Meas. Sci. Technol. **18** (2007) 3511–3522 doi:10.1088/0957-0233/18/11/035

Mode-ratio bootstrapping method for PIV outlier correction

Chan-Seng Pun, Andree Susanto and Dana Dabiri

Department of Aeronautics and Astronautics, University of Washington, Box 352400, Seattle, WA 98195, USA

Received 24 April 2007, in final form 23 August 2007
Published 4 October 2007
Online at stacks.iop.org/MST/18/3511

Abstract

This paper proposes a new method for correcting spurious displacement vectors obtained by using particle image velocimetry. Unlike methods that generate and use statistics from neighboring vectors toward outlier identification, a bootstrapping process is employed to generate statistics for each component. The mode-ratio criterion, defined as the normalized absolute value of the difference between the mode of the generated statistics and the actual value on the field, is used to identify and tag spurious components. The bootstrapping process is repeated to generate new statistics from the untagged components. This process is then repeated multiple times. The resulting mode field is used as the correction field. Two different displacement fields, each with a different error type, are artificially generated to evaluate the performance of the method. The evaluation is done by measuring the deviation between the resulting mode field and the perfect component field. The method is then tested on a real turbulent jet flow, obtained from www.pivchallenge.org, and the effect of the spatial gradient is discussed.

Keywords: PIV, particle image velocimetry, outlier detection, outlier correction, bootstrapping

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Particle image velocimetry (PIV) has been extensively used to measure the velocities and other kinematic properties of the observed fluid flows. The PIV method illuminates a cross-section of the desired flow that is seeded with reflective particles with a pulsed laser sheet, while sequentially imaging this illuminated flow field. The resulting images can then be processed by various algorithms to obtain the velocity flow fields [1, 2]. However, the results are almost always prone to spurious vectors which are mostly due to the seeding inhomogeneities, effects of turbulence, poor image quality, varying intensity light sheet, laser reflections, etc. These spurious vectors, or simply outliers, not only corrupt the velocity field, but also affect the differential and integral velocity quantities such as vorticity, streamlines, etc. Thus, it is indeed important to develop a method to accurately correct outliers to prevent such data corruption.

Many outlier detection methods have been proposed that determine their detection criteria based on neighboring vector statistics. Westerweel [2] first proposed three automated outlier detection methods: the local mean method, the global mean method and the local median method. The local mean method defines the residual as the difference between a vector in question and the mean of its eight surrounding vectors, the global mean defines the residual as the difference between a vector in question and the mean of the entire velocity field and the local median method defines the residual as the difference between a vector in question and the median of its eight surrounding vectors. A threshold is then chosen and compared with the residual of each vector to determine if it is spurious. Westerweel's results show that the local median method provided the best results. This method, however, requires choosing different optimal threshold values for different velocity fields, since no single threshold was found to be optimum for all flows.

Nogueira *et al* [3] proposed the method of local coherence. The residual, calculated for all vectors in the field, is defined as the sum of the differences between a vector in question and its eight surrounding vectors, normalized by the sum of the surrounding vectors. The location of the minimum residual field is then marked as the local coherence region. Based on the user-defined parameter, the method then starts including vectors into the local coherence region, thereby validating vectors. The vectors that are within the specified tolerance are accepted as good vectors, while those that are not are marked as spurious. The method involves picking two parameters, which are the number of vectors to include in the local coherence region before they are validated, and the tolerance within which a vector is considered a non-spurious vector. However, the tolerance by which a vector can be considered good assumes *a priori* knowledge of the flow in order to determine how much flow gradients can be tolerated.

Song *et al* [4] incorporated an outlier detection method into their Delaunay tessellation particle tracking velocimetry method. The detection method is based on whether the incompressible continuity equation is satisfied within a Delaunay triangle. They hypothesized that the total flux crossing all sides of the triangles will be minimal if a vector is not spurious. This method was found to be effective; however, the detection method is part of their Delaunay tessellation particle tracking velocimetry method, and cannot be applied separately as a post-processing algorithm.

Liang *et al* [5] used the cellular neural networks (CNN) to create a detection scheme by obtaining the stable states of neurons. Weights of the neurons are calculated by subtracting the residual between two vectors from the selected threshold level. Then, the CNN identifies outliers by using the calculated weight of the neurons and the neuron's outputs from the algorithm. Overdetection (valid vectors detected as outliers) and undetection (outliers not detected by the method) concepts were introduced for performance criteria. The results suggest a significant improvement on the robustness and accuracy of outlier detection by comparing with the local median test proposed by Westerweel [2]. However, the CNN fails to satisfactorily identify outliers when the velocity field gradient is large.

Shinneeb *et al* [6] proposed a variable threshold outlier detection method in 2004. The CNN method and the original local median test were tested, but with variable, instead of constant, threshold which is determined as a function of location in the field of view. At first, an aggressive local median test is applied so that any suspected outliers can be detected. These suspected outliers (which may include some overdetected vectors) are then replaced by applying a local Gaussian filter about the vector in question. The variable threshold outlier detection method is applied to two simulated fields and a real axisymmetric jet flow. For the simulated tests, the performance of the method is measured by counting the number of overdetections and undetections. The authors confirmed that the CNN method generally performs better than the local median test, and that varying the thresholds also gives better results. The variable thresholds are more independent of the velocity gradient than the constant thresholds. However, the Gaussian filter involves choosing the appropriate filter width. Varying such a parameter will have an impact on the number of overdetections, and the optimum value for the filter's width varies depending on the experimental conditions.

Young *et al* [7] described a general approach for validating PIV vectors, where vectors are compared to a smoothed vector field that reliably characterized the measured vector field. To maintain sufficient velocity gradient information, a thin-plate spline model is used within an iterative weighted routine to generate a smoothed representation of the displacement field. The local difference between the actual and smoothed vectors is then used toward determining if the vector is an outlier.

The most recent study of outlier detection was performed by Westerweel and Scarano [8]. They proposed a modified local median test by normalizing the original local median method's residual with the median of the residuals of the eight neighboring vectors of the vector in question. They also noted that the normalization factor tends to approach zero for a region with low turbulence intensity. This makes the residual very small and the modified local median test very large. To compensate, an empirically determined constant was added to the denominator. The method was tested on many experimental flows, and the probability density functions of their residuals were shown to collapse for values less than or equal to 2, and diverge for values greater than 2, thereby suggesting a universal method for detecting outliers.

As outlined above, many of the previous studies in outlier detection are based on statistics that are locally obtained, which do not provide solutions for outlier correction. The purpose of this paper is to therefore propose a new method that uses statistics generated from the entire data population to identify and correct outliers, while reducing the effect of spatial gradients. This paper will start by explaining the statistics generation method that uses a bootstrapping method in section 2, the detection criteria will be described in section 3, the testing methods done on the simulated fields will be explained in section 4, section 5 will describe the displacement correction approach, section 6 will provide an analysis of optimum parameters' determination based on the results of the simulated field testing, section 7 will validate these results through the application of the correction method to a turbulent jet flow as well as discuss the associated required computational effort, the effects of spatial gradients and interrogation resolutions will be elaborated in section 8 and, finally, section 9 will conclude this paper.

2. Statistics generation by bootstrapping

In general, the bootstrapping process [9–11] allows for statistical inference (i.e. mean, standard deviations, modes, etc) for a data set. This is achieved by sampling the data set *with replacement*. Sampling *with replacement* means that after data points are randomly sampled to create a subsample, they are reinserted into the original data set for the next subsequent random sampling. For each sampling, the desired statistics are calculated. This process is repeated multitudinous times, after which, the distributions of the statistics are examined. This *bootstrap* distribution represents the distribution of the original data set. It has the advantage of working even when theory fails, not requiring that the distribution be specific, i.e. normal, or that the sample sizes be large, and finally providing greater accuracy than other methods [12]. The presently incorporated bootstrapping process shares the fundamental procedure of generating data, but is slightly different because of its implementation on two-dimensional data and collaborative use of an interpolation routine, which has previously been

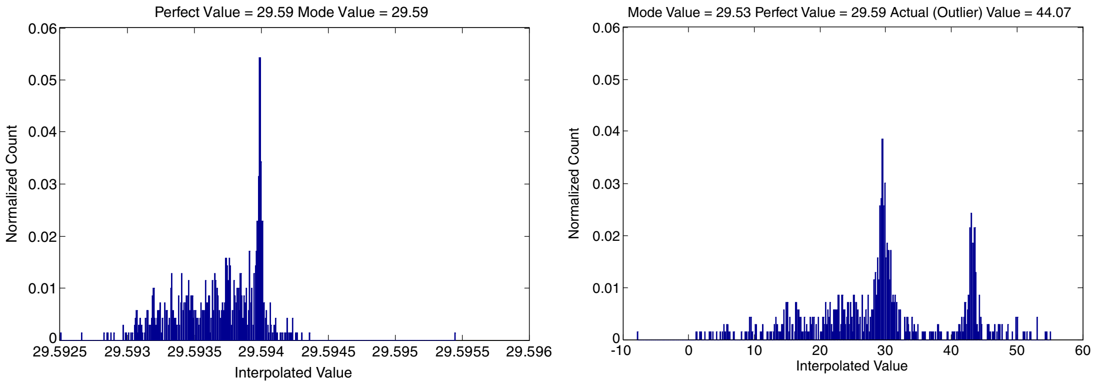

Figure 1. The histogram of a point after the bootstrapping process is performed. Left: histogram of a point from a component field without outliers. Right: histogram of the same point from the same vector field as shown in the left, where this point has been replaced by a spurious component.

used by Rignot and Spedding [13], Dabiri and Gharib [14] and Spedding and Rignot [15]. The bootstrapping process begins with randomly sampling a specified percentage p of components from the component field. This sample is then used to interpolate the remaining field using *Gridfit*, which incorporates a cubic interpolation algorithm [16]. The sampling and interpolation processes repeat until a specified number, ℓ, of interpolated component fields are obtained. The samples used to interpolate this field are returned to the original data set, after which, new samples are randomly taken to interpolate a new component field. In this manner, the interpolated fields are achieved by sampling *with replacement*. Since either component field can be independently prone to outliers, the bootstrapping process is applied separately to each of the component fields. The mode-ratio bootstrapping process can be sensitive to spatial gradients, which are a function of the velocity magnitudes, as well as the spatial resolution of flow structures. To eliminate any magnitude effects, each component field is normalized by its maximum magnitude before being processed. The implementation of the bootstrapping method is further discussed in section 5.

3. Filtering criterion

Figure 1 (left) shows the histogram resulting from applying the bootstrapping method to a component field without outliers at a specific location. The distribution is unimodal, and its mode and actual component values, shown on top of the figure, are identical. Figure 1 (right) shows the histogram at the same point resulting from applying the bootstrapping method to the same field but with outliers randomly distributed through the field, where this particular point is an outlier. The distribution is now bimodal and the difference between the mode and the actual component value is significant (14.54 pixels). The higher peak is due to the successive accumulation of the interpolated values, and the lower peak corresponds to the number of times that the actual value, which is an outlier, was taken as the seed for interpolation. Also noteworthy from figure 1 (right) is that the value of the mode calculated from the outlier field is very close to the perfect component value (a difference of 0.06 pixels).

3.1. Definition of the mode and the mode-ratio criterion

In order to identify outliers similar to those shown in figure 1, the mode-ratio criterion is developed. The mode m within each histogram is defined as the midpoint of the highest frequency bin. The residual is defined as the absolute value of the difference between the mode of a component in question and the actual component value:

$$|m_{i,j} - x_{i,j}|. \tag{1}$$

This residual is normalized by the mode, henceforth referred to as the mode ratio and compared with a threshold t to form the detection criterion,

$$\left|\frac{m_{i,j} - x_{i,j}}{m_{i,j}}\right| \leqslant t. \tag{2}$$

When the mode ratio is greater than the threshold, the component in question will be tagged as an outlier. The initial starting threshold assumes that an outlier will deviate more than 20% from its mode. Therefore, the starting threshold, t, is set to 0.2.

3.2. Implementation of the mode-ratio criterion

If the mode value is near zero, the mode ratio approaches infinity even when the residual is small, resulting in good components being identified as outliers. Thus, a tolerance level, *tol*, is introduced to overcome the aforementioned challenge. The tolerance is compared with the residual, which can be described by the following equation:

$$r_{i,j} = |m_{i,j} - x_{i,j}| < tol. \tag{3}$$

If the difference between the mode and actual values was within tolerance, then that component would be considered as good without further outlier verification with the mode-ratio criterion.

The mode bin at each (i, j) location for all interpolated fields is determined by

$$b_{i,j} = \frac{\max(x_{i,j,l}) - \min(x_{i,j,l})}{\ell/2}, \qquad l \in \ell. \tag{4}$$

In order to determine the mode accurately, the mode bin is adjusted such that it is less than twice the tolerance.

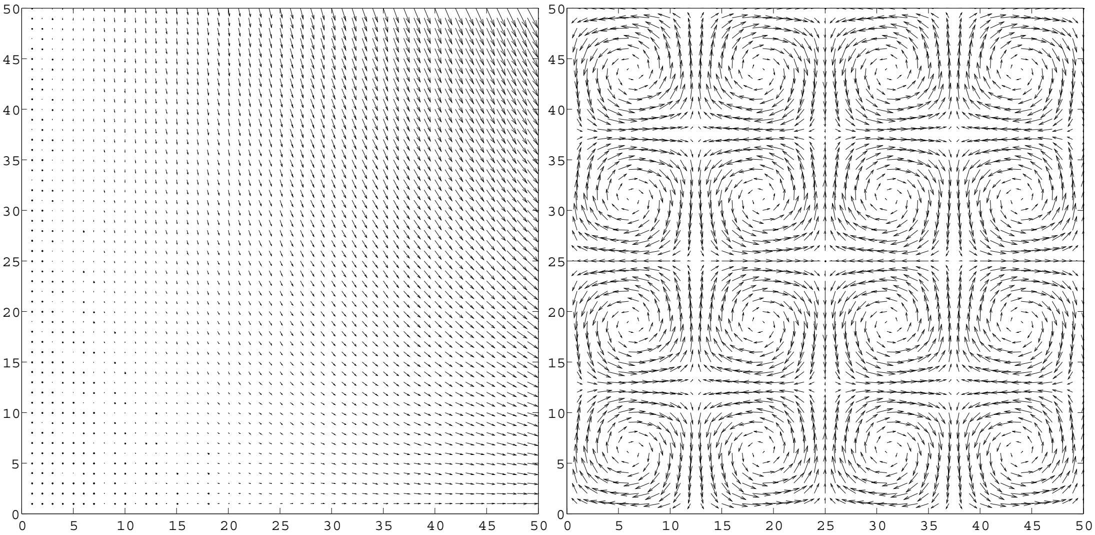

Figure 2. Left: the perfect first simulated displacement field with an A value of 400. The field comprises displacements in 50 nodes in each axis, with an inter-node distance of 16. Right: the perfect second simulated displacement field. It is a vortical cellular flow where N_x and N_y are chosen to be 4, the values of L_x and L_y are chosen to be 800, and the field comprises displacements in 50 nodes in each axis, with an inter-node distance of 16.

3.3. Multi-pass implementation

The multi-pass approach is implemented to improve the accuracy of the outlier detection. With the multi-pass approach, the number of passes, k, is first specified. Then, a sufficiently large threshold, t, is used to identify and remove the most severe outliers. The detection process is then repeated with an incrementally reduced threshold value in order to identify and remove the next series of less severe outliers. Note that higher passes imply smaller incremental intervals and vice versa. This process is repeated until k is reached and t becomes zero. The equation for determining the threshold value for each pass is

$$t - \left(\frac{t}{k-1}\right)(n-1), \tag{5}$$

where n implies the nth pass and can range from 1 to k.

4. Testing method

4.1. Simulated displacement fields

The correction method is tested on simulated displacement fields, where all information is known *a priori*, to measure its performance. Two simulated displacement fields are chosen. Following Shinneeb *et al* [6], the first displacement field is a potential flow, satisfying the Laplacian incompressible continuity equation

$$\nabla \cdot \overrightarrow{U} = 0, \tag{6}$$

where the displacement field is

$$u = Ax^2 \tag{7}$$

$$v = -2Axy \tag{8}$$

and the constant, A, which is set to 400, determines the maximum magnitude of the displacement field. The displacement field is plotted for x and y values ranging from 0 to 800. The number of nodes is chosen to be 50, with an inter-node distance of 16 (see figure 2 (left)).

The second field is a vortical cellular flow, where the displacement field is

$$u = V_{\max}\cos\left(\frac{xN_x\pi}{L_x} + \frac{\pi}{2}\right)\cos\left(\frac{yN_y\pi}{L_y}\right) \tag{9}$$

$$v = V_{\max}\sin\left(\frac{xN_x\pi}{L_x} + \frac{\pi}{2}\right)\sin\left(\frac{yN_y\pi}{L_y}\right). \tag{10}$$

L_x, L_y represent the size of the field, $V_{\max}$ is the maximum displacement magnitude and N_x, N_y are the number of vortices in the x and y directions respectively. L_x and L_y are set to 800. $V_{\max}$ is set to 10 and N_x and N_y are set to 4. Similarly, this field is divided into 50 nodes, with an inter-node distance of 16 (see figure 2).

4.2. Types of error

Two types of error are introduced into the simulated displacement fields. The first type of error (type 1) is a completely random error. The amount of deviation for both the vector direction and the magnitude are determined randomly, and the locations of the outliers are also randomized throughout the entire field. This type of error typically results in outliers that are most often surrounded by good vectors. Shinneeb *et al* [6] state that this type of error usually occurs in a practical PIV flow if the noise on a correlation plane is erroneously identified as the signal peak. While Westerweel [2] has suggested that the number of spurious vectors in properly acquired PIV data should be around 5% of the total number of vectors in the field, for this study, 10% of the vectors are replaced by outliers in order to more severely test the mode-ratio bootstrapping method.

Following Shinneeb *et al* [6], the second type of error (type 2) is designed to produce clustered outliers. Unlike the former study that allowed for larger variations, for this study,

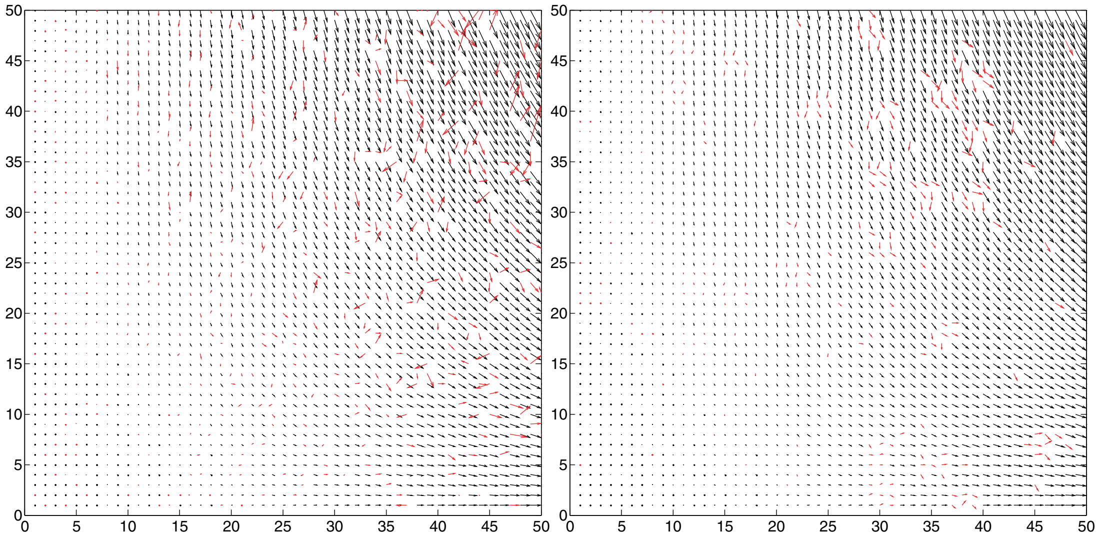

Figure 3. Left: the first simulated displacement field with 10% of its vectors replaced by random outliers. The locations, deviations and magnitudes of the outliers are chosen randomly. Right: the first simulated displacement field with 10% of its vectors replaced by clustered outliers. The locations of the outliers are chosen randomly, but they are made to form clusters of a specified size. The size of the cluster, in this case, contains at most six vectors. The deviation is also controlled; for this case, the vectors are deviating 10% in magnitudes from their corresponding true displacement values, and 15° in direction. The red vectors highlight the outliers.

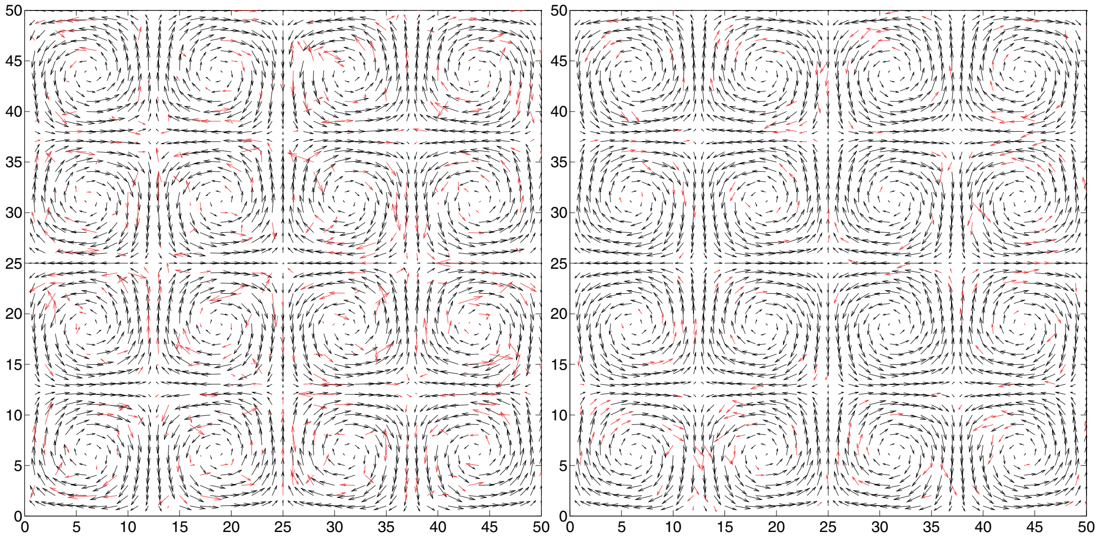

Figure 4. Left: the second simulated displacement field with 10% of its vectors replaced by random outliers. The locations, deviations and magnitudes of the outliers are chosen randomly. Right: the second simulated displacement field with 10% of its vectors replaced by clustered outliers. The red vectors highlight the outliers.

the magnitudes of the vectors are set to deviate at most 25% of the original displacement value, and the direction of the vectors is limited to deviate as far as 15°. In this manner, the outliers will be harder to detect, and therefore a better test for the mode-ratio method. The number of outliers within a cluster is chosen to be at most 6, as our observation of outliers within properly acquired data suggests that most clusters contain at most six vectors. This type of error happens in a practical PIV flow mostly because of imperfections in the PIV image or low seeding density. Figures 3 and 4 show the two types of errors on the two aforementioned simulated fields. As with the previous error, the number of spurious vectors in the field is set to be 10% of the total number of vectors in the field. Shinneeb *et al* [6] found that this type of error is in general harder to detect, as can be seen from these figures. Therefore in total, there are four displacement fields that are examined: field 1 with type 1 error (F1T1), field 1 with type 2 error (F1T2), field 2 with type 1 error (F2T1), field 2 with type 2 error (F2T2).

5. Approach toward correction and performance assessment

5.1. Parameters summarized

As discussed above, the mode-ratio bootstrapping method is dependent on several parameters as follows.

Table 1. Tested parameters.

Sampling percentages	5	15	25	35
Iterations	500	700	1000	
Tolerance	0.001	0.0025	0.005	
Mode-ratio passes	8	12	16	

(1) Sampling percentage (p): the fraction of the data that are taken from the component field, from which the remaining field is interpolated (section 2).
(2) Iterations (ℓ): the number of times the bootstrapping method is repeated (section 2).
(3) Threshold (t): the threshold to detect spurious components, initially set to 0.2 and then incrementally reduced to zero (sections 3.1 and 3.2).
(4) Tolerance (tol): the tolerated difference between the actual component value and the mode, above which the normalized residual of a point will be tested against the current threshold. If the residual is below the tolerance level, then the displacement component will be considered non-spurious, regardless of the value of the normalized residual (section 3.2).
(5) Passes (k): the number of passes the mode ratio has iteratively performed before the final threshold is reached (section 3.3).

In order to determine the optimal values for these parameters, the mode-ratio bootstrapping method is tested for a range of parameters, as is summarized in table 1, on the four displacement fields presented in section 4.

5.2. Approach for correcting outliers

In addition to accurately detecting outliers, the outlier detection process can incorrectly detect good components as outliers, as well as fail to detect certain outliers. Therefore, it is important to observe how many non-spurious components are mistakenly detected as outliers (overdetection), how many outliers are not detected (undetection) and how close the modes of the overdetected and accurately detected components are to the perfect values. The difference between the modes of the over/accurately detected components will be referred to as the *mode error.* While it is desirable to have small overdetected counts, it is also desirable that the overdetected components are replaced accurately. Therefore, for each field and type of error, the mode error of the accurately detected outliers, the mode error and number of the overdetected outliers, and the number of undetected outliers are statistically studied to determine the performance of the mode-ratio bootstrapping correction method.

5.3. Algorithmic flow and performance analysis methods

Figure 5 shows the flowchart of the algorithm's procedure, which is applied to each of the vector components separately. First, a certain percentage of the data, p, is randomly sampled from the component field. This selection is used to interpolate the rest of the field using *Gridfit*. This process repeats until the specified number of iterations, ℓ, is reached. Upon completion, at each point, interpolated values are collected among all the interpolated fields, and a histogram is set up to obtain the mode for outlier identification. If identified as outliers, the detected components are tagged and excluded from the original population. The remaining population is then used to repeat the entire process k times, and with each pass, the threshold value is incrementally reduced until zero. The final mode values are then used to correct the detected components within the field.

To assess the performance of the correction method, the mean and standard deviation of the mode error for each field, error type and component are calculated for overdetected and accurately detected components. Because of the limitation of the interpolation at the boundaries, boundary points are excluded in the analyses for a more accurate performance assessment of the correction method. The mean is used to measure the overall accuracy of the modes, and the standard deviation is used to indicate if the accuracy is consistent throughout the field. The goal is to, therefore, find which sets of parameters best minimize these statistics, the overdetections and the undetections. These parameters will henceforth be referred to as optimal parameters.

6. Results

In this section, the statistical analyses are discussed and the resulting optimal parameters are presented in terms of the robustness of the detection and correction results with respect to the tested fields and error variations. The parametric studies show that the results for the u and v components are very similar; therefore, only the results for the u component are discussed.

For easier comparison, the y-axes in figures 6–8 of the mode error statistics and overdetection counts are scaled. Toward this end, the y-axes of the mode error statistics for these plots are broken into two scales, from 0 to 0.3 and from 0.3 to the maximum. For the overdetection counts, the two scales range from 0 to 200 and from 200 to the maximum. In this manner, the variation of the data amongst these different plots can be easily compared.

6.1. Optimum parameters for the correction method

The algorithm replaces the detected components with their corresponding modes. Components that are not tagged as outliers consist of non-spurious and spurious components. The undetected spurious components are referred to as undetected components and treated as good components which will not be replaced by the modes. The detected components comprise correctly detected outliers, and non-outliers incorrectly detected as outliers, or overdetected components. It is then of interest to observe how close the modes of the accurately detected and overdetected components are to their perfect component values. Consequently, the set of parameters resulting with the fewest undetected outliers will be highly favorable. Furthermore, it is desirable to minimize the number of overdetected components, because too many overdetections may cause insufficient points for proper interpolation of the fields. A fairly large number of overdetected vectors, however, might be tolerable if the associated mode errors are acceptably small. Therefore, the optimum parameters should minimize the overdetections and

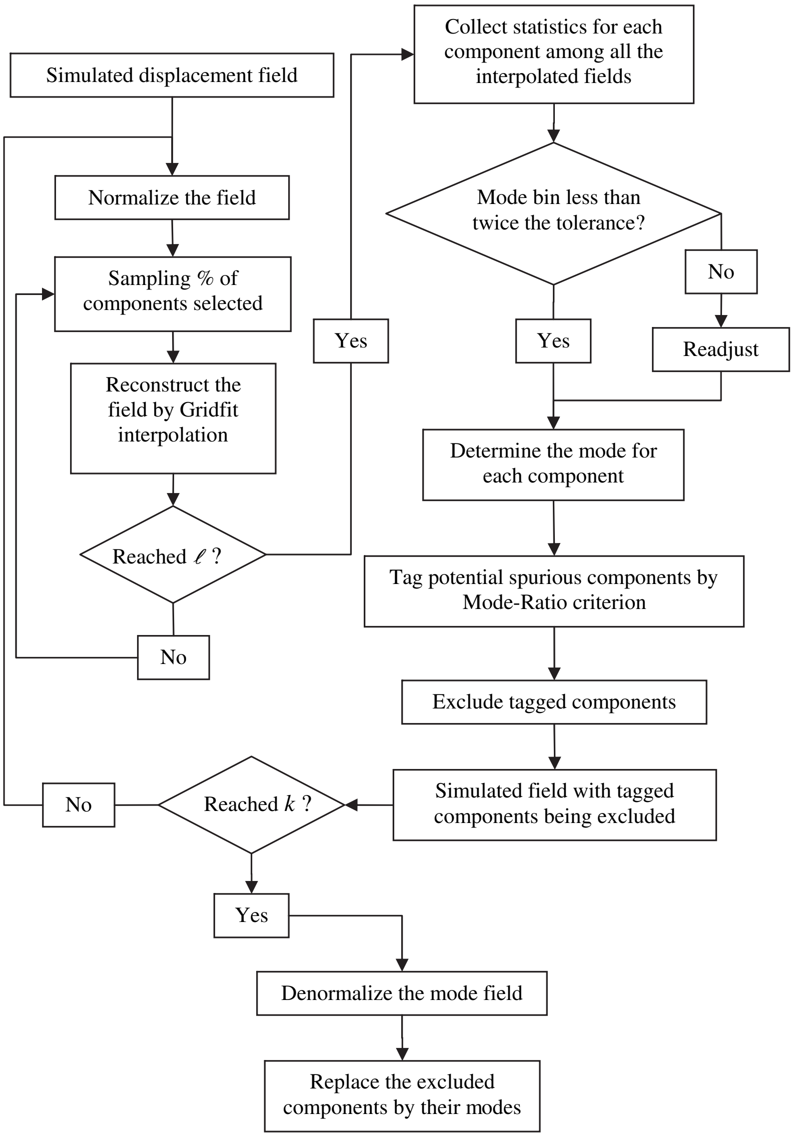
Simulated displacement field
Normalize the field
Sampling % of components selected
Reconstruct the field by Gridfit interpolation
Reached ℓ ?
No
Yes
Collect statistics for each component among all the interpolated fields
Mode bin less than twice the tolerance?
No
Readjust
Yes
Determine the mode for each component
Tag potential spurious components by Mode-Ratio criterion
Exclude tagged components
Simulated field with tagged components being excluded
Reached k ?
No
Yes
Denormalize the mode field
Replace the excluded components by their modes

Figure 5. Flowchart describing the algorithmic procedure for outlier detection and correction.

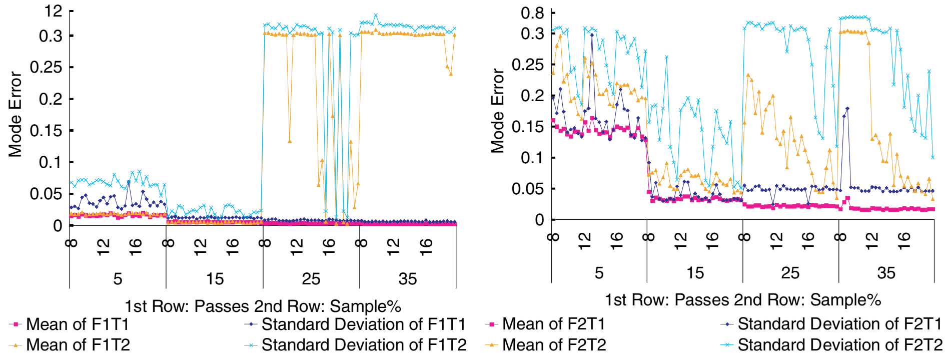

Figure 6. Statistics of the mode error for accurately detected u components for field 1 (left) and field 2 (right).

undetections, while also minimizing the mode error statistics among the tested fields.

Figure 6 shows the mode error statistics of the accurately detected outliers for all parametric studies for field 1 (left) and field 2 (right). In general, 15% data sampling can be seen to produce the lowest mode error statistics, thus yielding the most accurate correction results. While 25% data sampling produces good results for type 1 errors, in general, they

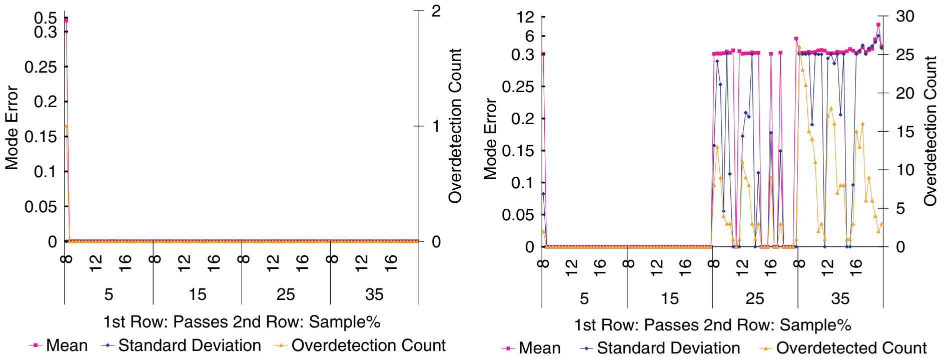

Figure 7. Statistics of the mode error of the overdetected *u* components for field 1 with type 1 errors (left) and type 2 errors (right).

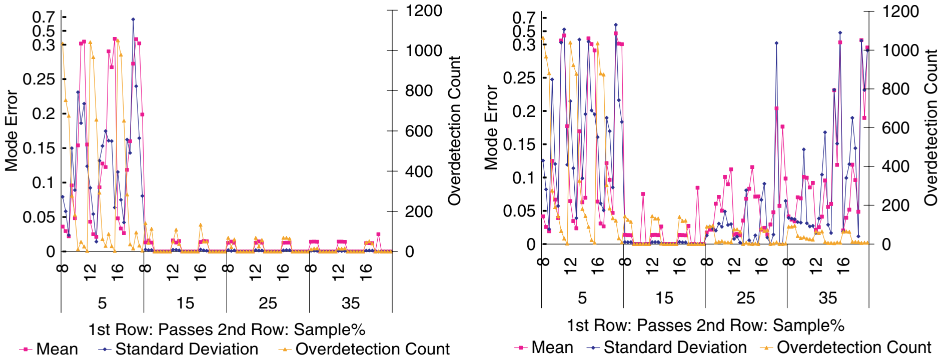

Figure 8. Statistics of the mode error of the overdetected *u* components for field 2 with type 1 errors (left) and type 2 errors (right).

produce large mode error statistics for type 2 errors. Also noticeable is that for the type 2 errors, as the number of passes increases, the mode error statistics decrease.

In addition to investigating the accurately detected statistics, it is also necessary to investigate the mode error statistics for overdetected components as well as the number of undetections in order to properly determine the performance of the correction method. Figure 7 (left) shows the mean and standard deviation of the mode errors and the number of the overdetected components for field 1. No overdetections are observed for type 1 errors except for one case with the 5% sampling size. Figure 7 (right) shows that the 35% data sampling produces the poorest results for type 2 errors. The 25% data sampling performs slightly better than the 35% except for some 16 passes cases, which produce only slightly worse results than the 15%. No components are overdetected for the 5% and 15% data sampling, except for a single case within the 5% sampling size. Figure 8 shows the mean and standard deviation of the mode errors, and the count of the overdetected components for field 2. The 5% overdetection results are statistically and quantitatively worse than other sample percentages regardless of the type of error. Figure 8 (right) clearly shows that 15% data provide the smallest mode error statistics without a significant increase on the number of overdetections.

Since there are no overdetections for field 1 for the 15% data sampling, field 2 is analyzed to determine the optimal parameters that minimize the number of overdetections and undetections. Figure 9 shows the overdetection mode error statistics for field 2 with type 1 errors (left) and type 2 errors (right) for 15% sampling size. While the effect of the passes is not obvious, both figures show that tolerance levels at 0.0025 with iterations higher than 500 provide no overdetections. Figure 10 shows the undetection counts for the 15% data sampling results. Clearly, the 0.001 tolerance provides the lowest undetection count for both fields, with increasing counts as the tolerance value increases. This trend is opposite to that seen for the overdetections shown in figure 9, where the tolerance value of 0.001 results in overdetection counts which are around 140. Therefore, the tolerance value of 0.0025 is selected as a good balance between the number of overdetections and undetections. Figures 9 and 10 also show that for the tolerance value of 0.0025, iteration values of 700 and 1000 best minimize the overdetections without affecting the number of undetections. Lastly, for these parameters, the results are insensitive to the number of passes. In summary, the number of iterations at 700 is chosen for computational efficiency; the tolerance level is chosen at 0.0025 to balance the number of undetections and overdetections, and the number of passes is unspecified. These optimal parameters are summarized in table 2.

Table 2. Sets of optimal parameters.

Sampling (%)	Tolerance	Iteration	Passes
15	0.0025	700	8
15	0.0025	700	12
15	0.0025	700	16

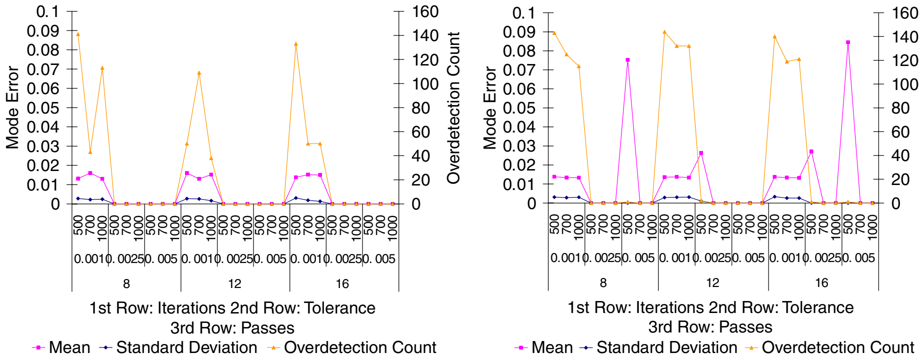

Figure 9. Statistics of the mode error of the overdetected *u* component for field 2 with type 1 errors (left) and type 2 errors (right) and 15% sample size.

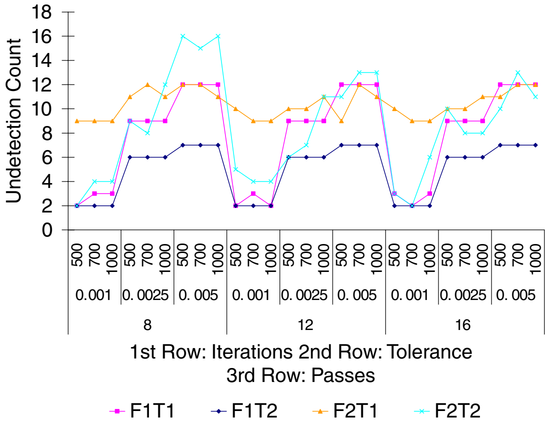

Figure 10. The undetection counts for 15% sampling data.

7. Application to a turbulence jet and computational effort

To identify a single optimal parameter set from table 2, the parameters determined in the prior sections are tested on an experimental turbulent jet flow provided by Westerweel for the second international PIV challenge in Busan, South Korea[1]. The image pairs are 992 × 1004 pixels, and are processed with a 32 × 32 interrogation window using a standard cross-correlation routine with a 50% window overlap [17], resulting in the displacement field shown in figure 11. Similar flows were also used by Shinneeb *et al* [6] and Westerweel [8] to assess the performance of their outlier detection methods on experimental fluid flows. This particular displacement field is most useful as it contains random errors, small outlier clusters (2–6 outliers) and large outlier clusters (>6 outliers), as shown by 1–3, and will therefore be a good testing field for the mode-ratio bootstrapping method. Since this is an experimentally obtained flow and actual locations of the outliers are unknown, the performances of the correction methods are determined visually. This data field is processed with the optimal parameters listed in table 2, the results of which are discussed below.

With eight passes (figure 12), random and small outlier clusters (regions 1 and 2) are easily identified and corrected.

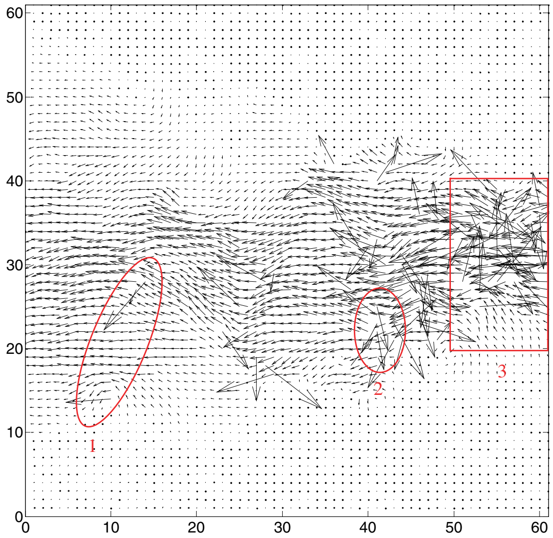

Figure 11. A real turbulence jet flow field. This field is a good example of an experimentally obtained fluid flow, with both random and clustered errors [17].

There are, however, at least two large groups of visibly identifiable cluster outliers (region 3) that have not been identified and corrected. Twelve passes (figure 13) reduce the size of these two clusters, though not completely eliminating them. For comparison, 16 passes (figure 14) clearly show that the random, small and large outlier clusters are identified and corrected, except for one suspect component which does not seem to be accurately corrected. This thereby suggests that the detection scheme better identifies outliers at smaller incremental steps, i.e. higher passes for large cluster sizes. It should also be noted that the large cluster shown in 3 is indicative of improperly acquired PIV images. It is therefore recommended to use 8 passes for appropriately acquired PIV images that result in outlier clusters no larger than about six vectors, 12 passes for random and medium outlier clusters and 16 passes for large sized outlier clusters.

While highly effective, the computational effort for the mode-ratio bootstrapping method is substantial. The present results were obtained using a 2.6 GHz Dell PowerEdge server, with 32 GBytes of RAM. Table 3 lists the computational times for the tested simulation fields and turbulence jet data. It should be further noted that the computing language used was

[1] http://www.pivchallenge.orgpub03/CaseA/.

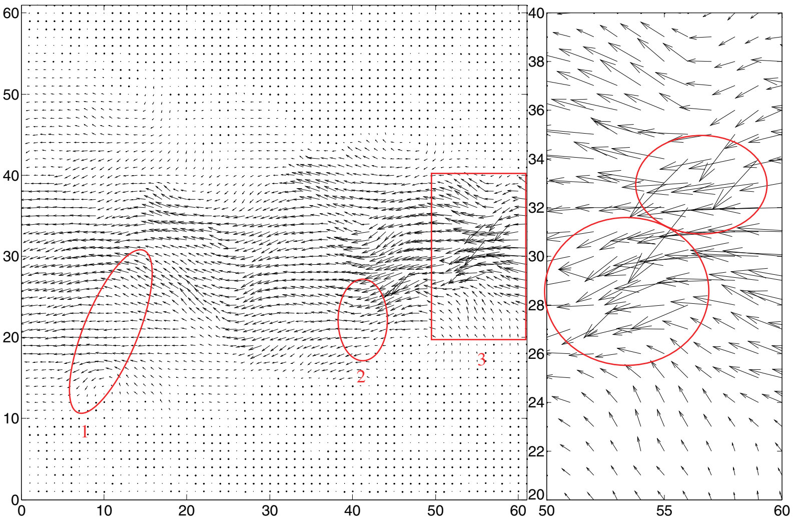

Figure 12. The mode field for the turbulence jet flow processes with 700 iterations, 15% data sampling, 0.025 tolerance and eight passes for the second approach. The plot to the right shows a close-up view of the large cluster outlier region enclosed in the rectangular region identified by the red box shown in the figure to the left.

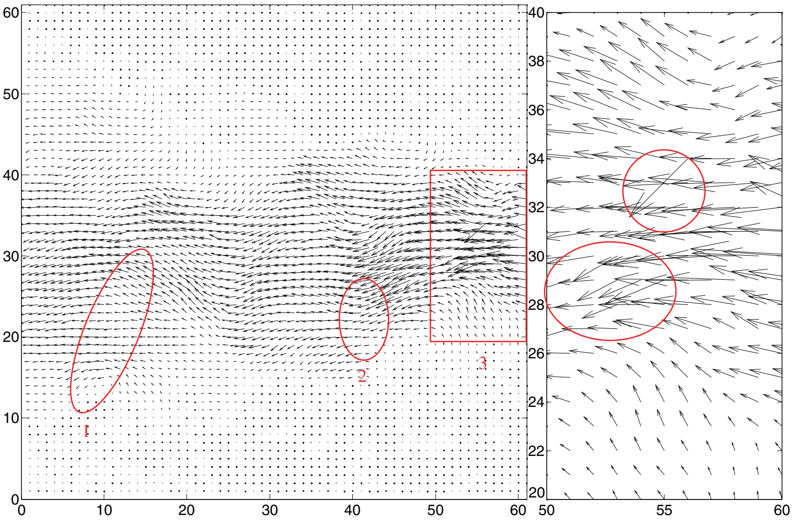

Figure 13. The mode field for the turbulence jet flow processes with 700 iterations, 15% data sampling, 0.0025 tolerance and 12 passes for the second approach. The plot to the right shows a close-up view of the large cluster outlier region enclosed in the rectangular region identified by the red box shown in the figure to the left.

Table 3. A summary of computational expenses.

Passes	Fields 1 and 2 (min) (50 × 50 vectors)	Turbulence jet (min) (61 × 61 vectors)
8	14.12	22.15
12	21.40	32.90
16	27.86	44.10

MATLAB, and no efforts were made toward optimizing the algorithm for speed.

8. The effects of the spatial gradient on the detection and correction results

To address the spatial gradient effects, two factors are considered: (1) the displacement magnitude and (2) the resolution of the vortex cell. The displacement magnitude of the fields is dependent on the field constants, C and $V_{\max}$, for field 1 and field 2, respectively. The resolution is quantified in terms of the number of nodes per vortex for the present study. Because the tested component fields are normalized before the algorithm is processed, the detection results are not dependent on the magnitudes of the field but, rather, only on the spatial gradients.

Since field 2 provided a greater challenge to the mode-ratio bootstrapping method, the effects of the spatial gradient on this field will be studied. To assess the effects of spatial gradients, the total number of vortices is varied while keeping the number of nodes per axis equal (N_x and N_y in equations (9) and (10) are identical), which, in general, leads to the question of flow structure resolution. In this study, the number of nodes in field 2 is held at 50 in each of the x- and y-axes. The optimal

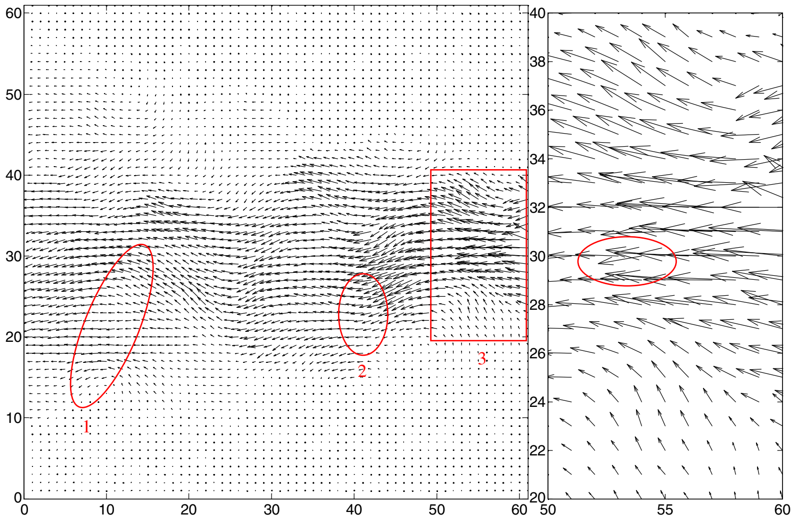

Figure 14. The mode field for the turbulence jet flow processes with 700 iterations, 15% data sampling, 0.0025 tolerance and 16 passes for the second approach. The plot to the right shows a close-up view of the large cluster outlier region enclosed in the rectangular region identified by the red box shown in the figure to the left.

Table 4. Resolution conversion.

Number of vortex cells/axis	2	3	4	5	6	7	8
Number of nodes/vortex	625	277	156	100	69	51	39

parameter set with 16 passes (see table 2) is chosen to process the vortex field with different numbers of vortices in each axis for each run. The processed cases and their resolutions are summarized in table 4.

Figure 15 illustrates how the detection and the correction results change with respect to the number of vortex cells in each axis. As can be observed from the figures, the mode error statistics for the accurately detected components figure 15 (right) and the number of detection errors, figure 15 (left), increase exponentially with decreasing vortex resolution. This is due to fewer nodes per vortex being used as interpolation seeds, thus resulting in decreasing interpolation accuracy. Figure 15 (right) shows that the undetection and overdetection counts for vortex cells ranging from 2 to 4 per axis are at most 10 (4% with respect to the total number of outliers) and zero, respectively, for both type 1 and type 2 errors. Correspondingly, their mode error statistics, figure 15 (left), are seen to be within 0.1 for type 1 errors and around 0.1 for type 2 errors, which is the typical rms noise level observed within PIV data [8, 18]. For vortex cell numbers larger than four per axis, the number of undetections and overdetections, as well as their mode errors, are unacceptably large. It is therefore recommended that if the flow structure resolution is not satisfied, the experiment be repeated with a smaller interrogation window to better achieve detection and correction results.

9. Conclusions and future work

A PIV post-interrogation outlier detection and correction method was developed based on generating statistics using a bootstrapping procedure in order to reduce the effects of gradients within the field. A mode ratio was defined toward identifying outliers, and other parameters affecting this procedure were also defined.

Parametric studies of this outlier correction method have been performed to identify optimum parameter sets. The method was tested on two simulated fields and two types

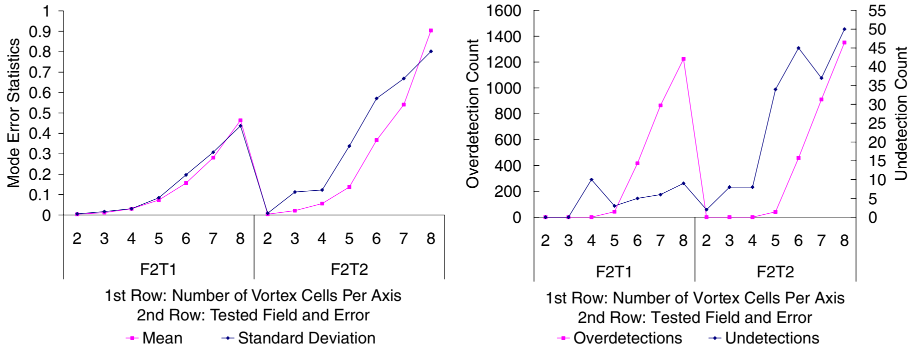

Figure 15. Left: the correction results for the accurately detected vector class as a function of the number of vortex cells in each axis. Right: the detection results as a function of the number of vortex cells in each axis.

of outliers, where 10% of the components are replaced with the outliers. The mean and the standard deviation of the mode errors with respect to the perfect displacements of the fields were calculated and used to determine several parameter sets that were optimal and most robust among the tested fields. The optimal parameter sets were then tested on an experimentally obtained turbulent jet flow inflicted with severe outliers for validation. Visual observations revealed that the proposed parameter sets show that eight passes are sufficient to detect random and small-sized cluster errors, 12 passes are sufficient to detect random, small- and medium-sized cluster errors and 16 passes are sufficient to detect random, small-, medium- and large-sized cluster errors. To determine the effect of the outlier percentage on the optimum parameters, identical data processing of these synthetic fields with 5% outliers revealed that the optimum parameters were identical to those in the present study. This suggests that the optimum parameters found in the present study can be used for any outlier percentages below 10%.

The effects of the distribution of the spatial gradient are examined by varying the number of vortices for field 2 in both the x- and the y-axes, while keeping the total number of nodes constant, thereby quantifying the flow structure resolution by the number of vortices per axis. The set of optimum parameters determined from the parametric studies are also applied on these different vortex fields to assess the effects of spatial gradients on detection and correction. The results show that the number of undetections and overdetections, as well as the mode error statistics, increase exponentially with an increasing number of vortices per axis, a fact which is due to the interpolation difficulties occurring over large spatial gradients with low resolution.

It is postulated that the mode-ratio method would be dependent upon the performance of the interpolation method, especially at the boundary points, which was not investigated in this paper. For future work, it is recommended that the effects of various interpolation methods be investigated in order to minimize their effects and determination on the optimum parameters that result in fast and accurate PIV outlier detection and correction.

Acknowledgments

The authors gratefully acknowledge the support of the Mary Gates Endowment Scholarship program at the University of Washington. The authors also acknowledge the comments of the reviewers whose insights have added to the quality of this paper.

References

[1] Raffel M, Willert M and Kompenhans J 1998 *Particle Image Velocimetry, A Practical Guide* (Berlin: Springer)

[2] Westerweel J 1994 Efficient detection of spurious vectors in particle image velocimetry data *Exp. Fluids* **16** 236–47

[3] Nogueira J, Lecuona A and Rodriguez P 1997 Data validation, false vectors correction and derived magnitudes calculation on PIV data *Meas. Sci. Technol.* **8** 1493–501

[4] Song X, Yamamoto F, Iguchi M and Murai Y 1999 A new tracking algorithm of PIV and removal of spurious vectors using Delaunay tessellation *Exp. Fluids* **26** 371–80

[5] Liang D, Jiang C and Li Y 2003 Cellular neural network to detect spurious vectors in PIV data *Exp. Fluids* **34** 52–62

[6] Shinneeb A-M, Bugg J D and Balachandar R 2004 Variable threshold outlier identification in PIV data *Meas. Sci. Technol.* **15** 1722–32

[7] Young C, Johnson D and Weckman E 2004 A model-based validation framework for PIV and PTV *Exp. Fluids* **36** 23–35

[8] Westerweel J and Scarano F 2005 Universal outlier detection for PIV data *Exp. Fluids* **39** 1096–100

[9] Efron B 1979 Bootstrap methods: another look at the jackknife *Ann. Stat.* **7** 1–26

[10] Efron B and Tibshirani R J 1993 *An Introduction to The Bootstrap* (New York: Chapman and Hall)

[11] Diaconis P and Efron B 1983 Computer-intensive methods in statistics *Sci. Am.* **248** 116–30

[12] Moore D S, McCabe G, Duckworth W and Sclove S 2003 Bootstrap methods and permutation tests, http://bcs.whfreeman.com/pbs/cat_140/chap18.pdf

[13] Rignot E J M and Spedding G R 1988 Performance analysis of automatic image processing and grid interpolation techniques for fluid flows *USC Aerospace Engineering Internal Report*

[14] Dabiri D and Gharib M 1991 Digital particle image thermometry: the method and implementation *Exp. Fluids* **11** 77–86

[15] Spedding G R and Rignot E J M 1993 Performance analysis and application of grid interpolation techniques for fluid flows *Exp. Fluids* **15** 417–30

[16] D'Errico J 2005 Surface fitting using gridfit, http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=8998&objectType=FILE

[17] Willert C and Gharib M 1991 Digital particle image velocimetry *Exp. Fluids* **10** 181–93

[18] Westerweel J 2000 Theoretical analysis of the measurement precision in particle image velocimetry *Exp. Fluids* **29** S3–12