A New Textual/Non-textual Classifier for Document Skew Correction

Xiaoyan Zhu, IEEE Member, Xiaoxin Yin
State Key Lab of Intelligent Tech. & Systems,
Dept. of Computer Science & Technology, Tsinghua University,
Beijing 100084, P.R.China
(Email:zxy-dcs@tsinghua.edu.cn)

Abstract:
A robust approach is proposed for document skew detection. We use Fourier analysis and SVM to classify textual areas from non-textual areas of documents. We also propose a robust method to determine the skew angle from textual areas. Our approach achieves good performance on documents with large area of non-textual contents.

I INTRODUCTION
Skew correction is the first step of OCR to remove the influence of improperly feeding documents into scanners. Many approaches of skew detection can process pure textual document images successfully. But it is a challenging problem to process documents with large areas of non-textual contents. Some approaches have been proposed to address this problem. Le et al. [3] select a square region dominated by text from the document image and calculate the skew angle by this area. Avanindra and Subhasis Chaudhuri [1] divide the document into blocks and use the median of the cross-correlations [4] of all blocks to determine the skew angle.

In this paper we propose a robust approach for skew correction based on classification of textual and non-textual areas. We divide the document image into blocks and use Fourier Transform and Support Vector Machines [2] to determine whether each block is a textual one or not. Then we calculate the skew angle of the document according to the textual blocks.

II Textual/non-textual classification
The skew angle of a pure textual document can be easily calculated by projection profile analysis. For a document image with large area of non-textual contents such as images and graphs, it is crucial to distinguish the textual areas from non-textual areas.

We divide the document image into blocks. For each block, we calculate the Fourier transform of the projection profile, and use it as the feature. By this feature, we use SVM to determine whether a block is a textual one or not. The performances of SVMs of different kernels are compared in Table 1. We choose linear kernel in our system.

Table 1 Comparison between different kernels for textual/non-textual classification

<table>
<thead>
<tr>
<th>Kernel</th>
<th>SVs</th>
<th>dimension</th>
<th>Pr-train</th>
<th>Pr-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>179</td>
<td>&lt;46220</td>
<td>94.6%</td>
<td>81.1%</td>
</tr>
<tr>
<td>Poly</td>
<td>161</td>
<td>&lt;1.9M</td>
<td>97.6%</td>
<td>72.8%</td>
</tr>
<tr>
<td>Gaussian</td>
<td>168</td>
<td>&lt;107361</td>
<td>97.1%</td>
<td>77.5%</td>
</tr>
</tbody>
</table>

Here SVs is the number of SVs, dimension is the dimension of VC, Pr-train is the precision on training set and Pr-test is the precision on training set, and Poly is Polynomial.

Suppose there are $D$ blocks in a document image, and $O_i$ is the output of the SVM for the $i^{th}$ block, indicating whether it is a textual block. It is noticed that a large $O_i$ value does not guarantee a pure textual block from which the correct skew angle can be calculated. Therefore we set an upper bound and a lower bound for $O_i$. Define $O_i^*$ as this:

$$O_i^* = \begin{cases} U, & O_i > U \\ O_i, & L \leq O_i \leq U \\ L, & O_i < L \end{cases}$$

(1)

For calculating the skew angle of the document image, we only need to consider the several blocks with larger $O_i^*$ values. In practice we use the blocks with $O_i^*$ values larger than the average $\overline{O}^*$ value.

$$O_i^* = \begin{cases} O_i - \overline{O}^*, & O_i^* \geq \overline{O}^* \\ 0, & O_i^* < \overline{O}^* \end{cases}$$

(2)

$\overline{O}^*$ is the average of all $O_i^*$.

The weight of each block $w_i$ is calculated from $O_i^*$:

$$w_i = \frac{O_i^*}{\sum_{j=1}^{D} O_j^*}$$

(3)

---

1 This work is supported by National Nature Science Foundation of China (69982005) and Projects of Development Plan of the State Key Foundation Search (G199803050703)

1051-4651/02 $17.00 (c) 2002 IEEE
can be considered as the importance of the \(i^{th}\) block for skew detection of the whole document image. Figure 1 shows an example document image.

![Figure 1](image)

**III Skew correction**

We use a simple but accurate approach to calculate the skew angle of a block. For each block, we project its content to lines of different angles, and calculate the standard deviation for the projection profile of each angle. The angle with the maximum standard deviation corresponds to the skew angle of this block.

After calculating the skew angles of all the blocks, we need to calculate the skew angle of the whole document page.

Suppose there are \(D\) blocks in the document page. The weights of the blocks are \(w_1, w_2, \ldots, w_D\). And the skew angles of the blocks are \(a_1, a_2, \ldots, a_D\).

\(a^*\) is the skew angle of the page. Suppose the expectation and standard deviation of \(a^*\) is \((\mu, \sigma)\). We assume that and only one skew angle from \(a_1, a_2, \ldots, a_D\) is the correct estimation of \(a^*\) (which means that \(a^*\) is very close to one and only one skew angle from \(a_1, a_2, \ldots, a_D\)). Suppose \(a_i\) \((i \leq c \leq D)\) is the correct estimation, use \(w_i\) as an estimation that \(i = c\) (assume that \(P(i = c) = w_i\)). By assuming the expectation and standard deviation of \(a^*\), we can calculate the probability that \(a^*\) lies in the close neighborhood of \(a_i\).

\[
P(a^* \in (a_i - \frac{1}{2} \Delta a, a_i + \frac{1}{2} \Delta a) | c = i) = \frac{\exp\left(-\frac{(a^* - a_i)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma} \cdot \Delta a
\]

(4)

And we can calculate the probability that \(a^*\) lies in the close neighborhood of the skew angle of any block.

\[
P(a^* \in \bigcup_{i=1}^{D} (a_i - \frac{1}{2} \Delta a, a_i + \frac{1}{2} \Delta a)) \approx \sum_{i=1}^{D} \left[ P\left(a^* \in \left(a_i - \frac{1}{2} \Delta a, a_i + \frac{1}{2} \Delta a \right) \right) \cdot P(i = i) \right]
\]

(5)

For a skew angle \(a\), define a function \(E(a)\) to evaluate this angle,

\[
E(a) = \sum_{i=1}^{D} \left[ w_i \frac{\exp\left(-\frac{(a - a_i)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma} \cdot \Delta a \right]
\]

(6)

The angle \(a\) that maximizes \(E(a)\) can be calculated as the skew angle of the whole page. However, since the skew angle of the whole page should be very close to the skew angle of a certain block, it is unnecessary to use heuristic approach to search for the angle \(a\) that maximizes \(E(a)\). Instead, we only need to calculate \(E(a_i), i = 1, \ldots, D\), and take the \(a_i\) that with the maximum \(E(a_i)\) as the skew angle of the page.

\[
a^* = a_{\text{max}}, \quad i_{\text{max}} = \arg \max_i \left(E(a_i)\right)
\]

(7)

With this approach, the skew angle of the whole page will not be affected if the skew angle of a certain block is incorrectly calculated. This adds to the robustness of our system. After the skew angle is detected, the document image is rotated according to this angle.

**IV Experiment results**

The proposed approach is tested on 54 documents of different languages from newspapers, magazines and academic papers. All documents are scanned at a resolution of 200 dpi. And the maximum skew angle is
restricted to ±10°. We compare the performance of our approach and the approach of [1] in Table 2.

<table>
<thead>
<tr>
<th>Percentage of documents within error range</th>
<th>Our approach</th>
<th>Approach of [1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0° - 0.1°</td>
<td>83.3</td>
<td>14.8</td>
</tr>
<tr>
<td>0.1° - 0.2°</td>
<td>14.8</td>
<td>20.4</td>
</tr>
<tr>
<td>0.2° - 0.5°</td>
<td>1.85</td>
<td>38.9</td>
</tr>
<tr>
<td>&gt;0.5°</td>
<td>0</td>
<td>25.9</td>
</tr>
</tbody>
</table>

Table 2: Comparison between our approach and approach of [1] (The tests are conducted on a computer with a CPU of AMD Athlon 900MHz).

Figure 2 shows an example of skew correction. Experiments show that the proposed approach achieves more robustness and higher precision than the previous approach of robust skew correction.

References

Figure 2: The skew angle of the original document is 4.17° and the skew angle detected is 4.13°.