1 Methodology

What follows in this Chapter, is a description of the Empathic Visualisation Algorithm (EVA). The theory behind EVA is presented formally, the method being based on Slater [1] that automatically maps data to visual structures. A statement of the research problem will be presented initially.

1.1 Statement of the problem

As stated previously, the problem presented here is that of visualising multi-dimensional data sets. The overall objective is to construct a visualisation such that the salient features of the data can intuitively be recognised by an observer and where the representation gives a holistic view of the data set. In other words, it can be described as a technique for the visualisation of complex data in a naturalistic form.

Complex data in the sense that it is presented here, is data that is relatively large both in terms of the amount of data present, and the number of variables (dimensionality) that the data encompass. The variables themselves are usually correlated and hence cannot be treated separately.

The most interesting data are multi-dimensional. Multiple dimensions refer to the difficult problem of information visualisation where data tables (matrices of data with cases as rows and attributes as columns) have so many variables that an Orthogonal Visual Structure (such as a graph) is not sufficient. They have too many variables to be directly encoded using 1, 2 or 3d dimensional structures. For these kind of data, graphs and charts lose their effectiveness. Numerous techniques have been described in the literature that attempt to visualise such data with both advantages and disadvantages. However, no method claims to achieve the overall objective of this study.

An example of such multi-dimensional data set is that of accounting (financial) data. The data shown in Figure 1 represent a Balance sheet and profit and loss accounts for a single company over a period of 5 years. Apart from the fact that there is a great amount of data, the data components are correlated and their values (or range) affect each other with respect to the decision analysis process. The assessment depends on the simultaneous effect of several of these variables in different spheres of activity. One, must also take into consideration the fact that this is data from a single company. Imagine having hundreds of companies. How do you visualise such data? How do you get a better understanding of your data set? The prospect of visualising and understanding data, therefore from multiple companies is,

in the least, a daunting task.

Balance Sheet					
Capital & Reserves					
ORDINARY SHARE CAPITAL	131000	131000	132000	141000	141000
SHARE PREMIUM A/C	836000	840000	856000	1441000	1460000
OTHER RESERVES	39000	104000	163000	230000	308000
PROFIT & LOSS A/C	2380000	2629000	2690000	2381000	2644000
EQUITY CAP. AND RESERVES	3386000	3704000	3841000	4193000	4553000
PREFERENCE CAPITAL	0	0	100000	200000	325000
TOT. SHARE CAPITAL &RESERVES	3386000	3704000	3941000	4393000	4878000
Fixed Assets					
INTAGIBLE					
TANGIBLE	509000	534000	585000	715000	698000
INVESTMENTS	25312000	32374000	35298000	39782000	46558000
OTHER	1415000	1433000	2074000	2038000	2908000
	27236000	34341000	37957000	42535000	50164000
Current Assets					
STOCKS	0	0	0	26000	14000
DEBTORS	50857992	53668000	57512992	71361992	86548992
INVESTMENTS	0	0	0	0	0
OTHER	1080000	1620000	1753000	1880000	2023000
CASH	757000	598000	391000	339000	1957000
	52694992	55886000	59656992	73606992	90542992
Current Liabilities					
PROVISION FOR TAX	442000	498000	496000	299000	266000
PROVISION FOR DIVIDENDS	129000	158000	191000	245000	290000
CREDITORS <1 YEAR	55757992	60526000	64129992	72659000	91462992
OTHER	11373000	11372000	15334000	22796000	25594000
	67701992	72554000	80150992	95999000	1.18E+08
Net Current Assets	-15007000	-16668000	-20494000	-22392008	-27070000
Total Asset Less Current Liabilities	12229000	17673000	17463000	20142992	23094000
Long Term Liabilities					
PROVISIONS	296000	399000	630000	970000	1144000
LOAN CAPITAL	8547000	13570000	12892000	14780000	17072000
OTHER					
	8843000	13969000	13522000	15750000	18216000
	3386000	3704000	3941000	4392992	4878000
Profit & Loss					
OPERATING PROFIT-ADJ	-446000	-504000	-1169000	-611000	-394000
TOTAL NON-OPERATING INCOME	1247000	1514000	2337000	2019000	2011000
TOTAL INTEREST CHARGES	66000	80000	134000	167000	193000
PROFIT BEFORE TAX	735000	930000	1034000	1241000	1424000
TAX	310000	319000	337000	403000	480000
PROFIT AFTER TAX	425000	611000	697000	838000	944000
ORDINARY DIVIDENDS	184000	233000	288000	360000	434000
TO SHAREHOLDERS FUNDS	241000	378000	409000	478000	510000

Figure 1: Sample of data for a single company,

In situations such as the above, information visualisation helps in "understanding" a set of data (which may be dynamically unfolding in time) by allowing users to visualise representations of the data, thus using vision to build "understanding", and allowing the formation of hypothesis for later statistical analysis.

There are two main problems involved in this, the first being, the choice of a suitable mapping from the data to the chosen visual representation (structure) and disinguishing between arbitrary and automatic mapping. The possibility of taking into account the user's emotions during visualisation in an automatic mapping, is the focal point of this method. The second

problem highlighted is the choice of a suitable paradigm for the representation of the data, whether abstract (e.g. using colour) or more realistic. Different representations may lead to quite different understandings. The possibilities of naturalistic representations, something encountered in everyday life, is also addressed in the present study.

Figure 2 shows the placement of *"Empathetic Visualisation Algorithm (EVA)"* in the information visualisation classification presented in the literature.

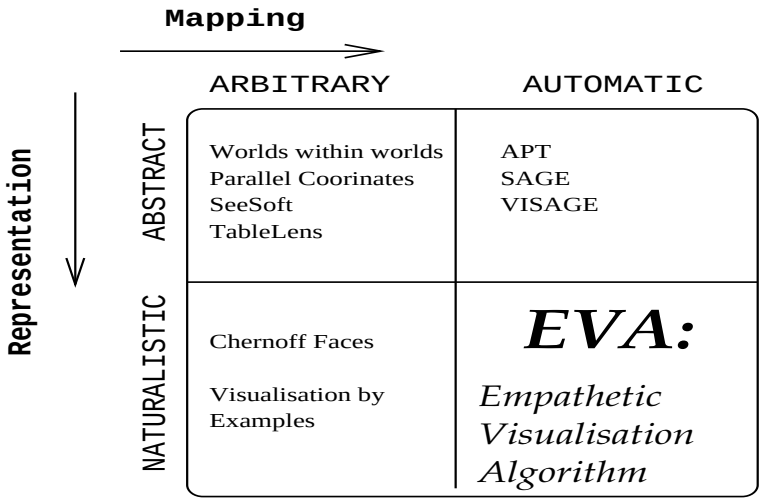

Figure 2: Classification of EVA,

The specific aim of this study is the construction of a system for such a representation, and then to test this in an experimental setting. The system should be such, that it can be used with as many different data sets and visual structures as possible - i.e. that is, a generic system rather than one tied to a particular form of data or visual structure.

1.2 Fundamentals

Throughout this thesis the representation of multivariate data in an $n \times k$ data matrix X, consisting of n cases on k "quantifiable" variables $x_1, x_2, ..., x_k$ is presented. Each row in the data matrix typically represents an individual and there are k observations made for each individual. As mentioned above, the objective is to construct a visualisation of the data matrix such that the salient features of the data can intuitively be recognised by an observer and the representation gives an overall view of the data set. Within this overall

objective there are two further fundamental objectives:

1. Naturalistic visual representation. It should be something encountered in everyday life, something that does not require special knowledge for interpretation by a normal human observer.

2. Automatic mapping. The mapping should be that semantically "important" features in the data are mapped to "important" features of the visual structure that are significant to human perception or emotion.

Examples of (1) include faces, buildings, body posture, scenery and other. No human needs to be an expert to recognise the emotional content of another human face - it is recognisably "happy", "sad", "angry", "relaxed", "scared", "neutral" in addition various combinations of these basic emotions. Frequently used throughout this thesis, is an example of a "face" because it is the epitome of a naturalistic visual structure in the sense that is presented here.

Given a set of data and a visual structure, it is trivially easy to construct a mapping from the data to a visual structure - for example map variable x_i to the ith facial feature like in Chernoff faces described in the literature. However, such a mapping is arbitrary - it does not take into account the impact of the face on the emotions of the observer. The data is of interest to the observer for some reason; associated with the data is some "value system" reflecting the interest, importance or consequences of aspects of the data for the situation of the observer. A fundamental goal, reflected in (2), is that the perceptually or emotionally significant features of the visual structure directly reflect the value system over the data. This is termed *visual homomorphism*. Hence the mapping from data to visual structure cannot be arbitrary, but must be constructed in such a way that this visual homomorphism is realised.

What follows is an explanation of how such a mapping can be constructed. Two different types of visualisation problem are considered: the first is the representation of the data matrix as a whole - one visual structure representing the entire data matrix, i.e. a face. The second, is where each row (i.e., individual) in the data is to be represented by a different instance of the same type of visual structure - so using faces again, each row is mapped to a different face. In the first case, the problem is to capture overall features of the same data set. In the second case the problem is to examine the differences in the individuals, or to examine one individual changing through time. In fact the method is very similar in both cases

and does not affect the computation. The difference is on the focus and interpretation. The overall representation is first considered.

1.3 Assumptions and Notation

Let $\nu_s(X)$, $s = 1, 2, ..., p$ be p functions over the data representing "values" over the data matrix.

Consider the example where the data matrix represents a set of customers of a telephone company, and the variables are quantities such as age, gender, marital status, income, number of years with the company, number of telephone calls made per week, monthly phone bill, and so on. One value might be a function of the overall age distribution of the population - such as the mean age, the percentage over 65, or the percentage under 20. Another value might be the "flatness"of the data - for example the ratio of the variance of the first principal component to the total variation in the data. Another value might be the quality of service given by a telephone company and so on, and many other quantities that characterise the interests or "value system" of the observer.

Consider a visual structure (Ω). Similarly there are p importance aspects, characteristics, of Ω that are measurable and significant to human perception or emotions, $e_s(\Omega)$, $s = 1, 2, ..., p$. In the example of the face these might be the degree of anger, happiness, boredom, fear - or characteristics such as age, beauty, gender and so on.

The fundamental goal, in terms of any X and any Ω, is to produce a mapping $\mu(X) \longrightarrow \Omega$ such that 'values' over the data matrix are reflected in characteristics of the visual structure. In particular that $e_s(\Omega)$ is a monotonically increasing function of $\nu_s(X)$ for each $s, s = 1, 2, ..., p$. For example, an increase in profitability of a company should result in an increase of happiness of the corresponding visual structure.

A *characteristic* is a measurement of some aspect of Ω as a whole (such as the emotions on a face) rather than some individual *feature* (such as the shape of the mouth). It is some measure representing the totality of the face i.e. the degree of "happiness". The appearance of "happiness" depends on many different individual features of the face - specific configurations of muscle tensions, for example. Similarly, the appearance of beauty, age or gender is derived from many different features - such as size of the eyes, inter-ocular distance, shape of the mouth, symmetry, and so on. In other words, features are the individual components that make up a face - such as the specific configuration of muscle tensions for a specific face, or the geometric and material properties of the actual features (eyes, colour of the

eye, mouth, nose, lips) that make up a face. Knowing its features enable us to *render* a face. Once rendered the face will have a set of measurable "characteristics" (qualities).

Suppose that there are r features of the visual structure: $\phi_t(\Omega)$, $t = 1, 2, ..., r$. Knowing these features Ω can be rendered. Once rendered, we can measure it to determine its characteristics $e_s(\Omega)$, $s = 1, 2, ..., p$.

Finally, we introduce *feature functions* to the data matrix: $f_t(X)$, $t = 1, 2, ..., r$. These functions completely determine the features of the visual structure, in fact

$$\phi_t(\Omega) = f_t(X),\ t = 1, 2, ..., r \tag{1}$$

The values of these functions are *interpreted* as the values of the features of the visual structure. The aim is to choose these functions f in order to attain the required correspondence (association) between the value system over the data and the characteristics of the visual structure and therefore, to attain the visual homomorphism.

Let $\nu = (\nu_1, \nu_2, ..., \nu_p)$ and $e = (e_1, e_2, ..., e_p)$. Suppose $\parallel \nu - e \parallel$ is a measure of the 'distance' between these two *vectors*. Then the specific goal is to choose $f_t(X)$, $t = 1, 2, ..., r$ such that $\parallel \nu - e \parallel$ is minimised. If this is achieved it means that the characteristics of the visual structure produced by the *feature function* best represents the "value system" of the data set.

1.3.1 Using Genetic Progamming (GP)

The minimisation problem introduced above can be tackled with a Genetic Program (GP). Let $F_i = (f_{1i}, f_{2i}, ..., f_{ri})$ be a specific set of feature functions which, when applied to X, produce the visual structure features. A large collection of such sets of functions F_i, $i = 1, 2..., N$ is chosen at random. This collection defines a *population* of sets of feature functions. The ith member of the population produces a specific visual structure Ω_i. This visual structure has characteristics $e_s(\Omega_i)$, $s = 1, 2, ..., p$. These characteristics can be used to produce the distance measurement $\parallel \nu_i - e_i \parallel$. The distance measurement can be used to compute a "fitness" for the ith member of the population. Hence, each member of the population has an associated fitness, which can be expressed as a probability. An example might be to set minimum fitness of population to 1, maximum fitness to 100 and interpolate for values in between. These probabilities determine survival into the next generation and selection for mating - thus producing a second generation. The process continues until (possible) convergence. The most fit member of all populations is chosen for the required mapping.

It is expected that at each successive generation, the average fitness would increase, until a generation is reached such that subsequent iterations produce only negligible increments in fitness. This is interpreted as convergence to a local minimum (solution). Different runs of the Genetic Program result to different solutions due to randomness of the technique and absence of an optimal solution. The most fit member of each population is chosen for the required mapping.

Visualising Individuals in a Data Matrix The method above produces a visualisation for an entire data matrix. Instead the observer may be interested in visualising each individual(row) of the data matrix - in order to look for "special" individuals (eg. companies to add to his investment portfolio), or where the rows represent the changing of one individual through time, representing an evolving situation (showing the changes in company shares during time).

The method is fundamentally unchanged, and only involves a reconsideration of the domain of the feature functions f. Previously, the domain of these functions was the whole of the data matrix X. Instead, is now the domain over the variables represented by the columns of the data matrix. So each individual row of the data produces a set of feature function values, which therefore determines (renders) a visual structure for each row.

Similarly, the domain of the value functions ν is restricted also to the variables represented by the columns. Hence each row now produces a distance measure - that between the characteristics of the visual structure for that row, and the values over the data for the row. The overall distance for the ith member of the population of sets of feature functions can therefore be taken as a combination (e.g., a sum) of the distances over all rows of the data. The method then proceeds as before.

1.4 Overview

A summary of the steps required in order to implement the method is described below, using the notation above. Figure 3 gives a graphical visualisation of the overview of the method.

1. Decide on Ω, the visual structure. Determine the number of features Ω has. Assuming there are r, $f_t(X)$, $t = 1, 2..., r$ feature functions are required in order to render an individual.

2. The user (or observer), identifies the p important values of the data set $\nu_s(X)$, $s = 1, 2, ..., p$.

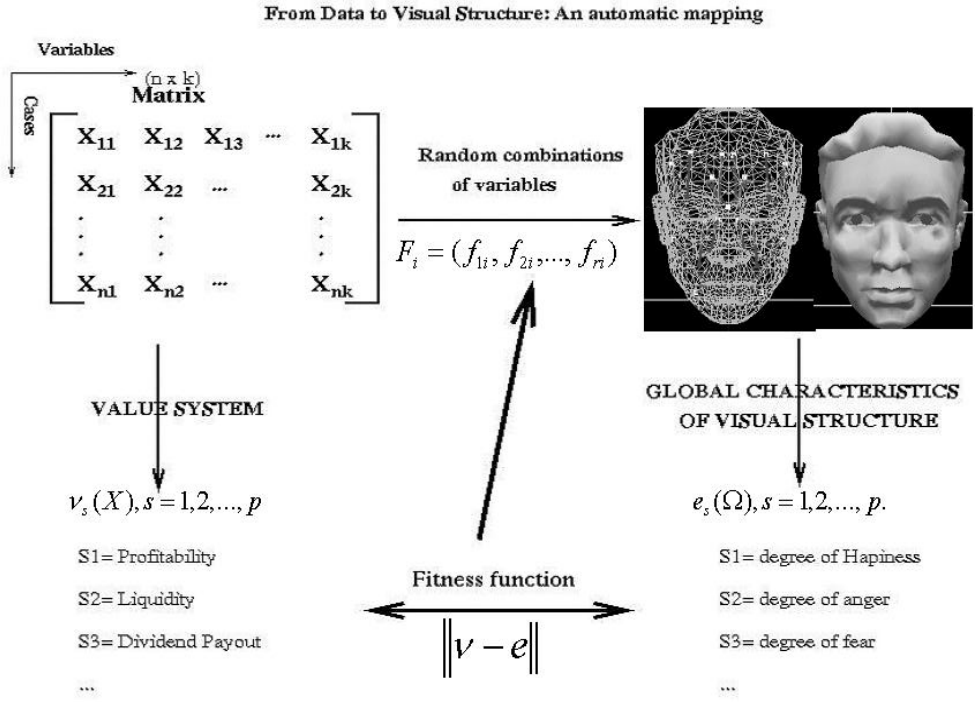

Figure 3: Overview of the method.

3. Identify $e_s(\Omega)$, $s = 1, 2, ..., p$ the p characteristics of the visual structure that measure its totality and are significant to human emotions and perception. This can be selected by the user or automatically selected by the system.

4. Identify the fitness function i.e. the function we are trying to minimise, which is, for example:

$$\sum_{d=1}^{n} \sum_{s=1}^{p} (\nu_{ds} - e_{ds})^2 \tag{2}$$

5. Identify the GP parameters and run the GP.

The steps shown above achieve the visual homomorphism described earlier. In other words, it can be defined as the 'extraction and visualisation of qualitative data from quantitative one'.

1.5 Discussion

This section has presented a method for automatic determination of a mapping from data to visual structure. It requires a user (someone interested in the data) to construct a set of value functions of interest over the data. The designer of the visualisation must decide on a type of visual structure, and a set of perceptually or emotionally significant characteristics of this visual structure that matches the number of value functions inserted by the user. The visual structure must be determined by a set of quantifiable features. A GP is then used to construct the mapping that minimises some given measure of distance between what the user inserted as value functions over the data and what the designer provided as characteristics of the visual structure. The minimisation criterion is the only factor that specifies the exact nature of the mapping.

For the method to work it is assumed that the GP will converge. In such a case we can make the hypothesis that the observer will pick up out important features of the data from this mapping. We can later test the validity of this hypothesis.

References

[1] M.Slater. From data to visual structure: An automatic mapping. Research note, Department of Computer Science, University College London, 1999.