University of Bern
Institute of Computer Science and Applied Mathematics
Computer Vision and Artificial Intelligence (FKI)

Description of the Distance Matrices

Barbara Spillmann
spillman@iam.unibe.ch

December 2004

Contents

1 File Format of the Distance Matrices Files (.dm files)

Distance matrices of five different datasets have been calculated and stored: Chicken Pieces, Chromosomes, Toolset, Pendigits and Sea Animals. They were saved as .dm files, whose names are composed of the original database name (db), a non–mandatory normalization value (nv) (see Sections 2 and 4) and the applied cost function (cf) with its constant value (k) as follows:

db[_norm nv]_cf k.dm

The file format of a .dm file for a set of n strings gives information about the type of the strings (angles, vectors, characters,...), the applied cost function (dependent on the string representation), the list of the class names of the n strings, and finally the n×n distance matrix. The .dm files look as follows:

```
.CHARACTERBAND TYPE
string_representation

.COST FUNCTION
cost_function k

.CLASS MEMBERSHIP
list_of_n_classes

.DISTANCE MATRIX
nxn_matrix
```

The possible values of string_representation and where they were used is shown in the following table:

Value of string_representation	Description
CharCharacterband	Chromosomes – difference code (section 3) Chromosomes – absolute code (section 3)
NoRotAngleCharacterband	Pendigits – angle string representation (section 5)
RotInvAngleCharacterband	Chicken Pieces (section 2) Toolset (section 4) Sea Animals (section 6)
VectorCharacterband	Pendigits – vector (segment) string representation (section 5)

The value of `cost_function` can be one of the following:

Value of `cost_function`	Description
`AngleCostFunction`	Pendigits – angle string representation (section 5) Chicken Pieces (section 2) Toolset (section 4) Sea Animals (section 6)
`CharCostFunction`	Chromosomes – difference code (section 3)
`IntCostFunction`	Chromosomes – absolute code, (section 3)
`VectorCostFuction`	Pendigits – vector (segment) string representation (section 5)

Parameter `k` is the constant double-precision value k of the cost function c_k.
The `list_of_n_classes` is a blank-separated list of the classes of the strings in the same order as listed in the distance matrix.
The values `nxn_matrix` of the distance matrices are blank-separated. After `n` values there is a new line.
An example for the Chicken Pieces dataset is:

```
.CHARACTERBAND TYPE
RotInvAngleCharacterband

.COST FUNCTION
AngleCostFunction 60.0

.CLASS MEMBERSHIP
BREAST BREAST BACK BACK WING WING WING ...

.DISTANCE MATRIX
0.0         9.748284    18.034143   14.478384   18.297358   ...
9.748286    0.0         15.302798   13.20809    ...
17.65857    15.745487   0.0         ...
...         ...         ...
```

2 Chicken Pieces Silhouettes Database

2.1 Relevant Information

This dataset consists of 446 images of chicken pieces. Each piece belongs to one of five categories, which represent specific parts of the chicken: wing (117 samples), back (76), drumstick (96), thigh and back (61), and breast (96). Each image is in binary

format containing the silhouette of a particular piece. Pieces were placed in a natural way without considering orientation.

2.2 Extraction of the Strings

Figure 1 shows an example image of the wing class. To extract a string representation out of such binary images, some preprocessing steps had to be done. First, edge detection was performed (Figure 2). Secondly, the edges were approximated by straight line segments of fixed length. Figure 3 shows the results for segment lengths of 7, 10, 15 and 20 pixels. The applied normalization value is in the file name of the .dm file. E.g. *_norm30_*.dm means that the contour has been normalized to segments of length 30. These segments could have been chosen as symbols for the strings. But due to the following two facts, we tried to find a better string representation. As the pieces were placed in a natural way without considering orientation,

1. the figures are rotation invariant, and
2. mirror symmetry occurs.

As a consequence, the sequence of angles between the segments were chosen as the string representation.
Additionally, the approximate algorithm of Bunke and Bühler (BBA) [BB93], which handles rotation invariance and axis symmetry, was applied. Accordingly, the `.CHARACTERBAND TYPE` in the .dm file is named `RotInvAngleCostFunction`.

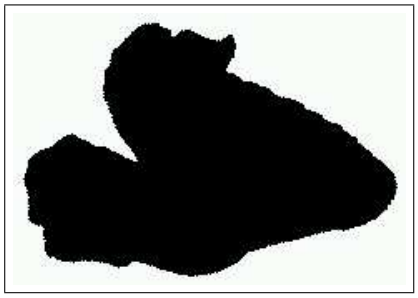


Figure 1: Original binary image

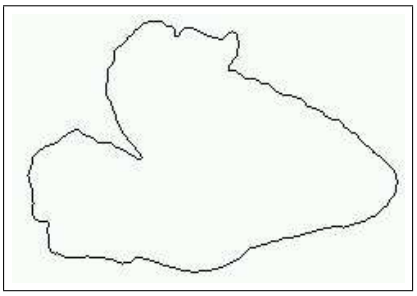


Figure 2: Edge detection

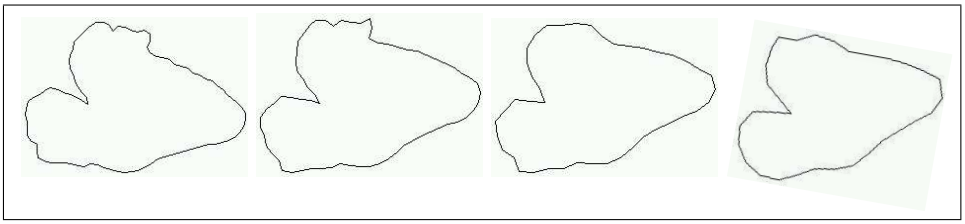


Figure 3: Segment lengths of 7, 10, 15, 20 pixels

2.3 Cost Functions

The cost functions are defined as the angle difference in case of substitution and as a constant k in case of inserting or deleting a symbol. In the following equation, α and β are arbitrary angles, and ε stands for the empty symbol.

$$\begin{aligned} c_k(\alpha \to \beta) &= |\alpha - \beta| \quad \text{(angle difference)} \\ c_k(\varepsilon \to \alpha) &= k \\ c_k(\alpha \to \varepsilon) &= k \end{aligned} \tag{1}$$

Distance matrices were calculated for the values $k = 45, 60, 90$ and 120. The applied function can be found in the .dm file under `.COST FUNCTION`.
`AngleCostFunction 60.0` means that cost function c_{60° has been applied.

2.4 Download and References

The Chicken Pieces Silhouettes Database is available at

http://algoval.essex.ac.uk:8080/data/sequence/chicken/chicken.tgz

The original binary images in the folder /CHICKEN/images were used to calculate the distance matrices.

References are:

- The original source paper: [ACV97]
- Other papers: [MVC00] [PM02], [MVC02b], [MVC02a]

3 Copenhagen Chromosome Database

3.1 Relevant Information and String Extraction

The database consists of 44 files, e.g., dif22da, each consisting of 100 lines of the form `/ 5467 119 22 27 9 / AA==a==E===d==A==a=Aa=A=a=b` where 5467 is a unique chromosome identifier, 119 refers to the metaphase the sample came from (1..180), 22 is the chromosome type, 27 is the overall string length, and 9 is the length of the p-arm, i.e., the centromere position. The slashes are only delimiters and should be ignored, i.e., the alphabet consists of the letters a-f, A-F. The string itself is a difference–coded six level nonlinear profile of the chromosome, whose extraction is illustrated in Figure 4. The difference code is defined as shown in Figure 5.
The dataset partition is fixed, with test and training sets of the same size. The used training set is given by the files ending in a (dif{1|...|22}da), whereas the files belonging to the test set end in b (dif{1|...|22}db). The order of the elements in the

.dm file corresponds to the lexical order of all dif* files [1]. I.e. the first element in the .dm file is the first element listed in dif10da , . . . , the 100^{th} element is the last listed in dif10da , the 101^{st} corresponds to the first listed in dif10db , and so on.

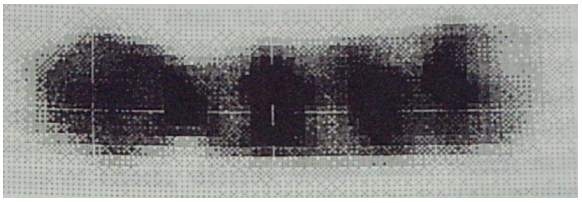
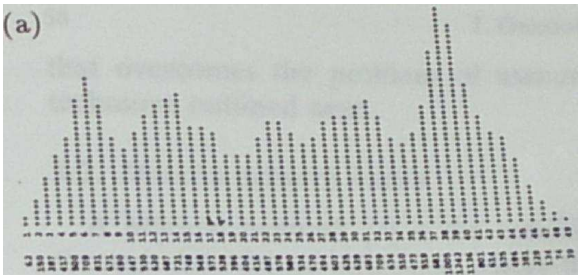
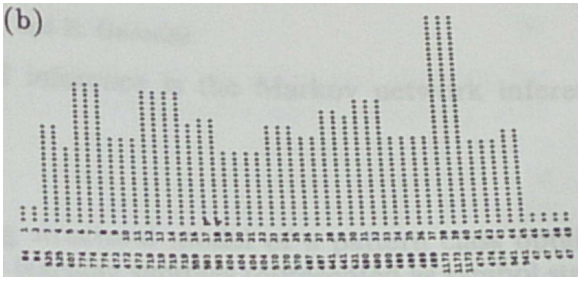
Original chromosome image	
Raw density profile, where peaks correspond to dark bands	
Idealized profile	
Six level string representation of the profile using absolute coding	1133244422233332332222333223323333222266622233111 1
Six level string representation of the profile using difference coding	A=B=aB==b==A===aA=a===A==a=A=aA==a===D==d==A=b===a

Figure 4: String extraction of a chromosome

[1] The lexical order is: dif10da < dif10db < dif11da < dif11db < dif12da < dif12db < dif13da < dif13db < . . . < dif19da < dif19db < dif1da < dif1db < dif20da < dif20db < dif21da < dif21db < dif22da < dif22db < dif2da < dif2db < dif3da < dif3db < dif4da < dif4db < dif5da < dif5db < dif6da < dif6db < dif7da < dif7db < dif8da < dif8db

Difference	-5	-4	-3	-2	-1	0	1	2	3	4	5
Code	e	d	c	b	a	=	A	B	C	D	E

Figure 5: Difference code

3.2 Cost Functions

Cost functions were defined on both the absolute code and the difference code. These functions are

$$\begin{aligned} c_k(x \to y) &= |x - y| \\ c_k(\varepsilon \to x) &= c_k(x \to \varepsilon) = \begin{cases} |x| & \text{if } k < |x|, \\ k & \text{else.} \end{cases} \end{aligned} \tag{2}$$

in case of difference code, and

$$\begin{aligned} c_k(x \to y) &= |x - y| \\ c_k(\varepsilon \to x) &= k \end{aligned} \tag{3}$$

in case of absolute code.

Values for k:
1.5, 2 and 2.5 (diffcode), and 4.5 and 5 (absolute code).

3.3 Download and References

The Copenhagen Chromosome Database is available at

http://algoval.essex.ac.uk:8080/data/sequence/copchrom/chrom.tgz

References:

- The original paper:[LPG80]
- Other papers: [GTG89], [GT90], [GG91]

4 Rutgers University Tool Image Database

4.1 Relevant information

The Rutgers University Tool Image Database is a small set of color images of tools, i.e. it consists of 47 images, of which 8 are brushes, 15 pliers, 12 screws, 2 Swiss army knives, 2 cutters, 3 hammers, 2 wrenches and 3 miscellaneous. Due to its small size, the Toolset Database might not lead to very good results.

4.2 Extraction of the Strings and Cost Functions

The extraction of the strings as well as the definition of the cost functions are exactly the same as for the Chicken Pieces (see Section 2). An example is illustrated in Figures 6 to 9.

Figure 6: Example tool "HAMMER"

Figure 7: Binarized image

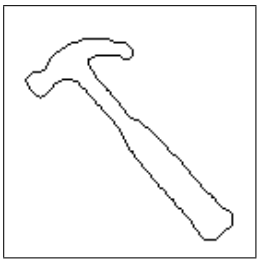


Figure 8: Edge detection

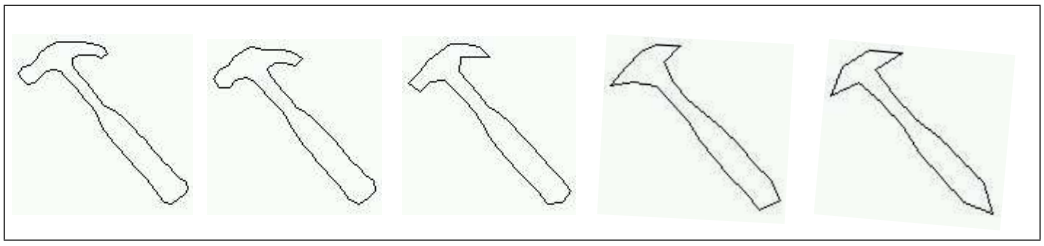


Figure 9: Segment lengths 5, 7, 10, 15, 20 pixels

4.3 Download and References

The Rutgers University Tool Image Database is available at

http://www.cs.rutgers.edu/pub/sven/rutgers-tools/

5 Pen–Based Recognition of Handwritten Digits Original, unnormalized version

5.1 Relevant Information

The unnormalized version of this digit database consists of 250 samples from 44 writers. The samples written by 30 writers are originally used for training, cross–validation and writer dependent testing, and the digits written by the other 14 are used for writer independent testing.

This database is available in the UNIPEN format ([Guy94]).
Distances were calculated for training and test set. The first 7494 elements listed in the .dm file are the elements of the training set (in the same order), the remaining 3498 elements belong to the test set.

5.2 Extraction of the Strings

For the extraction of the strings, a normalization of the curve was done again, analogously to the Chicken Pieces Database. Pen–Ups were ignored.
In one case, the string representation is given by the sequence of normalized segments, and in the other case by the sequence of angles between the segments.
Although there is an analogy to the procedure with the Chicken Pieces, one has to be aware of the fact that the digit strings are not cyclic. Consequently, a different algorithm to calculate the edit distance was applied.

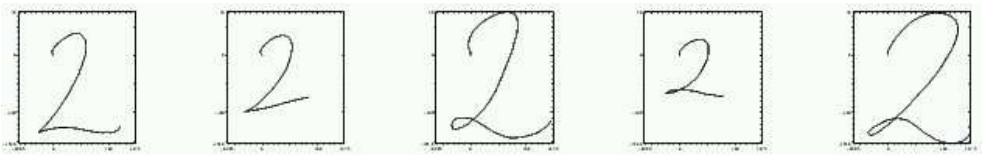


Figure 10: Examples of digit 2

5.3 Cost Functions

The used cost functions are for the angle strings

$$\begin{aligned} c_k(\alpha \to \beta) &= |\alpha - \beta| \quad \text{(angle difference)} \\ c_k(\varepsilon \to \alpha) &= k \\ c_k(\alpha \to \varepsilon) &= k \end{aligned} \tag{4}$$

and for the segments (considered as vectors of a constant length l)

$$\begin{aligned} c_k(\vec{x} \to \vec{y}) &= |\vec{x} - \vec{y}|^k \\ c_k(\varepsilon \to \vec{x}) &= c_k(\vec{x} \to \varepsilon) = 2^{k-1} |\vec{x}|^k = 2^{k-1} l^k \end{aligned} \tag{5}$$

5.4 Download and References

The used database is available at:

Test set:
ftp://ftp.ics.uci.edu/pub/machine-learning-databases/pendigits/pendigits-orig.tes.Z

Training set:

ftp://ftp.ics.uci.edu/pub/machine-learning-databases/pendigits/pendigits-orig.tra.Z

References:

- ftp://ftp.ics.uci.edu/pub/machine-learning-databases/pendigits
- [AA98], [Ali95], [FA96]

6 Sea Animal Database

6.1 Relevant Information

The Sea Animal Database consists of 1100 binary images of sea animals. Some examples are shown in Figure 11. As there are no class labels assigned to the shapes, this dataset might be used for clustering.
Again the elements in the .dm file are lexically sorted by the file name.

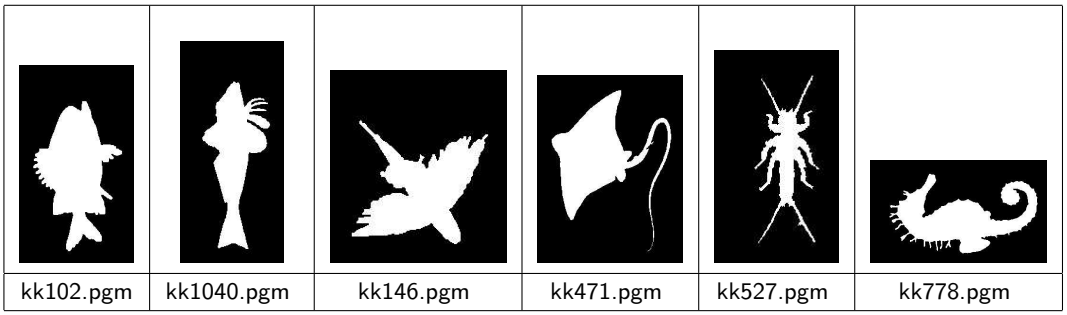


Figure 11: Examples of Sea Animal Database

6.2 Download

The Sea Animal database is available at:

http://www.ee.surrey.ac.uk/Research/VSSP/imagedb/

See also:

http://www.ee.surrey.ac.uk/Research/VSSP/imagedb/demo.html

6.3 Extraction of the Strings and Cost Functions

The extraction of the strings as well as the definition of the cost functions are exactly the same as for the Chicken Pieces (see Section 2) and Toolset (see Section 4).

7 Folder Structure

The structure of the folder is shown in Figure 12.

```
|-- chickenpieces
|   |-- chickenpieces_norm10.0
|   |-- chickenpieces_norm15.0
|   |-- chickenpieces_norm20.0
|   |-- chickenpieces_norm25.0
|   |-- chickenpieces_norm29.0
|   |-- chickenpieces_norm30.0
|   |-- chickenpieces_norm31.0
|   |-- chickenpieces_norm35.0
|   |-- chickenpieces_norm40.0
|   |-- chickenpieces_norm5.0
|   `-- chickenpieces_norm7.0
|-- chromosomes
|   |-- chrom_abscode
|   `-- chrom_diffcode
|-- pendigits
|   |-- pendigits-orig_angle
|   |   |-- orig
|   |   `-- small
|   `-- pendigits-orig_vector
|       |-- orig
|       `-- small
|-- seaanimals
|   |-- seaanimals_norm10
|   |-- seaanimals_norm15
|   |-- seaanimals_norm20
|   |-- seaanimals_norm25
|   `-- seaanimals_norm30
`-- toolset
    |-- toolset_norm10.0
    |-- toolset_norm15.0
    |-- toolset_norm20.0
    |-- toolset_norm25.0
    |-- toolset_norm5.0
    `-- toolset_norm7.0
```

Figure 12: Folder structure

Chickenpieces Separated with respect to the normalization value.

Chromosomes One folder for the difference code, one for the absolute code.

Pendigits One folder for each angle representation and vector representation. `orig` contains the distance matrix of the original database, whereas `small` is a subset of the training set (700 digits).

Toolset Separated with respect to the normalization value.

Sea Animals Separated with respect to the normalization value.

References

[AA98] E. Alpaydin and F. Alimoglu.
Department of Computer Engineering
Bogazici University, 80815 Istanbul Turkey,
September 1998.

[ACV97] G. Andreu, A. Crespo, and J.M. Valiente. Selecting the toroidal self-organizing feature maps (TSOFM) best organized to object recognition. *Proceedings of ICNN'97*, 2:1341–1346, June 1997. Houston, Texas (USA). IEEE.

[Ali95] F. Alimoglu. Combining multiple classifiers for pen-based handwritten digit recognition. Master's thesis, Institute of Graduate Studies in Science and Engineering, Bogazici University, 1995.
http://www.cmpe.boun.edu.tr/∼alimoglu/alimoglu.ps.gz

[BB93] H. Bunke and U. Bühler. Applications of approximate string matching to 2D shape recognition. *Pattern Recognition*, 26(12):1797–1812, 1993.

[FA96] E. Alpaydin F. Alimoglu. Methods of combining multiple classifiers based on different representations for pen-based handwriting recognition. *Proceedings of the Fifth Turkish Artificial Intelligence and Artificial Neural Networks Symposium (TAINN 96)*, 1996.
http://www.cmpe.boun.edu.tr/∼alimoglu/tainn96.ps.gz

[GG91] J. Gregor and E. Granum. Finding chromosomes centromeres using band pattern information. *Comput. Biol. Med.*, 21(1/2):56–57, 1991.

[GT90] E. Granum and M. G. Thomason. Automatically inferred Markov network models for classification of chromosomal band pattern structures. *Cytometry*, 11:26–39, 1990.

[GTG89] E. Granum, M. G. Thomason, and J. Gregor. On the use of automatically inferred Markov networks for chromosome analysis. In C Lundsteen and J Piper, editors, *Automation of Cytogenetics*, pages 233–251. Springer-Verlag, Berlin, 1989.

[Guy94] I. Guyon. UNIPEN 1.0 Format Definition, 1994.
ftp://ftp.cis.upenn.edu/pub/UNIPEN-pub/definition/unipen.def

[LPG80] C. Lundsteen, J. Phillip, and E. Granum. Quantitative analysis of 6985 digitized trypsin G-banded human metaphase chromosomes. *Clinical Genetics*, 18:355–370, 1980.

[MVC00] R.A. Mollineda, E. Vidal, and F. Casacuberta. Efficient techniques for a very accurate measurement of dissimilarities between cyclic patterns. *SSPR&SPR 2000, LNCS 1876*, pages 337–346, 2000.

[MVC02a] R.A. Mollineda, E. Vidal, and F. Casacuberta. Cyclic sequence alignments: Approximate versus optimal techniques. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 16(3):291–299, 2002. World Scientific.

[MVC02b] R.A. Mollineda, E. Vidal, and F. Casacuberta. A windowed weighted approach for approximate cyclic string matching. *Proceedings of Int. Conf. on Pattern Recognition 2002*, IV:188–191, August 2002. Quebec city, Canada. IEEE.

[PM02] G. Peris and A. Marzal. Fast cyclic edit distance computation with weighted edit costs in classification. *Proceedings of Int. Conf. on Pattern Recognition 2002*, IV:184–187, August 2002. Quebec city, Canada. IEEE.