Grapheme based Retrieval and Analysis for PaleograpHic Expertise of Medieval manuscripts

Graphem

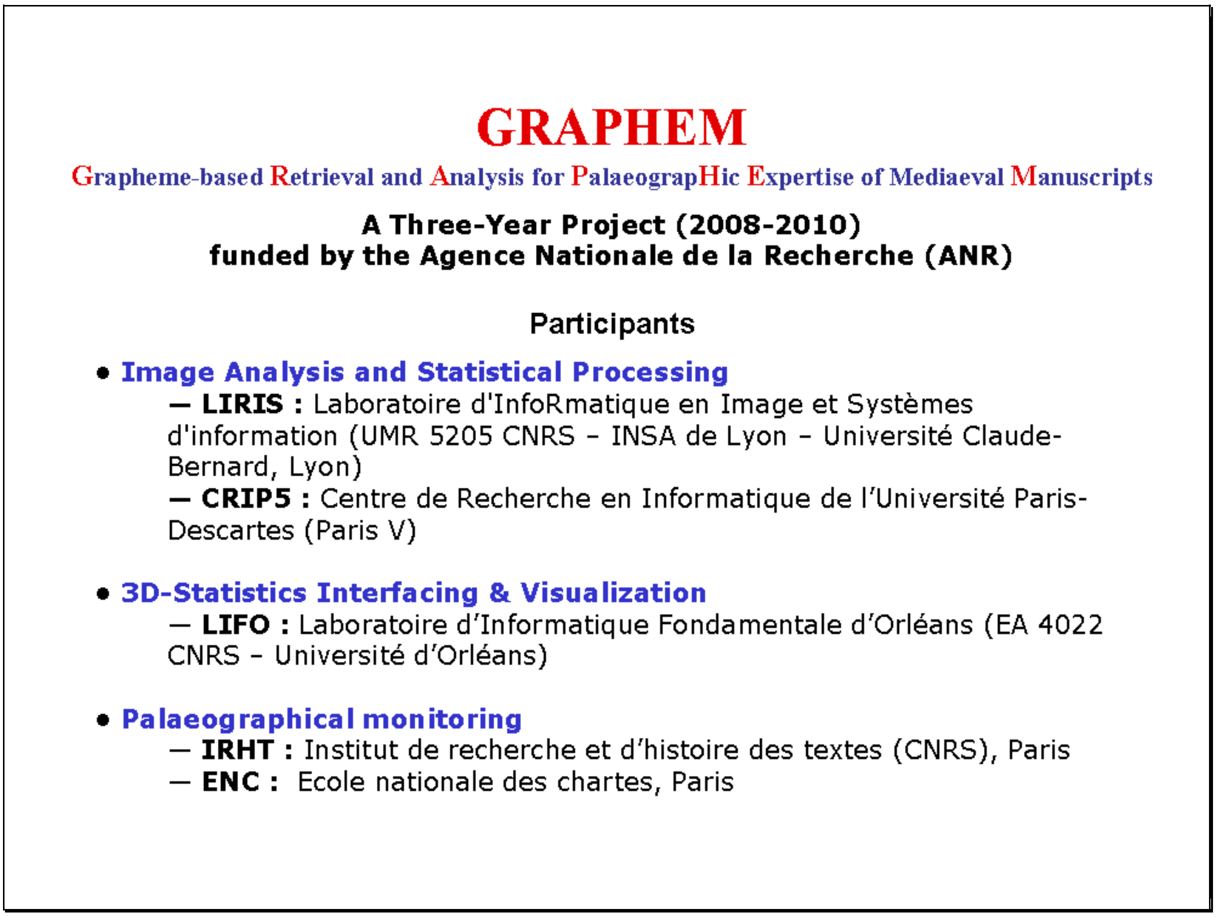
http://liris.cnrs.fr/graphem/

ANR AGENCE NATIONALE DE LA RECHERCHE
cnrs
LIRIS
irht
CRIP 5
LABORATOIRE D'INFORMATIQUE FONDAMENTALE D'ORLEANS LIFO

Graphem For Dummies

Denis Muzerelle, IRHT (CNRS), Paris

International Workshop
The Manuscript Triangle France-England-Scandinavia
1100-1300
University of Bergen, 4-6 November 2009



GRAPHEM

Grapheme-based Retrieval and Analysis for PalaeograpHic Expertise of Mediaeval Manuscripts

A Three-Year Project (2008-2010)
funded by the Agence Nationale de la Recherche (ANR)

Participants

- **Image Analysis and Statistical Processing**
  - **LIRIS :** Laboratoire d'InfoRmatique en Image et Systèmes d'information (UMR 5205 CNRS - INSA de Lyon - Université Claude-Bernard, Lyon)
  - **CRIP5 :** Centre de Recherche en Informatique de l'Université Paris-Descartes (Paris V)

- **3D-Statistics Interfacing & Visualization**
  - **LIFO :** Laboratoire d'Informatique Fondamentale d'Orléans (EA 4022 CNRS - Université d'Orléans)

- **Palaeographical monitoring**
  - **IRHT :** Institut de recherche et d'histoire des textes (CNRS), Paris
  - **ENC :** Ecole nationale des chartes, Paris

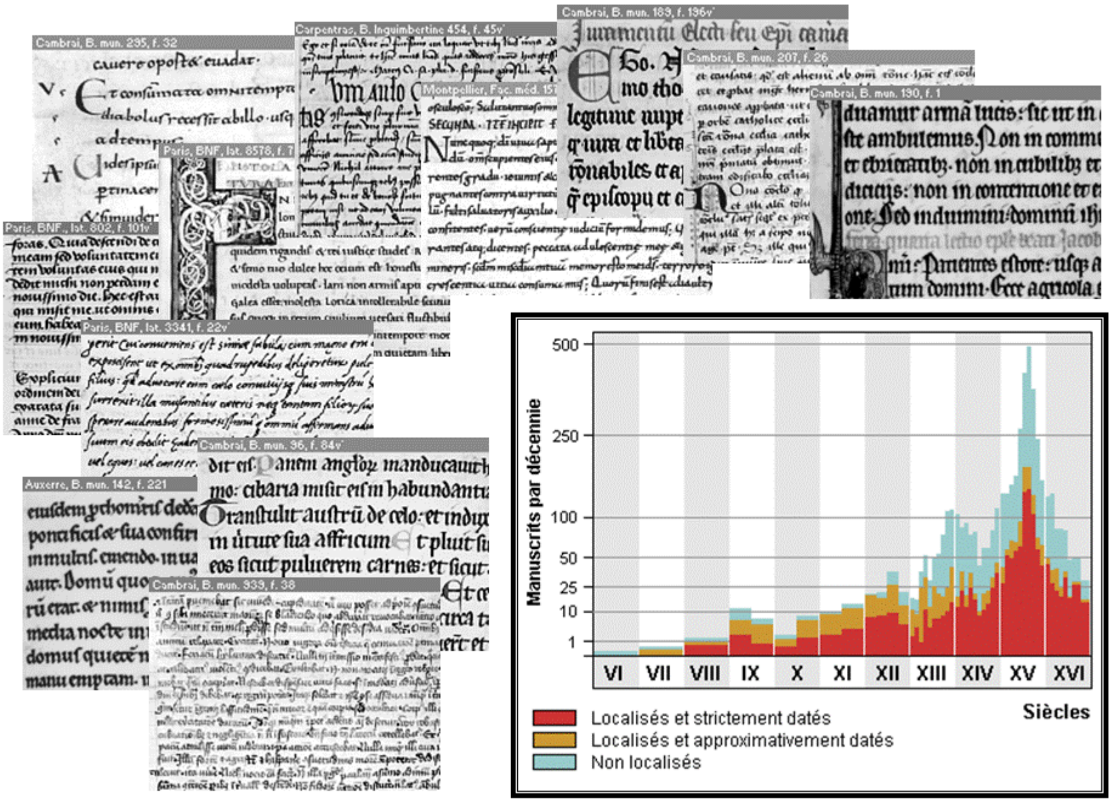
I will not dwell on the institutional aspects of the project: everything you would like to know appears on the screen. Just one precision should be added: the preposterous acronym stands simply for "computer-assisted palaeography".

Of course, Graphem is not the first project that endeavours to deal with palaeographical matters using computerized methods. A number of similar projects are currently underway, and some have now been running for quite a number of years. Four or five of these were recently presented in a conference held last July in Munich in connection with the publication of the book *Codicology and Paleography in the Digital Age*.

The projects just referred to have in common that they are oriented towards practical palaeography; more precisely, towards the recognition of single hands or characteristic styles – and for the most daring of them, towards automated transcription.

Graphem is aiming at something else, namely: fundamental palaeography – that is: the typology of medieval scripts and the history of their development.

The Catalogues of Dated Manuscripts: an Overflow of Palaeographical Data



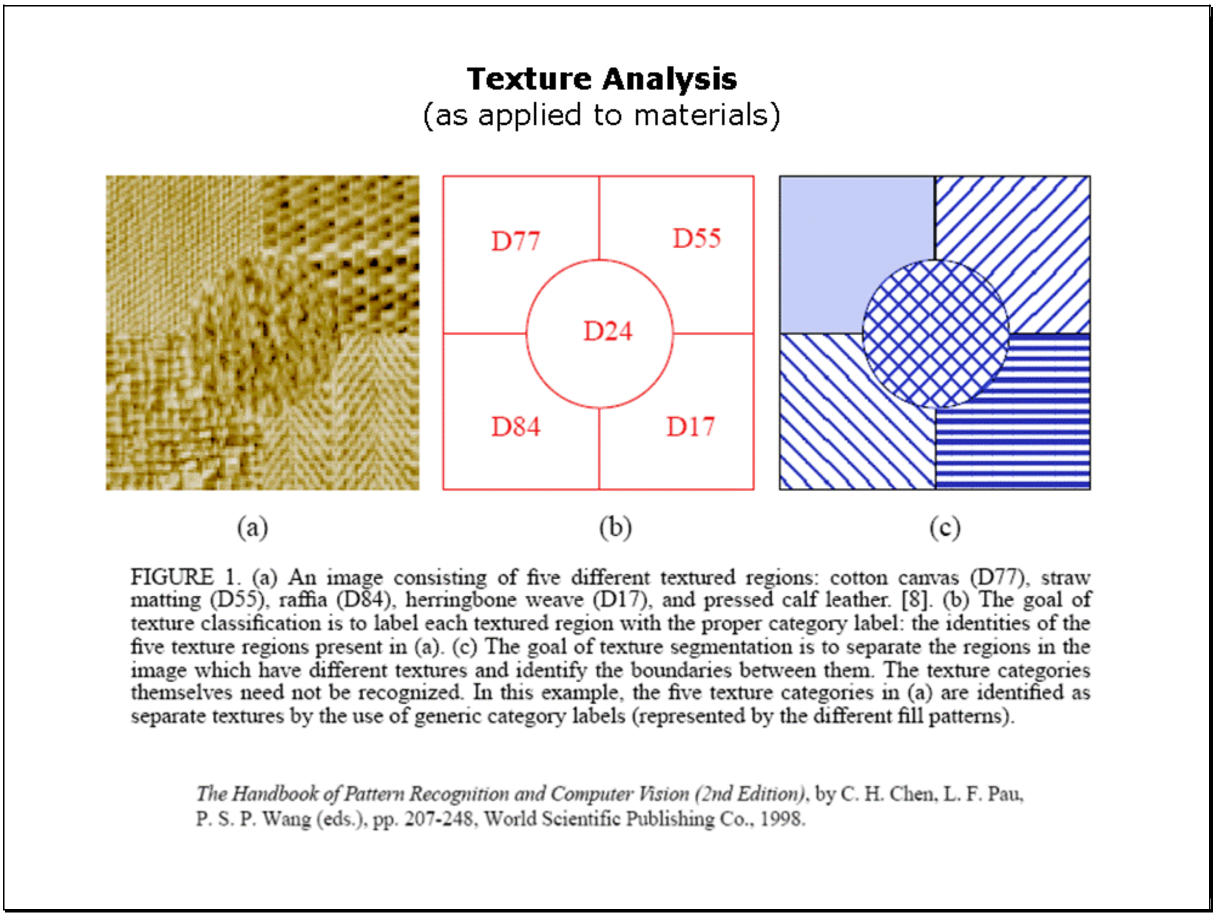
Indeed, there is an urgent need for new tools in this field, because palaeography is currently in a critical situation. The reason is that we have not yet developed the appropriate means to manage the consequences of the fantastic growth of our documentation over the last fifty years. Throughout this period, palaeographers have amassed a gigantic collection of specimens, published in the various series of Catalogues of Dated Manuscripts. We currently have at our disposal thousands and thousands of specimens of scripts that are exactly dated and/or located. But we are still unable to methodically exploit this wealth. Not only because such quantities cannot be handled without the help of a computer, but because we have been unable so far to set up a typological frame proportionated to such a mass.

If we leave out the earliest periods, where some typological classes contain no more than half a dozen specimens, all we can use to classify this material is about twenty classes or subclasses, comprising those recently introducted by Albert Derolez for the later Middle Ages. This is obviously insufficient.

To remedy this poverty, we usually resort to an endless repertoire of trivial adjectives, such as elongated, constricted, fluent, flourished, slim, sharp, fat, etc. But this can hardly help as long as such terms have no precise definition, and the qualities they involve cannot be accurately measured. Who can say when a Caroline minuscule becomes a fat Caroline minuscule?

We clearly need a new approach and new categories: not to substitute them for the current ones, but to use them as a second dimension in some kind of cross-table.

What may be expected from the methods recently developed in the field of Image Analysis is precisely that they should help us build a set of descriptors of a different, non-genetical nature.



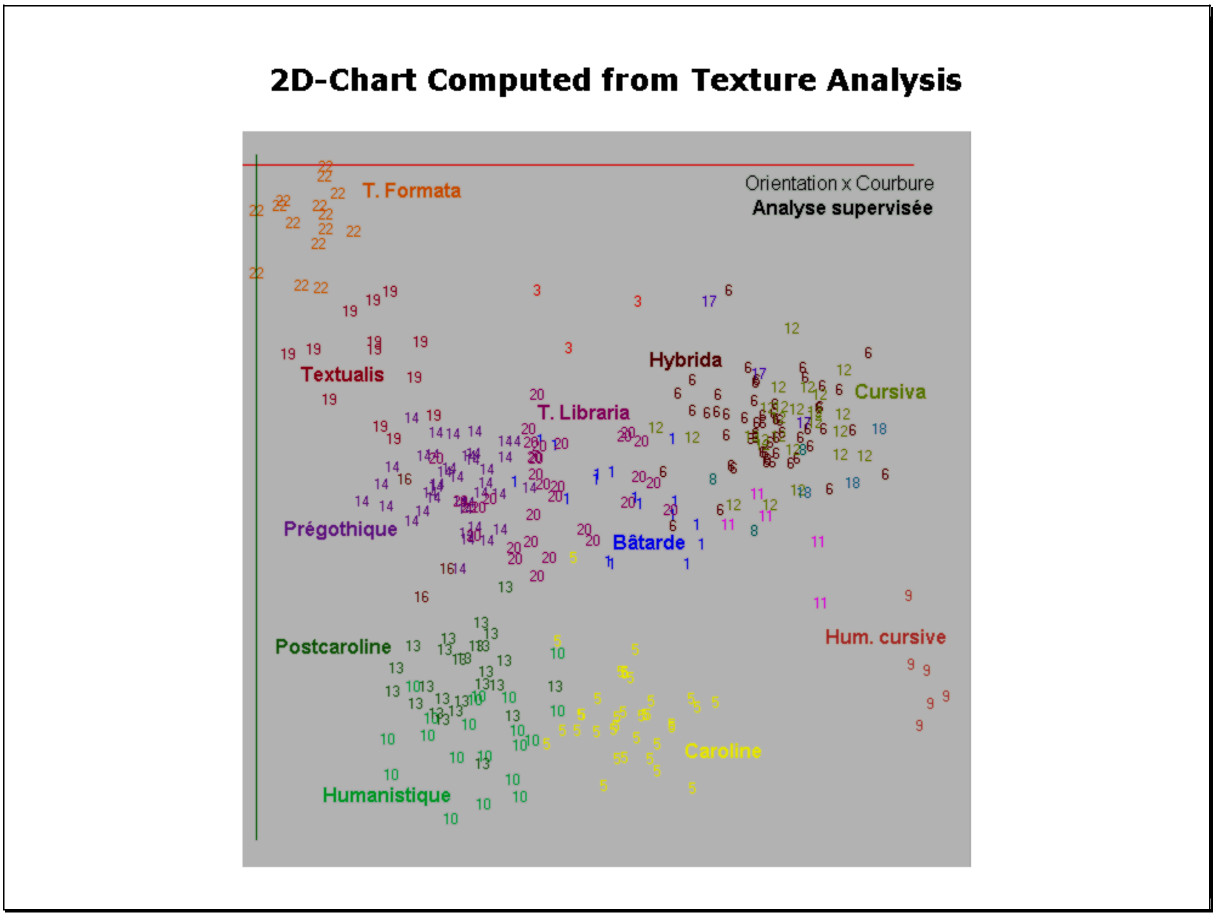
Texture Analysis
(as applied to materials)

FIGURE 1. (a) An image consisting of five different textured regions: cotton canvas (D77), straw matting (D55), raffia (D84), herringbone weave (D17), and pressed calf leather. [8]. (b) The goal of texture classification is to label each textured region with the proper category label: the identities of the five texture regions present in (a). (c) The goal of texture segmentation is to separate the regions in the image which have different textures and identify the boundaries between them. The texture categories themselves need not be recognized. In this example, the five texture categories in (a) are identified as separate textures by the use of generic category labels (represented by the different fill patterns).

The Handbook of Pattern Recognition and Computer Vision (2nd Edition), by C. H. Chen, L. F. Pau,
P. S. P. Wang (eds.), pp. 207-248, World Scientific Publishing Co., 1998.

One of these methods is Texture Analysis, a technique widely used in the industrial sphere to identify materials, and also in medicine, where it serves to recognize abnormal cellular tissues. The basic principle is to evidentiate the recurrence of certain small, elementary shapes scattered over a wide area.

It seemed interesting to test what would happen when it was applied to a whole page of writing, thus considered as a texture.



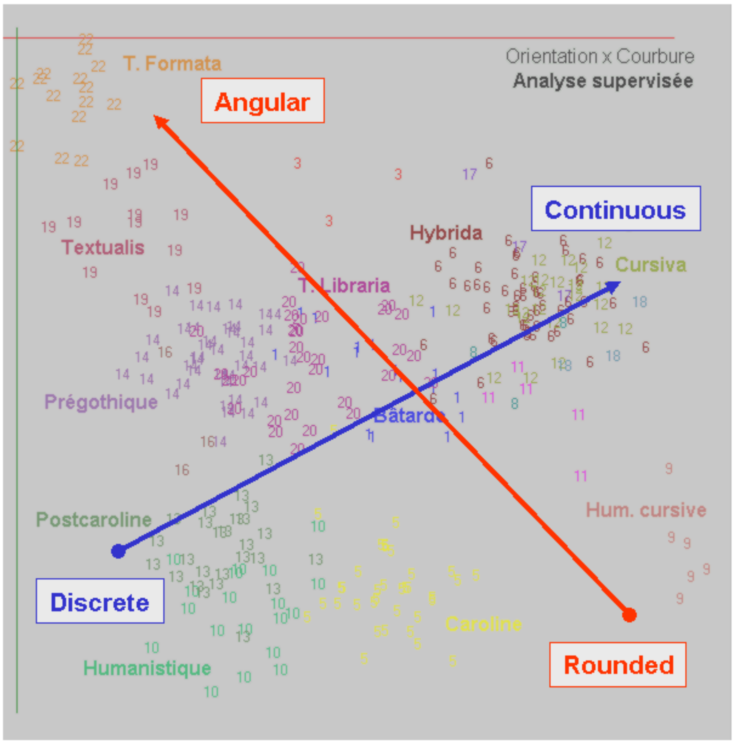
2D-Chart Computed from Texture Analysis

At first glance, the result looks outstanding: on this chart, the various types of scripts are identified by numbers in different colours – not as distinct as one would hope, because it is not easy to select 22 different colours.

However it can be seen that each type is well grouped together and sufficiently independent from others – with the exception of a few minor groups that have only a few representatives, or that are just subspecies of a major group. But it would take too long to detail these peculiarities.

2D-Chart Computed from Texture Analysis

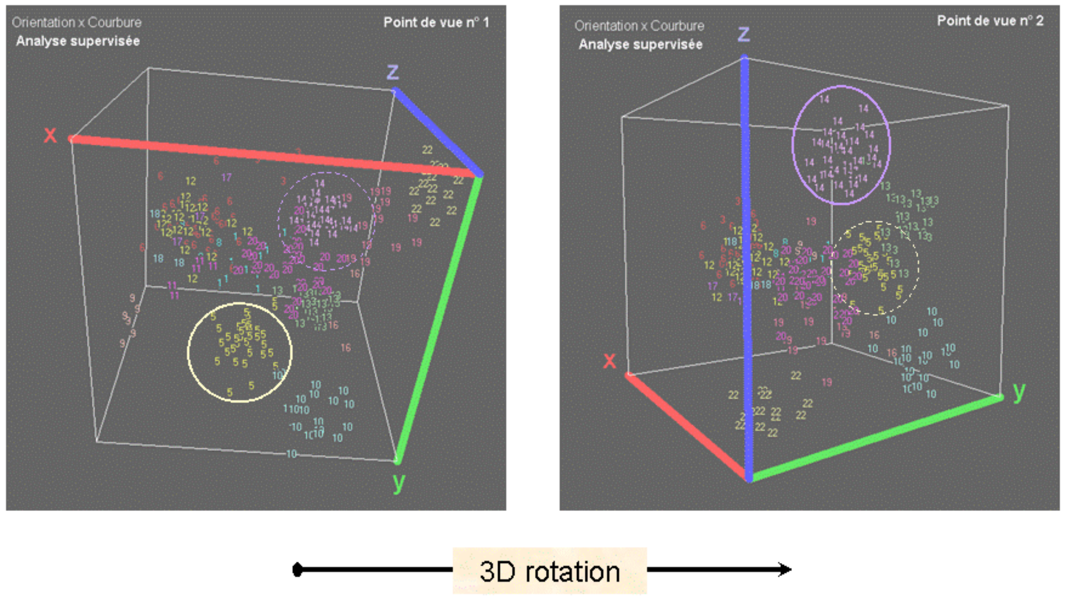
One Possible Interpretation



It is even possible to find a meaning for the position of each type in the graph: from bottom left to top right they are displayed in order of increasing angularity, from Caroline to Textualis Formata. In the perpendicular direction, they spread from those where letters are clearly separate, to those where they are tightly linked: the extreme positions are occupied by Later Caroline at bottom left, and Cursiva at top right.

This is only <u>one</u> possible interpretation, because what we are looking at is two-dimensional only, whereas the calculated scatterplot is multidimensional.

3D Viewing
with perspective artifacts



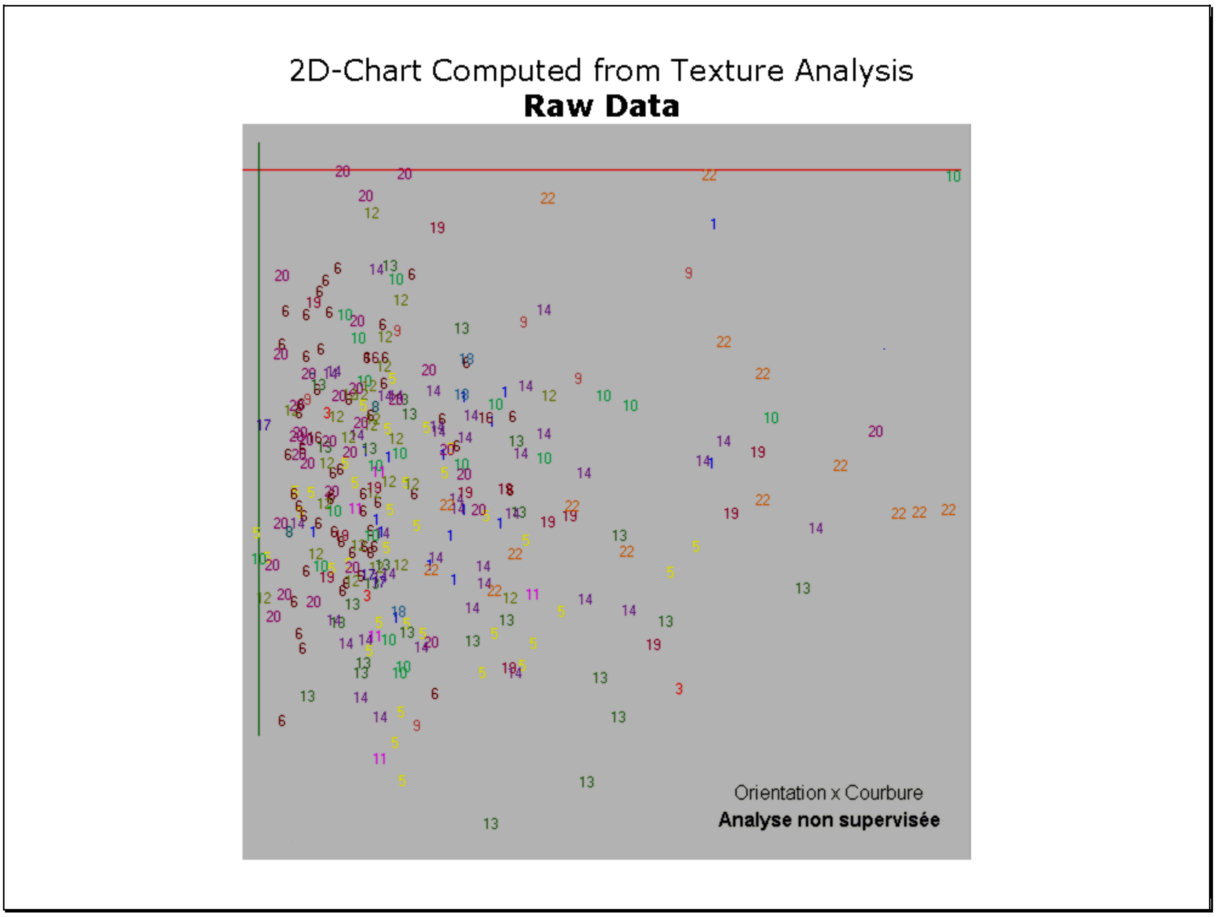
Three-dimensional viewing allows to inspect the dispersion in one more dimension, which can be thought of as the depth of the previous chart.

These two graphs are two pictures of a single analysis, viewed from different angles. By rotating the cube in the appropriate direction we may, for instance, bring to the foreground groups that were previously in the background and partly hidden by those closer to the observer. Here you can see that group no. 14, which seemed mixed up with group no. 20, is in fact completely independent and clearly individuated.

As a counterpart, interpreting becomes more complex, as the situation seems to change every time the system is rotated in one direction or another.

You will probably think that such a result is miraculous – and too good to be true. And you are right: there has been some cheating.

It is untrue because information was supplied beforehand about the palaeographical type of each specimen, and this information was taken into account while processing the data. Mathematical tricks were applied in order to minimize the distance between individuals of the same group, and to keep each group as far as possible apart from others.



2D-Chart Computed from Texture Analysis

Raw Data

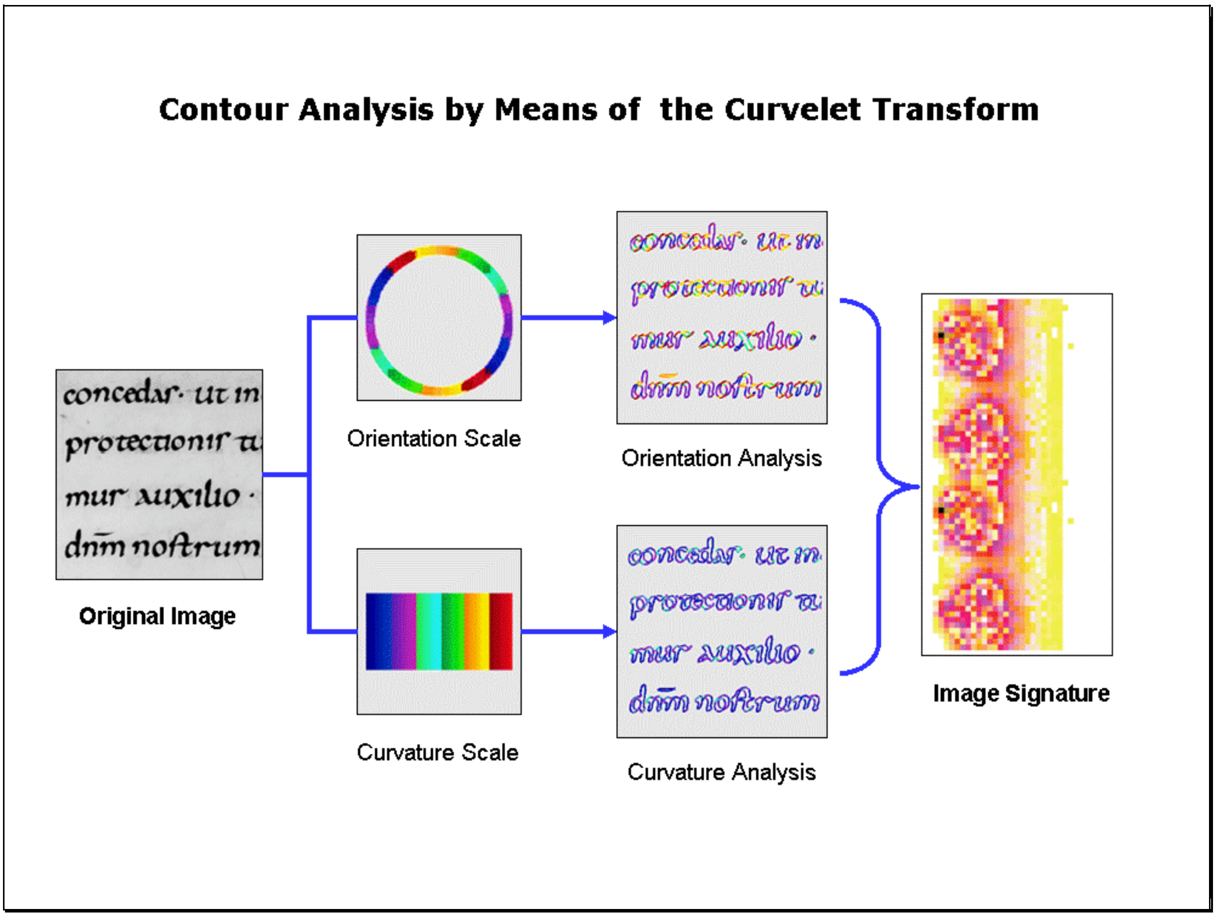
When this information is not provided beforehand, our wonderful scenery vanishes like a mirage.

All the groups explode, and their components are scattered all over the diagram. It becomes impossible to detect any structure or even any trend in this jumble, and we are left with desperate chaos.

The conclusion is that this method can be a powerful tool for illustrating a pre-defined typology. But it does not provide any great help if we wish to test whether the theory is reliable, or to build up a new, factual classification, or even to refine the pre-existing one.

The overall distribution of types in the previous, assisted result did make some sense, in a way which could be apparent only to palaeographers and not to computer scientists, thus proving that it was not entirely an arbitrary construction. Nonetheless, it remains practically impossible to decide how any one of a collection of possible factors influences the end result, and thus to analyze the scatterplot in historical terms.

At least, these graphs make it clear that a degree of likeness or difference between two scripts can be expressed as a distance. And you may imagine that there are several ways to calculate such a distance. Thus, we have been able to test two further methods.



Contour Analysis by Means of the Curvelet Transform

Original Image

Orientation Scale

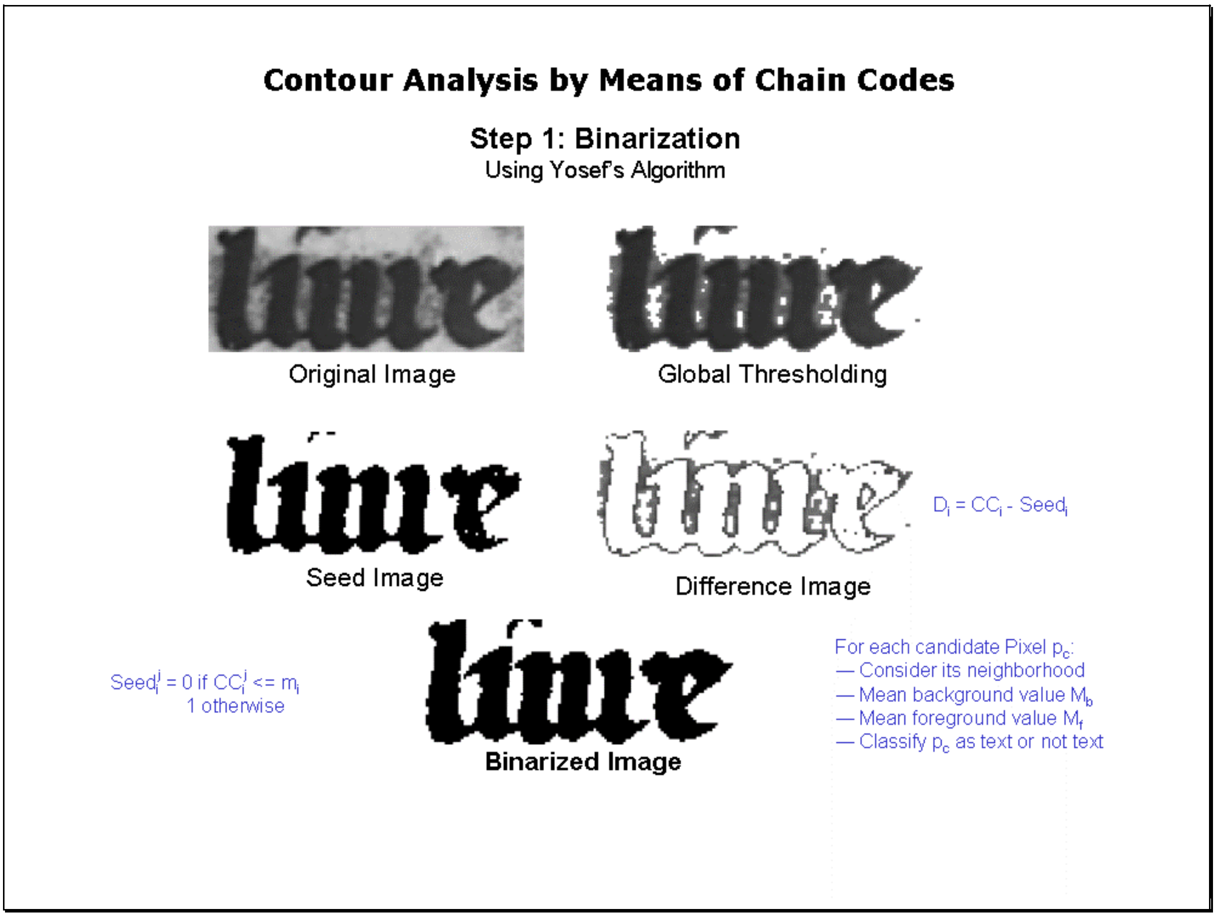
Curvature Scale

Orientation Analysis

Curvature Analysis

Image Signature

The first one is really too absconse to be exposed here. It is based on Curvelet Analysis, a new-born descendant of Fourier Transform Analysis, and involves mathematical concepts that are far beyond the reach of ordinary brains.



Contour Analysis by Means of Chain Codes

Step 1: Binarization

Using Yosef's Algorithm

Original Image

Global Thresholding

Seed Image

Difference Image

$D_i = CC_i - Seed_i$

$Seed_i^j = 0$ if $CC_i^j <= m_i$

1 otherwise

Binarized Image

For each candidate Pixel p_c:

— Consider its neighborhood

— Mean background value M_b

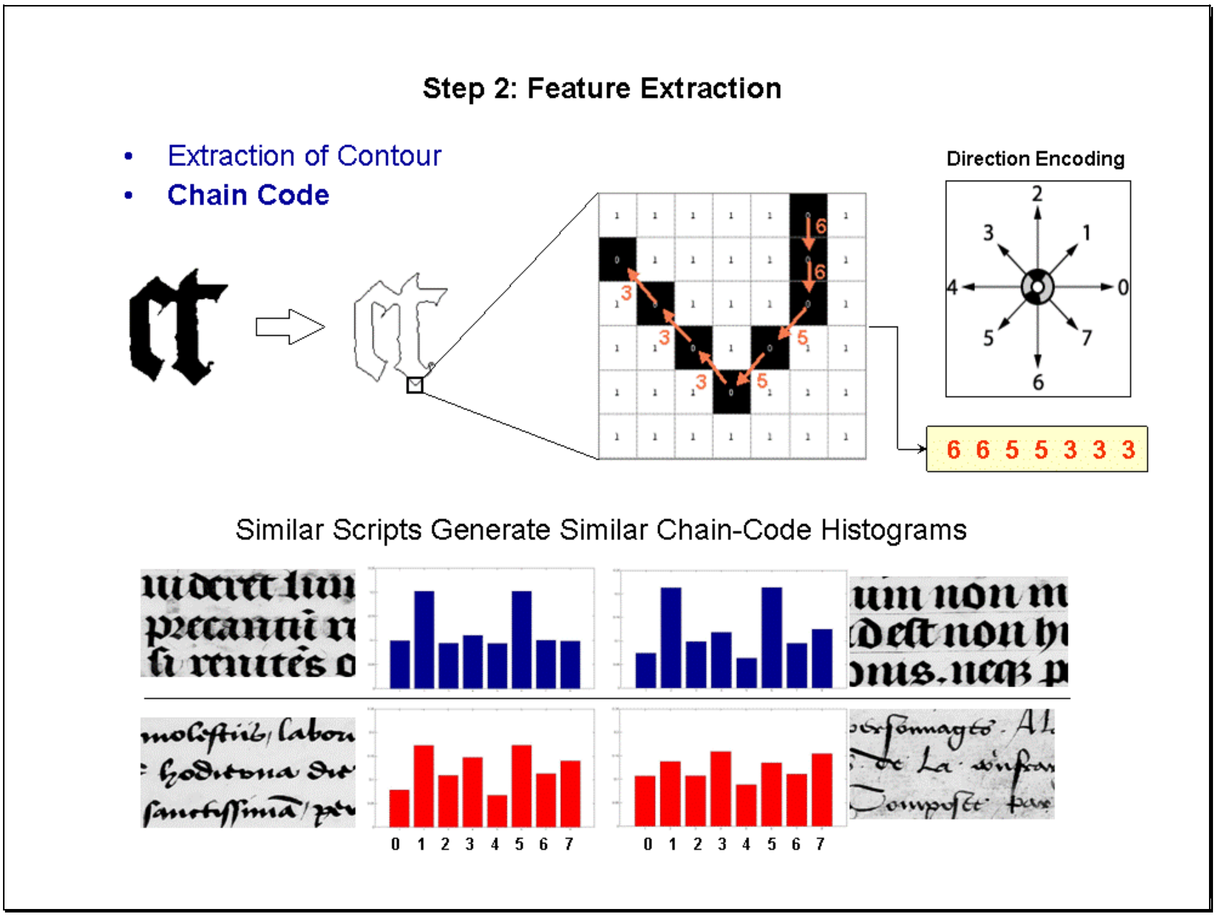
— Mean foreground value M_f

— Classify p_c as text or not text

The second one, on the other hand, is very accessible, at least at the level of its principles.

Everything starts with binarization, that is to say converting the grey-level picture into strict black and white – in other words, separating the script from the background. This is not so easy with ancient documents written on darkish and uneven material, defaced by later accidents such as stains or tears – not to mention blurred photographs. Simple thresholding is unsufficient: some computing has to be added to take the surroundings of each pixel into account.

Several methods have been experimented; but none is perfect, and we are still working to improve this operation.



Step 2: Feature Extraction

- Extraction of Contour
- **Chain Code**

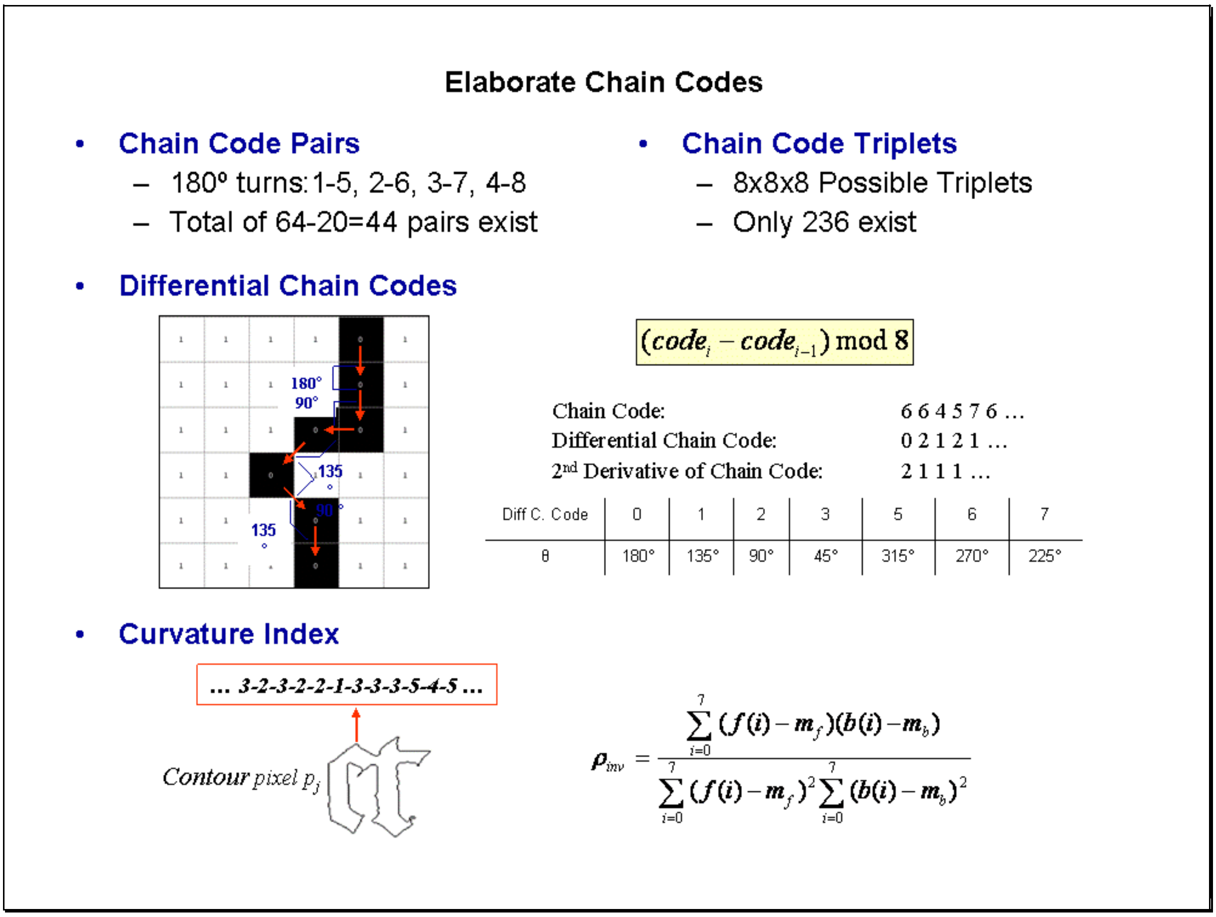
Direction Encoding

6 6 5 5 3 3 3

Similar Scripts Generate Similar Chain-Code Histograms

The basic process is very straightforward – I would even say childish. From any point of the outline you move to the next point, and you make a note of the direction in which the move was made, using a number from 0 to 7. Then you repeat this step until the loop is looped.

A simple statistic of these elementary moves, in the form of a histogram, already shows significant differences between scripts belonging to different types.



Elaborate Chain Codes

- **Chain Code Pairs**
  - 180° turns: 1-5, 2-6, 3-7, 4-8
  - Total of 64-20=44 pairs exist
- **Chain Code Triplets**
  - 8x8x8 Possible Triplets
  - Only 236 exist
- **Differential Chain Codes**

$$\boxed{(code_i - code_{i-1}) \bmod 8}$$

Chain Code:	6 6 4 5 7 6 …
Differential Chain Code:	0 2 1 2 1 …
2nd Derivative of Chain Code:	2 1 1 1 …

Diff C. Code	0	1	2	3	5	6	7
θ	180°	135°	90°	45°	315°	270°	225°

- **Curvature Index**

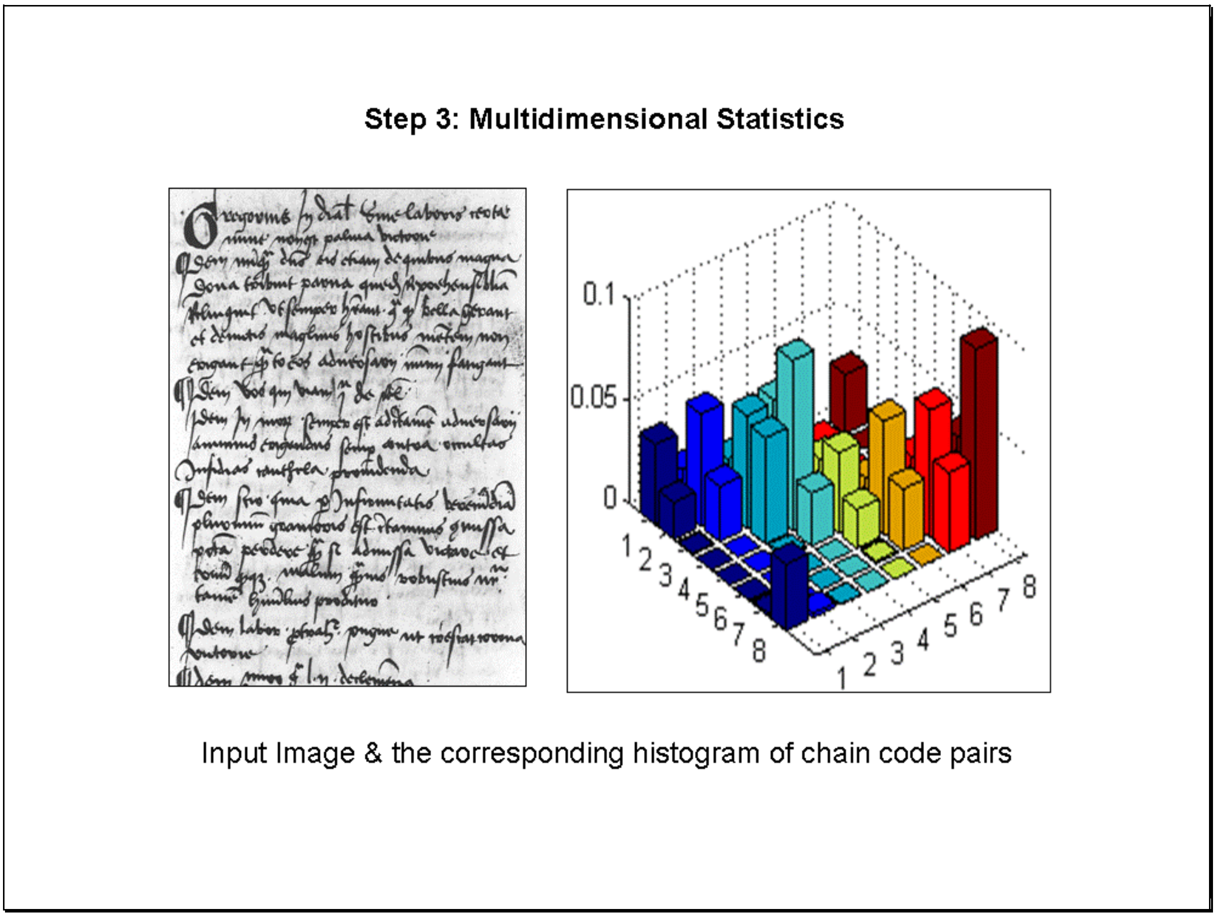
… 3-2-3-2-2-1-3-3-3-5-4-5 …

$$\rho_{inv} = \frac{\sum_{i=0}^{7}(f(i)-m_f)(b(i)-m_b)}{\sum_{i=0}^{7}(f(i)-m_f)^2 \sum_{i=0}^{7}(b(i)-m_b)^2}$$

Of course, this computation is too rudimentary to take us very far. If we wish to make more refined comparisons, we need more. But that won't be a great challenge.

From the elementary codes it is possible to build up more elaborate descriptors that account for what happens in a sequence of three, five, or seven pixels, and thus provide information about the global slope and curvature of the shape.

The mathematical formulas displayed on the screen may look alarming; but most of them involve only the repetition of very simple calculations.



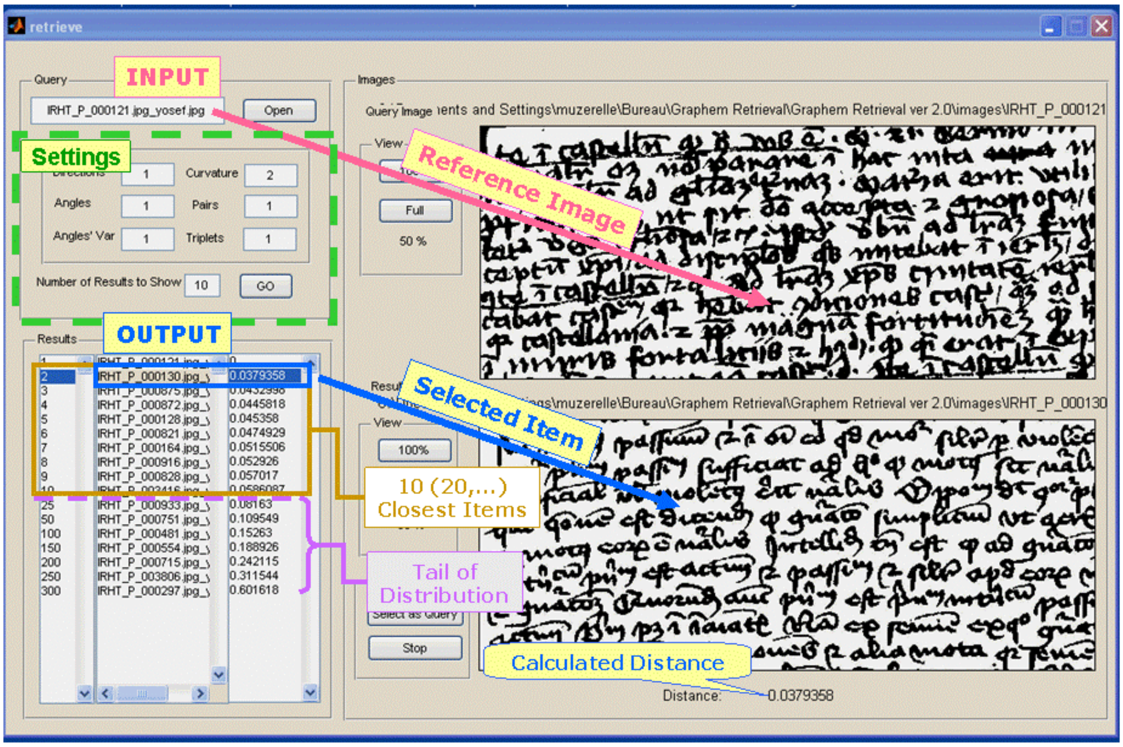
Step 3: Multidimensional Statistics

Input Image & the corresponding histogram of chain code pairs

For every sample of script the distribution of each of these descriptors is summarized as a frequency histogram. The sets of histograms thus obtained are each peculiar to one single script and identify it with as much certainty as a fingerprint. Then, the differences between such "fingerprints" are evaluated to serve as a distance.

This requires a complicated and time-consuming computational business. Just to give an idea, six to eight hours are necessary for a powerful computer to calculate the distance of each item to every other in a set of some 300 samples.

Retrieving Scripts Using their Distance



Query: IRHT_P_000121.jpg_yosef.jpg

Directions: 1 | Curvature: 2
Angles: 1 | Pairs: 1
Angles' Var: 1 | Triplets: 1

Number of Results to Show: 10

Results

1	IRHT_P_000121.jpg_\	0
2	IRHT_P_000130.jpg_\	0.0379358
3	IRHT_P_000875.jpg_\	0.0432998
4	IRHT_P_000872.jpg_\	0.0445818
5	IRHT_P_000128.jpg_\	0.045358
6	IRHT_P_000821.jpg_\	0.0474929
7	IRHT_P_000164.jpg_\	0.0515506
8	IRHT_P_000916.jpg_\	0.052926
9	IRHT_P_000828.jpg_\	0.057017
10	IRHT_P_003416.jpg_\	0.0586087
25	IRHT_P_000933.jpg_\	0.08163
50	IRHT_P_000751.jpg_\	0.109549
100	IRHT_P_000481.jpg_\	0.15263
150	IRHT_P_000554.jpg_\	0.188926
200	IRHT_P_000715.jpg_\	0.242115
250	IRHT_P_003806.jpg_\	0.311544
300	IRHT_P_000297.jpg_\	0.601618

Images

Query Image: ients and Settings\muzerelle\Bureau\Graphem Retrieval\Graphem Retrieval ver 2.0\images\IRHT_P_000121

View: 50 %

Resu ings\muzerelle\Bureau\Graphem Retrieval\Graphem Retrieval ver 2.0\images\IRHT_P_000130

View: 100%

Distance: 0.0379358

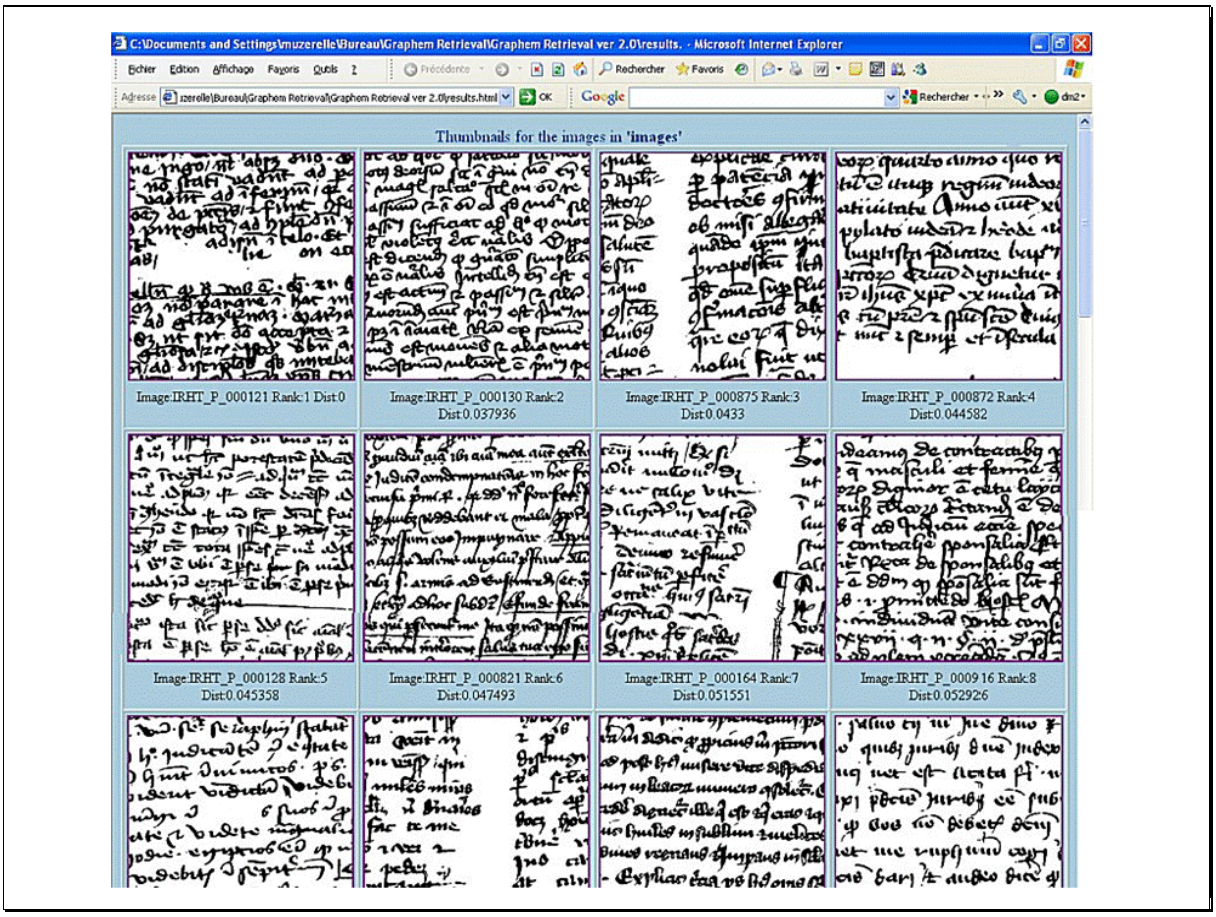
This is how the results may be visualized using a specific interface (which is still undergoing daily improvements).

At top-left of the screen is the input field, allowing to select any item currently registered in the database. The corresponding image is displayed in the top window.

After launching a short computation, a list of the ten, twenty, or thirty closest specimens is displayed in the bottom frame. Clicking on any one of these items makes it appear in the second window.

At this stage, a complementary list is provided with items at regular intervals from the top of the distribution – just as a means of control.

In addition, a number of parameters may be adjusted so as to have the distance calculated in accordance with the nature of the script currently dealt with – I will say more about that in a minute.



C:\Documents and Settings\muzerelle\Bureau\Graphem Retrieval\Graphem Retrieval ver 2.0\results. - Microsoft Internet Explorer

Adresse: ızerelle\Bureau\Graphem Retrieval\Graphem Retrieval ver 2.0\results.html

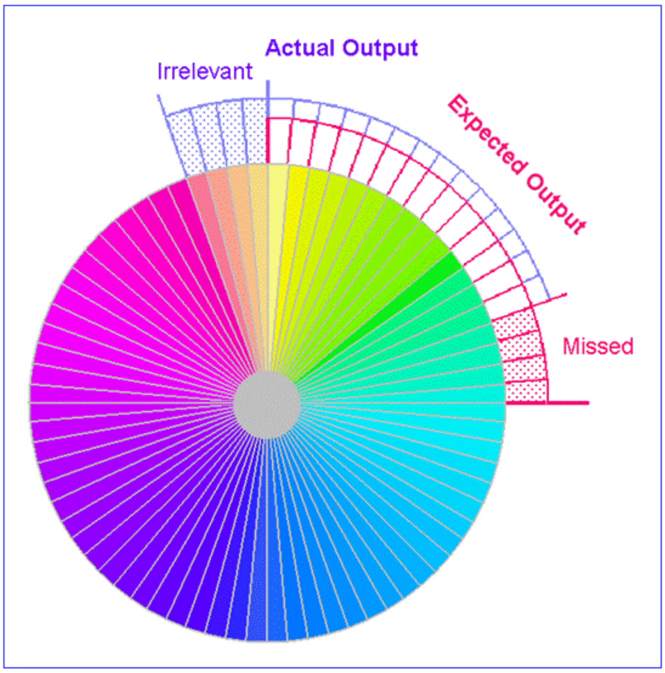
Thumbnails for the images in 'images'

Image:IRHT_P_000121 Rank:1 Dist:0	Image:IRHT_P_000130 Rank:2 Dist:0.037936	Image:IRHT_P_000875 Rank:3 Dist:0.0433	Image:IRHT_P_000872 Rank:4 Dist:0.044582
Image:IRHT_P_000128 Rank:5 Dist:0.045358	Image:IRHT_P_000821 Rank:6 Dist:0.047493	Image:IRHT_P_000164 Rank:7 Dist:0.051551	Image:IRHT_P_0009 16 Rank:8 Dist:0.052926

It is also possible to save the results as small pictures for comprehensive inspection. I don't think this view needs any further comment.

Once again please refrain from being over-enthusiastic. We have had some results indeed, but not as conclusive as we might have expected, and in any case too casual to allow this method to be used as an everyday palaeographical tool.

Performance Ratio of Contour Analysis
(estimated)



On the whole, the performance ratio does not rise above 75%. This means that among the top best, one out of four is irrelevant – which is not tragical as long as the software is used by competent scholars. But at the same time, the same number of items that should have been in the list do not appear; and that is a major hindrance, because you have one chance out of four of missing what you were looking for.

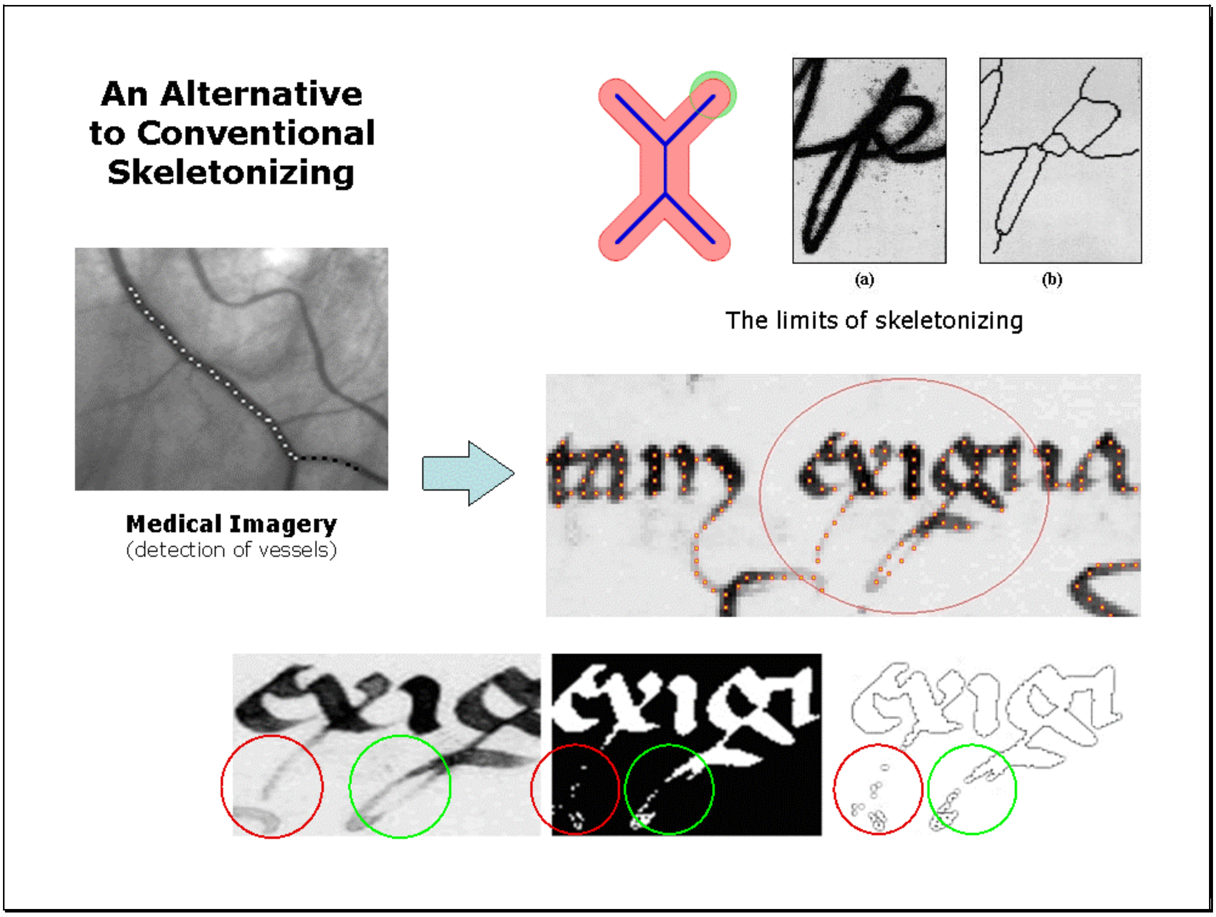
Another inconvenience originates in the adjustment of parameters. By this operation more or less weight can be given to one or another of those elaborate descriptors that are supposed to account for angularity, slope and the like.

But it has been impossible so far to define the ideal combination, not only for the whole corpus, but even for a certain type of script. In each single case, repeated trials must be made until the optimal tuning is reached. The situation is thus reversed, since we have to help the computer do what it was supposed to help us to do.

At this stage there seems to be scarce hope for great improvements, since the computer scientists are already at the top of their ability.

Besides, it should be noted that the Curvelet method meets with exactly the same limitations in spite of its sophistication, and that in both cases problems arise under the same conditions. Now, the only feature that these two methods have in common is that they deal exclusively with the outlines. Consequently the problem must lie there.

This problem could have been predicted from the start (and it was), since we know from palaeographical experience that the outline of letterforms is just an epiphenomenon, and a highly unreliable one. However, it could hardly be avoided that computer science would experiment with outlines first, since detecting and analyzing the outlines of shapes is the primary approach of Image Analysis. But it is now clear that other methods must be brought into practice for the specific case of medieval scripts.



An Alternative to Conventional Skeletonizing

Medical Imagery
(detection of vessels)

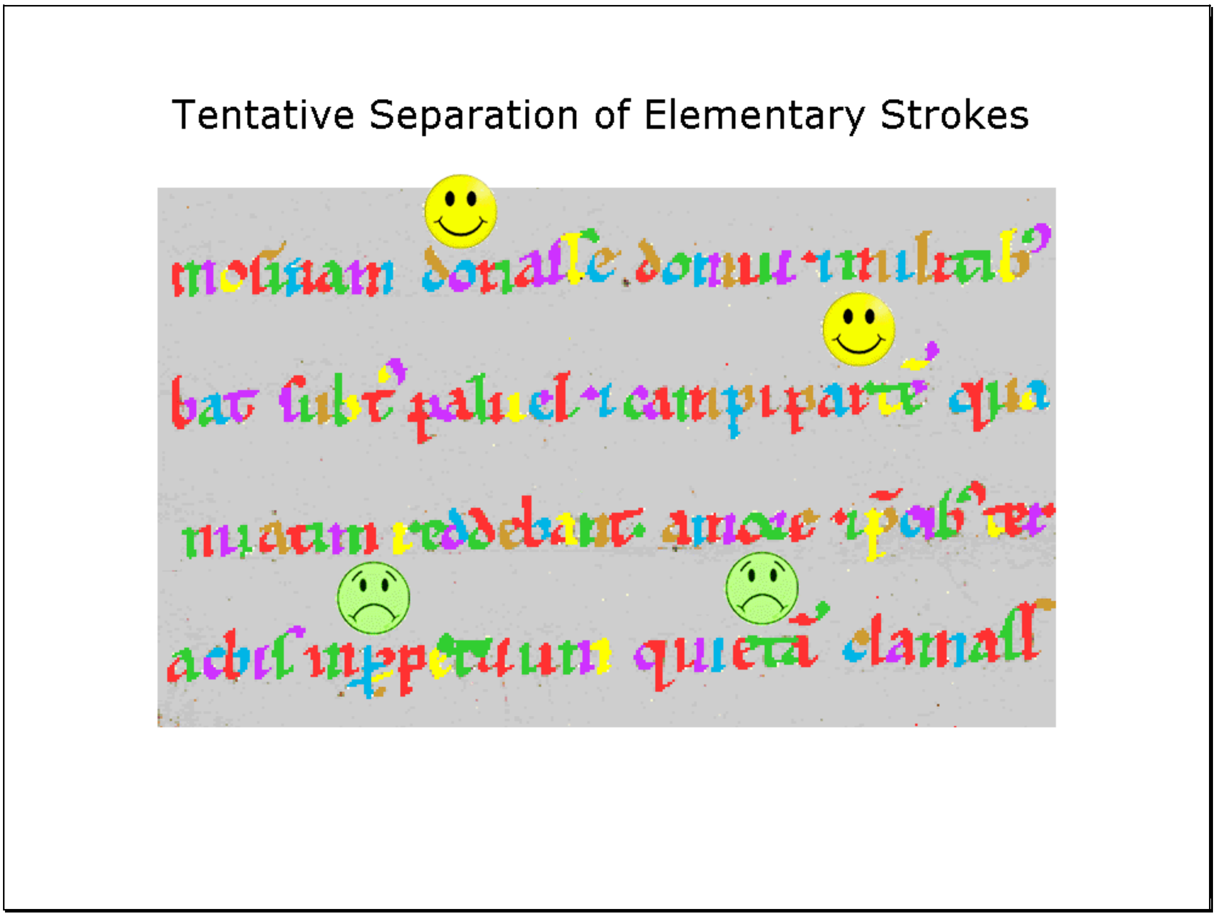
The limits of skeletonizing

What can we do if we leave the outlines out?

The usual alternative in Image Analysis is skeletonizing – that is: thinning the shape until it is reduced to a simple line.

However, the classical methods used in such an operation do not yield satisfactory results in the case of scripts. This has been experienced previously on modern scripts, as you can see on the screen; and what we get with medieval scripts is even worse.

Lately we have started trying out a new method inspired from medical imagery. And it seems to work fine, since it is able to follow a penstroke even where the trace becomes so faint as to be completely lost in the binarization process.

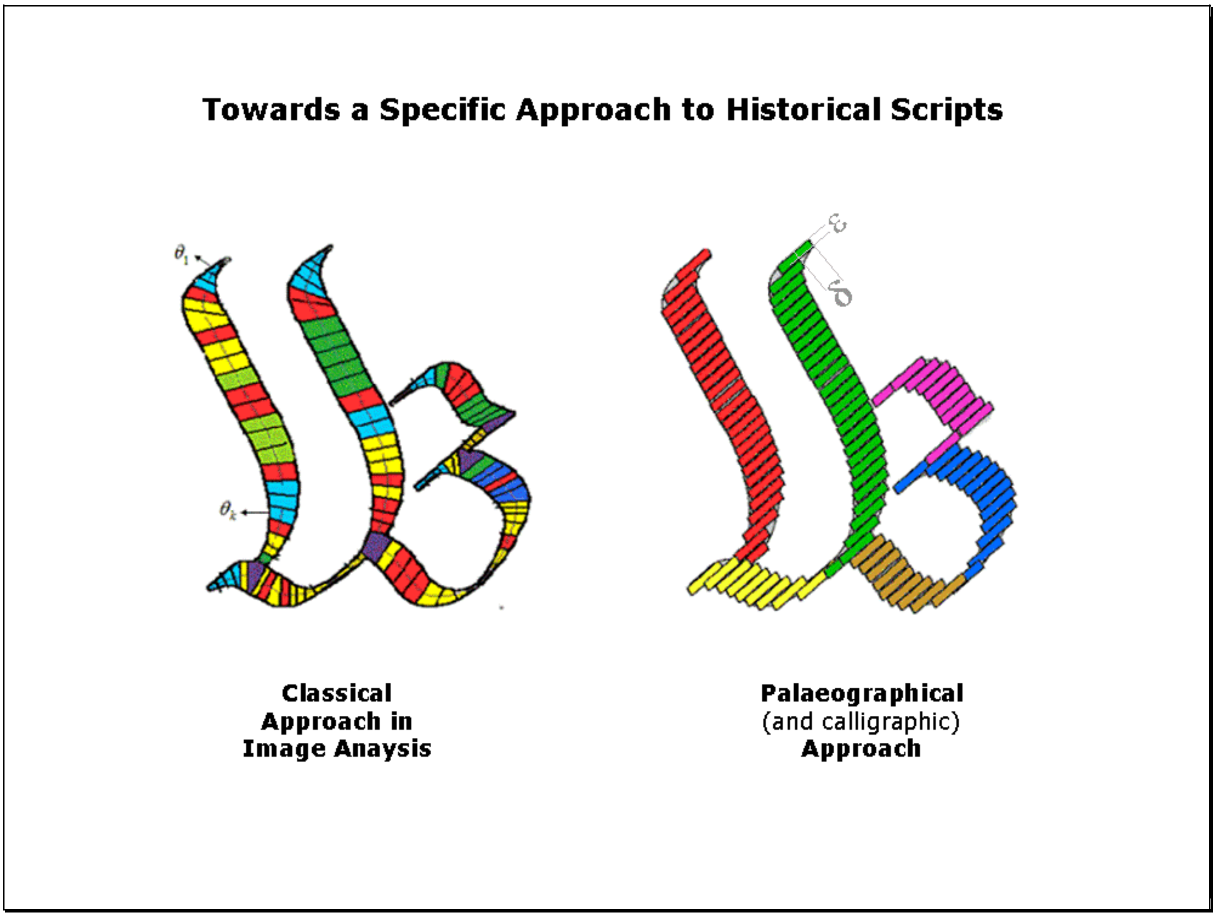


Tentative Separation of Elementary Strokes

The other direction we are presently exploring is reconstructing the path followed by the pen; and to begin with, recognizing and separating the individual strokes that constitute each letter or sequence of letters. Although we are just starting, some interesting advances have already been made.

This very first attempt shows many encouraging features – as, for instance, a correct decomposition of the letters 'o' and 'd', and of the abbreviations 'us' and 'ur'. On the other hand, some other details are completely erroneous – as for instance the decomposition of the abbreviation stroke in 'per'. In any case, it seems that the computational algorithm will necessarily have to be complemented with a set of rules of a palaeographical nature.

Once we have reached a more reliable stage, it will be possible to set up the list of elementary strokes each type of script is using, following the spirit of Arabic or Chinese calligraphy. Comparing the resulting sets or “code-books” on a large scale will undoutedly lead to interesting observations, and possibly to new criteria for classification.



Towards a Specific Approach to Historical Scripts

Classical Approach in Image Anaysis

Palaeographical (and calligraphic) **Approach**

As a conclusion, I would say that the major difficulty in this project is for computer science to adapt its usual methods to an unfamiliar and highly specific material. Conventional Image Analysis regards every shape as a surface delimited by specific contours, or as a combination of small areas with specific geometrical features – while palaeography would decompose the script into a sequence of elementary moves of the pen. In theory, there is no reason why these elementary moves could not be individuated and statistically treated. We just have to convince the technicians that scripts are not shaped, but written, in other words that strokes are the result of linear movements, not outlined surfaces.

Two corollaries ensue from this statement. The first one is almost trivial to us: it is that scripts cannot be studied without reference to the instrument that produced them and how it was handled.

The second one is considerably more pregnant with consequences. It is that the trace left by the scribe is less important than what he intended to trace, and what the reader is trained and accustomed to imagine it was. This side of the question will escape forever any geometrical approach; it belongs to the field of neuro-psychological science. It is therefore a different story – or hopefully, the subject for another project.

Finis