Experiments in Fluids (2005) 39: 1096–1100
DOI 10.1007/s00348-005-0016-6

LETTER

Jerry Westerweel · Fulvio Scarano

Universal outlier detection for PIV data

Received: 28 January 2005 / Revised: 9 May 2005 / Accepted: 17 June 2005 / Published online: 12 August 2005

Abstract An adaptation of the original median test for the detection of spurious PIV data is proposed that normalizes the median residual with respect to a robust estimate of the local variation of the velocity. It is demonstrated that the normalized median test yields a more or less 'universal' probability density function for the residual and that a single threshold value can be applied to effectively detect spurious vectors. The generality of the proposed method is verified by the application to a large variety of documented flow cases with values of the Reynolds number ranging from 10^{-1} to 10^{7}.

The so-called 'median test' is the most widely used method for outlier detection in post-interrogation validation of PIV data (Westerweel 1994). The principle and effectiveness of the original method was demonstrated for homogeneous and isotropic turbulence, for which a single outlier detection threshold can be applied to the entire data set on the basis of the velocity fluctuation intensity. The original paper contains an appendix that describes the application of the method for general (turbulent) flow fields, but this recipe is quite elaborate and requires a priori information of the flow field; this procedure has—to the best of the authors knowledge—never been actually applied in practice. Instead, it is quite common to apply a single detection threshold in the evaluation of (strongly) inhomogeneous flow data. For example, in a PIV measurement of a submerged turbulent jet (see Fig. 1) the observed flow region contains both high-velocity turbulent data (inside the jet) and low-velocity laminar data (outside the jet); the average value of the median vector residual correlates with the mean velocity (Fig. 1). This means that a single detection threshold applied to the entire flow domain will generally tend to reject part of the valid measurement data in the turbulent flow region and to accept part of the spurious measurement data in the laminar flow region. This problem even occurs in experiments where the assumption of (near) homogeneous and (near) isotropic turbulence appears applicable, i.e., grid-generated turbulence. Figure 2 shows histograms of the median vector residuals from PIV measurements in grid turbulence at three distances from the grid (Poelma 2004); the mean residual clearly decays in correspondence to the decay of the turbulent kinetic energy. Again, a single detection threshold would tend to accept an increasing fraction of spurious data as the measurement location moves away from the grid.

This effect, illustrated in the two examples above, can be avoided when the residual is normalized with an estimate of the local instantaneous flow fluctuations physically expected. The most straightforward choice is to adopt the root-mean-square velocity fluctuation u' evaluated within a close neighborhood of the vector under consideration. However in general two problems occur: (1) for the estimation of u' only a small number of data is available (i.e., typically only 8 to 24—strongly correlated—data points in a 3×3 or 5×5 neighborhood), and (2) the local neighborhood can contain spurious measurement data which makes the estimation of u' unreliable. Effectively, the presence of a spurious measurement will result in an over-estimated value for u' which will *reduce* the value of the 'normalized residual' and possibly will make the estimated residual drop below the chosen threshold value; hence, this will make it more difficult to detect spurious data in the presence of other spurious data.

Shinneeb et al. (2004) use a filter on the PIV data to determine a local threshold value that should account

J. Westerweel (✉)
Laboratory for Aero and Hydrodynamics,
Delft University of Technology,
Delft, The Netherlands
E-mail: j.westerweel@wbmt.tudelft.nl

F. Scarano
Department of Aerospace Engineering,
Delft University of Technology, Delft, The Netherlands

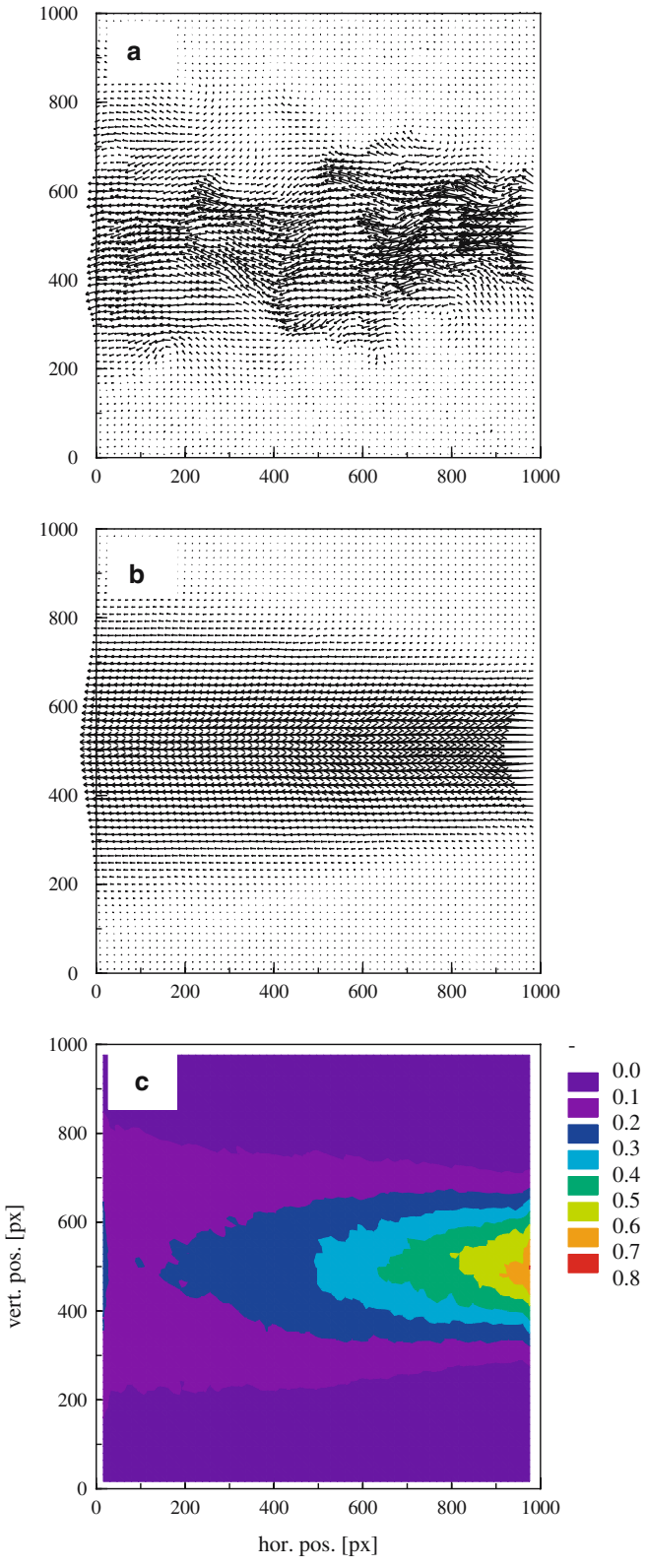

Fig. 1 The instantaneous (**a**) and averaged (**b**) displacements for a turbulent jet (Fukushima et al. 2002), and (**c**) the corresponding averaged residual (in pixel units) based on the vector median

for the local variation of u' and for local gradients. Although this method enhances the detection efficiency, it still relies on a (stringent) initial outlier detection and an appropriate choice of a filter length; both will depend on the interrogation resolution and experimental conditions.

The authors therefore propose an adaptation of the original median test that uses a median estimate of u' that is robust with respect to the presence of spurious measurement data in the neighborhood. Consider a displacement vector denoted by U_0, its 3×3-neighborhood data, denoted by $\{U_1,U_2,...,U_8\}$, and U_m as the median of $\{U_1,U_2,...,U_8\}$ (following the procedure for vector data (Westerweel 1994); note that U_0 is excluded). A residual r_i, defined as: $r_i = |U_i - U_m|$ (Westerweel 1994), is determined for each vector $\{U_i \mid i = 1,...,8\}$, and the median r_m of $\{r_1,r_2,...,r_8\}$ is used to normalize the residual of U_0:

$$r'_0 = \frac{|U_0 - U_m|}{r_m}. \tag{1}$$

The algorithm is represented in *pseudo code* and as a Matlab macro in the Appendix. This method is quite general in outlier detection (Barnett and Lewis 1978) and can be used to process a large variety of inhomogeneous data, including e.g. traffic data (Shekhar et al. 2001; Wouters et al. 2005, in press).

When the residual defined in Eq. 1 is applied to the grid turbulence data, it is found that the histograms of the residuals approximately collapse on a single curve, i.e., the residuals for the normalized median test become independent of the turbulence level, as shown in Fig. 2b. When the normalized median test is applied to the jet data of Fig. 1, the correlation between the mean residual and the mean displacement has been substantially attenuated, as is evident by comparing Fig. 1c with Fig. 3a. However, a weak correlation of the mean residual and the turbulence level remains visible, and it was found that r_0' shows elevated values for regions with very low turbulence intensities (e.g., the laminar outer flow regions of jets and boundary layers). In fact, for purely uniform flow, the normalization factor r_m' tends to zero. This can be compensated by assuming a minimum normalization level ε, i.e.:

$$r_0^* = \frac{|U_0 - U_m|}{r_m + \varepsilon}, \tag{2}$$

where ε may represent the acceptable fluctuation level due to cross-correlation. Evidently, Eq. 2 with $\varepsilon \equiv 0$ yields Eq. 1. It was found that a suitable value for ε is about 0.1 px, which would correspond to the typical rms noise level of the PIV data (Westerweel 2000).

In Fig. 3b is also shown the mean residual as defined in Eq. 2 with $\varepsilon = 0.1$ px; this has further attenuated the correlation between the mean residual and turbulence level.

We now demonstrate that the normalized median test yields residuals with a more or less 'universal'

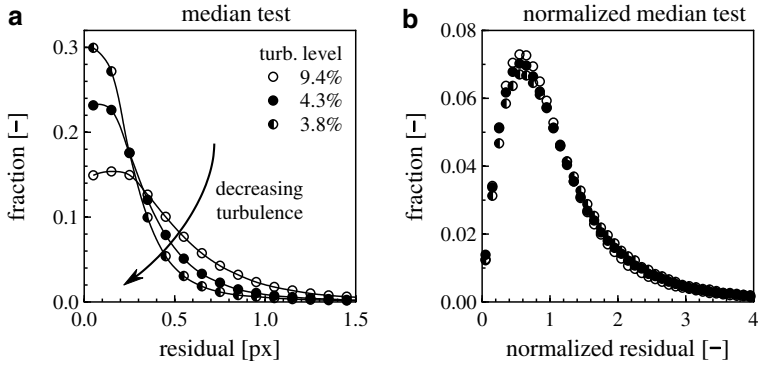

Fig. 2 The histograms of the residual obtained with the conventional median test (**a**) and the normalized median test (**b**) for the grid turbulence data at decreasing turbulence levels. The *histograms* represent at least 99.7% of the vector data

probability density function and that a single threshold value can be used for the detection of spurious vector data.

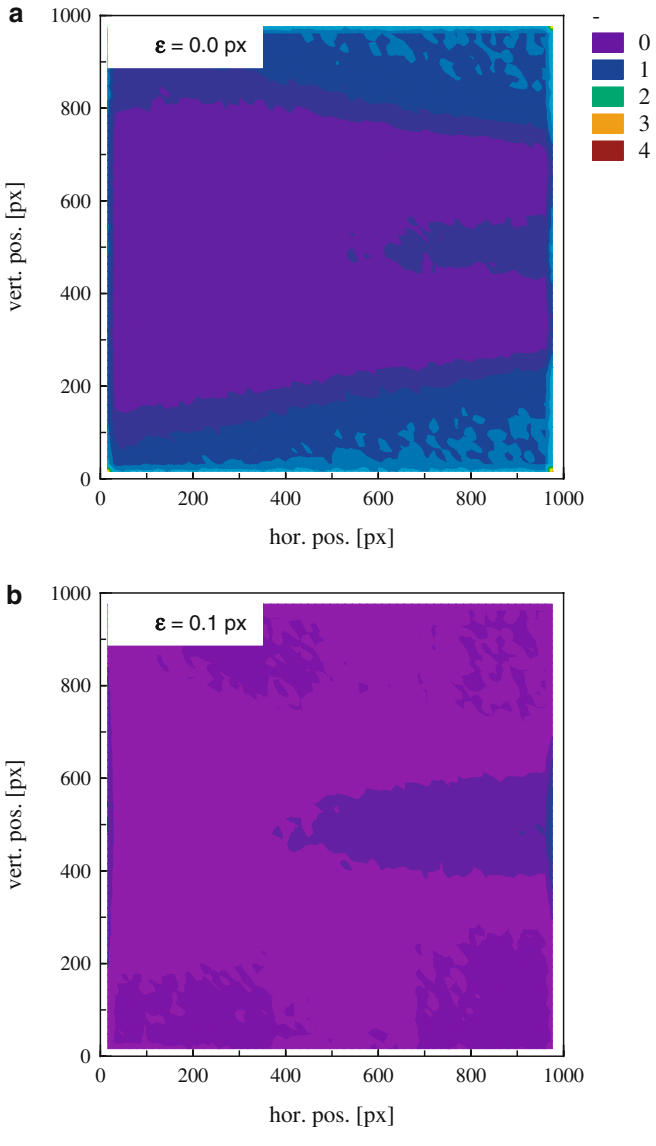

Fig. 3 Contour plot of the averaged residual based on the normalized median defined in Eq. 2 with $\varepsilon = 0$ (**a**) and $\varepsilon = 0.1$ px (**b**) respectively. The *range* of contour levels corresponds to the one in Fig. 1c

To demonstrate this, the normalized median test is applied to a variety of (documented) PIV experiments, listed in Table 1. These experiments cover flows with

Table 1 Overview of PIV data and corresponding references

Description	Reference
Grid turbulence	(Poelma 2004)
Turbulent pipe flow	(Westerweel et al. 1996)
Turbulent jet	(Fukushima et al. 2002)
Microchannel flow	(Westerweel et al. 2004)
Backward-facing step	(Scarano et al. 1999)
Supersonic wake	(Scarano and Van Oudheusden 2003)
Von Kármán wake	(Van Oudheusden et al. 2005)

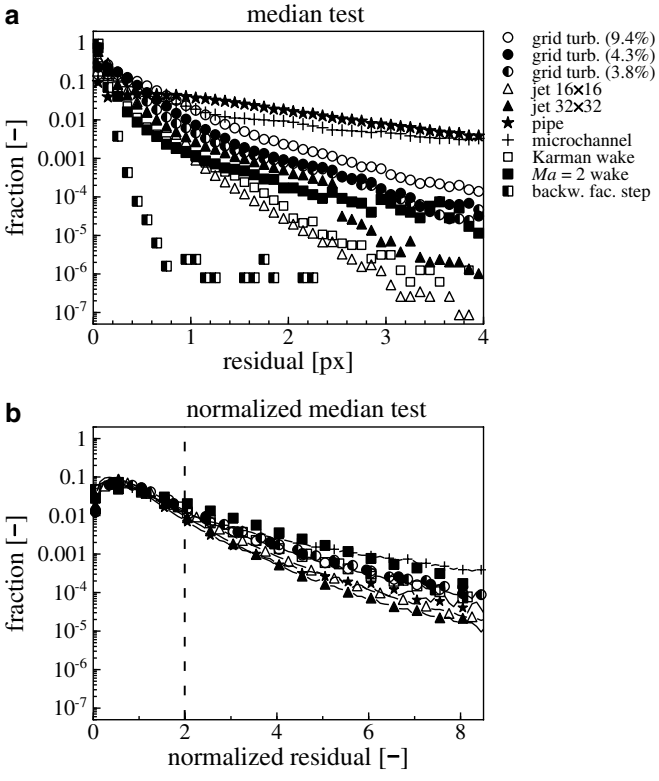

Fig. 4 The histograms of the residuals using the conventional median (**a**) and the normalized median (**b**) for the experimental data listed in Table 1

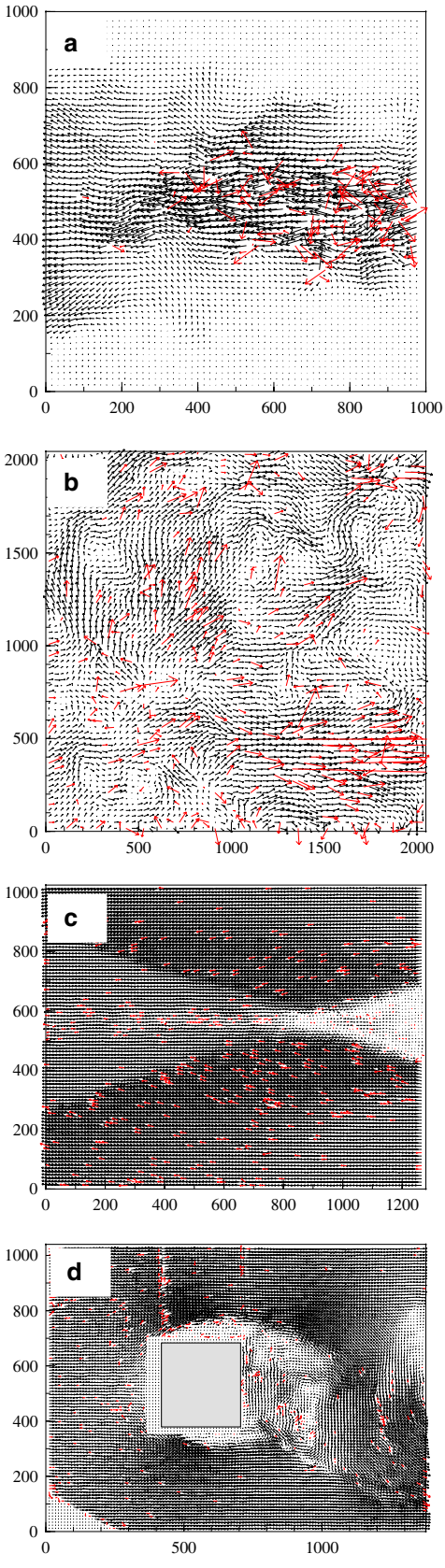

Fig. 5 Examples from the data listed in Table 1 where vectors with $r^* > 2$ are shown in *red*: **a** turbulent jet, **b** grid turbulence (with mean displacement subtracted), **c** supersonic ($Ma = 2$) wake, **d** von Kármán wake

Reynolds numbers from 0.1 (for micro-channel flow) to 10^7 (for the supersonic wake). Histograms of the residuals for the conventional median test and the normalized median test are shown in Fig. 4. The histograms of the residuals for the conventional median test strongly depend on the experimental conditions, whereas the histograms for the residuals of the normalized median test appear to more or less collapse on a single curve. This graph also shows how the residuals for the conventional median test depend on the interrogation resolution (e.g., compare the 32×32-px and 16×16-px jet data), which implies that each pass in a multi-grid or multi-scale PIV interrogation would require its own (optimal) detection level for the identification of spurious data; the same data for the normalized median test practically coincide, so that for each pass the same detection level can be used.

When the histograms of the residual for the normalized median test in Fig. 4 would be integrated, it is found that the 90-percentile occurs for $r' \approx 2$. This means that in *all* cases a *single* detection threshold can be used that labels the largest 10% of residuals. A value larger than 2 would yield a less stringent detection, whereas a value smaller than 2 would yield a more stringent detection. Hence, we now have a detection threshold that is more or less independent of the (local) level of the velocity fluctuations, and that even appears valid for different experimental conditions.

In Fig. 5 are shown the vector plots from four arbitrary examples of the data listed in Table 1. Each of these examples clearly shows a small fraction of spurious vectors, and in each example vectors with a residual $r^* > 2$ have been indicated by a red color, which essentially captures all the spurious vector data in each of these four examples.

In conclusion, it has been demonstrated that a small adaptation of the original algorithm for the median test of PIV data yields a 'universal' distribution of the normalized residual. Instead of a detection threshold for spurious vector data that is specific to each experiment, or different flow regions within a PIV measurement domain, it is now possible to use a single detection threshold that would be applicable to a variety of flow conditions without any a priori knowledge of the flow characteristics (e.g., turbulence level). The universal character makes it also possible to use a single detection threshold in transient flows (e.g., in laminar-turbulent transition).

A threshold value of about 2 seems to be an appropriate choice, with smaller and larger values leading to a more stringent and less stringent outlier detection respectively. The use of a normalized median test makes the implementation of multi-pass or multi-grid PIV

interrogation more straightforward, as the normalized median test eliminates a dependence of the detection criterion on the interrogation domain size. It is also a great help to inexperienced users who can use a threshold value of 2 as a convenient starting point for outlier detection.

Appendix

The normalized median test in *pseudo code* and implemented as a Matlab macro.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Matlab macro for detection of outliers in cross-correlation analysis of PIV recordings   %
% Authors: F. Scarano, J. Westerweel, Delft University of Technology                       %
% Date 29-April-2005                                                                       %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% The macro input is the displacement vector field components "U" and "V" and the
% fluctuation threshold "Thr". "U" and "V" are two dimensional arrays with I columns and
% J rows. The routine output is the two-dimensional array containing the error indicator:
% Info1=0 if the data point is not an outlier, Info1=1 otherwise
%
% Pseudo Code:
%
% Input the displacement vector field components "U" and "V"
% Input the value of the selected threshold (commonly 2)
%
% Set the value of the neighborhood radius "B"; (commonly 1 or 2)
% Set the value of the noise threshold level "Eps"; (e.g. eps=0.1)
%
% repeat for each velocity component
%    loop over all data-points
%       select data neighborhood excluding the center-point: "Neigh"
%       evaluate the median value: "NeighMed"
%       evaluate the fluctuation w.r.t. median: "MedFluct"="U"-"NeighMed";
%       evaluate neighborhood fluctuation w.r.t. median: "NeighMedFluct"="Neigh"-"neighMed"
%       evaluate median of absolute value of neighbourhood fluctauation: "MedNeighMedFluct"
%       normalised fluctuation: "NormFluct"=abs("MedFluct"/("MedNeighMedFluct"+"Eps"));
%    end of loop
%
% combine the fluctuation from each of the velocity components (e.g. sum or maximum)
% apply detection criterion: if "NormFluctComb" > "Thr" then the vector is an outlier

function Info1=normres(U,V,Thr)

[J,I]=size(U);          % size of the displacement field

Medianres=zeros(J,I);   % initialise median residual
Normfluct=zeros(J,I,2); % initialise normalised fluctuation
b=1;                    % data-point neigborhood radius (commonly set to 1 or 2)
eps=0.1;                % estimated measurement noise level (in pixel units)

for c=1:2 % loop over the two velocity components
   if c==1; VelComp=U; else; VelComp=V; end;
   % loop over all the data-points (excluding border)
   for i=1+b:I-b
      for j=1+b:J-b
         Neigh=VelComp(j-b:j+b,i-b:i+b); % data neighborhood with center-point
         NeighCol=Neigh(:);              % in column format
         NeighCol2=[NeighCol(1:(2*b+1)*b+b); NeighCol((2*b+1)*b+b+2:end)];
                                         % neighborhood excluding center-point
         Median=Median(NeighCol2);       % median of the neighborhood
         Fluct=VelComp(j,i)-Median;      % fluctuation with respect to median
         Res=NeighCol2-Median;           % residual: neighborhood fluctuation w.r.t. median
         MedianRes=Median(abs(Res));     % median (absolute) value of residual
         NormFluct(j,i,c)=abs(Fluct/(MedianRes+eps));
                                         % normalised fluctuation w.r.t. neighborhood
                                         % median residual
      end;
   end;
end;

info1=(sqrt(NormFluct(:,:,1).^2 + NormFluct(:,:,2).^2) > Thr); % detection criterion
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

References

Barnett V, Lewis T (1978) Outliers in statistical data. Wiley, Chichester

Fukushima C, Aanen L, Westerweel J (2002) Investigation of the mixing process in an axisymmetric turbulent jet using PIV and LIF. In: Adrian RJ et al (eds) Laser techniques for fluid mechanics. Springer, Berlin Heidelberg New York, pp 339–356

Poelma C (2004) Experiments in particle-laden turbulence. PhD Thesis, Delft University of Technology

Scarano F, Benocci C, Riethmuller ML (1999) Pattern recognition analysis of the turbulent flow past a backward facing step. Phys Fluids 11:3808–3818

Scarano F, Van Oudheusden BW (2003) Planar velocity measurements of a two-dimensional compressible wake. Exp Fluids 34:430–441

Shekhar S, Lu C, Zhang P (2001) A unified approach to spatial outlier detection. Tech Rep 01-045 Univ Minnesota

Shinneeb A-M, Bugg JD, Balachandar R (2004) Variable threshold outlier identification in PIV data. Meas Sci Technol 15:1722–1732

Van Oudheusden BW, Scarano F, Van Hinsberg NP, Manna L (2005) Phase–resolved characterization of vortex shedding in the near wake of a square-section cylinder at incidence. Exp Fluids 39:86–98

Westerweel J (1994) Efficient detection of spurious vectors in particle image velocimetry data sets. Exp Fluids 16:236–247

Westerweel J (2000) Theoretical analysis of the measurement precision in particle image velocimetry. Exp Fluids 29:S3–S12

Westerweel J, Draad AA, Van der Hoeven JGT, Van Oord J (1996) Measurement of fully-developed turbulent pipe flow with digital particle image velocimetry. Exp Fluids 20:165–177

Westerweel J, Geelhoed PF, Lindken R (2004) Single-pixel resolution ensemble correlation for micro-PIV applications. Exp Fluids 37:375–384

Wouters JAA, Chan K-F, Kolkman J, Kock RW (2005) Customized pre-trip prediction of freeway travel times for road users. Proc Trans Res (in press)