Digitizing of Analogue Data-Charts from Thermohygrographs

Jan Radon

Engineering Consulting & Software Development
ul. Balicka 140D
30-149 Kraków/Poland
jradon@kki.pl

Abstract. *A thermo-hygrograph is a chart recorder that measures and records both temperature and humidity in an analogue form. Converting the results into digital form is necessary to enhance evaluation of measurements archived on paper charts. Coordinate system of the charts is different than common Cartesian rectangular axis. Temperature and relative humidity axis are curved according to recorder pen holder ray. Different chart types, depending on particular thermohygrograph model, were used so far. In every case, however, the axis grid depicted in the background is geometrically precisely defined. So pattern recognition techniques could be used to find its location automatically. Temperature and relative humidity patterns can be retrieved automatically assuming that the color group of plotted lines is different than the background including the grid. Due to storing conditions, manual descriptions, discoloration and other factors, the plotted pattern cannot be unambiguously identified. Many randomly scattered pixels are interpreted as being part of a plotted line. Among additional measures, like excluding particular areas from analysis, manually removing and including points, a vector analysis of retrieved courses makes digitizing more effective.*
Keywords: *thermohygrograph chart, digitizing, software algorithm .*

1. Introduction

In the cultural heritage sector the traditional use of thermo-hygrographs is still standard. The data charts are valuable for instant inspection of the actual temperature and relative humidity, but cannot be easily evaluated for logistical reasons since the conservator has to check and collect the sheets on a daily or weekly or monthly basis. Many current and historical data charts have been archived and are not effectively used for scientific analysis. Precise evaluation of microclimate parameters is only possible if measured parameters are available in continuous digital form.

On the internet a few digitizing programs for converting hard copy graphs and plots back to data can be found [1]. They mostly refer to plots with Cartesian, rectangular coordinate system. The algorithm applied for digitizing is not described. However, the way of their usage suggest, that digitizing is based on color definition of plotted lines. Axis have to be located manually by pointing using mouse and defining the scaling parameters. Additionally all programs enable inserting missing points by mouse clicking or line tracing like using digitizer tablet.

The edge following technics [2] was recently used for automatic digitization of pluviograph strip charts [3]. The authors estimated that in about 76% of test cases the detection of the curve was accurate enough for direct storage of the resulting precipitation intensities. For the rest of the cases, an interactive tool that would allow some manual adjustment of the rainfall curve was needed.

The very few available computer programs for digitizing thermohygrograph charts, like "Gscanth" [4] indeed enhance digitizing in comparison to digitizer tablet. Nevertheless they are limited to particular paper chart type and require lot of manual work both for axis location and extracting of curves.

Against this background a computer program called DigitChart (Digitizing Chart) has been developed. Great development effort was made to make digitizing of charts as simple as possible. The software should be highly intuitive, with short training time. As the software tries to make most of the work automatically, only a few mouse clicks are needed to digitize, if charts are of good quality. The paper presents an algorithm used for digitizing scanned thermo-hygrograph charts applied by the software.
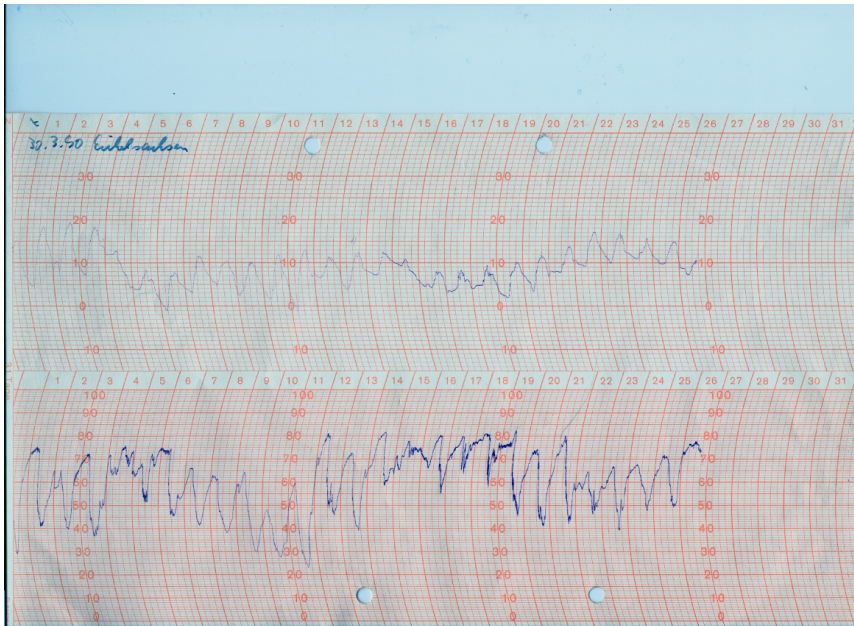
Figure 1: Example of plotted thermo-hygrograph chart.

2. Materials and methods

Scanned images, so called raster graphics, include only matrix of colored points (pixels) without any functional connection. The software has to retrieve both pixels, which are part of plotted patterns, as well as location of grid (axis system) to be able to rescale the pixel position to real temperature and relative humidity. The algorithm is built on the assumption that drawn lines have different color groups than the background and the grid. The most common way to store colors is Red, Green, Blue coding (RGB). The second parameter needed for picture analyzing is DPI (dots per inch). Real physical size (in inches or cm) can be obtained using this parameter.

Some well-known technics to recognize pictures were used so far [5]. Due to discontinuities, discoloration and color mixing by crossing grid and plotted line edge following and edge recognition were time consuming and not very effective. Grouping of colors with a different number of subranges turned out to be a viable solution. Low division by 3, i.e. 85 range for every RGB part proved very effective.
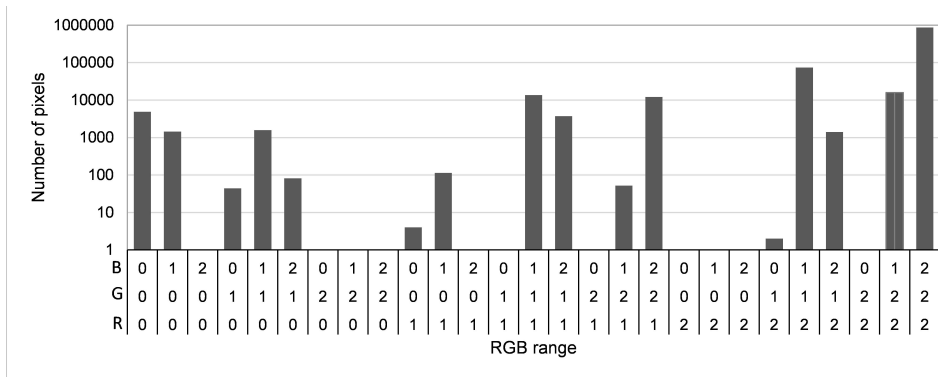
Figure 2: Number of pixels of RGB color range (in logarithmic scale). The color range is $[n \cdot 85 - (n + 1) \cdot 85] \cdot (R, G, B))$, *n*-number below the diagram.

The best results could be obtained with good quality scans with resolution of 150 – 200 DPI and no color reduction. The exemplary scan (Fig. 1) has 47901 unique colors Fig.2. shows pixel numbers according to RGB range for the exemplary chart (see Fig. 1).

Fig. 2. shows that one color group includes about ten times more pixels than any other color group. This group is regarded representative for the background. The remaining groups include axis, plotted lines and other elements (e.g. description). This regularity can be observed in all thermo-hygrograph charts.

To locate the coordinate system automatically a virtual grid is first created in memory. There are a number of different layouts used for the chart axis system. Temperature and relative humidity may have constant or variant interspaces. They can be scaled from bottom to top or from top to bottom. Both diagrams can be scaled differently. Only time is always scaled linearly, mostly for 1, 7 or 31 days. To ensure covering for most chart types detailed information about chart layout must be defined. Fig. 3. shows data input required for any chart. It seems to be a lot of work. However only a few or only one chart type is usually used at one place. So the input is done once and used many times.

To obtain the radius of pen arm and other parameters, including time shift caused by axis curvature, some simple recalculations must be made (Fig. 4). Once the virtual grid is ready it is moved from top to bottom over the scanned picture and pixels with colors excluding the background are calculated. Only horizontal
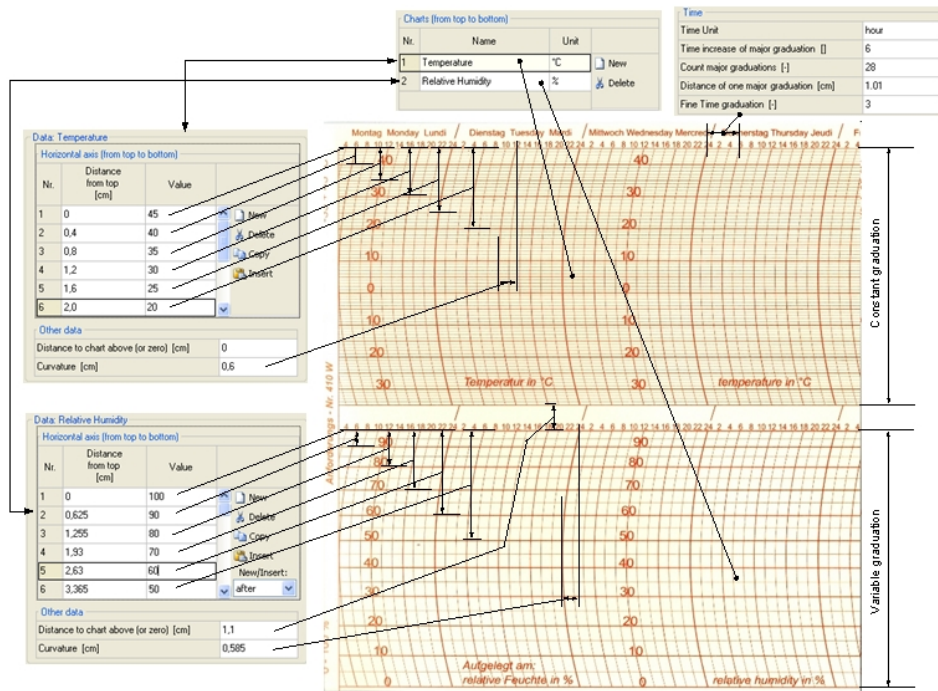
Figure 3: Information necessary to create virtual layout of any chart type.

lines of the virtual grid are regarded. Fig. 5. shows the number of pixels dependent on top distance obtained in this way. The highest number of pixels points to top offset of the grid matching best the background. Obviously pixels not belonging to the grid are also added, but their relatively small number does not change this general rule.

The left offset can be obtained in a similar way. Fig 6. shows the number of pixels obtained when moving the virtual pattern horizontally, from left to right. This time horizontal axis lines are omitted. The grid is vertically already positioned with vertical offset obtained previously.

To save calculation time only a few cm around a particular chart's vertical center can be taken. Due to not optimal storing or measurement environmental conditions some paper deformation can take place. So 2 pixels around virtual grid in every direction (± 2) were added. If geometry parameters of charts are precisely specified automatic positioning works very well.
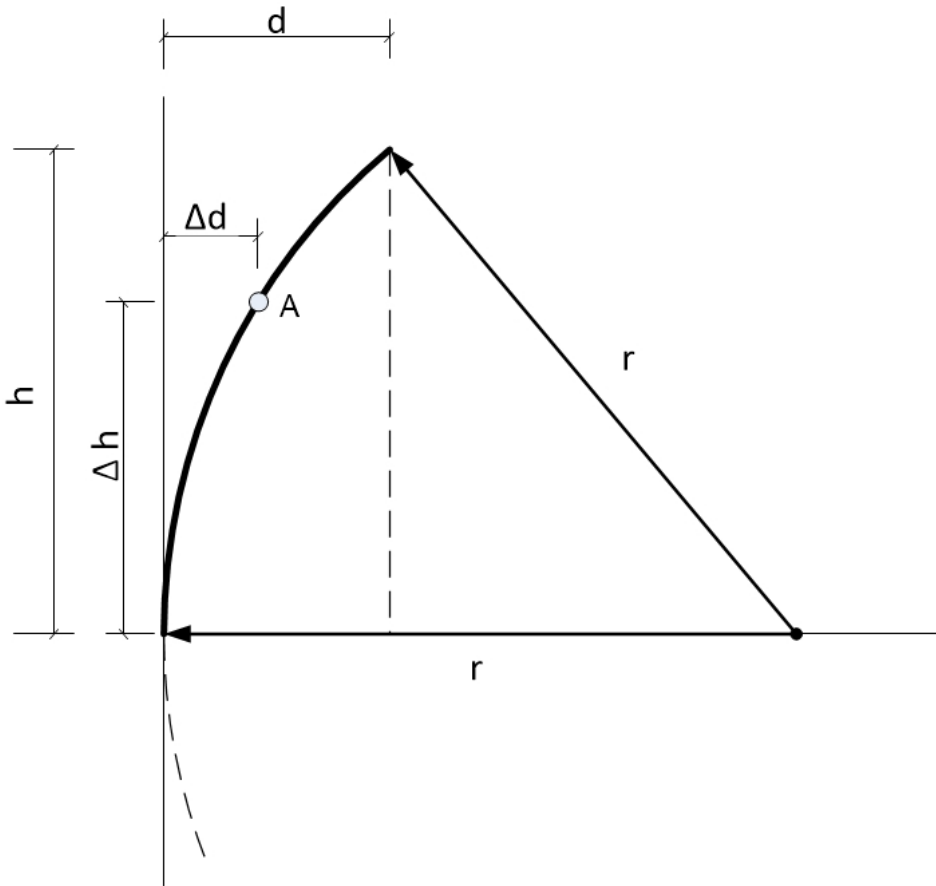
Figure 4: Geometry and formulas used to generate virtual coordination system.

$$h^2 + (r - d)^2 = r^2 \tag{1}$$

$$r = \frac{h^2 + d^2}{2d} \tag{2}$$

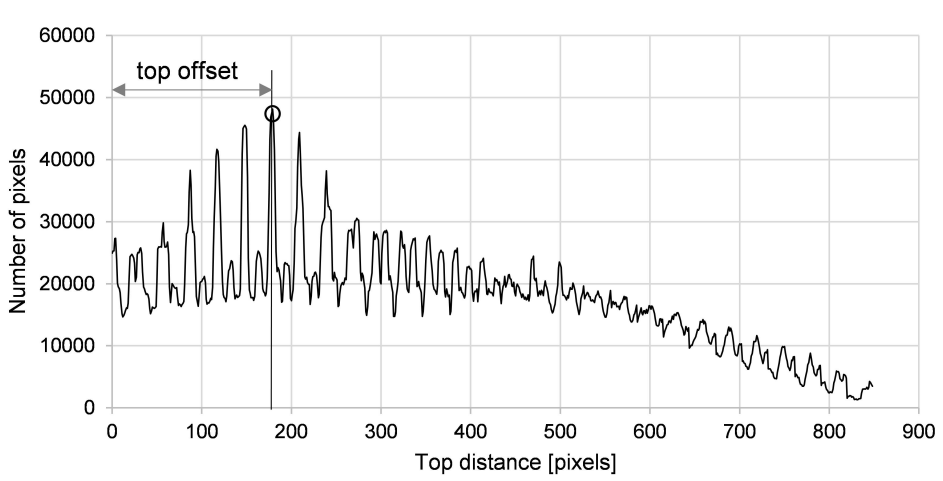$$\Delta d = r - \sqrt{r^2 - \Delta h^2} \tag{3}$$

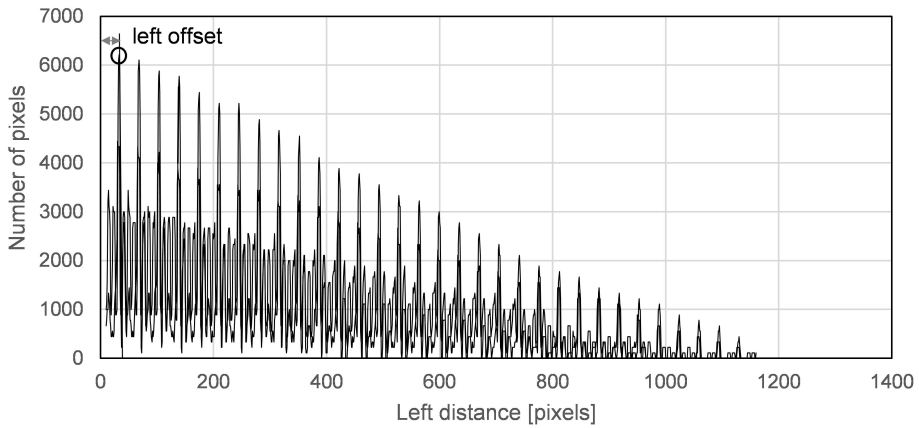Figure 5: Number of pixels below virtual grid with other then the background colors depending on top offset.

Figure 6: Example of plotted thermo-hygrograph chart.

After the position of the coordinate system is located it's time for retrieving plotted temperature and relative humidity patterns. Pens used in thermo-hygro-
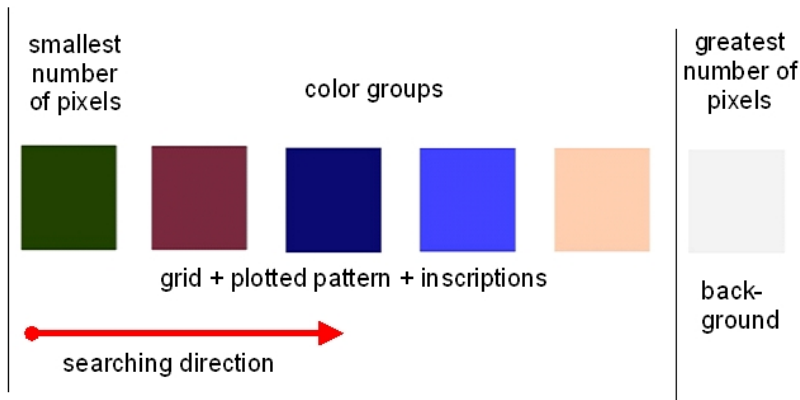
Figure 7: Searching for pixels within plotted pattern.

graphs always have a different color than the chart grid, even if it looks similar. So the idea is to order color groups from the smallest number of pixels to the highest. Than starting from the group with the smallest pixel number move towards groups with the greatest pixel number depicting pixels belonging to selected range (Fig. 7). The process must be visually controlled by the user. Particular pixels are ordered according to time taking into account time shift resulting from axis curvature (see Fig. 4). Unfortunately the effect is usually not rewarding. Even sharp color differences alone do not guarantee clear differentiation between pattern, grid and the background. Often gridline crossings or inhomogeneous background (dust, descriptions) are interpreted as a drawn pattern. Also due to storing conditions, discoloration and other factors, the plotted pattern cannot be unambiguously identified. Many randomly scattered pixels are interpreted as being part of the plotted line. Additional measures have to be applied to filter out unwanted pixels. The simplest method is to exclude some areas from searching for pixels or zooming out of total search area to the smallest range defined by minimal and maximal temperature or RH values.

Excluding areas from pixel search yields a better outcome but still many single pixels cause annoying jumps out of plotted lines (Fig. 8). They still can be excluded by clicking or drawing excluding rectangles, but this means a lot of additional work and concentration. Therefore another solution was applied – vector analysis. Due to mechanical thermo-hygrograph inertia and analysis of real changes in temper-
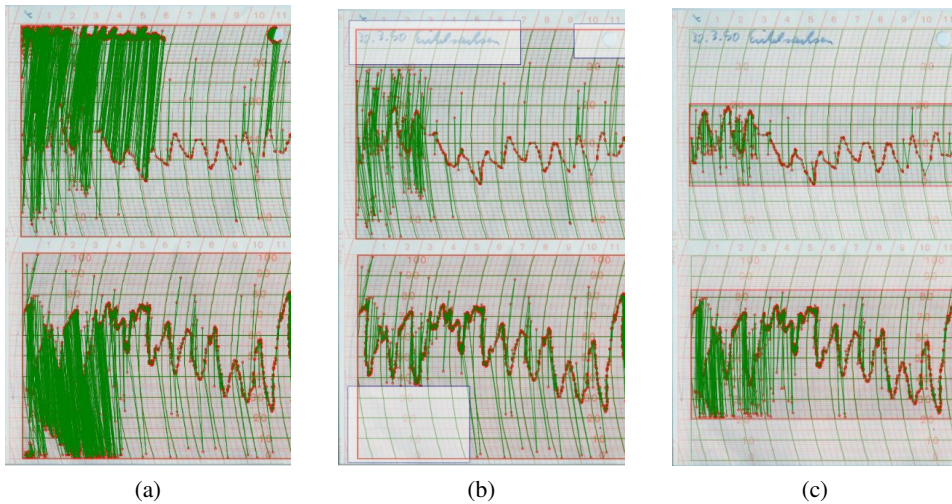
Figure 8: Simple way to exclude unwanted pixels, a-original pattern, b-excluding areas, c-zooming out of total search area.

ature and relative humidity some patterns can be excluded, e.g. a sudden jump to and from a single pixel. This can be checked by means of dot product. If the angle between lines touching at particular pixel is smaller than 5o and value jump higher than 20% of vertical axis range the pixel is ignored. The angle and amplitude were obtained experimentally and so can be further optimized.

Applying vector analysis, zooming out of search area and setting excluding rectangles makes the digitizing procedure very effective. Fig. 9. shows the digitizing of exemplary chart. Temperature and relative humidity are finally obtained by averaging all pixel-values within 1 hour time-step. As can be seen still manual intervention is necessary but it is reduced to a necessary minimum.

3. Summary and Conclusion

Digitizing of thermo-hygrograph charts cannot be fully automated, but user input can be reduced to a few mouse clicks. Chart geometry data must be exactly defined to enable automatic location of coordinate system and precise rescaling
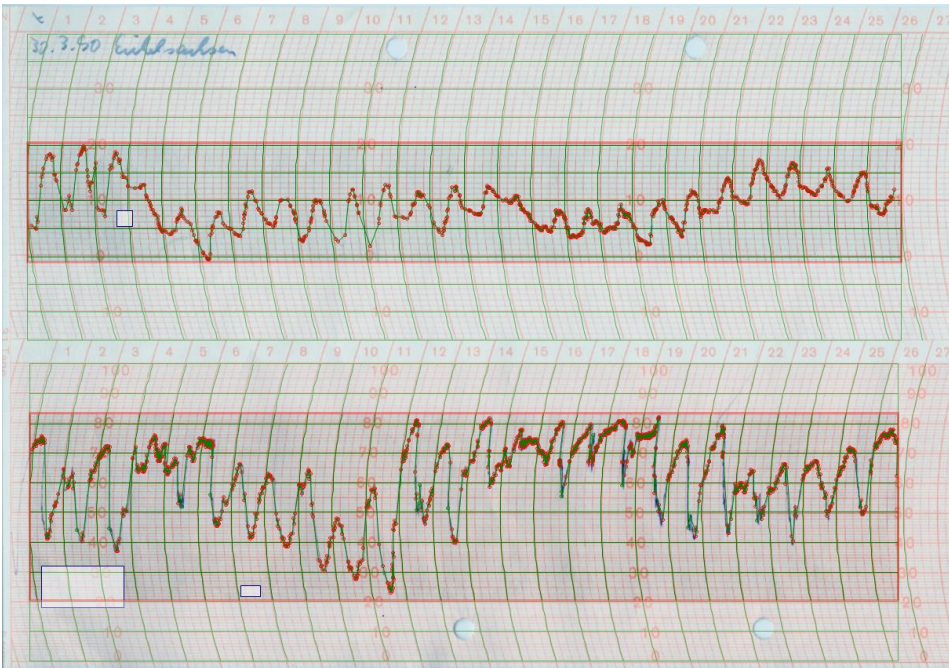
Figure 9: Final effect of digitization, mean hourly values.

of temperature and relative humidity including time shift. Application of vector analysis proved very effective in digitizing process.//

   The proposed algorithm does not define any specific color for background, grid or pens. Instead grouping colors is used. Hitherto, charts with beige and bluish grid colors were tested with satisfying outcome. So presumably the algorithm will work with any grid color.
   The best format supported by the scanner should be used, provided that RGB color coding and DPI information is included. A good quality picture should be obtained by using the lowest possible resolution. A higher resolution is not impediment but requires more computational time.
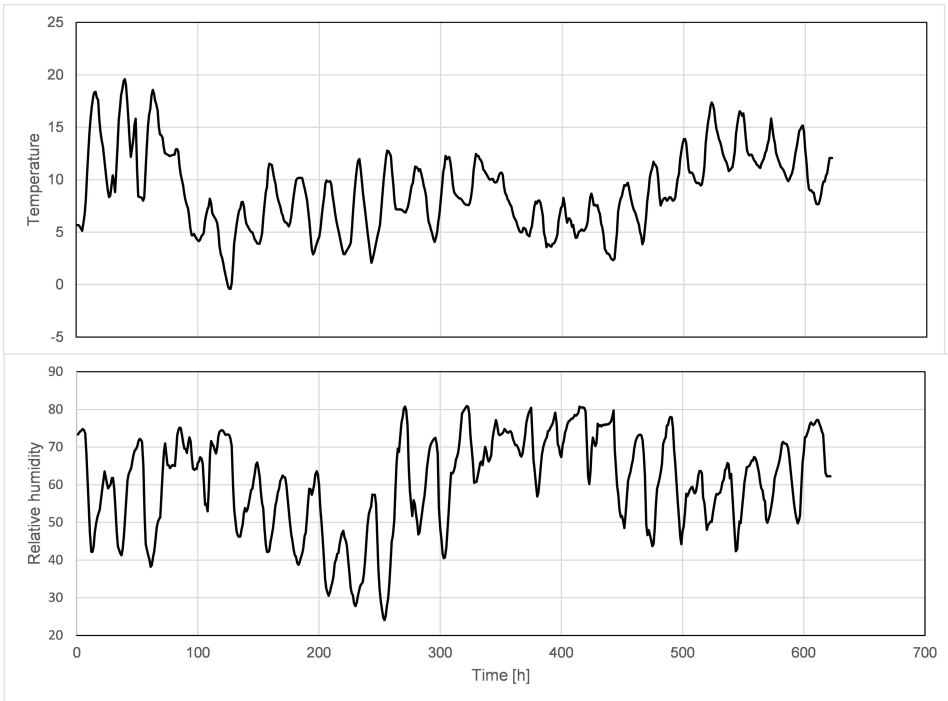
Figure 10: Final effect of digitization, mean hourly values.

3.1. Acknowledgement

References

[1] *Digitizing programs for Converting Hard Copy Graphs and Plots back to Data*, accessed 30.06.2016.

[2] Ziou, D. and Tabbone, S., *Edge Detection Techniques - An Overview*, International Journal of Pattern Recognition and Image Analysis, 1998.

[3] A. Jaklic, L. Sajn, G. D. P. P., *Automatic digitization of pluviograph strip charts, Meteorological Applications*, Meteorological Applications 23, 2016, pp. 57–64.

[4] *Gscanth software*, Internet, http://dufters.de/Alfred/Xari/gscanth/, accessed 12.07.2016.

[5] Jahne, B., *Digital Image Processing*, Springer-Verlag, 2002.