ISSN 1750-9823 (print)
International Journal of Sports Science and Engineering
Vol. 01 (2007) No. 02, pp. 129-136

Clustering and display of elite swimmers' race patterns across various comparable criteria at the same time [*]

Ilene Chen [1 +], Hiroto Homma [2], Craig Jin [1], Hong Yan [1, 3]

[1] School of Electronic and Information Engineering, University of Sydney, Australia
[2] North Sydney Aussie Swimming Inc., Australia
[3] Department of Electronic Engineering, City University of Hong Kong, Hong Kong

Abstract. In addressing the practical difficulty of surveying elite swimmers' race patterns within an individual event from multiple swimming championships across various comparable criteria at the same time, a system of cluster analysis is described that uses standard statistical algorithms to arrange elite swimmers according to similarity in their race patterns. The outputs are displayed graphically using heat maps and tree-like dendrograms, transmitting the information to our human brains by means of a high-bandwidth channel than a "number-reading" channel. Since the end product is a visual representation of elite swimmers' race patterns within an individual event from multiple swimming championships across various comparable criteria at the same time, it would meet with a great deal of enthusiasm if coaches could develop an integrated knowledge into elite swimmers' race patterns being studied from such images. As an illustration, a combined dataset relating to the finalists' race patterns for the men's 200m freestyle finals from 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid are used to represent the hidden group structures in connection with their race results at any one time.

Keyword: cluster analysis, heat map, tree-like dendrogram, race pattern

1. Background

Throughout this paper, the reader is assumed to have undergone rudimentary statistical training such as one might acquire in an elementary course in multivariate analysis. A range of standard statistical algorithms will be evolved to enable coaches to monitor, rapidly and efficiently, elite swimmers' race patterns within an individual event from multiple swimming championships across various comparable criteria at the same time. Emphasis will be placed on the identification of the hidden group structures with reference to elite swimmers' race results. A combined dataset relating to the finalists' race patterns for the men's 200m freestyle finals from four major swimming championships will be used for the representation of the hidden group structures in connection with their race results at any one time. In this paper, we focus on the practical difficulty of surveying elite swimmers' race patterns within an individual event from multiple swimming championships across various comparable criteria at the same time.

In a sense it is naturally obvious that a coach investigating each first-rate elite swimmer's race patterns during major swimming championships and we must also ask the question: why are those particular first-rate elite swimmers present [1]? The ready answer seems to be available for the complementary question: what are the hidden group structures in connection with their race results within an individual event from multiple major swimming championships across various comparable criteria at the same time? To answer this question, sport scientists have identified with related work since the last two decades of the past century [2-9]. Unfortunately, their attempts were unsuccessful since the methods they developed did not enable coaches to derive any insight into any of the elite swimmers' race patterns intuitively [1]. Moreover, the outputs they presented were merely summary statistics of various quantities relating to elite swimmers' race patterns; consequently, a detailed knowledge of the rapid motion of elite swimmers' race patterns would be quite superficial [1].

[*] This work is in part supported by a grant from City University of Hong Kong (Project CityU 9610034).

[+] Correspondence should be addressed to Ilene Chen ilene@ee.usyd.edu.au *Tel*: +61 2 93517221 *Fax*: +61 2 93513847

In our opinion, the nature of the difficulty seems to lie in the inadequacy of available statistical machinery to appreciate the quantitative information relating to elite swimmers' race patterns within an individual event from multiple swimming championships at any one time. Since we have very little a priori knowledge of the absolute repertoire of the actual factors which determine the winner, a natural way is to find the hidden rudiments with reference to elite swimmers' race results at the same time [1]. To do this, a natural basis for the extraction of this information is to first construct a combined dataset relating to elite swimmers' race patterns within an individual event from multiple swimming championships, and then to arrange swimmers into clusters according to similarity in their race patterns. The first step to this end is to adopt a mathematical description of similarity [10]. For any series of quantities, several similarity metrics in the behaviour of two swimmers can be employed, such as the Euclidean distance, Pearson correlation coefficient, or Spearman correlation coefficient of the two *n*-dimensional vectors representing a series of *n* measurements [9]. we have fund here that the Euclidean distance assimilates adequately to the insightful mechanisms of elite swimmers' race patterns within an individual event from multiple swimming championships at the same time; this may be so since this metric captures similarity in the "distance" of the two series of quantities.

The purpose of this paper is to show how such method can be intuitive to coaches in the identification of hidden information in connection with elite swimmers' race results on the basis of their race patterns within an individual event from multiple swimming championships at the same time. It is very often convenient to display the primary combined data tables relating to elite swimmers' race patterns using heat maps, which visually represent quantitative information relating to elite swimmers' race patterns over their entire race record times through each of the i-th laps. The computed tree-like dendrograms can be used to order elite swimmers with reference to their race results across various comparable criteria, so that swimmers or groups of swimmers with similar race patterns are adjacent [1]. Since the end product is a visual representation of elite swimmers' race patterns within an individual event from multiple swimming championships at the same time, it would meet with a great deal of enthusiasm if coaches could develop an integrated knowledge into elite swimmers' race patterns being studied from such images. To illustration this approach, a combined dataset relating to the finalists' race patterns for the men's 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid are used to represent the hidden group structures in connection with their race results at any one time.

2. Methods

Source of data. A total of four longitudinal data tables relating to the finalists' race patterns for the men's 200m freestyle finals from 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid were collected from a web site maintained by a sport scientist, Professor Rein Haljand, Department of Kinesiology, Tallinn University. These data tables are available from Professor Rein Haljand's web site at http://www.swim.ee/competition/.

In this analysis, we have constructed a combined dataset relating to the finalists' race patterns on swimming speeds, stroke lengths, turning speeds, and turning times, respectively, within the men's 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid (Supplementary Dataset).

Hierarchical clustering. Cluster analysis is a genetic term that attempts to determine whether or not a given data contain distinct groups, and, if so, to determine the groups [11]. Two general reference on the topic of cluster analysis are given by Hartigan [12] and Gordon [13]. In this analysis, a hierarchical clustering algorithm is used, which leads to a series of hierarchical grouping, reported by tree-like dendrograms [11]. This method starts from a similarity matrix calculated between the swimmers to be clustered on the strength of their race patterns. Similarity metric used here is the Euclidean distance, defined as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{d} \left(X_{ik} - X_{jk}\right)^2}\ ,$$

where X_{ik} and X_{jk} are the parameter values for swimmers i and j; the number of various comparable criteria is d. The similarity between two clusters is based on average linkage, which judges the inter-cluster distance as the average of the distance between all pairs of data points, where members of a pair are in a distinct cluster [11].

Displays. The primary combined dataset relating to elite swimmers' swimming speeds, stroke lengths, turning speeds, and turning times, respectively, are reported graphically using heat maps with columns corresponding to swimmers and rows corresponding to quantities on swimmers' swimming speeds, stroke lengths, turning speeds, and turning times, respectively. Here we used heat maps to plot the data relating to elite swimmers' swimming speeds, stroke lengths, turning speeds, and turning times, respectively, on a 2-dimensional grids and use the rank-by-feature color scale scheme to represent the magnitude of the data on the strength of their race patterns [14, 15]. In this analysis, medium data points are coloured black, data points with increasing ranks with reds of increasing intensity and decreasing ranks with greens of decreasing intensity. Then, outputs from average linkage hierarchical clustering are displayed using tree-like dendrograms that are appended to the corresponding heat maps to indicate the nature of the computed relationship among elite swimmers in the combined primary dataset using any ordering.

3. Results

We have used average linkage hierarchical clustering for the combined dataset on the finalists' swimming speeds (Fig. 1), stroke lengths (Fig. 2), turning speeds (Fig. 3) and turning times (Fig. 4), respectively, for the men's 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid at the same time. In particular, quantitative information relating to the finalists' swimming speeds, stroke lengths, turning speeds, and turning times, respectively, were recorded repeatedly over their entire race record times through each of the i-th laps. The order of the race record times commenced from the fastest within the men's 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid is: Hoogenband (01:44.9 in Berlin), Hoogenband (01:45.3 in Sydney), Thorpe (01:45.8 in Sydney), Davis (01:46.7 in Sydney), Rosolino (01:46.7 in Sydney), Brembilla (01:46.9 in Berlin), Rosolino (01:47.3 in Helsinki), Hoogenband (01:47.5 in Madrid), Hoogenband (01:47.6 in Helsinki), Palmer (01:47.8 in Sydney), Rosolino (01:48.0 in Berlin), Kapralov (01:48.3 in Madrid), Svoboda (01:48.4 in Berlin), Magnini (01:48.7 in Madrid), Salter (01:48.7 in Sydney), Rosolino (01:48.7 in Madrid), Barnier (01:48.8 in Berlin), Leveaus (01:48.8 in Madrid), Say (01:48.8 in Sydney), Svoboda (01:49.2 in Madrid), Wildeboer (01:49.2 in Berlin), Oikonomou (01:49.4 in Berlin), Palmer (01:49.5 in Helsinki), Hackett (01:49.5 in Sydney), Herbst (01:49.6 in Berlin), Svoboda (01:50.2 in Helsinki), Herbst (01:50.4 in Helsinki), Wildeboer (01:50.9 in Madrid), Kapralov (01:51.4 in Helsinki), Carstensen (01:51.4 in Helsinki), Carstensen (01:52.2 in Madrid), and Arnarson (01:52.3 in Helsinki).

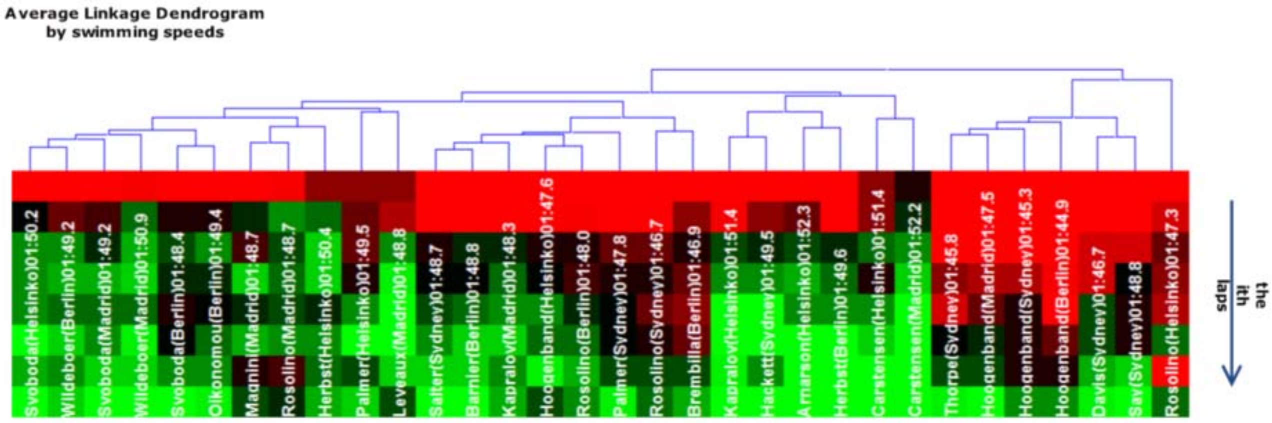

Fig. 1 Clustered display of data from the finalists' swimming speeds within the men's 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid.

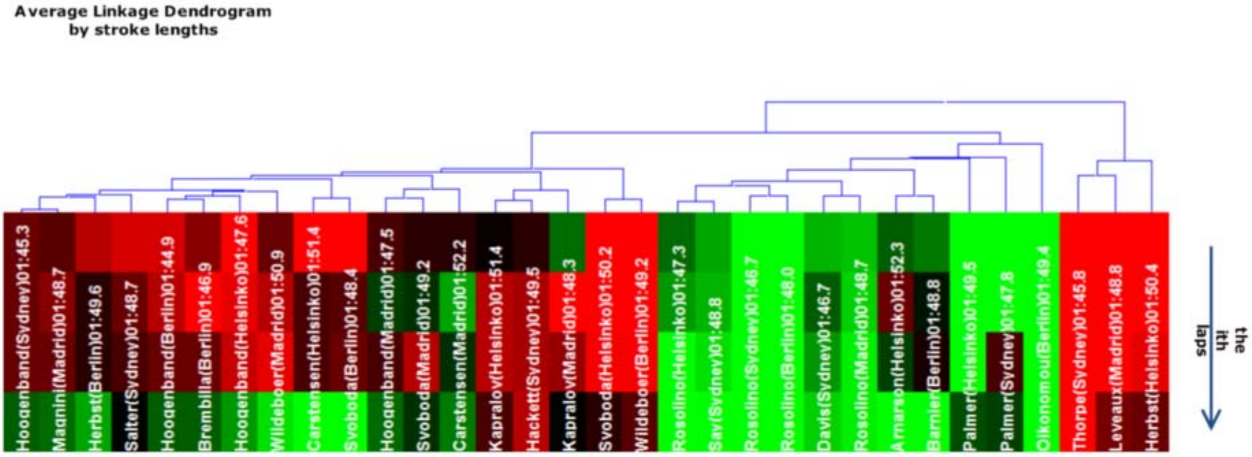

Fig. 2 Clustered display of data from the finalists' stroke lengths within the men's 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid.

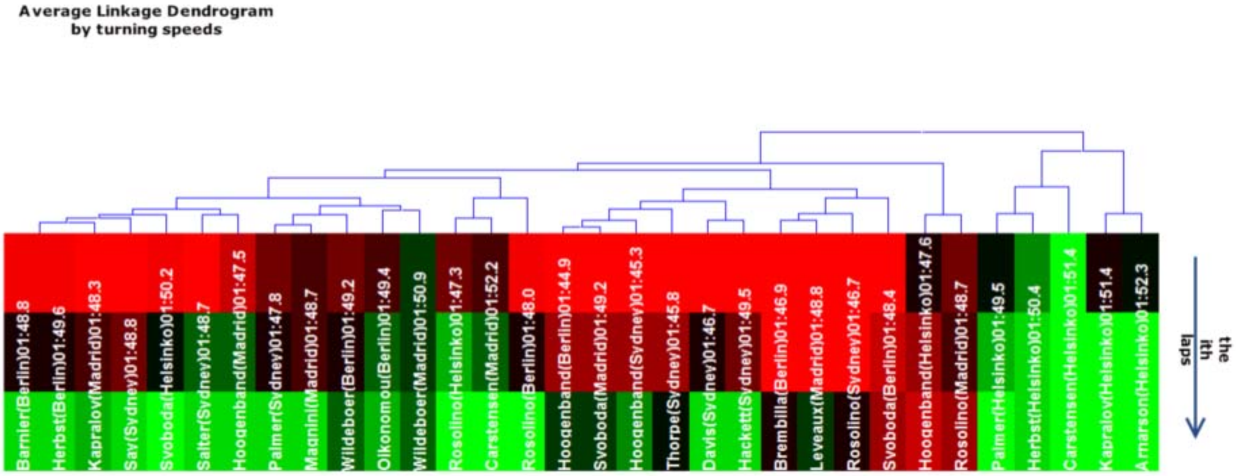

Fig. 3 Clustered display of data from the finalists' turning speeds within the men's 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid.

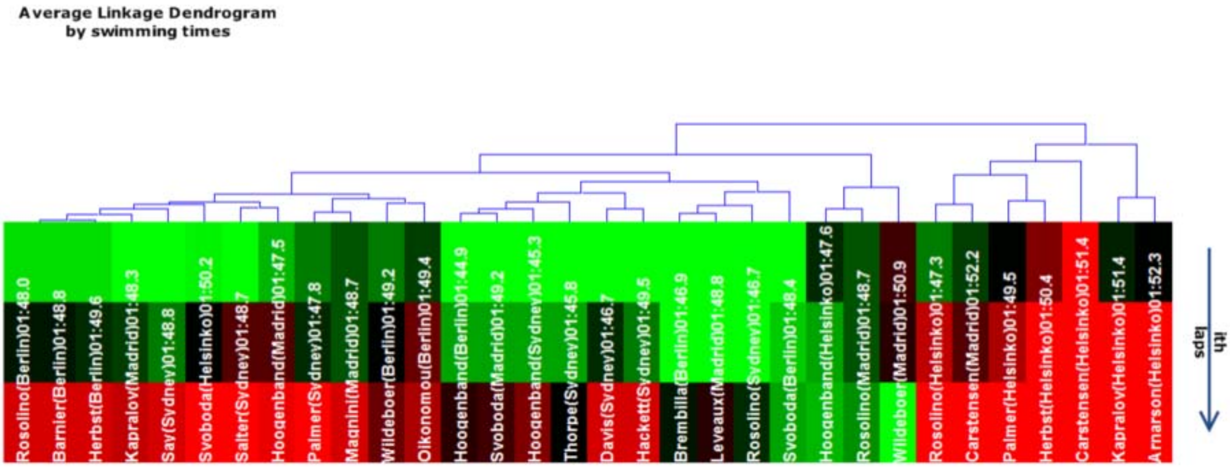

Fig. 4 Clustered display of data from the finalists' turning times within the men's 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid.

A central feature of Figs. 1-4 is that one can look at such images identifying finalists' race patterns of interest, readily zoom in on the detailed race patterns, and characterising the finalists contributing to those race patterns. Each of these tree-like dendrograms visually encapsulate the steps in the hierarchical group process in passing from each of the single-swimmer clusters to a single cluster containing all of the swimmers within the men's 200m freestyle finals from 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid. What we have found repeatedly is that swimmers represent more than one heat map next to which the grid elements are clustered or in the immediate vicinity of each other, indicating that swimmers share similar race patterns are tended to be grouped together. On the other hand, individual race patterns can also be distinguished from all others because of the substantial dissimilarities in their race patterns. Therefore, the visual representation of elite swimmers' race patterns using such images makes little difference in the observed race patterns.

We consider the three-group arrangements given by average linkage hierarchical clustering corresponding to the finalists' swimming speeds, stroke lengths, turning speeds, and turning times, respectively, within the men's 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid. Corresponding to the finalists' swimming speeds, the three-group arrangement given by average linkage hierarchical clustering is:

Cluster 1: Rosolino (01:47.3 in Helsinki);

Cluster 2: Thorpe (01:45.8 in Sydney), Hoogenband (01:47.5 in Madrid), Hoogenband (01:45.3 in Sydney), Hoogenband (01:44.9 in Berlin), Davis (01:46.7 in Sydney), Say (01:48.8 in Sydney);

Cluster 3: the other swimmers except those in Clusters 1 and 2.

Corresponding to the finalists' stroke lengths, the three-group solution is:

Cluster 1: Thorpe (01:45.8 in Sydney), Leveaus (01:48.8 in Madrid), and Herbst (01:50.4 in Helsinki);

Cluster 2: Rosolino (01:47.3 in Helsinki), Rosolino (01:47.7 in Sydney), Rosolino (01:48.0 in Berlin), Rosolino (01:48.7 in Madrid) Say (01:48.8 in Sydney), Davis (01:46.7 in Sydney), Arnarson (01:52.3 in Helsinki), Barnier (01:48.8 in Berlin), Palmer (01:49.5 in Helsinki), Palmer (01:47.8 in Sydney), and Oikonomou (01:49.4 in Berlin);

Cluster 3: the other swimmers except those in Clusters 1 and 2.

Corresponding to the finalists' race patterns, the three-group solution is:

Cluster 1: Palmer (01:49.5 in Helsinki), Herbst (01:50.4 Helsinki), Carstensen (01:51.4 in Helsinki), Kaprakov (01:51.4 in Helsinki), Arnarson (01:52.3 in Helsinki);

Cluster 2: Hoogenband (01:44.9 in Berlin), Svoboda (01:49.2 Madrid), Hoogenband (01:45.3 in

Sydney), Thorpe (01:45.8 in Sydney), Davis (01:46.7 in Sydney), Hackett (01:49.5 in Sydney), Bermbilla (01:46.9 in Berlin), Leveaus (01:48.8 in Madrid), Rosolino (01:46.7 in Sydney), Hoogenband (01:47.5 in Berlsinko), Rosolino (01:48.7 in Madrid);

Cluster 3: The other swimmers except those in Clusters 1 and 2.

Corresponding to the finalists' turning times, the three-group solution is:

Cluster 1: Kaprakov (01:51.4 in Helsinki), Arnarson (01:52.3 in Helsinki);

Cluster 2: Rosolino (01:47.3 in Helsinki), Carstensen (01:52.2 in Madrid), Palmer (01:49.5 in Helsinki), Herbst (01:50.4 in Helsino), Carstensen (01:51.4 in Helsinki);

Cluster 3: the other swimmers except those in Clusters 1 and 2.

To characterise the nature of the race patterns of interest, for example, the groups corresponding to the finalists' race patterns on swimming speeds reveal that Rosolino's swimming speeds at 2000 European Championships in Helsinki was arranged into an individual cluster containing only itself, implying the uniqueness of his swimming speeds at 2000 European Championships in Helsinki. Likewise, the groups corresponding to the finalists' race patterns on swimming speeds show that Thorpe's swimming speeds at Sydney 2000 Summer Olympics, Hoogenband's swimming speeds at Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid, respectively, Davis's swimming speeds at Sydney 2000 Summer Olympics, and Say's swimming speeds at Sydney 2000 Summer Olympics were analogous and unique when compared with the other finalists. As another dramatic example, the groups corresponding to the finalists' race patterns on stroke lengths show that Thorpe's stroke lengths at Sydney 2000 Summer Olympics, Leveaus' stroke lengths at 2004 European Championships in Madrid, and Herbst's stroke lengths at 2000 European Championships in Helsinki were arranged into an individual cluster containing themselves, pointing out that their stroke lengths have particular similarities.

In general, it is not necessarily that "best" *n*-group solution for these data in hierarchical cluster analysis [11]. One can explore relationships among swimmers on the basis of their race patterns by choosing different cut-off similarity points given by the average linkage hierarchical clustering.

4. Discussions

The assessment of elite swimmers' race patterns one by one has already given a wealth of coaching insight in the past. To establish an integrated knowledge of the elite swimmers' race patterns being studied, what we really need are innovative tools to enable coaches to survey the swimmers' race patterns from an initial state to final state through a succession of many intermediate states across various comparable criteria at the same time. To do this, a natural approach is to first scan and inspect the hidden group structures of elite swimmers' race patterns and then to address the details of interest [1]. In this paper, a different and much simpler concept is described to display the quantitative information relating to elite swimmers' race patterns across various comparable criteria at the same time. First, we represent the primary dataset relating to elite swimmers' race patterns by heat maps, which encode the magnitude of data on the strength of their race patterns using a colour scale scheme. Next, we represent the results from hierarchical cluster analysis by tree-like dendrograms, which are more intuitive to coaches in the discovery of advantageous competitive strategies. These visual displays preserve all the quantitative information relating to elite swimmers' race patterns, but convey the information to our human brains more effectively by means of a high-bandwidth channel than a "number-reading" channel [10].

We have used average linkage hierarchical cluster analysis to find dense naturalistic visual representation conveying efficiently the quantitative information relating to elite swimmers' race patterns for the men's 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid. On the basis of our observations here, it is probably that many swimmers share similar race patterns but are by no means identical, whereas a small number use individual race patterns. While not based on any presupposed factor to becoming the winner, the similarity of race patterns may be the earliest available information in the identification of insightful mechanisms into the elite swimmers' race patterns being studied.

What we have found to be the most valuable feature of the approach described here is that it allows the recognition of elite swimmers' race patterns within an individual event across multiple swimming championships at the same time, giving an analysis of the factors we can examine quantitatively and graphically, which may not be viewed easily otherwise. The above example has illustrated a feature of the quantitative information relating to elite swimmers' race patterns within an individual event from multiple swimming championships, namely, the tendency to arrange elite swimmers with respect to their order intrinsic on race results across various comparable criteria at any one time. It is, of course, not very surprising that elite swimmers that are grouped together share similar race patterns within an individual event from multiple swimming championships. The success of these natural approaches has given us confidence in the identification of hidden group structures with respect to elite swimmers race results in a natural intuitive manner.

Finally, we have seen that the approach described here can supply general tools to coaches in training promising elite swimmers, and that is has made possible a revolution in the face of swimming coaching.

5. Supplementary Dataset

The combined longitudinal dataset relating to the finalists' race patterns on swimming speeds, stroke lengths, turning speeds, and turning times, respectively, within the men's 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid.

6. References

[1] I. Chen, H. Homma, C. JIN, H. YAN. Identification of elite swimmers' race patterns using cluster analysis. *Journal of Sport and Coaching Science*, 1997, to appear.

[2] J. Cossor, B. Mason. Swim start performance at Sydney 2000 Summer Olympics, http: //coachesinfo.com/category/swimming/143

[3] J. Cossor, B. Mason. Swim turn performance at Sydney 2000 Summer Olympics, http: //coachesinfo.com/category/swimming/144

[4] R. Haljand. The new scientist way to analyze swimming technique models of swimming technique. *How to Develop Olympic Level Swimmers.* Helsinki, 1981

[5] R. Haljand. Competition analysis in swimming. *The World of swimming*. febr Nr4 FINA, 1993

[6] R. Haljand. Swimming technique aspects from the coach view, *FINA Medical Congress*. Goeteborg. pp. 28-29, 1997

[7] D. Daily, and Y. Vanlandewijck. Some criteria for evaluating swimming classification. *Adapted Physical Activity Quarterly*. pp. 271-289, 1999

[8] D. J. Daily. Swimming, impairment and classification. *Adapted Physical Activity Quarterly*. pp.251-270, 2005

[9] S. K. Wu, and T. Williams, Swimming, impairment and classification. *Adapted Physical Activity Quarterly*. pp. 251-570, 1999

[10] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Postein, Cluster analysis and display of genome-wide expression patterns. *Processing to National Academic of Science*. pp.14863-14868, 1998

[11] B. S. Everitt. *Statistical analysis using S-Plus*, Chapman and Hall, 1994

[12] J. A. Hartigan. *Clustering algorithm*. John Wiley and Sons, 1975

[13] A. D. Gordon. *Classification: methods for the exploratory analysis of multivariate data*, Chapman and Hall, 1981

[14] J. Seo, and B. Shneideman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projection. *Proc. IEEE Info Vis.* pp. 65-72, 2004

[15] Hierarchical Clustering Explorer 3.0, http://www.cs.umd.edu/hcil/hce/hce3.html