Vincent Mom – Hans Paijmans

SECANTO: A Retrieval System and Classification Tool for Simple Artefacts

Abstract: Secanto is a computer program that compares the shapes of artefacts like vessels, arrow points and axes by calculating dissimilarities. This comparison of objects leads into two interesting application areas: retrieval systems to obtain look-alikes of a specific object, and typologies of objects. However, just like human beings, automated classification systems are not perfect, but the type of mistakes they make are quite different, which may be one of the reasons that automatic classification systems are not so popular among archaeologists.

Introduction

About 30 years ago, Wilcock and Shennan used a computational method to try to establish a typology for Bell Beaker pottery (Wilcock / Shennan 1975). This method was called the "slice method" as the shapes of the objects under study were divided in horizontal "slices" before comparison took place *(Fig. 1)*.

But even according to the authors themselves, the method was not very successful: a simpler method, comparing height/width ratios, gave better results and the method dropped from sight. It may well be that part of the unsatisfactory performance was caused by the fact that computing at the time was expensive and that interfacing with the program was complicated, which discouraged experimenting with alternative settings and data.

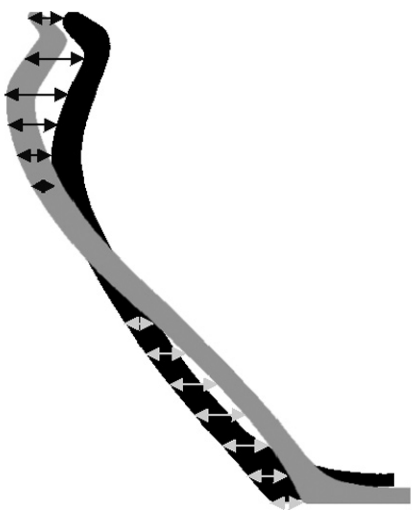

Fig. 1. A sliced profile. For each slice the distance between the profiles is measured. The sum of squared distances is the dissimilarity.

The slice method is basically a simple algorithm to calculate objectively and reproducibly the (dis)similarity between two two-dimensional shapes ("profiles"). This means that usually two objects are required to employ the method. However, if an object does not have exact rotational symmetry (e.g. hand-shaped vessels, arrow points, flint axes, etc.), one may obtain several different shapes by drawing a single object from different angles, in which case the calculation will produce a measure for the internal symmetry of the object (see *Fig. 2* for an example).

Also, the slice method is size invariant: the images of the objects are normalized to a common height in pixels. This means that two objects that have the same shape, but different sizes, are considered equal.

The dissimilarity between two shapes can be interpreted as a *distance* (see *Fig. 3* for an example). The greater the distance between two shapes, the less they look alike. If the distance is zero the shapes are exactly alike, which does not mean, however, that the *objects* are alike: they may differ in material, decoration, size, and so on. So shape is only one aspect in the process of comparing objects and other aspects may be at least as important, especially when the objects are being fit into typologies.

Of course the slice method is by no means the only method to compare objects. One only needs to do a quick scan of the CAA proceedings to obtain an impressive list of relevant publications. This paper, being a description of two application areas of a specific computer program, is not the place to go further onto these sidetracks.

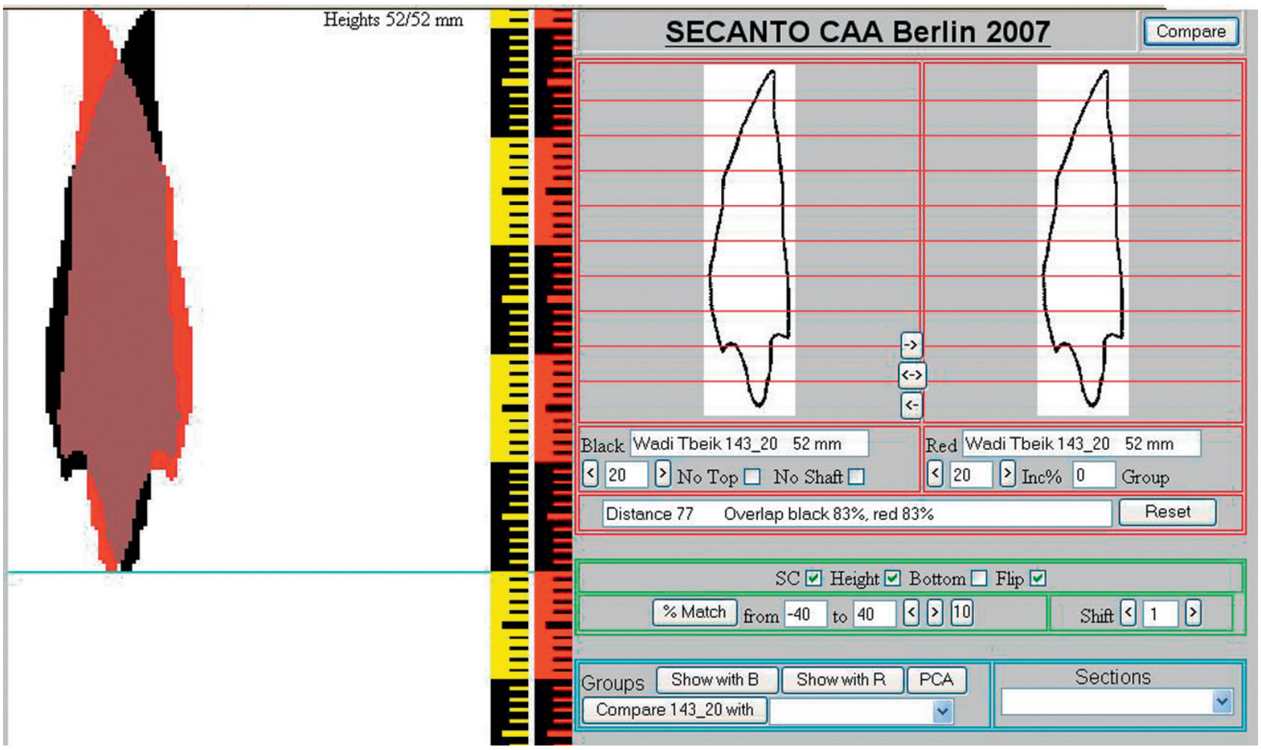

Fig. 2. A mirrored asymmetric profile is compared with itself.

Secanto as a Retrieval System

The first version of Secanto was a retrieval system based on two algorithms: the slice method and an "area-fitting" method (Mom 2005). The slice method proved very valuable for vessels but also for objects like arrow points and axes. The "area-fitting" method did not give good results for vessel profiles (as the influence of the thickness of the profiles interfered with the retrieval process) but gave good results for the "solid" objects (although not substantially better than the slice method). The initial data base of the Secanto system consisted of about 800 profiles of Iron Age handshaped vessels from the

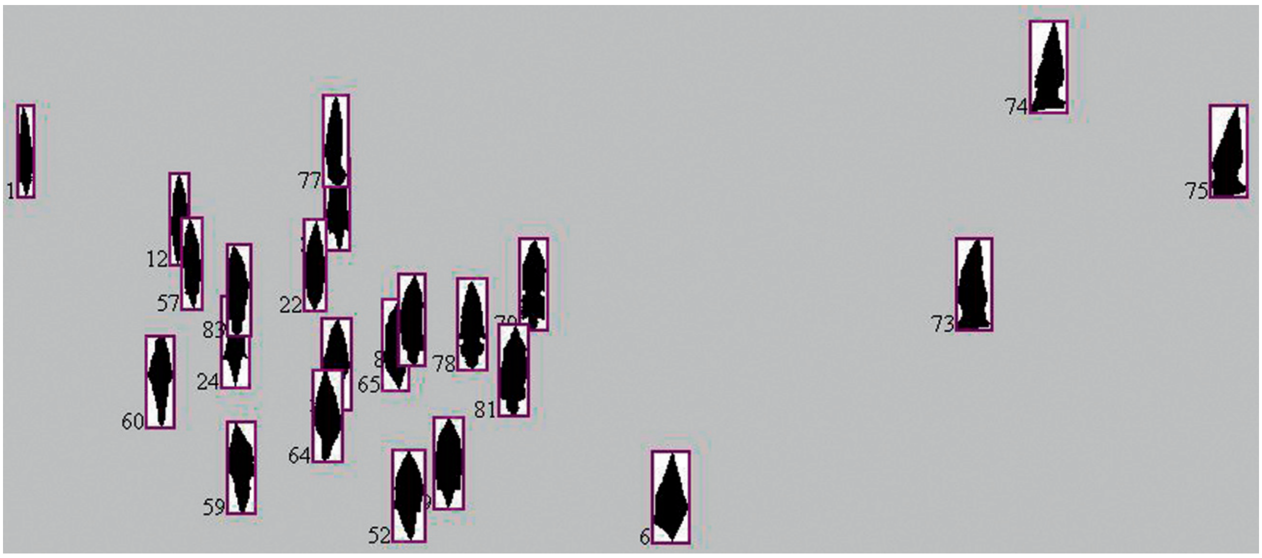

Fig. 3. A set of arrowpoints. The 2D distances between the profiles are optimized to reflect the calculated dissimilarities as good as possible.

north-western part of Europe. Using a point-and-click mechanism the users could enter their own profile and compare it with the vessels in the data base. All vessels with a dissimilarity value below a certain value would be retrieved and presented as “look-alikes” and it was up to the user to pick out the “best fit”, which often was not the first in line (i.e. the object that fitted most closely the original profile), but one of the other images in the list that were more like the object under study regarding features like size, decoration and coarseness.

In order to judge the strength of a retrieval system, two parameters are of interest: “recall” and “precision”. These measures are well known and are defined as follows:

	Relevant	Not Relevant
Retrieved	A	B
Not Retrieved	X	Y

Tab. 1. The results of a retrieval action.

The recall then is A/(A+X) while the precision is defined as A/(A+B). An ideal retrieval system would give results with both recall = 1 and precision = 1. This, however, is a goal that is seldom reached. There are several reasons for this, but perhaps the most important one is that “relevance” and “similarity” in real life are no binary properties. For this reason the so-called vector space model was introduced (Salton / McGill 1983) that allowed retrieved objects to be ranked to estimated relevance. For the Iron Age vessel data base, then, the results are good for “ordinary” shapes. However, if the height/width ratio of a shape is rather high or low, then to obtain a satisfactory recall it is required to retrieve more objects, automatically decreasing the precision as more “not relevant” objects are retrieved. With modern computing techniques, it is possible to experiment with the settings of the algorithm, which adds an extra dimension to the retrieval activity.

Secanto as a Classification Tool

Automated image recognition systems often give results that do not agree with (human) classifications. However, if a human assigns an object to a different class than an original, authoritative classification system that serves as a reference, it is generally clear (to the authoritative humans, at least) where the assignment process went wrong. “Misclassifications” by the computer, on the other hand, are as often as not opaque, depending on the algorithm that was used and the degree of feedback to the user. Related to this aspect is the question of what happens if the typology of the objects is partly based on non-visual properties, which in fact is usually the case. A computer program that uses image recognition cannot take such non-visual properties in account. The human classifier has no choice either, but in the opposite direction: they cannot make decisions on the visual data exclusively, as they always carry the burden of past history, associations and connotations. The problem with image recognition, as with every form of retrieval, is that such automated techniques therefore may have comparable performance in terms of precision and recall, but that they may return very different sets of retrieved objects. It is possible to create typologies that are based on such computerized recognition methods by the application of automatic clustering, but again the resulting typologies are very much dependent on the particular technique.

To investigate the strength of Secanto as a classification tool, our first task was to establish how many of the objects were assigned to their correct class (van der Maaten et al. 2006; Belongie / Malik / Puzicha 2001). The collection of objects used for this experiment was a set of medieval glass objects which is preserved by the Netherlands National Service for Archaeological Heritage. One of the reasons to choose this particular collection was that a human expert was available (J. Kottman) who provided us with the standard human classification given in *Tab. 2*.

The baseline (i.e. the number of correct classifications if the objects are selected at random) for this particular collection of objects is 80 or almost 25%. This rather high value is caused by the large number of beakers, goblets and bottles. If you would select an object “in the dark”, you will have a much bigger chance of taking a beaker, goblet or a bottle, but this does not mean that you have a “special ability” to recognize such shapes in the dark. The “sliced” method implemented in Secanto classified not 80, but 250 of the 311 objects correctly (80%).

As we already mentioned, the performance of a system in terms of correct classifications is only one facet of its potential in a working environment. The subjective “feel” of its performance is also an important factor for its acceptance, and (we assume) part of this is defined by the question whether misclassifications by the system are similar to those that the human would make. Therefore we asked Kott-

Class	No of Objects	Class	No of Objects
berkemeijer	1	flute	2
bird seed-dish	1	lamp glass	2
bowl	1	Maigelein	2
butter dish	1	pedestal disc	2
carafe	1	rod	2
dispensing pot	1	stopper	2
Keulen-glass	1	vase	2
lens	1	Stangenglas	3
mortar		Krautstrunk	5
show object	1	lid	6
spectacle glass	1	Römer	7
tazza	1	salt cellar	12
urinal	1	beaker	74
		goblet	79
		bottle	98

Tab. 2. A collection of 311 glass objects subdivided into classes with resp. names and numbers.

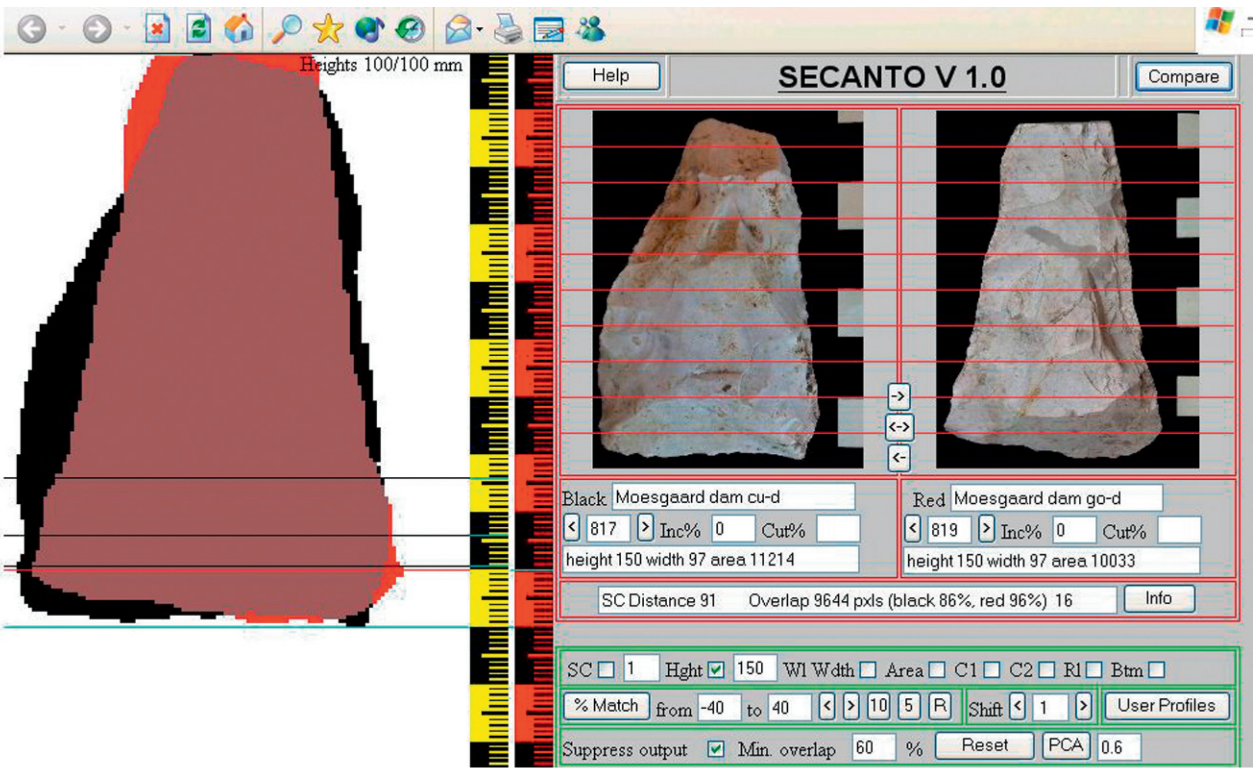

Fig. 4. Comparison of two Mesolithic flake axes.

man to indicate which class was considered nearest to every class in the system. This was taken as the human bias for misclassifications. Then we analysed the misclassifications of the Secanto results. In the 28 classes, the misclassifications agreed only once with the human bias. It must be stressed that the number of objects in this experiment was very small, so the results are only an indication, but nevertheless the outcome suggests that the algorithm has not much in common with human perception, and that computerized classification on visual data only is no "drop-in" for human typologies.

Current Status and Further Developments

The later versions of Secanto have been improved in several respects. The rather awkward two-step method to search the data base (which was caused by the fact that one comparison took about 3 seconds, so a full data base scan would take 40 minutes) was replaced by a faster implementation written in C++ which could complete the scan within about two minutes. Also, the current version uses an .xml file to keep the data separate from the application.

In these versions, however, the user still must convert image files into ASCII files with hexadecimal data, which remains a tedious job. In one of the newer versions, it is possible to enter .gif files with profiles directly. Secanto uses an image processing library which converts these gif files into XY-coordinates which are used in the dissimilarity calculations. The problem that arises here is the quality of the .gif files: small spots, hardly visible, may have disastrous effects on the coordinate sets which results in unpredictable results when calculating dissimilarities.

In addition, specific versions of Secanto are being developed: an example is the version for arrow points and stone axes (Mom 2007). This version has optimizing functions to compare objects that have been damaged and is currently being used to investigate a set of about 600 Mesolithic flake axes from Denmark *(Fig. 4)*.

References

Belongie / Malik / Puzicha 2001

S. Belongie / J. Malik / J. Puzicha, Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 2002, 509–522.

van der Maaten et al. 2006

L. P. J. Van der Maaten / P. J. Boon / A. G. Lange / J. J. Paijmans / E. O. Postma, Computer Vision and Machine Learning for Archaeology. In: J. T. Clark /E. M. Hagemeister (eds), Digital Discovery. Exploring New Frontiers in Human Heritage. CAA2006. Computer Applications and Quantitative Methods in Archaeology. (Budapest 2007) 361–367.

Mom 2005

V. Mom, Secanto: The Section Analysis Tool. In: A. Figueiredo / G. Leite Velho (eds.), The world is in your eyes. CAA. 2005. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the XXXIII Computer Applications in Archaeology Conference, Tomar, Portugal, March 2005 (Tomar 2007) 95–101.

Mom 2007

V. Mom, Where did I see you before … A holistic method to compare and find archaeological artefacts. In: R. Decker / H.-J. Lenz (eds.), Advances in Data Analysis. Studies in Classification, Data Analysis and Knowledge Organization (Berlin 2007) 671–680.

Salton / McGill 1983

G. Salton / M. J. McGill, Introduction to Modern Information Retrieval (New York 1983).

Wilcock / Shennan 1975

J. D. Wilcock / S. J. Shennan, Shape and style variation in Central German Bell Beaker: A computer-assisted study. Science and Archaeology 15, 1975, 17–31.

Vincent Mom

DPP Foundation
Oranjelaan 284
3312GN Dordrecht
The Netherlands
v.mom@wxs.nl

Hans Paijmans

University of Tilburg
Department of Computational Linguistics
PO Box 90153
5000 LE Tilburg
The Netherlands
j.j.paijmans@uvt.nl