Chapter 1

Handwriting Recognition Engine

1.0.1 Recognition of Straight Strokes

1.0.1.1 Identification of a Straight Stroke

The notion of *stroke* essentially means *point sequence* in a technical sense. That is, strokes are not mathematical objects like functions. In order to perform stroke recognition, it is necessary to determine what shape a stroke has. For the identification of a straight stroke, a linear regression line is constructed from the point sequence with the Gaussian method of least squares. The construction is done with the standard method for calculating least squares, as described in literature. The method seeks for the minimal sum of squares of residuals. If the sum of residual squares is below a certain threshhold that has been determined empirically, the system assumes that the point sequence generally follows a straight line. In that case, straight stroke matching will be performed. If the line appears to be curved, it will be matched with curve analysis methods as described in section **??**.

1.0.1.2 Straight Stroke Matching

For the matching of a straight stroke there are a number of features that could be considered for a feature vector. In order to describe a straight stroke and then match it with another one, a feature vector is constructed that has the following features:

- **Length**: The total lenght of the line. Sum of the distances of succeeding points.
- **Initial point**: The coordinates of the initial point
- **Endpoint**: The coordinates of the endpoint
- Total number of measured points: The number of sample points that were measured by the input device
- **Velocity**: The velocity with which the stroke was drawn
- **Gradient/Direction**: The slope of the linear regression line that can be constructed for the point sequence, or the general direction of the line, represented as a numerical vector.

However, some of the information seem redundant. In order to match two strokes by their feature vectors F_{DB} and F_{exp}, the following values are conidered:

$$F := \begin{pmatrix} \text{Length} & l \\ \text{Initial point} & p_I \\ \text{Endpoint} & p_E \\ \text{Velocity} & v \\ \text{Direction vector} & \vec{d} \\ \text{Sum of residual squares} & \sigma \end{pmatrix}$$

The features are used for matching with other straight strokes by comparison of their values. The length l is calculated as the sum of the euclidian distances d_i between the individual points p_i and p_{i-i} within a sequence of N points.

$$d_i := \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$$
$$l := \sum_{i=1}^{N-1} d_i$$

In the case of the initial point and endpoint comparison of the two strokes, the euclidian distance between the original point $P_{I,DB}, P_{E,DB}$ and the candidate point $P_{I,exp}, P_{E,exp}$ is used as a comparison value. The velocity v is calculated from the time difference δ_T between sampling time of the initial point t_{p_I} and the endpoint t_{p_E}:

$$\delta_T := t_{p_E} - t_{p_I}$$

$$v := \frac{n_p}{\delta_T}$$

The velocities v_{DB} and v_{exp} are both real numbers and can be compared directly. The direction vectors of the regression line $\vec{d}_{DB}$ and $\vec{d}_{exp}$ are compared by the deviation of their directions. The deviation is expressed as the angle α between the original and the candidate vector.

$$\alpha := \vec{d}_{DB} \quad \angle \quad \vec{d}_{exp}$$

The sum of residual squares σ is calculated from the measured point sequence with respect to the regression line of the database point sequence. Given the regression line f_{DB}, the sum of the squares of the residuals for the input sequence is:

$$\sigma := \sum_{i=1}^{N} (f_{DB}(x_{i,exp}) - x_{i,exp})^2$$

The smaller σ the closer the measured points to the regression line from the database. This is only done for strokes that have been calculated to be straight, since any set of points can have a regression line, even curved ones. In the case of curved strokes, a linear regression line as it is used here does not seem to bear the potential helping identify a point seqence to be a certain shape.