doi:10.25358/openscience-9636

Tobias Konrad*

Hieroglyphs in a Multidimensional Space

A Case Study on the Applicability of Digital Paleography to Cursive Hieroglyphs

This paper presents a case study of basic machine learning techniques applied to a selected set of cursive hieroglyphs. A processing pipeline is described including analyses with dimensionality reduction and cluster algorithms. The research topic is part of the author's PhD thesis, which deals with the paleographic analysis of cursive hieroglyphs during the Middle Kingdom and is associated with the AKU project.

1. METHODS AND AIMS OF DIGITAL PALEOGRAPHY

The application of digital methods for research questions dealing with paleographic purposes has received little attention in Egyptology so far. First attempts have been made to address the issue of paleography on a digital level, such as the Polychrome Hieroglyph Research Project,[1] the Demotic Palaeographical Database Project (DPDP),[2] the project Hieroglyphic "Hands",[3] and the project *Altägyptische Kursivschriften* (AKU).[4] If one compares the state of research with, for instance, medieval studies, one can see that digital paleography has already established itself as a fully-fledged discipline.[5] The fact that digital methodology is an integral part of this field also shows the methodological reflection that is now being carried out within digital humanities.[6]

Depending on what is understood by the term "digital" in general, a wide variety of methods is attested for digital paleography. As Arianna Ciula[7] pointed out, a lot of semantic facets exist for the

* The author is scientific researcher in the project *Altägyptische Kursivschriften: Digitale Paläographie und systematische Analyse des Hieratischen und der Kursivhieroglyphen (AKU)*, Academy of Sciences and Literature | Mainz, Germany.

1. https://www.phrp.be/.
2. http://demotischdemotisch.de/.
3. https://journals.openedition.org/baefe/996#tocto2n12/.
4. For further information and digital methods, see Gülden et al. 2020; Bermeitinger et al. 2021.
5. See for example Ciula 2005; Kestemont et al. 2017.
6. See for example Ciula 2017; Lit 2020, pp. 102–123.
7. Ciula 2017.

terminology of this field of study. First of all, the term "digital" seems to be used as an opposite to "analog"—one could also say "traditional".[8] The minimum requirement to be considered digital is that the object to be studied is available in a somehow digital form, such as raster or vector graphics.[9]

One aim of digital paleography is to supply traditional paleographic analysis using a computer-aided methodology. This lowermost level is achieved by annotating digital images with metadata and making them available online.[10] Going deeper into a digital methodology, there is the possibility of manipulating the data positively, such as contrast enhancement for image data or preprocessing for further analysis.[11] Data that is transformed even more can be analyzed by using methods from the domain of machine learning.[12] This computational or artificial paleography produces models of prototypical graphemes.[13] The most complex field is the use of artificial intelligence, for example, the application of artificial neural networks that simulate human visual perception. Doing paleography digitally instead of doing it traditionally leads to more transparency and reproducibility. Considering traditional methods, we might expect little descriptive expressions like "very similar", "less similar", and "not similar" if one has to estimate the similarity of allographs of one sign. Asking another scholar creates different statements concerning the distinctiveness of the signs. Different experts focus on different morphological features depending on prior knowledge and experience which leads to the fact, that the statements are hardly reproducible. In contrast, the use of digital methods makes it possible to compare these allographs objectively, since the data is transformed into numerical representations.[14] Applying distance measurements among these numbers, one can describe their similarity among each other mathematically. However, one should be aware of the fundamental epistemic impacts of creating and analyzing digital objects for research purposes.[15]

2. DIGITAL ANALYSIS OF CURSIVE HIEROGLYPHS

2.1. EXPERIMENTAL SETUP

The key question for this study was whether we can collect information about morphological features of a single cursive hieroglyphic grapheme supported by digital paleography. To create a controlled experimental frame, we will look at a closed environment, namely on one single textual

8. Ciula 2017, p. ii90.
9. For different approaches of paleographic visualizations and formats, see Gülden 2018, pp. 91–95.
10. A good approach was made by the project *DigiPal*, http://www.digipal.eu/.
11. Ciula 2017, pp. ii89–ii91, esp. ii90, table 1.
12. See for example Stutzmann 2016.
13. Ciula 2005.
14. On this topic see further Ciula 2005, § 4.
15. Peursen 2010, pp. 12–11. Nevertheless, the statements given there also apply to traditional research methods. Scholars are used to analyzing printed photographs or facsimile drawings as well as transcriptions of hieratic or hieroglyphic inscriptions.

witness. The objects of study are the decorated tomb chamber and the sarcophagus of Harhotep, which were part of the Theban tomb TT 314 dating to the early 12th Dynasty.[16] First excavated in 1883 by Gaston Maspero,[17] both were dismantled and transported to the Egyptian Museum in Cairo.[18]

Their decoration program consists of Pyramid Texts and Coffin Texts executed in cursive hieroglyphs, the assigned Coffin Text siglum is T1C.[19] Analyzing a specific hieroglyphic grapheme and its morphological structures, the focus lies on Sign List no. G17, the sign of the owl ().[20] The basis for the research of the author's PhD project—and also for this study—is the extensive photographic material of the de Buck's collection at the *Nederlands Instituut voor het Nabije Oosten* (NINO),[21] that was used during the publication process of Adriaan de Buck's Coffin Texts edition.

2.2. DATA ACQUISITION

The workflow carried out is similar to other computer vision workflows that deal with image recognition or classification tasks. The first step within the processing pipeline is data acquisition where the digital images are enriched with specific metadata. The annotation is stored externally in XML files that use the TEI (*Text Encoding Initiative*)[22] model, which allows you to locate specific regions on an image and to store the pixel coordinates into the XML file. Thus, it is possible to annotate a lot of hieroglyphs using a graphical user interface like an XML editor. For this study, the occurrences of the sign G17 were recorded.[23] On completion of this procedure, the annotated regions of the source images were cropped out and stored as separated image files by parsing the XML files.

2.3. PREPROCESSING

Once the image data has been extracted, an additional step of preprocessing is needed to clean it from unwanted noise and to scale it to a fixed size. The image processing can be achieved by using the library scikit-image[24] for the programming language *Python*.[25] At this stage, the color depth of the images is 8-bit grayscale and needs to be converted to binary images containing numerical values only of 0 and 1. After removing elements at the borders that are not part of the hieroglyphs the images are rescaled to a standard width and height of 88 pixels. Now each image is represented by a matrix of 88 × 88 pixels which yields a total of 7744 pixels per image.

16. Recent findings seem to shift the dating to the early 12th Dynasty. See Chudzik, Caban 2017, p. 223, n. 8. Compare Willems 1988, p. 113 who summarizes the dating approaches for the late 11th Dynasty.
17. Maspero 1885, pp. 134–180, pl. XII–XVIII.
18. CG 28023, see Lacau 1904, pp. 42–56. One fragment of the sarcophagus is located in New York, Brooklyn Museum (37.1507E). I would also like to refer to the Middle Kingdom Theban Project, which is going to re-document this tomb. See Morales et al. 2016, pp. 257–261.
19. Lesko 1979, p. 100.
20. Gardiner 1994, p. 469.
21. Egberts 1982. I want to express my gratitude to Olaf Kaper for allowing me to use the material for research purposes.
22. https://www.tei-c.org/.
23. Since this is only a case study, the annotation was not performed on the entire object, but only in a limited scope.
24. https://scikit-image.org/.
25. https://www.python.org/.

Since each image can be understood as a data point in a coordinate system containing 7744 axes—or dimensions—they must be considered high-dimensional and multivariate objects, whose dimensions have to be reduced for the analysis. One possible technique is called *Principal Component Analysis* (PCA). It transforms the data in such a way that the principal components of the whole dataset are identified—keeping as much variance as possible in the first components.[26] This means most of the information about the data is kept within the first components, which allows the use of only the first 30 principal components for analyzing this set of images.[27] Using 30 dimensions instead of 7744 reduces the amount of calculation time, but does not change the fact that the data points must still be considered high-dimensional.

2.4. ANALYSIS

2.4.1. Principal component analysis

Fig. 1 shows a common visualization of the first seven principal components or eigenvectors (v_0–v_6) including the calculated mean (μ) of the whole dataset. The second image (v_0) displays the directions within the first principal component. A dark blue silhouette is visible in the first direction, representing a sign-form that consists of only one stroke for the body of the owl. The yellow silhouette (opposite direction) describes hieroglyphs that are composed of two strokes for the body.[28]

If we plot the images on the data points of the first two dimensions, we can see the disposition of the single occurrences within the multidimensional space revealing two main clusters, the left one showing a wider expansion than the right one (fig. 2). As mentioned above, the first principal component (left-to-right) describes the difference between "one-stroke" (fig. 2, left) and "two-stroke" owls (fig. 2, right). It is harder to recognize the directions of the second component (top-to-bottom), but a closer look reveals that the orientation or rotation of the hieroglyphs has been captured here. It should be noted, that due to natural limitation, only the first two components can be displayed in the plot, although the dataset contains a total of 30 dimensions.

2.4.2. Cluster analysis

Now these feature vectors represent the high-dimensional space in which we can define the similarity of allographs based on their distance from each other. Two data points showing a small distance can be considered similar regarding their morphological structure. To group single occurrences, we can use a standard algorithm for unsupervised learning called *k*-means clustering. The algorithm uses the squared Euclidean distance between the data points and labels *k* clusters, each one having a center called centroid. Every cluster member is defined by its distance to the

26. A general introduction into PCA can be found in Jolliffe, Cadima 2016.
27. The calculation was performed using *scikit-learn*, https://scikit-learn.org/.
28. Another example of the creation of prototypes based on PCA is given in Gilliam et al. 2010, pp. 1882–1883. See also the similar approach using *tangent space* in Ciula 2005, §§ 32–37.

assigned centroid. As a result, visually similar hieroglyphs are grouped. However, since the number of clusters to be found (k) is not known, one can use the silhouette coefficient.[29] Applied to the present feature vectors, the silhouette coefficient determines the number of six clusters as optimal.

At the coordinates of each centroid, there is no real hieroglyphic data, but centroid assignment allows the creation of prototypes, which is performed by transforming the centroid coordinates back to the original space of 7744 dimensions. Each artificial prototype represents the members of its cluster and approximates the real-life examples (compare fig. 2 and fig. 3). However, this method also reveals structures that human perception hardly recognizes. Although the algorithm has found the optimum number of clusters, two sets of duplicate clusters can be seen (fig. 3, no. 0+4 and no. 1+3).

This effect can be reduced by using a different algorithm which is called hierarchical clustering.[30] In the beginning, every data point is treated as an independent cluster being agglomerated step-by-step into larger ones. The result is displayed as a dendrogram, which can be used to read off the individual groups and their distance from each other. Considering the full dendrogram (not shown here), it is possible to cut the cluster tree at a specific point uniting the clusters below to a total set of four (fig. 4) where each cluster is visualized by the average mean of its members.[31]

RESULT

Compared to the plot of the PCA, we can see that the general division into two main clusters is measurable with this method. Considering the dendrogram, one can recognize the bipartition of the two main clusters separating at a high distance. The left cluster (blue mark) represents the occurrences of the owl written with only one stroke for the body. The right branch summarizes three subclusters, that represent several morphological forms of the owl with the body executed using two strokes. The members of the green cluster consist of hieroglyphs showing a distinctive left-sided rotation. The red cluster depicts the most detailed hieroglyphic form of the owl whereas the yellow one contains hieroglyphs with a more or less horizontal ground line.

The presented case study shows, that by the use of digital methods, paleography can be enriched with reproducible models, that allow us to investigate the morphological structures of the signs in detail. The prototype generation allows distance measurements and cluster algorithms enabling the objective comparison of the similarity or the dissimilarity of specific signs. Four main types of cursive hieroglyphs could be identified in the dataset. Some of the algorithms lack robustness due to input-related image rotations. Further study is needed to implement more resistant feature descriptors, e.g., *Histogram of Oriented Gradients* (HOG) in combination with vector quantization as it has been successfully tested for Mayan hieroglyphs.[32] Overall, it should be stated that digital methods are not intended to replace traditional working methods, but can usefully complement them, depending on the research question and data availability.

29. Rousseeuw 1987.
30. James et al. 2013, pp. 390–399.
31. An introduction can be found in James et al. 2013, pp. 391–394.
32. Bogacz et al. 2018.

BIBLIOGRAPHY

Bermeitinger et al. 2021
Bermeitinger, B., Gülden, S.A., Konrad, T., "How to Compute a Shape: Optical Character Recognition for Hieratic", in C. Gracia Zamacona, J. Ortiz García (eds.), *Handbook of Digital Egyptology: Texts*, Monografías de Oriente Antiguo 1, Alcalá de Henares, 2021, pp. 121–138. https://doi.org/10.25358/openscience-6757.

Bogacz et al. 2018
Bogacz, B., Feldmann, F., Prager, C., Mara, H., "Visualizing Networks of Maya Glyphs by Clustering Subglyphs", in R. Sablatnig, M. Wimmer (eds.), *Eurographics Workshop on Graphics and Cultural Heritage*, 2018, pp. 105–111, doi:10.2312/gch.20181346.

Chudzik, Caban 2017
Chudzik, P., Caban, M., "Observations on the Architecture of the Tomb of Horhotep in Western Thebes", *EtudTrav* 30, 2017, pp. 221–229.

Ciula 2005
Ciula, A., "Digital Palaeography: Using the Digital Representation of Medieval Script to Support Palaeographic Analysis", *Digital Medievalist* 1, 2005, n. pag., doi:10.16995/dm.4.

Ciula 2017
Ciula, A., "Digital Palaeography: What is Digital about it?", *Digital Scholarship in the Humanities* 32, suppl. 2, 2017, pp. ii89–ii105, doi:10.1093/llc/fqx042.

Egberts 1982
Egberts, A., "The collection de Buck at Leiden", *GöttMisz* 60, 1982, pp. 9–12.

Gardiner 1994
Gardiner, A.H., *Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs*, Oxford, 1994 (3rd, repr. ed.).

Gilliam et al. 2010
Gilliam, T., Wilson, R.C., Clark, J.A., "Scribe Identification in Medieval English Manuscripts", in *Proceedings of the 2010 20th International Conference on Pattern Recognition*, Istanbul, 2010, pp. 1880–1883, doi:10.1109/ICPR.2010.463.

Gülden 2018
Gülden, S.A., "Paläographien und Hieratogramme: digitale Herausforderungen", in S.A. Gülden, K. van der Moezel, U. Verhoeven (eds.), *Ägyptologische „Binsen"-Weisheiten III: Formen und Funktionen von Zeichenliste und Paläographie*, Abhandlungen der Akademie der Wissenschaften und der Literatur in Mainz. Geistes- und Sozialwissenschaftliche Klasse. Einzelveröffentlichung 15, Mainz, 2018, pp. 83–109.

Gülden et al. 2020
Gülden, S.A., Krause, C., Verhoeven, U., "Digital Palaeography of Hieratic", in V. Davies, D. Laboury (eds.), *The Oxford Handbook of Egyptian Epigraphy and Paleography*, New York, 2020, pp. 634–646.

James et al. 2013
James, G., Witten, D., Hastie, T., Tibshirani, R., *An Introduction to Statistical Learning*, Springer Texts in Statistics 103, New York, 2013.

Jolliffe, Cadima 2016
Jolliffe, I.T., Cadima, J., "Principal Component Analysis: a Review and Recent Developments", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374/2065, 2016, p. 20150202, doi:10.1098/rsta.2015.0202.

Kestemont et al. 2017
Kestemont, M., Christlein, V., Stutzmann, D., "Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts", *Speculum* 92/S1, 2017, pp. S86–S109, doi:10.1086/694112.

Lacau 1904
Lacau, P., *Catalogue général des antiquités égyptiennes du Musée du Caire, No. 28001–28086: Sarcophages antérieurs au Nouvel Empire* I, Cairo, 1904.

Lesko 1979
Lesko, L.H., *Index of the Spells on Egyptian Middle Kingdom Coffins and Related Documents*, Berkeley, 1979.

Lit 2020
Lit, L.W.C. van, *Among Digitized Manuscripts: Philology, Codicology, Paleography in a Digital World*, Handbook of Oriental Studies, Section 1: The Near and Middle East 137, Leiden, 2020, doi:10.1163/9789004400351.

Maspero 1885
Maspero, G., "Trois années de fouilles dans les tombeaux de Thèbes et de Memphis", *MMAF* 1, 1885, pp. 133–242.

Morales et al. 2016
Morales, A.J., Falk, S., Osman, M., Casado, R.S., Shared, H., Yamamoto, K., Zidan, E.H., "The Middle Kingdom Theban Project: Preliminary Report on the Freie Universität Berlin Mission to Deir el-Bahari, First and Second Seasons (2015–2016)", *SAK* 45, 2016, pp. 257–282.

Peursen 2010
Peursen, W. van, "Text Comparison and Digital Creativity: An Introduction", in W. van Peursen, E.D. Thoutenhoofd, A. van der Weel (eds.), *Text Comparison and Digital Creativity: The Production of Presence and Meaning in Digital Text Scholarship*, Scholarly Communication: Past, Present and Future of Knowledge Inscription 1, Leiden, 2010, pp. 1–27.

Rousseeuw 1987
Rousseeuw, P.J., "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis", *Journal of Computational and Applied Mathematics* 20, 1987, pp. 53–65.

Stutzmann 2016
Stutzmann, D., "Clustering of Medieval Scripts Through Computer Image Analysis: Towards an Evaluation Protocol", *Digital Medievalist* 10, 2016, n. pag., doi: 10.16995/dm.61.

Willems 1988
Willems, H.O., *Chests of Life: A Study of the Typology and Conceptual Development of Middle Kingdom Standard Class Coffins*, MVEOL 25, Leiden, 1988.

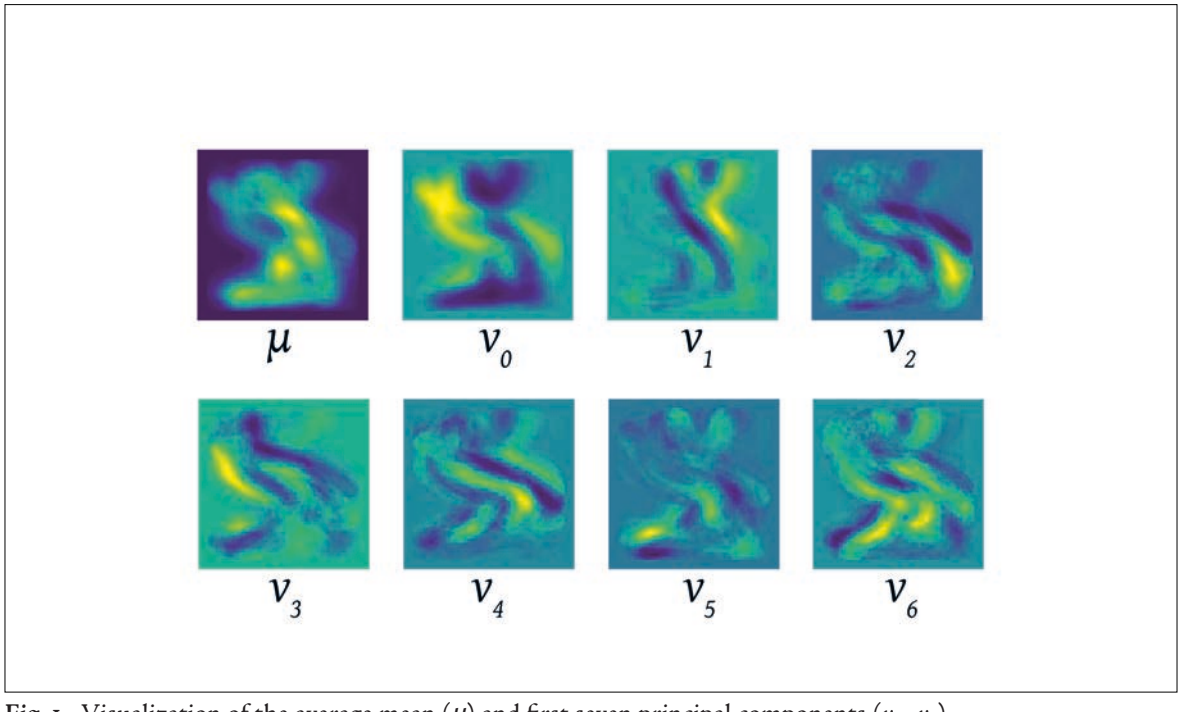

Fig. 1. Visualization of the average mean (μ) and first seven principal components (v_0–v_6).

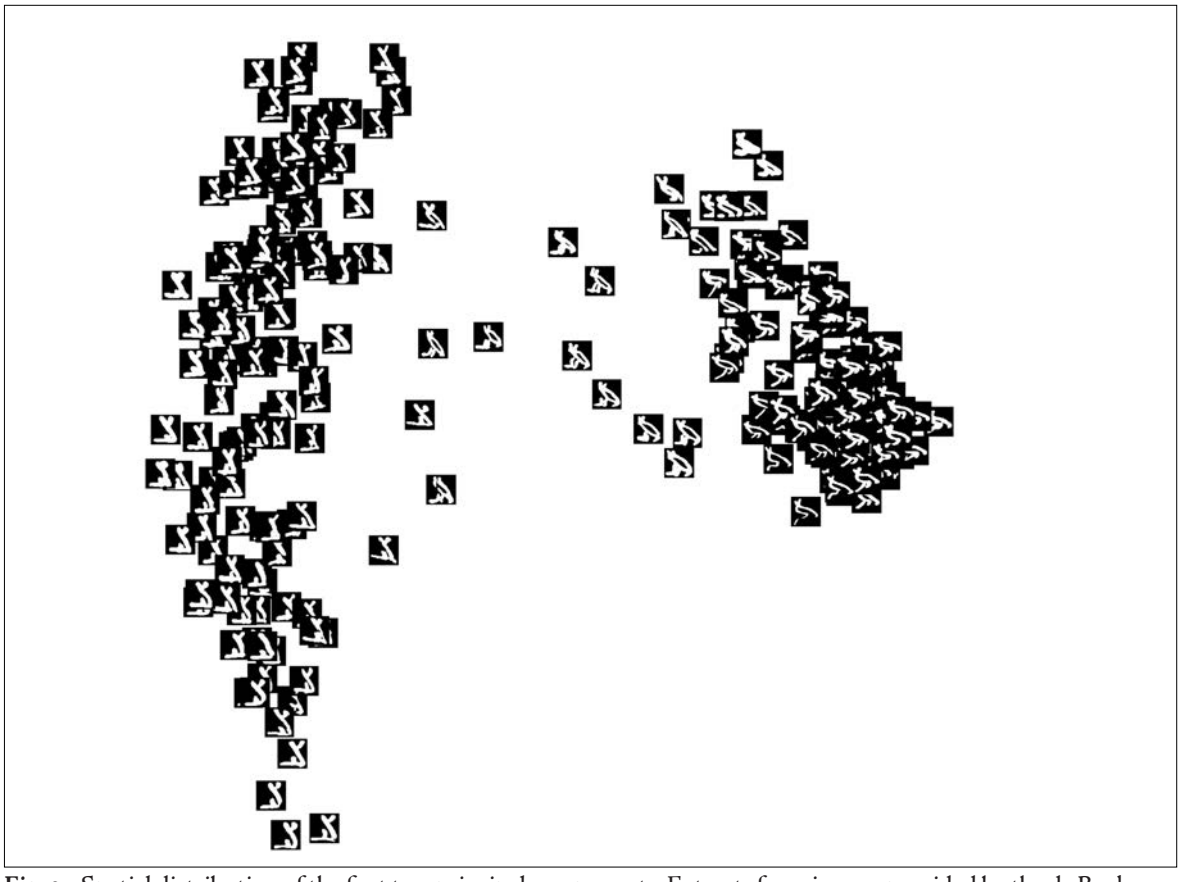

Fig. 2. Spatial distribution of the first two principal components. Extracts from images provided by the de Buck archive at the NINO. Permission of source images kindly provided by The Netherlands Institute for the Near East.

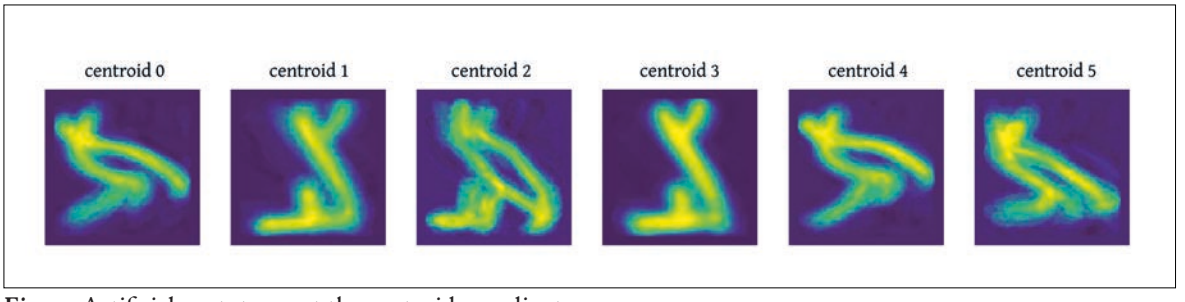

Fig. 3. Artificial prototypes at the centroid coordinates.

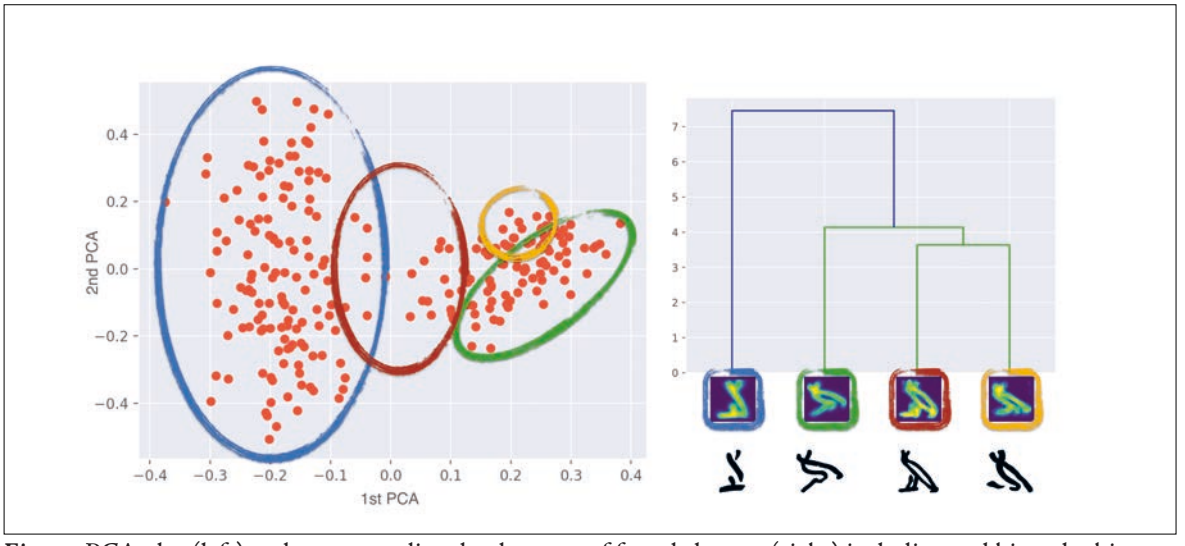

Fig. 4. PCA plot (left) and corresponding dendrogram of found clusters (right) including real hieroglyphic examples below.