A Statistics-based Dimension Reduction of the Space of Path Line Attributes for Interactive Visual Flow Analysis

Armin Pobitzer* University of Bergen, Norway
Alan Lež† VRVis Research Center, Austria
Krešimir Matković‡ VRVis Research Center, Austria
Helwig Hauser§ University of Bergen, Norway

ABSTRACT

Recent work has shown the great potential of interactive flow analysis by the analysis of path lines. The choice of suitable attributes, describing the path lines, is, however, still an open question. This paper addresses this question performing a statistical analysis of the path line attribute space. In this way we are able to balance the usage of computing power and storage with the necessity to not loose relevant information. We demonstrate how a carefully chosen attribute set can improve the benefits of state-of-the art interactive flow analysis. The results obtained are compared to previously published work.

Index Terms: I.3.8 [Computing Methodologies]: Computer Graphics—Applications; I.6.6 [Computing Methodologies]: Simulation and modeling—Simulation Output Analysis

1 INTRODUCTION

When analyzing the dynamics of unsteady flow, the investigation of particle movements is a canonical choice. In order to enable further analysis based on particle paths, these trajectories need to be characterized. Possible ways to describe the paths include deriving measures for their global and local behavior and properties of the field around the moving particles. A large number of such feature detectors is available and has been used in different contexts [12, 24, 25, 28].

Previous work by Bürger et al. [1], Shi et al. [28] and Lež et al. [17] has shown the great potential of the combination of *Interactive Visual Analysis* (IVA) and feature extraction. However, the question of how to choose an adequate attribute set to investigate is left open, although it is non-trivial. Feature detectors are usually designed to target one specific aspect of the flow behavior. An ad hoc choice of suitable attributes is therefore dependent on correct prior knowledge (or assumptions) on what type features to expect. This has the implication that unexpected behavior is possibly hard to detect. Therefore, a objective and complete investigation of the data set in question would require to look for "all possible" features (say vortices, vortex core lines, path lines with low average speed,...) and their detectors at once.

This leaves us with a large amount of possibly interesting features and even more detectors that should be considered. Computing all of them is tedious at least and results in a high computation time and storing a large number of attributes per trajectory. It can be expected that this brute force approach would generate a considerable information overhead, since many of the attributes are computed from the same velocity field. In general, different feature detectors may systematically correlated to each because they either describe the same aspect of the flow or are related to each other by physical principles (e.g., velocity and vorticity by the vorticity equation). From the practical side, a systematic analysis of a data set gets increasingly challenging the more dimensions it contains. Hence, a canonic question in this context is: Is there a common subset of the path line attributes that captures "all" complexity of the data sets? Or, in short, what is the intrinsic dimensionality of the path line attribute space?

The problem of analyzing high dimensional data sets is a classic challenge, that both statistics [13] and visualization [6] deal with, as well as others. Roughly speaking, the main distinction between these two approaches to multivariate data analysis, is the role of the user: while statistics relies on automatic methods, visualization-based approaches try to exploit a larger amount of user interaction [22, 27]. The benefits and drawbacks of the two approaches can be considered complementary. One should therefore aim to combine the strengths of both, namely rigidness and inherent objectiveness of statistical methods for determining the intrinsic dimensionality of the data and the flexibility and possibility of integration user knowledge of IVA in the analysis stage.

In statistics, a number of dimension reduction methods are available [5], two of the most prominent being *principal component analysis* (PCA) [21] and *exploratory factor analysis* (EFA) [29]. In short, the first finds orthogonal principal components (linear combinations of the observed variables) that account for the maximal amount of variance. Although efficient for the mere dimension reduction purpose, one of the draw-backs of this method is that the principal components are usually hard to interpret and the computation of them involves possibly all observed variables. Modifications of PCA that attempt to avoid this have been proposed [36], imposing additional constrains on the algorithm. Exploratory factor analysis gives the number of statistical variables, called *factors*, needed to explain the common variance between variables and how these factors *load on* (are correlated to) the observed variables. These factors are assumed to be not directly measurable, hence the actual interesting information being the *loadings* (correlation coefficients) on the observed variables. Taking the highly loaded variables for every factor yields a set of variables that are interpretable, account for the complexity of the data, and the set is (if the data allows for this) of considerable smaller dimensionality. Hence, exploratory factor analysis is a more promising choice for the dimension reduction for our purposes. For a more detailed overview of the similarities and differences between PCA and EFA, we refer to Suhr [30].

In this paper we investigate several CFD data sets with exploratory factor analysis, with the goal to find a common set of variables that can be used as a starting point for a deeper analysis of CFD data sets. This variable set should capture the underlying physical processes in fluids with a little as possible redundancy. The analyzed data sets span different geometries, constant/non-constant inflows as well as different simulation methods to prevent the variable set from being specific for one type of simulation/geometry/application.

We compare the results from IVA applied on the attribute set found in our investigation to previously published results. Our results match and also partially exceed previous ones. In contrast to previous work we have a predefined attribute set, which makes a more systematic analysis possible.

The remainder of this paper is organized as follows: first we briefly discuss previous work, and then we describe our statistical analysis and present its results. We give a demonstration of the

*e-mail: armin.pobitzer@uib.no
†e-mail: alanlez@vrvis.at
‡e-mail: kresimir@vrvis.at
§e-mail: helwig.hauser@uib.no

results achievable from the combination of our findings and IVA, comparing these results to previous ones.

2 Related Work

Because of their tight relation to the dynamic behavior of the flow, visualization by means of particle trajectories is a well established branch of flow visualization [19].

Theisel et al. [32] introduce the classification and segmentation of path lines according to attracting, repelling and saddle-like behavior for visualization purposes. This classification allows the authors to identify a path line-based topology for two-dimensional unsteady flows. Salzbrunn and Scheuermann [26] introduce a mathematical framework based on Boolean algebra that allows to define a topology based on so-called streamline predicates. These predicates are user chosen and can be seen as path line attributes with Boolean range. Later, Salzbrunn et al. [25] extend this approach to path lines.

Bürger et al. [1] investigate the opportunity to combine several feature detectors making use of interactive visual analysis, focusing on vortical features. Shi et al. [28] present a similar approach together with more general path line attributes, using both local and global descriptors for the path line behavior. Lež et al. [17] enhance the utility of path line based IVA by the possibility for direct path line brushing via projections.

The problem of dimension reduction in high-dimensional data sets is a well established research field within statistics. Pearson published his seminal work on *principal component analysis* [21] in 1901. Spearman laid the foundation for *factor analysis* with his 1907 article on the "true measurement of correlation" [29]. Since, a large number of related methods and algorithms has been presented. For an overview we refer to Fodor's survey on this topic [5].

In the context of information visualization, the possibility for user-guided dimension reduction has been investigated. Yang et al. present a method called *Visual Hierarchical Dimension Reduction* (VHDR) [35]. VHDR clusters the data dimensions according to similarity measures, generating a dimension hierarchy. The user selects clusters and specifies "representative dimensions" for those clusters. Finally, a projection step is applied.

Seo and Shneiderman present the *rank-by-feature* framework [27], that allows the user to rank the dimensions by some simple statistics for one-dimensional and two-dimensional representations of the original dimension and dimension pairs, respectively. Piringer et al. [22] extend this approach to the investigation of user-specified subsets of the original data set instead of dimensions only. Recently, Turkay et al. [33] presented a visualization model that allows for interaction in both *item* and *dimension space*. This eases the understanding of the relation of different data dimensions and the according analysis of high-dimensional data sets, yielding means of interactive dimension reduction.

This paper is targeting dimension reduction for interactive flow analysis along the lines of the works of Bürger et al. [1], Shi et al. [28], and Lež et al. [17].

3 Statistical Analysis

In this section we firstly give a description of the statistical methodology we use, and then describe the actual analysis of the data.

3.1 The Statistical Model

As explained in the introduction, we chose exploratory factor analysis (EFA) to investigate the dimensionality of the path line attribute space. It is worthwhile noticing that modern EFA is more a group of methods than one single algorithm. Since we want to find the minimal number of factors explaining the variation in the data set, we have to choose *Principal Factor Analysis* (also known as *Common factor analysis*) [8]. In order to increase the numerical stability, an iterative algorithm is used [8]. Since we are interested in factors that can be related back to one (or more) attributes that we can compute, we discharge the usual assumption of uncorrelated factors and use the so-called *varimax criterion* instead [8]. In short, this criterion tries to maximize the variation in the factor loadings onto the variables. This yields often the situation that each variable virtually loads one factor only [8]. For a thorough discussion of these algorithmic choices, and possible alternatives, we refer to Harman's book [8].

One crucial aspect of a factor analysis is the criterion that determines how many factors have to be retained. The eigenvalues of the respective factors give information on how much of the variance is explained by the single variable, compared to a uniform distribution of the variance. Hence, Kaiser [14] suggests to retain all factors associated to an eigenvalue greater than 1. Cattell suggests the use of the plot of the eigenvalues against their index to determine the right number of factors to retain [2]. The factors that lay on the *scree* (i.e., the base of a steep incline or cliff) of the plot are considered neglectable, therefore this criterion is commonly referred to the *scree plot test* [2]. Finally, if the goal is to guarantee that the retained contain a certain percentage of variance, one can simply include factors until their relative weight exceeds a desired threshold. Kaiser's criterion has the advantage of being objective, but has proven to be unreliable in extracting the true number of underlying factors [3]. Better results are obtained using the scree test [3]. Here the drawback lies in fact that this is a "soft" criterion that relies on the users interpretation of the scree plot. Finally, retaining factors accounting for more than 100% of the variance will not add information about the data set but noise. Hence, we consider the maximum of the factor number suggested by Kaiser's criterion and the scree plot test, with 100% of explained variance (or proportion) as a limiting bound. For a more thorough discussion of how to choose the correct number of factors to retain, we refer the reader to the article of Costell and Osbourne [3].

All statistical computations for this paper have been carried out with SAS© software.

3.2 The Data Sets

In total we analyzed 5 different data sets with different geometries and simulation methods and 1 analytic data set. For the greatest possible generality, we use only the velocity fields to calculate the path lines and their attributes. This means that the similar factor patterns across the data sets are due to the common underlying principles of fluid dynamics and not due to similarity in the data sets. The data sets investigated are the following:

Flow through a box: This data set is the simulation of flow through a box. The data set consists of 100 time steps. The inlet is on the top of the box. The data set consists of 17120 cells organized in a Cartesian grid.

T-junction: This data set is the simulation of flow through a T-junction with two inlets and one obstacle inside. The data set consists of 100 time steps. One inlet is in horizontal direction, another one in vertical direction. The obstacle is placed under the vertical inlet. The fluid flows through the horizontal inlet first, while the inflow from the top begins after some time. The data set consists of 30930 cells organized in a Cartesian grid.

Breaking dam: This data set is a flow simulation of a bursting dam with a box-shaped obstacle. The data set consists of 48 time steps. The burst occurs in the first time step. The data set consist of 76505 cells, organized in a Cartesian grid.

Exhaust manifold: This data set is a flow simulation of an exhaust manifold. The data set consists of 69 time steps, covering one inflow from every of the three exits from the cylinders. The data set consists of 36524 cells organized in an unstructured grid.

data set	param.	nr. of factors (Kaiser/scree/prop.)	factor: var. with highest loading (loading)
flow through a box	time	6/7/11	1:λ_2 (0.95), 2:inv (0.99), 3:startEnd (0.97), 4:vel (0.87), 5:usH (0.85), 6:windang (0.65), 7:dpos (0.81), 8:vel (0.71)
	space	5/5/9	1:λ_2 (0.96)/Hunt's Q (-0.94), 2:inv (0.98), 3:vel (0.86), 4:uSH (0.86)/SH (0.85), 5:startEnd (0.92), 6:pos (0.68)
t-junction	time	5/8/6	1:inv (0.98), 2:startEnd (1.00)/avspeed (0.99), 3:λ_2 (0.94)/Hunt's Q (-0.93), 4:SH (1.00), 5:pos (0.88), 6:avspeed (0.71)
	space	5/7/7	1:startEnd (0.99)/avspeed (0.97), 2:inv (0.89), 3:SH (0.96)/uSH (0.94), 4:λ_2 (0.95)/Hunt's Q (-0.93), 5:inv (0.78), 6:pos (0.86), 7:avspeed (0.58)
breaking dam	time	5/5/7	1:vort (0.95), 2:SH (0.94)/uSH (0.93), 3:inv (1.01), 4:startEnd (0.95)/ avspeed (0.94), 5:λ_2 (0.89)/Hunt's Q (-0.85)
	space	4/4/8	1:vort (0.97), 2:inv (1.01), 3:dpos (0.67)/vel (0.63), 4:uSH (0.89)/SH (0.87), 5:Hunt's Q (0.94), 6:pos (0.73)
exhaust manifold	time	4/8/8	1:vort (0.94), 2:SH (0.92)/uSH (0.92), 3:inv (0.92), 4:startEnd (0.94)/avSpeed (0.93), 5:λ_2 (0.89)/Hunt's Q (-0.85)
	space	4/8/7	1:vort (0.87), 2:avspeed (0.97)/dpos (0.94), 3:avspeed (0.97)/dpos (0.94) 4:inv (0.78), 5:avspeed (0.98)/dpos (0.93)
turb. chan. flow	time	5/7/5	1:inv (0.99), 2:λ_2 (1.00), 3:vel (0.93), 4:vel (0.92), 5:vel (0.94)
	space	5/7/6	1:inv (0.99), 2:λ_2 (0.96)/Hunt's Q (-0.95), 3:vel (1.00), 4:vel (1.00), 5:vel (1.00), 6:windang (1.00)
rot. vortex rope	time	7/6/7	1:inv (0.98), 2:vort (0.99), 3:SH (0.99)/normhel (0.99), 4:λ_2 (0.99)/ Hunt's Q (-0.99), 5:avspeed (1.00)/vel (0.96), 6:avspeed (1.00)/vel (0.95)
	space	7/7/9	1:inv (0.97), 2:vort (0.99), 3:SH (0.99)/normhel (0.99), 4:λ_2 (0.99)/Hunt's Q (-0.99), 5:vel (1.00)/avspeed (0.96), 6:avspeed (1.00)/vel (0.95), 7:windang (0.91)

Table 1: Summative result of the statistical analysis on the different data sets, according to their parametrization. The found patterns are discussed in Sec. 3.4.

Turbulent channel flow: This data set is a direct numerical simulation (DNS) of a fully developed turbulent channel flow at frictional Reynolds number Re_τ of 180. The flow domain is bounded by two infinitely large parallel solid walls, and the flow is driven by a constant mean pressure gradient in the stream-wise (x) direction. The boundary conditions are non-slip on the solid walls and periodic else. The data are produced by a *Spectral Element Method* (SEM) solver. The data set consists of 2097152 cells organized in a rectilinear grid.

Rotating vortex rope: This data set is an analytic model of a rotating vortex tube as used by Fuchs et al. [7] with parameters $R = 0.25$, $k = 2$, $\omega = 0.5$ and $s = 3$. The data set consists of 100 time steps and 504063 cells organized in a regular grid.

3.3 The Path Lines and their Attributes

For all data sets, we seed a path line at every cell center and integrate it until the particle leaves the flow domain or the time span described by the data set elapses. The particles are saved at the same time steps the original data sets consist of. Besides the positions we compute both attributes depending on these positions and attributes depending on the velocity field itself, evaluated at the particle positions. The fact that we save particle positions (and attributes) at the original time steps, avoids temporal interpolation of these fields. The investigated attributes with their dimensionality are:

Attributes from positions: position (3), quadratic statistical invariants (3) [18], temporal derivative of position (3), torsion (1), curvature (1), winding angle (1), arc length (1), average speed (3), distance actual position to end position (1)

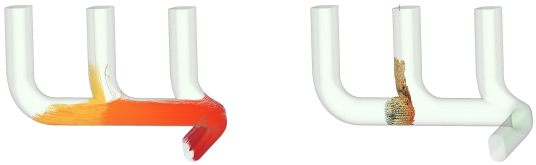
Attributes from velocity: velocity (3), λ_2 (1) [11], Hunt's Q (1) [10], normalized helicity (1) [16], scalar field corresponding to the *Eigenvector method* (Sujudi and Haimes) [31] (1), scalar field corresponding to the *Eigenvector method for unsteady flow* [7] (1), scalar field corresponding to the *Cores of swirling particle motion* [34] (1), vorticity (3)

It is worthwhile noticing that all attributes that would be constant along the path line (average speed, arc length,...), have been computed from the actual position in the time step to the last time steps. For example, the average speed at time step 0 is the average over the whole path line, at time step i the average over the part of the path line starting at its position in time step i to its end. This means that we have a time series for all of the attributes. The information of the usual definition is stored at time step 0 and is therefore easily retrievable. In the statistical analysis we consider the components of attributes independently, since no assumptions on the dependencies of the dimensions of the same attribute can be made. If one of the dimensions is characteristic for the data set, however, we include all of them since the meaning of a dimension can change from data set to data set. For example, the x-velocity may be the stream-wise velocity in one data set and the span-wise in another.

Finally, we investigate all data sets in the above described configurations, as well as sampled evenly with respect to the arc length. This is achieved by a re-parametrization. The arc length parametrized representation has the advantage that certain shape descriptors become unique, e.g., the combination of curvature and torsion (c.f. *Frenet-Serrat formulae* [9]).

3.4 Results

The main results of our statistical analysis are summarized in Table 1. The first column gives the data set in question, the second the parametrization type (time or space). The third column gives the number of retained factors according to Kaiser's criterion, the scree plot test, and the 100% proposition criterion, respectively. The last column gives the factors we consider, according to the principle explained in Sec. 3.1. Next to the index of the factor we give the attribute with the highest loading and in brackets the numerical value of the loading. If there is another attribute that is in a 5%-range, we include it as well. For the full statistical output we refer to the included extra material. The abbreviations used in the table as well as in the full output (in alphabetic ordering) are: arc (arc length), avspeed (average speed), curv (curvature), dpos (time-derivative of the position), HuntsQ (Hunt's Q), inv (quadratic statistical invariants), l2 (λ_2), normhel (normalized helicity), pos (position), SH (eigenvector method according to Sujudi and Haimes), swirl (cores of swirling particle motion), tors (torsion), uSH (eigenvector method for unsteady flow, "unsteady Sujudi and Haimes"), vel (velocity),



(a) path lines (b) seeding positions

Figure 1: (a) path lines meeting the analytic condition defined in Sec. 4.2 and (b) their seeding positions. The color coding gives the temporal evolution of the path line (from yellow to red).

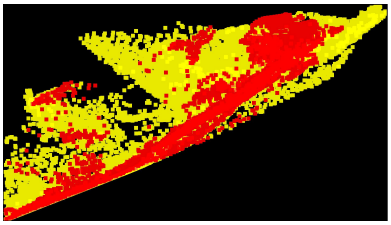


Figure 2: Scatter plot of the start-end distance (horizontal) and path line length (vertical). Red points represent path lines starting in or in the vicinity of the middle pipe. For further discussion see Sec. 4.2.

vort (vorticity), windang (winding angle). The statistical analysis has to answer two questions: first, which dimensionality has the attribute space, and second, which attributes represent these dimensions best?

The dimensionality The average number of factors retained by Kaiser's criterion is 5.25 ($SE = 0.28$, $CV = 0.18$), while the scree plot test retains 6.4 ($SE = 0.34$, $CV = 0.18$) factors, on average. From the proportion criterion we see that 7.5 factors ($SE = 0.47$, $CV = 0.22$) are on average sufficient to explain 100% of the variance in the data set. Our criterion, which balances maximum dimension reduction (Kaiser), a soft user-influenceable criterion (scree plot), and the goal of explaining "all" variance of the data set, retains on average 6 factors, being more stable than the other criteria ($SE = 0.28$, $CV = 0.16$).

The representative attributes For our final suggestion of 6 factors we order the attributes according to their frequency. The four most frequent attributes are inv (12), λ_2 (10), Hunt's Q (8) and avspeed (7). We include all but Hunt's Q, since this attribute is coupled to λ_2 in 7 out of 8 occurrences, and λ_2 is known to outperform Hunt's Q [11]. Two attributes have frequency 6: startEnd, and vel. We include both. From the now retained five attributes, two, namely inv and startEnd, depend on pos (frequency 3), so we decide to include this attribute to make the 6 attributes we investigate as self-contained as possible.

Hence, 6 good candidates for representing the path line attribute space are: the quadratic statistical invariants (inv), λ_2, the average speed (avspeed), the start to end distance (startEnd), the velocity (vel) and the positions (pos).

We evaluate our factor choice by rerunning the analysis with the number of factor to retain fixed to 6 and checking the obtained factor loading pattern for crossloadings and the amount of variance explained as suggested by Costello and Osborne [3]. On average 2.4 ($SE = 0.76$, $CV = 1$) attributes out of 28 exhibit crossloadings, and the average variance explained is 0.95 ($SE = 0.02$, $CV = 0.07$) which shows that the proposed factor structure is both expressive and stable for the investigated data sets. The full output of the statistical evaluation can found in the supplemental material.

A remark on time- vs. arc length-parametrization We observe that the arc length-parametrized data set allows a representation with the same number of factors, or fewer, following Kaiser's criterion or the scree plot test. With the 100% proportion criterion, no clear trend is observable. It is worthwhile noticing that an arc length parametrization represents the geometry of the path line more faithfully, but lacks information on the dynamics (uniform speed with respect to arc length!). Hence, the trend to be expressible by fewer factors may actually originate for that fact that this representation causes an information loss. On the other hand, we see that the shape descriptors inv perform well under both parametrization. Hence, we may conclude that the geometry-wise advantage of an arc length parametrized data set is too small to outweighed the possible risk of information loss.

4 Demonstration

After we determined both dimensionality and representative attributes, we now demonstrate how an interactive visual flow analysis based on our findings can look like. First, we describe the framework used. Then we analyze two different data sets. Both data sets have been investigated in previously published work, which allows us to assess the results we achieve.

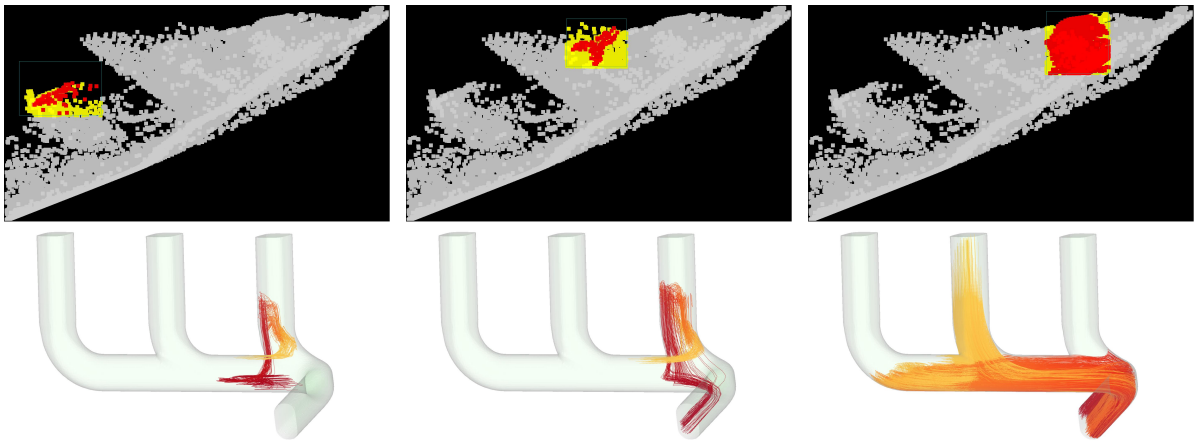
4.1 The framework

The framework used for this paper is the SimVis software [4]. This software is an *interactive visual analysis* environment, tailored to meet the special requirements of computational fluid dynamics. Apart from multiple linked views, consisting of different information visualization views (e.g., histograms, scatter plots, parallel coordinated), the system provides a passive 3D view for focus+context visualization of the flow domain. Besides this, the frame work offers the opportunity to derive new flow attributes on the fly. For further details we refer to Doleisch's paper on the SimVis software [4] and the references therein. One of the views, that makes the framework especially useful in the context of path line attributes, is the *curve view* [15, 20]. The curve view is used to display large families of function graphs at once (cf., e.g., Fig. 7) plotting the function values against time. Lines are selected by brushing a certain value range for a specific time step. Functions with multiple components can be analyzed component-wise.

4.2 Exhaust manifold

This data set has been investigated by Lež et al. [17]. Their paper is not targeting the question of which attributes to choose for a interactive flow analysis, however, the authors suggest several attribute combinations that they found useful for the case study. These attribute combinations are start to end distance and path line length (arc length), maximum velocity and mean velocity along path line, and maximal curvature and maximal torsion. All those attributes are constant along the path lines. As also mentioned in the original paper by Lež et al., one of the goals in the design of exhaust manifolds is the decrease of flow resistance (so-called back pressure). Hence, the detection of path lines/particles causing back pressure is a natural task in this context.

In order to make the visual analysis based on the different attribute sets comparable, we identify an analytically defined set of path lines that we try then try to retrieve using both the original variable combinations and the here proposed attribute set. We restrict the analysis to particles seeded in the middle tube and its imminent vicinity. We identify particles possibly causing back pressure as those which 1) move upstream (i.e. $\max_t(pos_x(t_0) - pos_x(t)) > 0$, x denoting the axis aligned with the stream and assuming the stream to have positive sign) and 2) are upstream from the middle pipe at some point in time (i.e. $\max_t(pipeboundary_x - pos_x(t)) > 0$, with $pipeboundary_x$ being the position on the x-axis where the inflow pipe is connected to the outlet and under the same assumptions as before). Hence, the path lines in question are those where both parameters are positive. See Fig. 1 for an overview over the path lines identified and their seeding positions. Obviously, an ad hoc



(a) cluster 1 (b) cluster 2 (c) cluster 3

Figure 3: Investigation of some of the clusters in Fig. 2. The top row shows the actual selections, while the bottom row gives the associated path lines in their 3D context (color coding according to temporal evolution from yellow to red). For the discussion of the figures, we refer to the main text (Sec. 4.2).

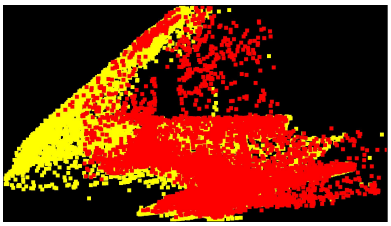


Figure 4: Scatter plot of the maximum velocity (horizontal) and the mean velocity along the path line (vertical). As in Fig. 2, red points represent path lines starting in or in the vicinity of the middle pipe. For further discussion see the main text in Sec. 4.2.

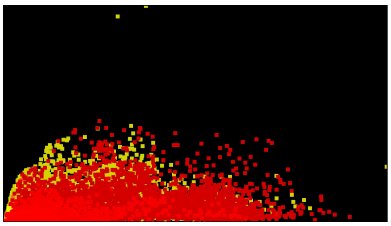


Figure 5: Scatter plot of the maximum curvature (horizontal) and the maximum torsion along the path line (vertical). As in Fig. 2 and 4, red points represent path lines starting in or in the vicinity of the middle pipe. For further discussion see the main text in Sec. 4.2.

analytic definition of interesting path lines is only possible in relatively clear and intuitive situation as this. We use this for the sake of comparability only.

First, we investigate the attribute combination start to end distance and path line length. We have preselected particles in the middle pipe and its immediate vicinity. In Fig. 2 we see a scatter plot of the two attributes. The red dots represent the path lines to investigate, the yellow dots give the context (i.e. the remaining path lines). In the scatter plots opacity scaling according to point density is used. In their paper, Lež et al. suggest investigating "unusual clusters", and we can visually identify several of them. We select those clusters one after the other and monitor the path lines associated to them (Fig. 3). We see that none of the visually distinguishable main clusters gives satisfactory results: on the one hand the clusters in Fig. 3(a) and Fig. 3(b) describe the same path line behavior, the cluster in Fig. 3(c) contains both path lines we are interested in (left branch) as well as path lines that seem not to be associated with back pressure (lighter path lines in the right branch). Further refinement of the query could help this, but no visual clues on how to do this are present in the scatter plot.

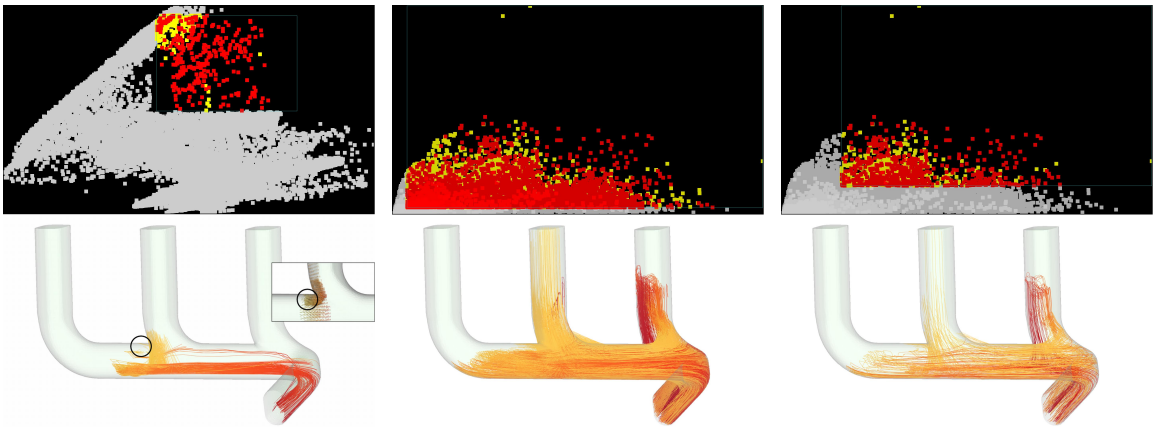
The next attribute combination investigated is maximum velocity and mean velocity along the path line. Fig. 4 shows a scatter plot of these two variables, the colors have the same meaning as before. In this case the visual detection of unusual clusters is harder. The most apparent abnormality seems to be the high share of path lines in question in the center of scatter plot. As Fig. 6(a) shows these path lines are indeed associated with the behavior we want to track. However, we systematically miss out on path lines seeded in a specific region (marked up with the circle).

Finally, we investigate the combination of maximum curvature and maximum torsion along the path line (see Fig. 5 for the respective scatter plot). Here, no clusters are visible. This means we would have to rely on thresholding. This thresholding gives, however, not the desired results, as seen in Fig. 6(b). Choosing a higher threshold refines the selection, but it fails to discriminate different types of flow behavior (Fig. 6(c)).

As a summary, we conclude that, following the state-of-the-art approach as described by Lež et al., we could find only a part of the path lines targeted.

Now we use the attribute set suggested by our statistical analysis. As remarked earlier, all of these attributes are time series. Hence, we make extensive use of the curve view. First we look at the streamwise position (in the same sense as used earlier). As in the first investigation, we select particles that originate from the middle pipe and its vicinity (Fig. 7(a) top). In order to cause back pressure, particles have to still be in the pipe, at the next stroke of the engine. Hence, we discharge ("not-selection") particles that are in the outlet at the time step the next stroke occurs (selection in Fig. 7(b) top). In the top of Fig. 7(c) we see the path lines corresponding to this selections. The particles that move "upstream" exhibit the same pattern as the once found by the analytic definition.

However, our selection is, at the current point, still containing a number of path lines with clearly different (so to say "correct") flow behavior. Hence, we move to a different attribute to refine our selection. In the bottom of Fig. 7(a) we see the time series for the second quadratic statistical invariant (in the following: inv2). We see (at least) two clearly distinguishable patterns: path lines with a medium-high value of inv2 in the beginning of the time series,

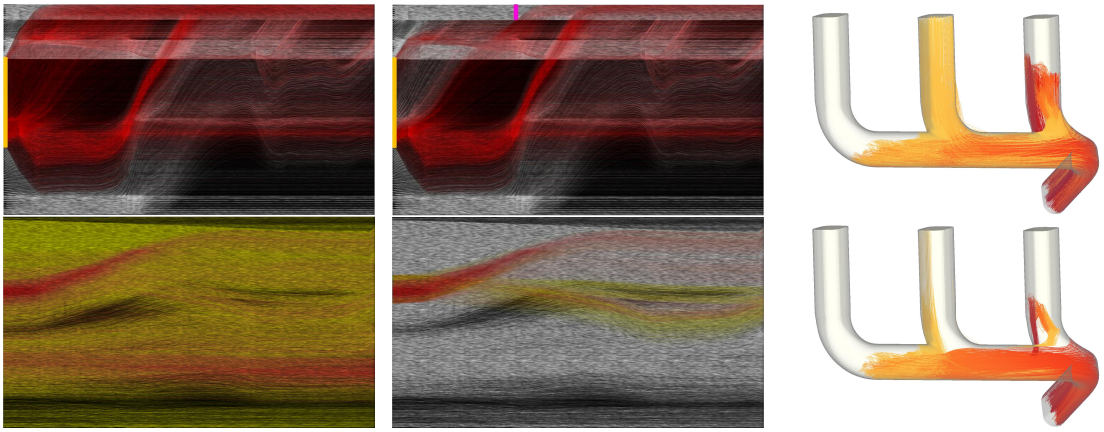


(a) investigation of maximum velocity vs. mean velocity

(b) investigation of maximum curvature vs. the maximum torsion

(c) investigation of maximum curvature vs. the maximum torsion

Figure 6: (a) The visually detectable "abnormality" in the scatter plot is selected. The path lines show the targeted behavior, but not all of them can be detected (cf. inset). (b) and (c) Due to the lack of visual clues, different thresholds are assessed, not showing the desired effect. For further discussion we refer to the main text (Sec. 4.2). For all three figures: Color coding of the path lines according to their temporal evolution (from yellow to red).



(a) (b) (c)

Figure 7: Intermediate steps of the interactive flow analysis based on the attribute set proposed in this paper. The time line is left to right, top to bottom. Regular selections are marked in orange, not-selections in pink. The final result can be found in Fig. 8. A detailed description of the analysis steps is given in Sec.4.2.

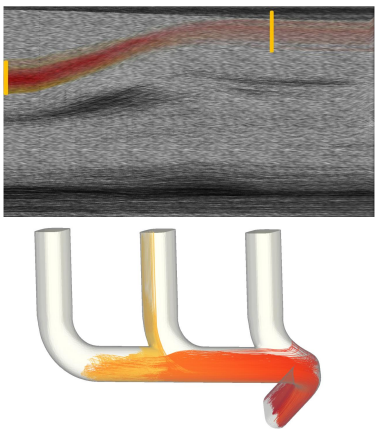


Figure 8: The upper branch in the inv2 line plot together with the respective path lines (together with the previous described selection, see Sec.4.2).

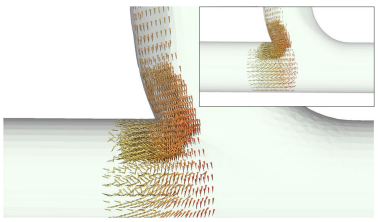
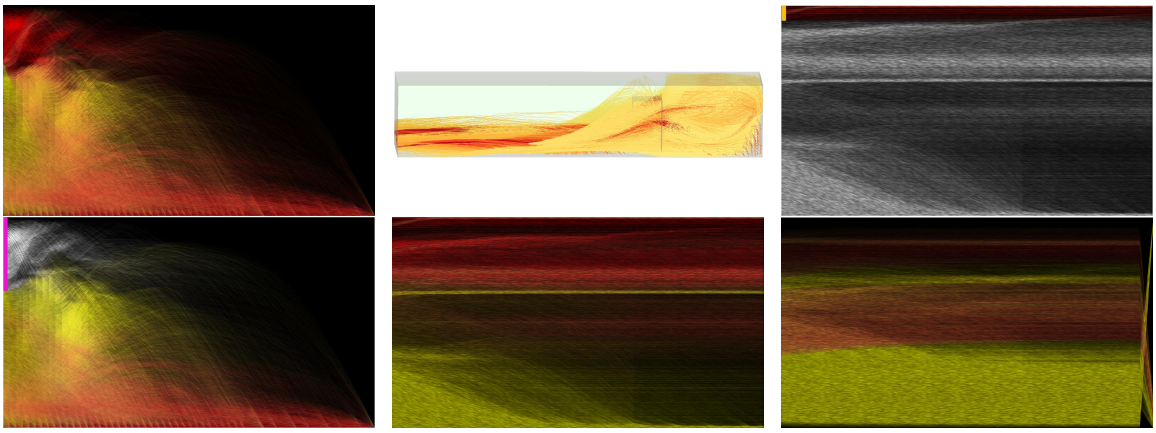


Figure 9: Close-up of the seeding positions of the path lines in Fig. 8 compared to the seeding points of the reference path lines. We see that the difference between those two sets is hardly detectable.

and others with a rather low value. We select the ones with the higher values and see (cf. Fig. 7(c) bottom) that now nearly all path lines exhibit the expected behavior. A small number of path lines is not of the expected type, representing particles being sucked in the rightmost tube. In fact, also the selected time series of inv2 have two branches (upper and lower, see Fig.7(b) bottom).

As Fig.8 shows the two branches are indeed associated to the



(a) (b) (c)

Figure 10: Intermediate steps of the interactive flow analysis based on the attribute set proposed in this paper. The time line is top to bottom, left to right. The final result can be found in Fig. 11 and Fig. 12. A detailed description of the analysis steps is given in Sec.4.3.

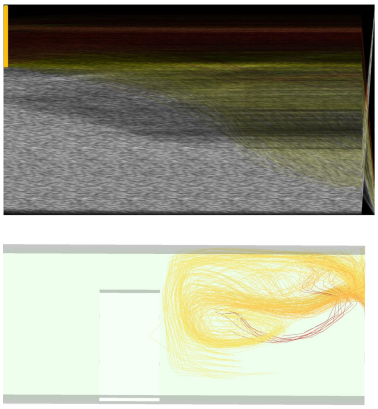


Figure 11: Final result of the interactive flow analysis based on the her proposed feature set. We are able to identify the recirculation area in front of the of the obstacle described by Pobitzer et al. [23]

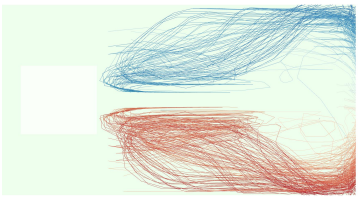


Figure 12: Top view on the path lines depicted in Fig. 11. The color coding is according to the attribute inv3, blue being low and red being high values.

different types of path line behavior present. Comparing the seed points of the path lines found by our analysis to the seed points of the reference path lines, we see a clear correspondence of the two sets (Fig. 9, in contrast to the situation in Fig. 6(a) bottom).

We see that the interactive flow analysis of the data set based on our suggestions is able to find the targeted path lines. In addition, the process is intuitive in the sense that different flow behavior is reflected by clearly distinguishable clusters in the attributes. We discussed our results with a domain expert, who confirmed the expressiveness of our results.

4.3 Breaking dam

This data set has been investigated by Pobitzer et al. [23] in the context of *finite-time Lyapunov exponents* (FTLE). One of the interesting features is a separation structure in front of the obstacle, separating particles passing at the two sides of the obstacle. Another structure is a recirculation zone in front of the obstacle. Due to its definition, the FTLE approach is not suitable to investigate the internal structure of the recirculation. We therefore investigate this question by means of interactive flow analysis of the attribute set proposed earlier in this paper.

Since we want to target recirculation behavior, we preselect particles seeded upstream from the obstacle. Since the recirculating particles do not pass the obstacle, we can assume that the distance from start to end is not too big. Hence, we exclude the path line cluster associated to high distances from our analysis, using a "not"-selection (Fig. 10(a)). The top of Fig. 10(b) shows the path lines corresponding to this selection.

Now we investigate one of the other attributes, namely the first quadratic statistical invariant (in the following: inv1). The attribute is chosen since it provides clearest clustering with the current selection (Fig. 10(b) bottom). The most distinct cluster is a rather small group of almost horizontal lines in the top. We recall that inv1 is a shape descriptor and ideal recirculation can be thought of as a circular motion. Hence, an almost constant shape descriptor could indicate the wanted behavior. After selecting this curve cluster, we investigate the second quadratic statistical invariant (in the following: inv2). Again, this attribute has been selected for analysis by the same principle as before. We detect two possible clusters and by the same reasoning as before, we select the family of more or less constant lines (cf. bottom of Fig. 10(c)). In Fig. 11 we see the selection and the resulting path lines. We conclude that we found the recirculation zone Pobitzer et al. found the boundary of in their paper [23]. Investigating the remaining attributes, we see a clear split in the second quadratic statistical invariant, color coding the path lines according to this attribute, yields Fig. 12, revealing that the left-right separation structure is also present inside the recirculation, an insight the FTLE-based analysis of Pobitzer et al. failed to convey.

5 Conclusions

In this paper we address the problem of selecting an expressive subset of the path line attribute space for interactive visual flow analysis. Investigating a number of CFD data sets using factor analysis

we found that there are common patterns both in dimensionality and attributes associated to them across data sets. We identify 6 path line attributes that represent those factors. The analysis based on the attributes suggested in this papers proves to match, and in part also exceed, previous work, showing how the benefit from the already proven concept of interactive flow analysis can be utterly increased by carefully selecting appropriate attributes. Prior knowledge of which attributes to investigate reduces both computational and storage overhead. In addition, a lower-dimensional data set is easier to handle in the context of IVA and allows for a systematic investigation.

Usually, one of the aims of factor analysis is to identify the underlying factors, at least qualitatively (as mentioned earlier, are the factors assumed to not be measurable directly). Looking at the attributes suggested, we can informally identify one attribute associated to shape (inv), one related to vortices (λ_2), and a bigger group of attributes related to motion (avspeed, startEnd, vel and pos). This may indicate that the motion is the most complicated aspect of path lines, or, more optimistically, better attributes for describing it could be found.

6 Acknowledgments

The authors want to thank Carl Erik Wasberg from the Norwegian Defence Research Establishment (FFI) for providing the DNS data set of the turbulent channel flow. The CFD simulation of a flow through a box, T-junction, breaking dam and the exhaust manifold are courtesy of AVL List GmbH, Austria. Thanks to Michael Mayer from AVL List GmbH, Austria, for his valuable feedback. Finally, the authors thank SimVis GmbH, Austria, for providing the framework.

The project SemSeg acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number 226042.

References

[1] R. Bürger, P. Muigg, H. Doleisch, and H. Hauser. Interactive cross-detector analysis of vortical flow data. In *Proceedings of CMV 2007*, pages 98–110. IEEE, 2007.

[2] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

[3] A. B. Costello and J. W. Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10:173–178, 2005.

[4] H. Doleisch. SimVis: Interactive visual analysis of large and time-dependent 3D simulation data. In *Proc. of WSC 2007*, pages 712–720, 2007.

[5] I. K. Fodor. A survey of dimension reduction techniques. Technical Report UCRL-ID-148494, Lawrence Livermore National Laboratory (LLNL), 2002.

[6] R. Fuchs and H. Hauser. Visualization of multi-variate scientific data. *Computer Graphics Forum*, 28(6):1670–1690, 2009.

[7] R. Fuchs, R. Peikert, H. Hauser, F. Sadlo, and P. Muigg. Parallel vectors criteria for unsteady flow vortices. *IEEE TVCG*, 14(3):615–626, 2008.

[8] H. H. Harman. *Modern Factor Analysis*. The University of Chicago Press, Chicago, IL U.S.A., 3. ed edition, 1976.

[9] C.-C. Hsiung. *A first course in differential geometry*. Wiley, New York, 1981.

[10] J. C. R. Hunt, A. A. Wray, and P. Moin. Eddies, stream and convergence zones in turbulent flows. In *2. Proc. of the 1988 Summer Program*, pages 193–208, 1988.

[11] J. Jeong and F. Hussain. On the identification of a vortex. *J. of Fluid Mech.*, 285:69–84, 1995.

[12] M. Jiang, R. Machiraju, and D. Thompson. Detection and visualization of vortices. In *The Visualization Handbook*, pages 295–309. Academic Press, 2005.

[13] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Saddle River, NJ U.S.A., 6. ed edition, 2007.

[14] H. F. Kaiser. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1):141–151, 1960.

[15] Z. Konyha, K. Matković, D. Gračanin, M. Jelović, and H. Hauser. Interactive visual analysis of families of function graphs. *IEEE TVCG*, 12(6):1373–1385, 2006.

[16] Y. Levy, D. Degani, and A. Seginer. Graphical visualization of vortical flows by means of helicity. *AIAA Journal*, 28:1347–1352, Aug. 1990.

[17] A. Lez, A. Zajic, K. Matkovic, A. Pobitzer, M. Mayer, and H. Hauser. Interactive exploration and analysis of pathlines in flow data. In *Proc. of WSCG 2011*, pages 17–24, 2011.

[18] C.-H. Lo and H.-S. Don. Invariant representation and matching of space curves. *J. of Intel. and Robotic Sys.*, 28:125–149, June 2000.

[19] T. McLoughlin, R. S. Laramee, R. Peikert, F. H. Post, and M. Chen. Over Two Decades of Integration-Based, Geometric Flow Visualization. *Computer Graphics Forum*, 29(6):1807–1829, 2010.

[20] P. Muigg, J. Kehrer, S. Oeltze, H. Piringer, H. Doleisch, B. Preim, and H. Hauser. A four-level focus+context approach to interactive visual analysis of temporal features in large scientific data. *Computer Graphics Forum*, 27(3):775–782, 2008.

[21] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[22] H. Piringer, W. Berger, and H. Hauser. Quantifying and comparing features in high-dimensional datasets. In *Proc. of InfoVis 2008*, pages 240–245, 2008.

[23] A. Pobitzer, R. Peikert, R. Fuchs, H. Theisel, and H. Hauser. Filtering of FTLE for Visualizing Spatial Separation in Unsteady 3D Flow. In R. Peikert, H. Hauser, H. Carr, and R. Fuchs, editors, *Topological Methods in Data Analysis and Visualization II*. Springer, 2012 (forthcoming).

[24] F. Post, B. Vrolijk, H. Hauser, R. Laramee, and H. Doleisch. The state of the art in flow visualization: Feature extraction and tracking. *Computer Graphics Forum*, 22(4):775–792, 2003.

[25] T. Salzbrunn, C. Garth, G. Scheuermann, and J. Meyer. Pathline predicates and unsteady flow structures. *Visual Computer*, 24(12):1039–1051, 2008.

[26] T. Salzbrunn and G. Scheuermann. Streamline predicates. *IEEE TVCG*, 12(6):1601–1612, 2006.

[27] J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Proc. of the IEEE InfoVis*, pages 65–72, Washington, DC, USA, 2004. IEEE Computer Society.

[28] K. Shi, H. Theisel, H. Hauser, T. Weinkauf, K. Matkovic, H.-C. Hege, and H.-P. Seidel. Path line attributes - an information visualization approach to analyzing the dynamic behavior of 3D time-dependent flow fields. In H.-C. Hege, K. Polthier, and G. Scheuermann, editors, *Topology-Based Methods in Visualization II*, Mathematics and Visualization, pages 75–88, Grimma, Germany, 2009. Springer.

[29] C. Spearman. Demonstration of formulæ for true measurement of correlation. *The Am. J. of Psychology*, 18(2):161–169, 1907.

[30] D. D. Suhr. Principal component analysis vs. exploratory factor analysis. In *Proceedings of the Thirtieth Annual SAS®Users Group International Conference*. SAS Institute Inc., 2005.

[31] D. Sujudi and R. Haimes. Identification of swirling flow in 3D vector fields. Technical Report 95-1715, AIAA, 1995.

[32] H. Theisel, T. Weinkauf, H.-C. Hege, and H.-P. Seidel. Topological methods for 2d time-dependent vector fields based on stream lines and path lines. *IEEE TVCG*, 11(4):383–394, 2005.

[33] C. Turkay, P. Filzmoser, and H. Hauser. Brushing dimensions – a dual visual analysis model for high-dimensional data. *IEEE TVCG*, 17(12):2591–2599, 2011.

[34] T. Weinkauf, J. Sahner, and H. Theisel. Cores of swirling particle motion in unsteady flows. *IEEE TVCG*, 13(6):1759–1766, 2007.

[35] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proc. of VISSYM '03*, pages 19–28. Eurographics Association, 2003.

[36] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *J. of Comput. and Graphical Stat.*, 15(2):265–286, jun. 2006.