Statistical pattern recognition for automatic writer identification and verification
Bulacu, Marius Lucian

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2007

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Part I

Texture-Level Approach
Chapter 2

Writer Identification Using Edge-Based Directional Features

Science is what we understand well enough to explain to a computer. Art is everything else we do.

Donald Knuth

Abstract

This chapter evaluates the performance of edge-based directional probability distributions as features in writer identification in comparison to a number of other texture-level features encoding non-angular information. We introduce here a new feature: the joint probability distribution of the angle combination of two "hinged" edge fragments. It is noted that the "edge-hinge" distribution outperforms all other individual features. Combining features yields improved performance. Limitations of the studied global features pertain to the amount of handwritten material needed in order to obtain reliable distribution estimates. A stability test is carried out showing the dependence of writer identification accuracy on the amount of handwritten material used for feature extraction.

2.1 Introduction

In the process of automatic handwriting recognition, invariant representations (features) are sought which are capable of eliminating variations between different handwritings in order to classify the shapes of characters and words robustly. The problem of writer identification, on the contrary, requires a specific enhancement of the variations which are characteristic to a writer’s hand. At the same time, such representations or features should, ideally, be independent of the amount and the semantic content of the written material. Slant represents a very stable characteristic of individual handwriting and gives the distinctive visual appearance of a handwritten text block under a general succinct view. The slant angle corresponds to the dominant direction in the handwritten script. In this chapter we will use the complete probability distribution of directions in the ink trace for writer identification. This distribution will be computed using edge fragments along the written ink. The edge-direction distribution will
constitute the building block for designing more complex features yielding increased performance.

Three groups of features can be identified in forensic writer identification:

- global measures, computed automatically on a region of interest (ROI)
- local measures, of layout and spacing features entered by human experts
- measures related to individual character shapes

We reiterate that, in this thesis, we analyze only features that are automatically extractable from the handwriting image without any human intervention. Furthermore, it is assumed that a crisp foreground/background separation has been realized in a pre-processing phase, yielding a white background with near-black ink. As a rule of thumb, in forensic writer identification one strives for 100% recall of the correct writer in a hit list of 100 writers, computed from a database of more than $10^4$ samples. This amount is based on the pragmatic consideration that a number of one hundred suspects is just about manageable in criminal investigation. Current systems are not powerful enough to attain this goal.

As regards the theoretical foundation of our approach, the process of handwriting consists of a concatenation of ballistic strokes, which are bounded by points of high curvature in the pen-point trajectory. Curved shapes are realized by differential timing of the movements of the wrist and the finger subsystem (Schomaker et al. 1989). In the spatial domain, a natural coding, therefore, is expressed by angular information along the handwritten curve (Plamondon and Maarse 1989). It has long been known (Maarse and Thomassen 1983, Maarse et al. 1988, Crettez 1995) that the distribution of directions in handwritten traces, as a polar plot, yields useful information for writer identification or coarse writing-style classification. It is the goal of this chapter to explore the performance of angular-distribution directional features, relative to a number of other features which are in actual use in forensic writer-identification systems. The edge-based probability distributions operate at the scale of the ink-trace width, they give a texture-level view of the handwritten sample and they are informative, in general, about the habitual pen grip and biomechanical makeup of the writing hand.

### 2.2 Experimental data

We evaluated the effectiveness of different features in terms of writer identification using the *Firemaker* dataset (Schomaker and Vuurpijl 2000). A number of 250 Dutch subjects, predominantly students, were required to write four different A4 pages. On page 1 they were asked to copy a text presented in the form of machine print characters. On
2.3 Features

In this section we describe the extraction methods for five texture-level features used in writer identification. The first two features are edge-based directional distributions. We will focus our attention on the second one of them which is a new feature proposed and analyzed by us in recent publications.

2.3.1 Edge-direction distribution

Feature extraction starts with conventional edge detection (convolution with two orthogonal differential kernels, we used Sobel, followed by thresholding) that generates a binary image in which only the edge pixels are "on". We then consider each edge pixel in the middle of a square neighborhood and we check (using logical AND oper-
Figure 2.2: Two handwriting samples from two different subjects. We superposed the polar diagrams of the edge-direction distribution $p(\phi)$ corresponding to pages 1 and 4 contributed to our dataset by each of the two subjects. There is a large overlap between the directional distributions extracted from samples originating from the same writer, while there is a substantial variation in the directional distributions for different writers.

In order to avoid redundancy, the algorithm only checks the upper two quadrants in the neighborhood because, without on-line information, we do not know which way the writer "traveled" along the found oriented edge fragment.

In the experiments, we considered 3, 4 and 5-pixel long edge fragments. Their orientation is quantized in $n = 8, 12$ and 16 directions respectively (Fig. 2.1 is an example for $n = 12$). Clearly, $n$ is also the number of bins in the histogram and the dimensionality of the final feature vector.

The distribution of the writing directions is characteristic of a writer’s style. The polar probability density function was used in an on-line study of handwriting (Maarse and Thomassen 1983) to describe differences between upward and downward strokes. It was also used off-line (Crettez 1995) as a preliminary step to handwriting recognition that allows a partition of the writers by unsupervised fuzzy clustering in different groups.

While in the mentioned studies the directional histogram was computed on the written trace itself, for the present work we computed it based on the edges. Edges follow...
the written trace on both sides and they are thinner, effectively reducing the influence of trace thickness.

We must mention an important practical detail: our generic edge detection does not generate 1-pixel wide edges, but they can usually be 1-3 pixels wide and this introduces smoothing into the histogram computation because the ”probing” edge fragment can fit into the edge strip in a few directions around a central main direction. This smoothing taking place in the pixel space has been found advantageous in our experiments.

As can be noticed in Fig. 2.2, the predominant direction in $p(\phi)$ corresponds, as expected, to the slant of writing. Even if idealized, the example shown can provide an idea about the "within-writer" variability and "between-writer" variability in the feature space.

By analyzing the data, we found out that differentiation of the feature vector ($dp(\phi)$) results in a significant performance improvement. Besides removing the DC component, the differentiated directional probability distribution conveys information about the changes in writing direction. Along this line of thinking came the idea of a more complex feature capable of bringing forth more information about the local writer specificities by computing locally on the image the probability distribution of changes in direction.

### 2.3.2 A new feature: edge-hinge distribution

Our goal is to design a feature characterizing the changes in direction undertaken during writing with the hope that it will be more specific to the writer and consequently making possible more accurate identification. The method of feature extraction is similar to the one previously described, but it has added complexity. The central idea is to consider in the neighborhood, not one, but two edge fragments emerging from the central pixel and, subsequently, compute the joint probability distribution of the orientations of the two fragments.

To have a more intuitive idea of the feature that we are proposing, imagine having a hinge laid on the surface of the image. Place its junction on top of every edge pixel, then open the hinge and align its legs along the edges. Consider then the angles $\phi_1$ and $\phi_2$ that the legs make with the horizontal and count the found instances in a two dimensional array of bins indexed by $\phi_1$ and $\phi_2$. The final normalized histogram gives the joint probability distribution $p(\phi_1, \phi_2)$ quantifying the chance of finding in the image two ”hinged” edge fragments oriented at the angles $\phi_1$ and $\phi_2$.

As already mentioned, in our case edges are usually wider than 1-pixel and therefore we have to impose an extra constraint: we require that the ends of the hinge legs should be separated by at least one "non-edge" pixel. This makes certain that the hinge is not positioned completely inside the same piece of the edge strip. This is an important
2. Writer Identification Using Edge-Based Directional Features

detail, as we want to make sure that our feature properly describes the shapes of edges
(and implicitly the shapes of handwriting) and avoids the senseless cases.

If we consider an oriented edge fragment AB, the arrangement of the hinge is different
whether a second oriented edge fragment attaches in A or in B. So we have to span
all the four quadrants (360°) around the central junction pixel when assessing the angles
of the two fragments. This contrasts with the previous feature for which spanning the
upper two quadrants (180°) was sufficient because AB and BA were identical situations.

Analogously to the previous feature, we considered 3, 4 and 5-pixel long edge frag-
ments. This time, however, their orientation is quantized in $2n = 16$, 24 and 32 directions
respectively (Fig. 2.3 is an example for $2n = 24$). From the total number of combinations
of two angles we will consider only the non-redundant ones ($\phi_2 > \phi_1$) and we will also
eliminate the cases when the ending pixels have a common side. Therefore the final
number of combinations is $C(2n, 2) - 2n = n(2n - 3)$ and, accordingly, our “hinge”
feature vectors will have 104, 252 and 464 components.

Figure 2.4 shows a 3D plot of the bivariate edge-hinge distribution $p(\phi_1, \phi_2)$. Ev-
ery writer has a different “probability landscape” and this provides the basis for very
effective writer identification.

For the purpose of comparison, we evaluated also three other features widely used
for writer identification:
2.3. Features

Figure 2.4: Graphical representation of the edge-hinge joint probability distribution. One half of the 3D plot (situated on one side of the main diagonal) is flat because we only consider the angle combinations with $\phi_2 > \phi_1$.

2.3.3 Run-length distributions

Run lengths, first proposed for writer identification by Arazi (Arazi 1977), are determined on the binarized image taking into consideration either the black pixels corresponding to the ink trace or, more beneficially, the white pixels corresponding to the background. Whereas the statistical properties of the black runs mainly pertain to the ink width and some limited trace shape characteristics, the properties of the white runs are indicative of character placement statistics. There are two basic scanning methods: horizontal along the rows of the image and vertical along the columns of the image. Similarly to the edge-based directional features presented above, the histogram of run lengths is normalized and interpreted as a probability distribution. Our particular implementation considers only run lengths of up to 100 pixels (the height of a written line in our dataset is about 120 pixels).
2.3.4 Autocorrelation

Every row of the image is shifted onto itself by a given offset and then the normalized dot product between the original row and the shifted copy is computed. The maximum offset (‘delay’) corresponds to 100 pixels. All autocorrelation functions are then accumulated for all rows and the sum is normalized to obtain a zero-lag correlation of 1. The autocorrelation function detects the presence of regularity in writing: regular vertical strokes will overlap in the original row and its horizontally shifted copy for offsets equal to integer multiples of the local wavelength. This results in a large dot product contribution to the final histogram.

2.3.5 Entropy

The entropy measure used here focuses on the amount of information, normalized by the amount of ink (black pixels) in the regions of interest. This was realized by using the normalized file size of ROI files after Lempel-Zif compression. The size of the resulting file (in bytes) is divided by the total number of black pixels, which closely estimates the amount of ink present on the page. The obtained feature gives an estimate of the entropy of the ink distribution on the page.

2.4 Results

2.4.1 Evaluation method

The efficacy of the considered features has been evaluated using nearest-neighbor classification (Cover and Hart 1967) in a leave-one-out strategy. Explicitly, one page is chosen and extracted from the total of 500 pages (notice that the experimental data contains 2 pages written by each of 250 subjects). Then the Euclidean distances are computed between the feature vector of the chosen page and the feature vectors of all of the remaining 499 pages. These distances are ranked starting with the shortest one (Press et al. 1992). Ideally, the first ranked page should be the pair page produced by the same writer: an ideal feature extraction making classification effortless and a remapping of the feature space unnecessary. If one considers, not only the nearest neighbor (rank 1), but rather a longer list of neighbors starting with the first and up to a chosen rank (e.g. rank 10), the chance of finding the correct hit increases with the list size. The curve depicting the dependency of the probability of a correct hit vs. the considered list size gives an illustrative measure of performance. Better performance means higher probability of correct hit for shorter list sizes which is equivalent to a curve drawn as much as possible toward the upper-left corner.
Table 2.1: Writer identification accuracy (in percentages) on the Firemaker dataset (250 writers, 2 pages / writer, page 1 vs. page 4). The numbers in the second row of the table header denote the dimensionality of the feature vectors, i.e. the number of bins in the feature histograms. The rightmost column shows the performance obtained by concatenating the edge-hinge PDF and the horizontal run-length PDF.

<table>
<thead>
<tr>
<th>List</th>
<th>p(φ)</th>
<th>dp(φ)</th>
<th>p(φ₁, φ₂)</th>
<th>comb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>size</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>1</td>
<td>26</td>
<td>30</td>
<td>35</td>
<td>45</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
<td>39</td>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>47</td>
<td>52</td>
<td>62</td>
</tr>
<tr>
<td>4</td>
<td>45</td>
<td>52</td>
<td>57</td>
<td>66</td>
</tr>
<tr>
<td>5</td>
<td>49</td>
<td>57</td>
<td>62</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>53</td>
<td>60</td>
<td>65</td>
<td>72</td>
</tr>
<tr>
<td>7</td>
<td>58</td>
<td>63</td>
<td>68</td>
<td>74</td>
</tr>
<tr>
<td>8</td>
<td>60</td>
<td>64</td>
<td>69</td>
<td>75</td>
</tr>
<tr>
<td>9</td>
<td>62</td>
<td>65</td>
<td>71</td>
<td>76</td>
</tr>
<tr>
<td>10</td>
<td>64</td>
<td>68</td>
<td>72</td>
<td>78</td>
</tr>
<tr>
<td>11</td>
<td>66</td>
<td>69</td>
<td>74</td>
<td>79</td>
</tr>
<tr>
<td>12</td>
<td>68</td>
<td>72</td>
<td>76</td>
<td>81</td>
</tr>
<tr>
<td>13</td>
<td>70</td>
<td>73</td>
<td>77</td>
<td>82</td>
</tr>
<tr>
<td>14</td>
<td>71</td>
<td>74</td>
<td>78</td>
<td>83</td>
</tr>
<tr>
<td>15</td>
<td>72</td>
<td>76</td>
<td>79</td>
<td>84</td>
</tr>
<tr>
<td>16</td>
<td>74</td>
<td>77</td>
<td>80</td>
<td>84</td>
</tr>
<tr>
<td>17</td>
<td>76</td>
<td>79</td>
<td>82</td>
<td>84</td>
</tr>
<tr>
<td>18</td>
<td>77</td>
<td>80</td>
<td>82</td>
<td>85</td>
</tr>
<tr>
<td>19</td>
<td>78</td>
<td>81</td>
<td>82</td>
<td>86</td>
</tr>
<tr>
<td>20</td>
<td>79</td>
<td>81</td>
<td>83</td>
<td>87</td>
</tr>
</tbody>
</table>

We point out that we do not make a separation between a training set and a test set, all the data is in one suite. This is actually a more difficult and realistic testing condition, with more distractors: not 1, but 2 per false writer and only one correct hit. Error rates are approximately halved when using the traditional train/test set distinction. Note also the added fact that we only have 2 samples per writer (more labeled samples increasing the chance of a correct identification of the author - see reference (Said et al. 1998) for results on 10 writers, 15 documents / writer). As a consequence of these circumstances, our results are more conservative.

It is also important to mention that the text (ASCII) content is different in the two
samples originating from the same writer: page 1 contains copied text, while page 4 contains self generated text describing a cartoon. The proposed features give a content-independent description of the texture of handwriting.

### 2.4.2 Analysis of performances

We present the performance curves of the edge-based directional features $p(\phi)$ and $p(\phi_1, \phi_2)$ for different direction quantizations (features are ordered with most effective at the top).

Confirming our initial expectations, the improvement in performance yielded by the new feature is very significant despite the excessive dimensionality of the feature vectors (verified by PCA analysis). As a second-order feature, the hinge angular probability distribution captures larger range correlations from the pixel space and therefore it characterizes more intimately the handwriting style providing for more accurate writer identification.
Figure 2.6: Performance curves for the evaluated features (features are ordered with the most effective at the top).

Examination of the family of curves in Fig. 2.5 attests that finer quantized directions result in improved performance at the expense of an increase in feature vector dimensionality (much more sizeable for the edge-hinge feature $p(\phi_1, \phi_2)$).

Figure 2.6 gives a general overview of the comparative performance for all the features considered in this chapter.

The edge-based directional features perform significantly better than the other features because they give a more detailed and intimate information about the peculiarities of the shapes that the writer produces (slant and regularity of writing, roundness or pointedness of letters).

An interesting observation is that the vertical run lengths on ink are more informative than the horizontal ones. This correlates with an established fact from on-line handwriting recognition research stating that the vertical component of strokes carries more information than the horizontal one (Maarse and Thomassen 1983).

The presented features are not totally orthogonal, but nevertheless they do offer different points of view on our dataset. It is therefore natural to try to combine them for
improving the accuracy of writer identification. This topic will be analyzed more thoroughly later in this thesis. We will present here though, for exemplification, the results obtained by concatenating the edge-hinge distribution with the horizontal run lengths on white into a single feature vector that was afterward used for nearest-neighbor classification (last column in Table 2.1).

2.4.3 Stability test

An important question arises: what is the degradation in performance with decreasing amounts of handwritten material? We provide three reference points: whole page (w), half page (top (t) and bottom (b)), and the first line (l). The answer to this question has major bearing for forensic applications (where, in many cases, the available amount of handwritten material is sparse, e.g. the filled-in text on a bank invoice or the address on a perilous letter).

We consider writer identification accuracies for hit lists up to rank 10 (deemed as a more reliable anchor point). Our results from Table 2.2 show significant degradation of performance when very little handwritten material is available. However, it is interesting to observe that the performance standings of the different features with respect to each other remain the same, independent of the amount of text.

2.5 Conclusions

In this chapter, a number of texture-level features have been described and evaluated on the task of text-independent writer identification. The edge-based directional fea-

Table 2.2: Feature performance degradation with decreasing amounts of written text (writer identification accuracy in percentages for list size = 10). The PDFs are extracted from samples containing diminishing amounts of handwritten ink: whole page (w), half page (top (t) and bottom (b)), and the first line (l).

<table>
<thead>
<tr>
<th>Feature</th>
<th>w</th>
<th>t</th>
<th>b</th>
<th>l</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(\phi_1, \phi_2)$</td>
<td>88</td>
<td>81</td>
<td>84</td>
<td>53</td>
</tr>
<tr>
<td>$p(\phi)$</td>
<td>72</td>
<td>66</td>
<td>69</td>
<td>36</td>
</tr>
<tr>
<td>run-length horiz. white</td>
<td>57</td>
<td>42</td>
<td>42</td>
<td>18</td>
</tr>
<tr>
<td>run-length vert. white</td>
<td>51</td>
<td>39</td>
<td>42</td>
<td>16</td>
</tr>
<tr>
<td>run-length vert. black</td>
<td>36</td>
<td>33</td>
<td>33</td>
<td>13</td>
</tr>
<tr>
<td>entropy</td>
<td>8</td>
<td>4</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>
2.5. Conclusions

Features give an overall better performance than run-length, autocorrelation and entropy features.

We described here a new edge-based feature for writer identification that characterizes the changes in direction undertaken during writing. The edge-hinge feature performs markedly better than all the other evaluated features.

Our stability test show that the best performing features when a large amount of text is available still perform best compared to the others when little text is available, despite having considerably higher dimension.

The next chapter of this thesis will focus on increasing the discriminatory power of the feature vectors by including also location information. We will also study how further improvements in performance can be obtained by combining different features in order to exploit their intrinsic degree of orthogonality.