partial x}\right) \\ \frac{1}{2}\left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right) & 0 & -\frac{1}{2}\left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}\right) \\ -\frac{1}{2}\left(\frac{\partial v_x}{\partial z} - \frac{\partial v_z}{\partial x}\right) & \frac{1}{2}\left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}\right) & 0 \end{pmatrix},$$

$$A = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix},$$

where S comprises isotropic scaling and rotation information, and A anisotropic information about the stretching of the velocity vectors in the field, based on which the following scalar features can be calculated:

- Velocity vector magnitude: $M_v = \left|\left(v_x, v_y, v_z\right)^T\right|$,
- Tensor magnitude: $M_t = \sqrt{\frac{1}{2}\sum_i^3 \sum_j^3 T_{ij}^2}$,
- Dilatation magnitude: $M_d = \sqrt{\sum_i^3 \varepsilon_i^2}$,
- Magnitude of shear strain rate: $M_s = \sqrt{\sum_i^3 \theta_i^2}$,
- Magnitude of vorticity: $M_\omega = \sum_i^3 \omega_i^2$.

Given those five features we define the feature space as:

$$\boldsymbol{f}(\boldsymbol{p}, \boldsymbol{v}): \boldsymbol{p}, \boldsymbol{v} \mapsto (M_v, M_t, M_d, M_s, M_\omega)^T.$$

The defined feature vector is considered to comprise all problem relevant information within the simulated flow domain. If this is not the case any subsequent analysis and modeling step might fail to reveal the desired information, requiring the revision of the calculated feature values.

Featured Streamlines

Streamlines are one of the most important flow field visualization techniques, see e.g. [19]. Streamlines describe the trajectories (tangential to the velocity vector) that particles would travel through the flow starting from an initial seeding position. Streamlines are represented by discrete sample points along the trajectory, where the distance between adjacent sample points is defined by the density of the CFD mesh. Assigning the feature vector to each sample point along each streamline, a single featured streamline is defined as:

$$l(u): u \mapsto \boldsymbol{p}, \boldsymbol{v}, (M_v, M_t, M_d, M_s, M_\omega)^T,$$

where u defines the position of each sample along the streamline l. A set of featured streamlines is denoted as $L = \{l_i\}, i \in [1, N_l]$, with N_l defining the number of streamlines in the set.

Beside the purpose of visualization, the transformation of the flow field into a set of streamlines can be applied for data reduction, where the entire flow and feature data is condensed into information along individual trajectories. Without a priori knowledge about the flow field, streamlines are seeded uniformly at the inlet. In order to capture all relevant flow phenomena, often a huge number of streamlines is required, what complicates the subsequent visual investigation of the flow and goes along with an increase in memory consumption.

Automatic Streamline Reduction

In the following a new algorithm for automatic streamline reduction is introduced, targeting to reduce the amount of flow data to be processed by humans or computer programs. Given a large set of streamlines, the algorithm targets to remove individual streamlines from the set, which contributes least to the overall information content. Thus, we target the creation of an efficient streamline based flow field representation with a minimum number of streamlines. Therefore, streamlines are ranked according to their information contribution, which is quantified based on the formulation of the Shannon entropy [17]. Given discrete feature data for each streamline, the expected entropy accumulated over all features defines the considered measure of relevance, formally defined as:

$$\overline{H_l} = E(H_{l_i}) = -\frac{1}{N_f}\sum_i^{N_f}\sum_j^{N_b^i} p(\hat{f}_i^j)\log p(\hat{f}_i^j),$$

with N_f defining the number of features, N_b^i defining the number of discrete states and $\hat{f}_i^j$ defining the discrete feature value of feature i, e.g., the velocity magnitude M_v. The probability $p(\hat{f}_i^j)$ denotes the likelihood of a discrete feature value to be observed along the considered streamline. On the one hand, if all discrete feature values are equally probable, the feature values are uniformly distributed and the entropy reaches its maximum value. On the other hand, if the feature value along the streamline does not change within a certain boundary, the value of the entropy vanishes. Under the assumption that streamlines with none or small feature value variations are of less relevance and capture least information about the flow field, a heuristic has been defined that removes streamlines with the smallest value of $\overline{H_l}$. Ranking all streamlines according to their expected entropy, a subset of streamlines L' can be derived, which captures the majority of the overall flow feature information. The information content of the reduced subset is defined by the relative expected entropy:

$$IC(L') = \frac{\sum_j^{N_k}\overline{H_J}}{\sum_i^{N_l}\overline{H_l}},$$

where N_k defines the number of streamlines in the reduced set. For an automatic streamline reduction procedure a threshold λ needs to be applied to $IC(L')$. The threshold guarantees that at least λ percent of the flow information remain in the subset.

Flow Field Distance

After applying the automatic streamline reduction step, the streamline subset L' defines a compact representation of the raw 3d flow. Consequently, the difference of two 3d flow fields can be estimated by a quantity that evaluates the distance between two compact (reduced) featured streamline sets.

Given that a measure for evaluating the distance $d_{m,n}$ between two corresponding streamlines m and n with respect to two compact streamline representations A and B is given, the overall distance can be defined by accumulating all individual distances $d_{m,n}$:

$$D(L'_A, L'_B) = \sum_k^{N_A} d_{m,n},$$

where N_A defines the number of streamlines of representation A, m refers to the index of a streamline l_m of the compact streamline set L'_A and n denotes the corresponding streamline l_n of the reduced streamline set L'_B.

The overall distance measure D, first, requires the identification of corresponding streamlines between two representations and, second, the definition of an appropriate streamline wise distance measure $d_{m,n}$.

Identification of Corresponding Streamlines

The task of identifying corresponding streamlines is to assign a streamline of one flow field representation A to its comparable streamline of flow field representation B. Under the assumption that corresponding streamlines are those which origin at the same or a nearby seeding point, the following algorithm can be applied. Given the initial point p_m^0 of one streamline $l_m \in L'_A$ with $u = 0$, the corresponding streamline in B is defined by the Euclidean closest point p_n^0 over all streamline seeding positions in L'_B, with $n \in [1, N_B]$. If the minimal distance is within a vicinity r the streamline l_n is said to be the corresponding streamline of l_m. The vicinity r is defined as the minimal distance between the seeding positions of neighboring streamlines considering all streamlines in L'_A and L'_B. If the minimal distance is not within the vicinity r a corresponding streamline to l_m is said to not exist.

Streamline Wise Distance

Once corresponding streamlines are identified, the distance between two streamlines can be evaluated on the basis of a streamline wise distance quantification. Given N_f different features and related weighting factors w_i, the streamline wise distance in its general form is denoted as:

$$d_{m,n} = \sum_i^{N_f} w_i \cdot d(P_i^m, P_i^n)$$

with $\sum_i^{N_f} w_i = 1$, and P_i^m, P_i^n defining the discrete feature distribution for feature i with respect to streamline l_m and the corresponding streamline l_n, respectively. Probabilistic distance measures, used to quantify the difference between feature distributions, can be categorized into bin-wise distance measures like L_1 and χ^2 [18], and cross-bin distance measures like Earth Mover's Distance EMD [15] and the Quadratic-Chi distance family QC [12]. In [8], the difference between the two types of distance measures has been investigated and evaluated concerning the quantification of streamline wise distances. The cross-bin distances measures, which are computational more expensive, are only preferable over the bin-wise distance measures if the number of sample points and feature evaluations along each individual streamline is small.

Flow Field Data Mining Applications – Experimental Studies

In order to investigate the introduced algorithms for its applicability, and to provide the reader with insights on the respective fields of application, experimental studies with flow field data from the CFD simulation of the external flow around passenger car shapes have been carried out. Typically, the shape and simulation data result from various diverse engineering design processes, where different design processes might follow different strategies to reach a pre-defined design goal, see [7] for further information. In the described

experimental studies an artificial flow field data set has been generated, which could equally result from a realistic computational design and optimization scenario.

Passenger Car Streamline Data

First, the process and techniques for generating the featured streamline sets, representing the flow of various passenger car shapes, are described, wherein standard CAE methods and tools have been used. Given the geometrical representation of a baseline car shape, *Figure 1* depicts the overall process for generating the featured streamline set of one deformed car geometry. The overall dataset generated is the basis for all successive experimental studies.

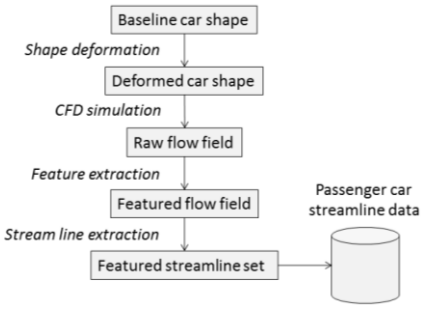


Figure 1, Illustration of the process flow for generating featured streamline sets of deformed baseline geometries.

First, in order to generate different flow fields and streamline sets, shape deformation techniques together with a Latin Hyper Cube sampling have been applied to derive variations of the baseline car geometry. Second, running CFD simulations for all deformed geometries results in different raw flow field datasets, from which feature based streamline representations can be extracted.

Shape Deformation

To create variations of the baseline design, Free-Form Deformation (FFD) [16], a space deformation technique, has been adopted to represent and modify the passenger car's chassis geometry. In contrast to more classical geometric representations, which use an absolute parameterization of the design surface, FFD represents variations of a chosen baseline object. Applying the FFD representation to the passenger car design requires setting up one or multiple three dimensional deformation grids, which are made of a fixed number of control points. The deformation grids need to be designed so that they are embedding parts of the point based representation of the baseline object. *Figure 2* shows the side view on the used baseline car geometry, visualized together with three deformation grids that serve as parameterization for generating car shape variations. All in all 208 control points have been defined.

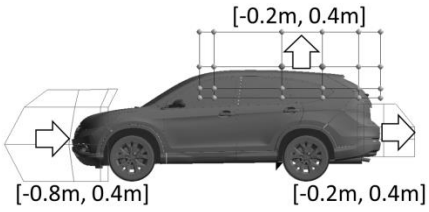
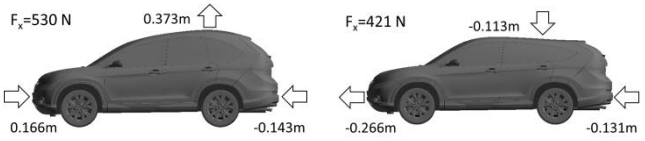


Figure 2, Illustration of the deformation grid, used to generate car surface modifications with FFD.

The control points of the deformation grids define handles for the deformation of the embedded car shape, which is represented by an unstructured triangular surface mesh. Once all points of the surface mesh are "frozen"[1] to the respective control points, the variation of the control point parameters leads to a smooth deformation of the car surface. Exemplary, in *Figure 3* two deformed car shapes are shown, where different deformations with FFD have been applied.



(a) Design 41 (b) Design 49

Figure 3, Illustration of two deformed car surfaces (side view) and the corresponding deformation parameters. Design **41** (left) and design **49** (right) are the one linked to highest and lowest drag force, respectively.

CFD Simulation

For each car shape, the OpenFOAM[2] computational fluid dynamics simulation tool has been applied to determine the flow field around the car. Therefore, a CFD mesh has been generated, i.e., the domain occupied by the air flow is subdivided into discrete cells using the OpenFOAM tool snappyHexMesh. The resulting CFD mesh consists of approximately 1.7 million cells. Given the CFD mesh together with pre-defined boundary conditions, e.g., defining the velocity of the incoming air stream passing the car shape, the simulator iteratively solves the Raynolds averaged Navier-Stokes equations including the SST $k-\omega$ turbulence model until a steady state has been reached. The different car shapes have been simulated at the speed of 110 km/h. From the resulting three dimensional flow fields, which defines the velocity of the air flow at each discrete position in the CFD mesh, global performance numbers like the overall drag force are derived, typically defining the objective for the engineering design process.

Featured Streamline Extraction

Each CFD simulation results in 3d flow field, which contains flow velocity information at about 1.7 million locations in the vicinity of the car surface. The investigation or computational processing of the entire raw flow field is typically a time-consuming process, especially when targeting to integrate information from multiple CFD simulations. For a more efficient processing of the flow field information, a compact feature based streamline representation is extracted by adopting state of the art streamline extraction algorithms, which are typically used for visualizing the flow. Therefore, a fixed number of streamlines are extracted from each flow field by uniformly seeding virtual particles at the flow inlet plane. In this study, $N_l = 2450$ streamlines have been derived by a greedy seeding strategy, in order to not lose relevant flow information. Finally, the flow features $M_v, M_t, M_d, M_s, M_\omega$ have been

---

[1] The freezing step transfers all object points from the Cartesian into a trivariate spline coordinate system.

[2] OpenFOAM is an open source CFD software package, see http://www.openfoam.com.

derived at the location of each sample point along the individual streamlines to retrieve the featured streamline representation ($L = \{l_i\}, i \in [1, N_l]$) of the 3d velocity vector field.

Latin Hyper Cube Sampling

In order to generate a variety of different flow fields and their respective featured streamline set representations, various car shapes need to be generated and the flow needs to be simulated on the computer. Therefore, we consider the FFD representation as depicted in *Figure 2*. The parameterization of the control point groups is the basis for the deformation of the baseline car geometry. The control point group defined for the grid along the roof allows for deformations in z-direction, whereas the other two control point groups allow for deformations in x-direction, modifying the front and rear of the car, respectively. Finally, with three parameter values we can generate various car surfaces with slightly different front, rear and roof shape. The variation in the shape will lead to a different flow field after running the CFD simulation, and hence different featured streamline sets.

For an efficient generation of the design and flow field data set, an optimized Latin Hyper Cube Sampling (LHS) [4] technique has been used to result in a uniform sampling of the deformation parameters. The optimized LHS method applies the optimality criteria of Morris and Mitchel [10] to achieve a space filling sampling in the deformation parameter space. Deformation parameters are sampled within the parameter range as depicted in *Figure 2*, e.g., parameter values that are related to the deformations of the roof of the car are restricted to a control point displacement in z-direction between 0.2m and 0.4m.

For generating the artificial car dataset, the LHS method has been applied to generate 100 parameter configurations, based on which FFD has been used to deform the baseline car surface. The different car surfaces undergo the CFD simulation, resulting in 100 different flow fields. From the different 3d vector fields streamlines are extracted and features along the streamlines have been calculated. In addition to the streamline sets the overall drag force has been calculated for each car from the raw flow field. *Figure 3* visualizes two deformed car shapes and the corresponding deformation parameters for the highest (left) and the lowest (right) drag force, respectively.

Automatic Streamline Reduction

Each of the 100 different featured streamline sets in the passenger car dataset contains flow feature data at the position of the seeding points for $N_l = 2450$ streamlines. Due to the greedy seeding strategy, in order to not lose relevant flow information, the number of streamlines and the amount of data that needs to be investigated or processed remains large. The current streamline sets are likely to contain streamlines that do not provide any information about the flow characteristic.

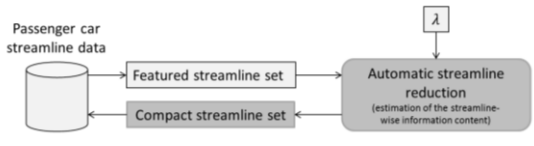
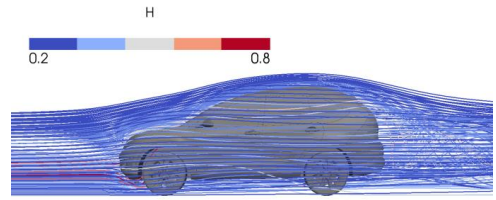


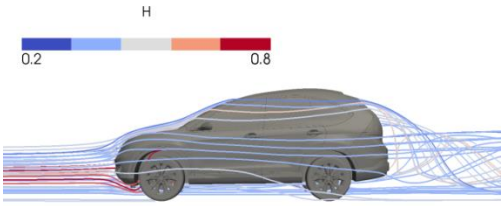
Figure 4, the automatic streamline reduction procedure transforms the featured streamline set into a reduced streamline set.

We apply an explicit streamline extraction step, as depicted in *Figure 4*, to reduce the flow field representation to a set of the most informative streamlines. Whether, a streamline remains in the reduced streamline set depends on the variation of the different feature values along the streamlines. To quantify the information content based on the flow features, for each streamline the expected entropy $\bar{H}_i$ has been calculated, given the feature distributions of the features: $M_v, M_t, M_d, M_s, M_\omega$.

All streamlines are ranked according to the expected entropy. Setting up the threshold λ and keeping only the streamlines with the highest expected entropy results in a compact streamline representation. Exemplary, the compact set of streamlines, with $\lambda = 0.5$ and $\lambda = 0.2$ are shown in **Figure 5**. The value of the expected entropy for each streamline is mapped on the color scale. For the resulting reduced streamline sets with $\lambda = 0.5$ and $\lambda = 0.2$ only 161 and 46 out of 2450 streamlines remain to describe the most relevant flow patterns in close vicinity around the car surface.



(a) $\lambda = 0.5$



(b) $\lambda = 0.2$

Figure 5, Illustration of the compact set of streamlines representing the flow data of design **41** for different values of $\boldsymbol{\lambda}$.

In order to investigate the relationship between λ or better the information content $IC(L')$ and the number of streamlines that remain in the compact streamline representation, the method for automatic streamline reduction has been applied to all of the 100 flow fields. The results for $\lambda = \{0.2, 0.5, 0.7, 0.9\}$ are summarized in *Figure 6*. The box plot visualizes the statistics of the remaining number of streamlines for all flow fields. It can be seen, e.g., that only around one fourth of the number of streamlines is required to keep 70% of the overall expected entropy in the compact streamline set. Applying a larger data reduction with $\lambda = 0.2$ less than 5% of the streamlines are required to describe more than 20% of the information about the flow around the car.

However, in general the number of streamlines that remain in the reduced streamline data set strongly depends on the initial seeding strategy, when converting the raw flow field into a featured streamline set. The correct choice of λ strongly depends on the applications, which act on the reduced representation. In section *Flow Field Retrieval* the best choice for the parameter λ has been studied for the flow field retrieval application.

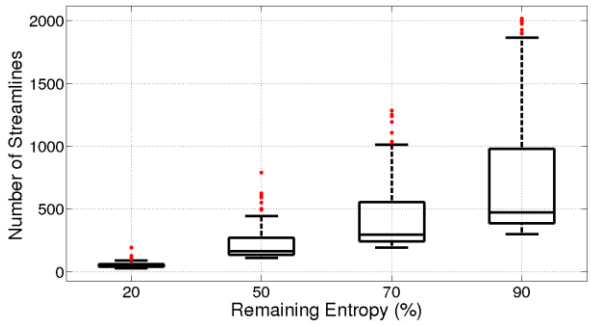


Figure 6, Boxplot visualizing the relationship between the remaining entropy, $\boldsymbol{IC(L')}$, and the number of remaining streamlines, over the data of **100** different flow fields.

In this study, the automatic reduction of the streamlines is carried out for two reasons. On the one hand for visualizing the most relevant streamlines to the aerodynamic engineer, and on the other hand to end up with a flow representation that allows a more efficient processing of the flow data. Concerning the streamline based representation, the memory consumption and computation time is directly linked to the number of streamlines that need to be processed.

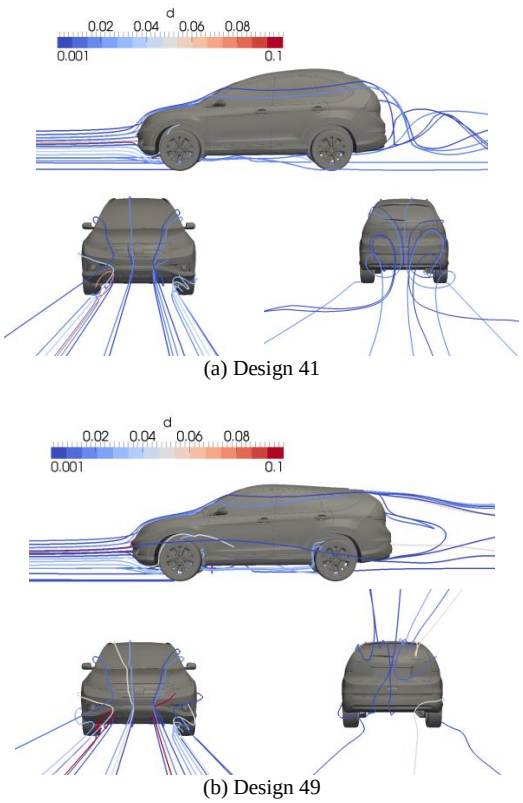
Streamline Wise Flow Field Distance

The reduced streamline representation is the basis for an efficient calculation of the similarity or difference between flow fields. Understanding the detailed differences between flow fields can provide knowledge for the cause of the change of the aerodynamic performance of a car. The overall difference between streamline sets is determined based on the individual distance values of corresponding streamlines, by means of estimating the similarity of the feature distributions of corresponding streamlines, as described above.

In the following, the compact streamline sets of design 41 and design 49, with $\lambda = 0.2$, are compared against the compact streamline set of the baseline design. The respective distances of the corresponding streamlines have been calculated using the X^2distance measure. The results are visualized in *Figure 7*. The color of the visualized streamlines depicts the actual distance value. In the figure, only those streamlines are visualized, where a corresponding streamline to the baseline streamline set has been found and the distance value is larger than zero. Thus, e.g., for the streamline set of design 49 correspondencies to the baseline set are found for 21 out of 32 streamlines are visualized. Analyzing the distances of the individual streamlines can give indications on the cause for the overall aerodynamic performance difference between designs. For example, the most different streamlines in *Figure 7 (a)* emphasize those flow field variations that are the potential cause for the increase in drag of design 41, with $F_x = 530N$ ($c_d = 0.3531$), with respect to the baseline flow field, with $F_x = 433N$ ($c_d = 0.3089$).

In addition to the distances of the corresponding streamlines, the visualization of the streamlines, which are unique to one or the other streamline set, where no correspondences were found, can be of interest for the analysis of the similarities as well.

The accumulation of the individual streamline distances results in the overall distance value that quantifies the similarity between the flow data sets, which is $D = 0.338$ and $D = 0.546$ for the streamline set of design 41 and 49 in comparison to the baseline set, respectively.



(a) Design 41

(b) Design 49

Figure 7, Illustration of the distance between corresponding streamlines, comparing the compact streamline representation of the baseline design with that of (a) design **41** and (b) design **49**, given compact streamline sets with $\boldsymbol{\lambda = 0.2}$.

This indicates that the aerodynamic flow characteristics of the baseline design is more similar to the one of design 41, rather than to the streamline set of design 49. Although, the flow characteristics is more similar to the one of design 41, the baseline design has with $F_x = 433N$ ($c_d = 0.3089$) a low aerodynamic drag which is closer to the drag force of design 49. This can be an indication that the change in the drag force is not directly correlated with the overall flow field distance, rather with the change of individual flow patterns or streamlines.

Flow Field Retrieval

The estimation of distances between flow fields defines the general basis for comparing CFD simulation results and as such the background for many data mining applications. One application is related to the retrieval of the most similar (or different) flow field or flow fields in a database. In the retrieval process, depicted in *Figure 8*, given a queried streamline set, the goal is to seek for the most similar streamline set over hundreds or thousands streamline sets stored in a database. Thus, the target of the flow field retrieval process is to identify those car related flow fields, which are most similar to the queried flow field.

The computational efficiency of the flow field retrieval algorithm strongly depends on the representation of the data that needs to be processed. Therefore, in the following studies the retrieval process

and the results for streamline sets with different information content are under investigation, comparing the retrieval accuracy and the related computation time.

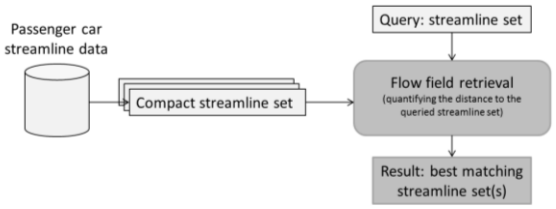


Figure 8, given a query by means of a streamline set the flow field retrieval application searches for the most similar streamline set in the database and outputs the best matching set or sets.

In the conducted experiments the generated streamline sets related to the 100 designs are stored in a database, except the streamline set related to the flow of design 41. In these experiments the streamline set related to design 41 is used to query the database. For each design the streamline sets with different information content are generated (by means of applying the automatic streamline reduction with $\lambda = \{0.2, 0.5, 0.7, 0.9\}$) and stored. For reference, also the initial streamline set with a fixed number of 2450 streamlines has been stored. In order to search for the most similar streamline set in the database, the distances of the query streamline set, L_Q, to all sets in the database are calculated and ordered accordingly.

The results for seeking the most similar streamline sets with respect to L_Q, are summarized in Table 1. The first column depicts the values chosen for λ, which implicitly defines the size of the streamline set. In addition the explicit size $|L_Q|$, by means of the number of streamlines is given in the second column. After that, the design indices of the three most similar streamline sets for different levels of reduction λ are also listed, followed by the time fraction required to compute all the distances compared to using raw set with 2450 streamlines. The last indicator, the fraction of correct matches quantifies how many of the streamline sets are ordered correctly based on the calculated distances, compared to the order when using the raw streamline set.

Table 1, Summary of the flow field retrieval results for design **41**.

λ	$\|L_Q\|$	Best matches id	Time fraction	Correct matches
1.0	2450	**77,73,95**	100.0%	100.0%
0.9	423	**77,73,95**	≈22.0%	90.0%
0.7	264	**77,73,95**	≈ 17.8%	90.0%
0.5	161	**77**,95,73	≈12.5%	50.0%
0.2	46	95,77,73	≈2.5%	10.0%

Investigating the best match, it can be noted that even with a small fraction of the most relevant streamlines, up to $\lambda = 0.5$, (with 161 out of 2450 streamlines) the same retrieval result with respect to the most similar streamline set (index 77) is obtained, while cutting down the computation time almost by a factor of 10. Thus, if only the best matching streamline set is of relevance for the aerodynamic engineer, only less than a tenth of the streamlines compared to the raw streamline set needs to be stored in the database for the retrieval task. Even with $\lambda = 0.2$ the three most similar streamline sets are identical. Only the order of the best matches changes, however, with a significant reduction in computation time. This shows that the defined compact representation (even for small values of λ) can serve as an appropriate representation for an efficient estimation of flow field similarities, thus, resulting in an efficient retrieval process.

Groups of Similar Flow Fields

So far, the experimental studies focused on the investigation of similarities or dissimilarities for individual streamline sets, respective flow fields. Calculating the distances over all combination of streamline sets can provide means to identify larger groups of similar flow fields from a large database. The automatic identification of such groups can be the source of inspiration for deriving new aerodynamic concepts and for supporting decision making in the engineering design process.

Given all 100 streamline sets, plus the streamline set of the baseline design (index 100), the complete distance matrix has been calculated, with $\lambda = 0.5$, visualized in *Figure 10(a)*. The distance matrix contains the distance value for any streamline set comparison. The extraction of valuable information by visual inspection of the distance matrix is mostly impractical. As for streamline set 5, often only strikingly different streamline sets can be identified, which show up a high dissimilarity with the majority of the compared streamline sets. A more detailed analysis or the identification of groups of similar or dissimilar streamline sets (e.g., identifying all those streamline sets, which are similar to the streamline set of design 45) is hardly possible from the inspection of the distance matrix solely.

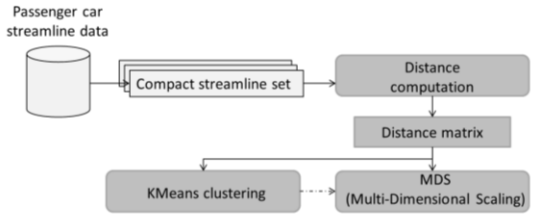
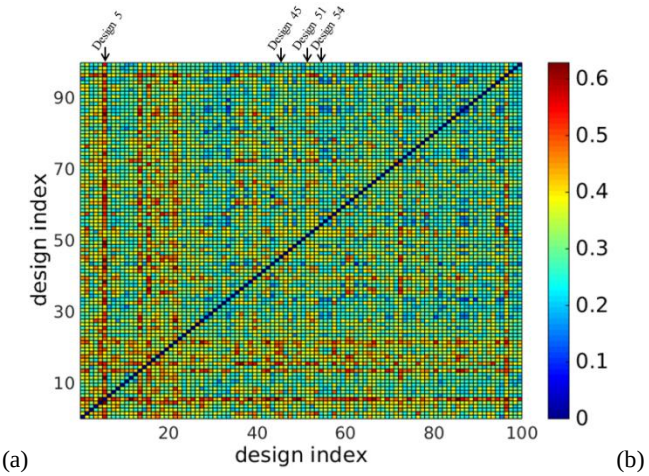


Figure 9, process for finding groups of similar flow fields based on the reduced streamline set representation. The results of the distance calculation and grouping are visualized on a two dimensional map using the MDS technique.

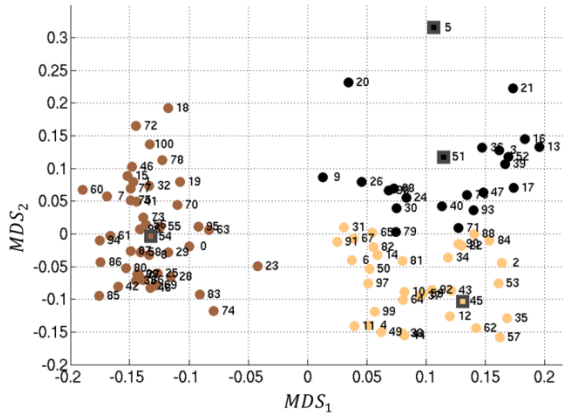
Therefore, multi-dimensional scaling (MDS), a classical data mining technique, has been adopted to visualize and analyze the level of similarity over all streamline sets, subsequent to the distance calculation, as conceptually shown in *Figure 9*. *Figure 10(b)* and *(c)* show the result of the non-metric MDS optimized for 50 iterations. Each point in the two dimensional plot, labeled with the related design index, represents one streamline set. The Euclidean distance in the two dimensional metric space, defined by MDS_1 and MDS_2, is proportional to the actual distance values in the distance matrix. From the resulting two dimensional projection it can be seen that point 5 is far away from all the others, what emphasizes the dissimilarity of the streamline set of design 5, already observed from the inspection of the distance matrix. The rear view on the streamline set linked to design 5 is shown in *Figure 11 (a)*.

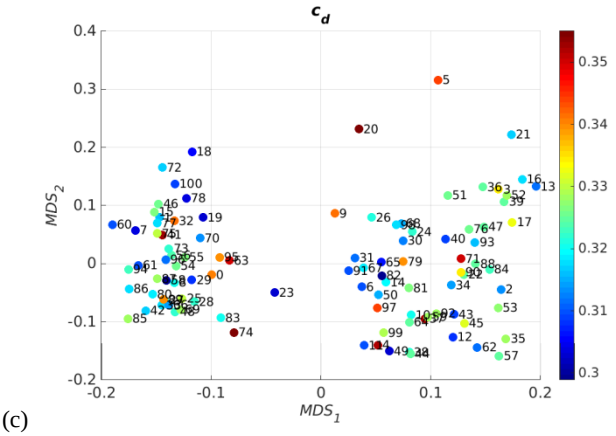
Furthermore, the two dimensional projection reveals two clearly separable groups of similar streamline sets. In order to automatically group the streamline sets into distinct groups, KMeans clustering with $k = 3$ has been applied. As depicted with the color value in *Figure 10(b)*, the clustering is able to automatically separate the two

groups on the left from the one on the right. Since k has been set to three, the large group on the right has been split into two separate clusters.



(a) (b)

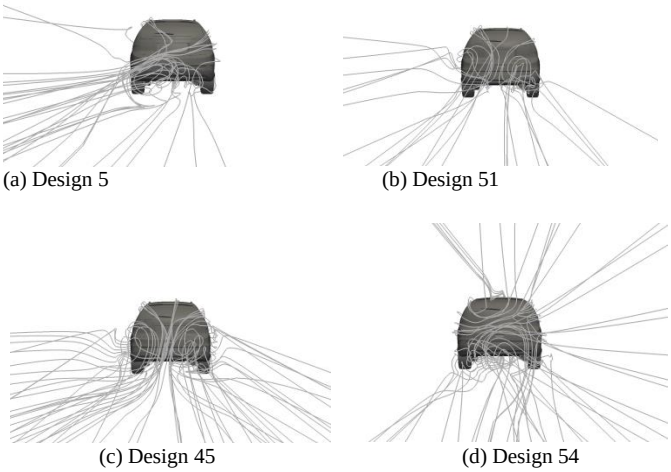




(c)

Figure 10, Illustration of (a) the distance matrix comparing the flow fields based on the streamline set representation, (b) and (c) the visualization of the distances after projecting the streamline sets on a two-dimensional map using MDS, where (a) shows the results of the automatic grouping of flow fields using KMeans clustering and (b) visualizes the drag coefficient values linked to the respective flow fields.

In order to give an impression about the differences between the individual groups, one streamline set out of each group has been selected and the rear view on the respective streamline sets are shown in *Figure 11 (b)* to *(d)*. They show the streamline set of design 51, 45 and 54, respectively. From a visual inspection, first, it can be seen that the streamline sets have a different characteristic, and second that the streamline sets of design 45 and design 51 are more similar compared to the set of design 54, which is confirmed by the fact that they are closer on the MDS projection.



(a) Design 5 (b) Design 51

(c) Design 45 (d) Design 54

Figure 11, Illustration of the compact streamline set marked on the two dimensional projection in Figure 10(b).

The three groups identified by the automatic clustering procedure might provide hints on common aerodynamic properties. In order to investigate the interrelation between the drag coefficient and the streamline set distance the drag coefficient, $c_d$has been calculated ($\rho = 1.225 \frac{kg}{m^3}$) and visualized for each design on the two dimensional projection in *Figure 10(c)*. It can be seen that there is no direct interrelation between the flow field distance and the drag coefficient. This emphasizes that the analysis of the flow field distance, as described in this paper, provides additional information about the flow, which is not captured by the drag coefficient.

Conclusion

In this paper a new attempt for deriving a reduced streamline wise flow field representation has been introduced, where streamlines are filtered based on the evaluation of its expected flow feature entropy. The reduced streamline set serves for visualizing the most relevant streamlines and for an efficient comparison of simulation results. In order to quantify the flow field similarity a streamline wise distance measure that estimates the differences of the feature distributions between corresponding streamlines has been introduced. The definition of an appropriate distance or similarity measure is the basis for several data mining tasks like flow field retrieval or for clustering flow fields to identify groups of similar flow fields. The theoretic concepts for quantifying the flow field distances based on a compact streamline representation have been verified by experimental results from the simulation of various passenger car shapes. It has been shown that the compact flow field representation results in an efficient distance calculation exemplified for the flow field retrieval task. With a significant drop in computation time and memory consumption an accurate retrieval result had been achieved. Furthermore, the use of MDS and k-means clustering techniques has shown to be a valid tool for visualizing the distances of flow field comparisons and to establish groups of car shapes with similar flow field. The interrelation of the flow field distances to other aerodynamic properties like noise and the transfer of those

technologies to time-varying flow fields is a big challenge and remains future work.

References

[1] A. Brambilla, R. Carnecky, R. Peikert, I. Viola, and H. Hauser. Illustrative flow visualization: State of the art, trends and challenges. In Marie-Paule Cani and Fabio Ganovelli, editors, *EG 2012 - State of the Art Reports*, pages 75–94, Cagliari, Sardinia, Italy, 2012. Eurographics Association.

[2] K. Chiba, S. Jeong, S. Obayashi, and H. Morino. Data mining for multidisciplinary design space of regional-jet wing. *IEEE Congress on Evolutionary Computation*, 3:2333–2340, 2005.

[3] S. Depardon, J. Lasserre, L. Brizzi, and J. BorÃ©e. Automated topology classification method for instantaneous velocity fields. *Experiments in Fluids*, 42:697–710, 2007.

[4] A. Forrester, A. Sobester, and A. Keane. *Engineering Design via Surrogate Modelling: A Practical Guide*. Wiley, 2008.

[5] C. Garth, H. Krishnan, X. Tricoche, T. Tricoche, and K. I. Joy. Generation of accurate integral surfaces in time-dependent vector fields. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1404–1411, 2008.

[6] C. Garth, R. Laramee, X. Tricoche, J. Schneider, and H. Hagen. Extraction and visualization of swirl and tumble motion from engine simulation data. *Topology-based Methods in Visualization*, pages 121–135, 2007.

[7] L. Graening and B. Sendhoff. Shape mining: A holistic data mining approach for engineering design. *Advanced Engineering Informatics*, 28(2):166 – 185, 2014.

[8] Lars Gräning and Thomas Ramsay. Quantifying vector field distances based on a reduced streamline representations. In H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, B. Li, T. Weise, and X. Yao, editors, *Int. Conf. on Intelligent Data Engeneering and Automated Learning (IDEAL)*, Lecture Notes in Computer Science, pages 445–454. Springer, October 2013.

[9] R. S. Laramee, D. Weiskopf, J. Schneider, and H. Hauser. Investigating swirl and tumble flow with a comparison of visualization techniques. In *Proceedings IEEE Visualization'04*, pages 51–58. IEEE Computer Society, 2004.

[10] M. D. Morris and T. J. Mitchell. Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43:381–402, 1995.

[11] S. Obayashi, S. Jeong, K. Shimoyama, K. Chiba, and H. Morino. Multi-objective design exploration and its applications. *International Journal of Aeronautical and Space Science*, 11(4):247–265, 2010.

[12] O. Pele and M. Werman. The quadratic-chi histogram distance family. In *Proceedings of the 11th European conference on Computer vision: Part II*, ECCV'10, pages 749–762, Berlin, Heidelberg, 2010. Springer-Verlag.

[13] F.H. Post, B. Vrolijk, H. Hauser, R.S. Laramee, and H. Doleisch. Feature extraction and visualization of flow fields. *Eurographics 2002 State-of-the-Art Reports*, pages 69–100, 2002.

[14] M. Rath and L. Graening. Modeling design and flow feature interactions for automotive synthesis. In *Int. Conf. on Intelligent Data Engeneering and Automated Learning (IDEAL)*. Springer, 2011.

[15] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[16] T.W. Sederberg and S.R. Parry. Free-form deformation of solid geometric models. *ACM Siggraph Computer Graphics*, 20(4):151–160, 1986.

[17] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, July 1948.

[18] L. Xu, H. Q. Dinh, P. Mordohai, and T. Ramsay. Detecting patterns in vector fields. In *AIAA Aerospace Sciences Meeting*, Orlando, Florida, USA, 2011.

[19] X. Ye, D. Kao, and A. Pang. Strategy for seeding 3d streamlines. In *Visualization, 2005. VIS 05. IEEE*, pages 471–478, 2005.

Acknowledgments

The authors would like to thank Stefan Menzel and Sebastian Schmitt from the Honda Research Institute Europe GmbH for their valuable comments and advice.

Definitions/Abbreviations

CFD	Computational Fluid Dynamics
FFD	Free-Form Deformation
LHS	Latin Hypercube Sampling
MDS	Multi-Dimensional Scaling
SOM	Self-Organizing Maps
ANOVA	Analysis of Variance
IC	Information Content of a streamline set
D	Distance between two streamline sets