Video text recognition using feature compensation as category-dependent feature extraction

Minoru Mori

NTT Communication Science Laboratories, NTT Corporation
3-1, Morinosato-Wakamiya, Atsugi-shi, Kanagawa 243-0198, Japan
mmori@eye.brl.ntt.co.jp

Abstract

When recognizing multiple fonts, geometric features, such as the directional information of strokes, are generally robust against deformation but are weak against degradation. This paper describes a category-dependent feature extraction method that uses a feature compensation technique to overcome this weakness. Our proposed method estimates the degree of degradation of an input pattern by comparing the input pattern and the template of each category. This estimation enables us to compensate the degradation in feature values. We apply the proposed method to the recognition of video text suffering from degradation and deformation. Recognition experiments using characters extracted from videos show that the proposed method is superior to the conventional alternatives in resisting degradation.

1. Introduction

The technical barriers to recognizing characters are mainly due to degradation, such as background noise, and deformation, like the variation in fonts. Most conventional character recognition methods have tackled either one or the other. For overcoming degradation, some methods design templates that reflect the degradation type anticipated [6, 19, 3]. Moreover a robust discriminant function for recognizing degraded characters was proposed in [16, 15]. Unfortunately, these methods are sensitive to shape distortion, since they employ image-based template matching. They fail to well handle multiple fonts and several style effects. On the other hand, geometric features are often used for recognizing multiple fonts. Stroke direction, e.g. [2, 17, 20], is particularly effective against character deformation. However, geometric features are not robust against degradation, because the geometric information is corrupted by degradation. Some methods try to compensate the corruption by assuming the type of degradation [14, 12]. They, however, are not effective when the assumption is invalid. This suggests the impracticality of determining stroke or noise from just the pixel distribution in the input pattern.

Our approach is to focus on a category-dependent method based on the top-down approach. One category-dependent method, a shape normalization method, has been proposed for tackling the deformation problem [18]. On the other hand, in an earlier paper, the author has proposed a basic concept of category-dependent feature extraction to achieve robustness against degradation [13]. In this paper, we improve this method and apply it to the recognition of video text that contains degradation (background noise and blur) and deformation (multiple fonts). Our method estimates the degree of degradation of the input pattern on the basis of the variation in run-length distribution. Exploiting the template of each category enables to extract the run-length variation. The fluctuation in feature values caused by degradation is then offset by the estimation of a compensation coefficient.

The paper is organized as follows: Section **2** describes related work and the key issues in the recognition of video text. Section **3** presents a directional feature and our proposed method. Experimental results are reported in section **4**. Section **5** summarizes this paper and lists future work.

2. Video text recognition

2.1. Needs and related work

It's increasingly common to receive video data through multi-channel broadcasts on TV, by DVD, and through the broadband Internet via PCs. However, it's very difficult to fully grasp the contents of the material and locate the contents that meet the user's need from among the huge video archives, so technologies to browse and retrieve video archives are strongly needed. In response to these needs, various video retrieval methods, e.g. [1, 9, 5], have been proposed. The video data contains several information

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

sources such as image and audio. Among them, text that is superimposed or appears in a video is important in understanding the video's contents. For example, headline and location names are common in news programs. In sports programs, captions often provide game scores and player names. Therefore, the recognition of video text provides semantic information and enables video content to be parsed, indexed, and retrieved.

Although several methods for video text recognition have been proposed, most, e.g. [9, 4, 8], are focused on locating and extracting text in/from video. Few papers have tackled the recognition of text extracted from video [11, 10]. In this paper we discuss the recognition of text extracted from video frames.

2.2. Characteristics of video text

Characters extracted from binarized video frames suffer from degradation such as background noise and blur, and deformation due to the variety of fonts and style effects. To investigate the attributes of characters extracted from video, we gathered samples of characters, many of which were extracted from news programs, using the method proposed in [7]. Table 1 shows the summation of image quality types for the samples[1].

Table 1. Analysis result of video text samples.

Type of Image quality	Number	Ratio[%]
clean / slight noise	7,779	69.2
background noise	2,055	18.3
blur	84	0.7
subtractive noise / lack of stroke	204	1.8
low spatial resolution	68	0.6
unique style font / style effect	523	4.7
fluctuation in height/width ratio	62	0.5
mis-normalization by noise	471	4.2
Sum	11,246	100

This paper tackles background noise and blur, the main causes of poor recognition accuracy, and deals with them in the same manner from the view point that both increase the pixel number. Background noise, often similar to blobs, is caused by misjudging background region as character region due to similar properties such as color or size. Blur is derived from the low spatial resolution of the image and inappropriate threshold used in binarizing the video frame. Figure 1 shows a part of a video frame and characters extracted from the binarized frame.

[1]Even though most samples exhibited several types in image quality, only the dominant type was counted.

A part of a frame with text "子供や孫と".

Characters extracted from the above frame.

Figure 1. A part of video frame with text and extracted characters.

3. Feature compensation

This section describes the directional feature used and the algorithm of our proposed method.

3.1. Directional feature

Geometric features that extract stroke direction are effective for discriminating multiple fonts. For example, the direction contributivity[2] [2] and LSD [17, 20], which are based on stroke run-length, are robust against deformation. The direction contributivity d_i $(i = 1, ..., 4)$ is given by

$$d_i = l_i / \sqrt{\sum_{j=1}^{4} {l_j}^2} \tag{1}$$

where l_1, l_2, l_3, and l_4 denote the run-lengths on the horizontal, right diagonal, vertical, and left diagonal directions at each black pixel, respectively. Let $d_{m,i}$ be the direction contributivity as components of feature vector for the m-th block obtained by partitioning a pattern. Let $l_{m,i}$ be the run-length yielded by averaging l_i on the m-th block. $d_{m,i}$ can be computed as follows;

Step 1: The input pattern is divided into $N \times N$ blocks.

Step 2: l_i is extracted at each black pixel.

Step 3: $l_{m,i}$ is calculated by averaging l_i on each block.

Step 4: $d_{m,i}$ is computed from Eq. (1) using $l_{m,i}$.

Here we use $N = 8$. Figure 2 shows an example of direction contributivity extraction from an input pattern.

3.2. Basic idea of compensation

As degradation and deformation models, one can express a feature vector extracted from an input pattern (input vector) F' as follows;

$$F' = \mathrm{A} \cdot F + B, \tag{2}$$

[2]Contributivity means the degree of contribution in stroke direction.

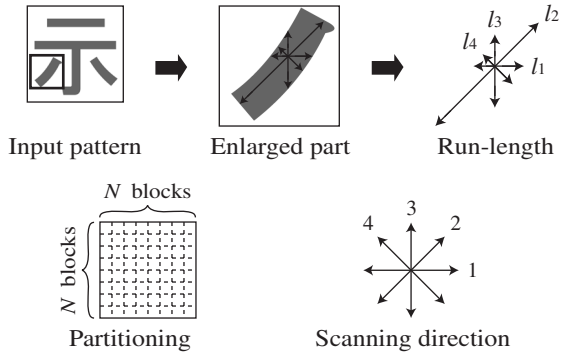


Figure 2. Direction contributivity feature.

where F is a feature vector extracted from a clean pattern without degradation or deformation. Operator A mainly represents deformation, while B is derived from degradation such as background noise. Because it's impossible to directly obtain F with no prior information, we approximate Eq. (2) as follows:

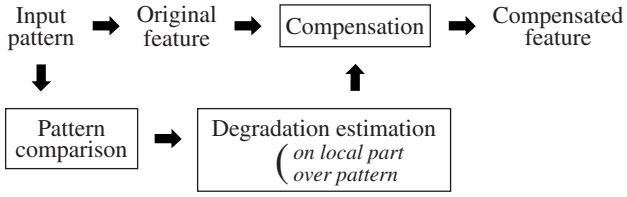
$$F' = \mathrm{A} \cdot F + B \fallingdotseq \mathrm{A}' \cdot F. \tag{3}$$

Here operator A' can be regarded as the degree of degradation of the input pattern. On the basis of Eq. (3), we try to acquire the compensation coefficient ($= 1/\mathrm{A}'$) by estimating the degree of degradation of the input pattern. This estimation enables us to obtain the approximate feature vector by compensating the degraded feature vector.

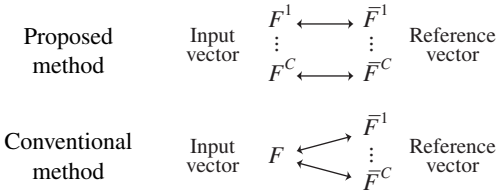
The key to estimating the degree of degradation is using the variation in run-length distribution of the input pattern. The variation of run-length depends on the degree of degradation, and increases with the pattern degradation. Therefore, the degree of degradation can be estimated by extracting the degree of variation in run-length distribution. As mentioned before, it's impractical to estimate the variant condition from just pixel distribution in the input pattern.

To realize this estimation, we exploit the template of each category. Comparing the input pattern to the template of each category enables us to calculate the variation of run-length against the focused category. However, it should be noted that using run-length distribution extracted from just the local part tends to result in failure. The reason is that many local parts of different categories have similar properties, particularly if the input pattern is degraded. Therefore, we estimate the degree of degradation not only from the local parts but also the total pattern. Adjusting the local estimation against the global estimation provides an appropriate compensation coefficient for the focused part. In summary, combining the local and global estimations based on the pattern comparison enables us to extract an approximate feature vector even from degraded characters by compensating the fluctuation in feature values. The sequence of feature compensation is shown in Figure 3 (a).

The compensation mentioned above produces feature vectors for every each category from the input pattern. The input pattern is classified by calculating distances between the input vector and the reference vector of each category. By comparison, conventional methods usually extract just one feature vector from the input pattern for classification. The procedure in classification is illustrated in Figure 3 (b).



(a) Sequence of feature extraction using compensation.



(b) Classification procedure.

Figure 3. Concept of the proposed method.

3.3. Procedure details

This part fleshes out the algorithm of the proposed method: the template of each category, the definitions of the degree of degradation on part and total pattern, and the feature compensation procedure.

As the template of each category, we use the directional stroke run-length that allows features on each direction to be compensated. The templates are obtained as follows: The averaged stroke run-length $l_{m,i}$ is calculated using steps $1 \sim 3$ in section **3.1**. The run-length vectors used as the template for the c-th category $\bar{L}^c$ ($\bar{l}^c_{m,i}$) is then obtained by averaging $l_{m,i}$ from all training samples of the c-th category; where $c\,(= 1, ..., C)$ denotes the category number.

The degree of degradation on the blocks as the local parts and over the complete pattern is defined as follows: The degree of degradation on the focused block is defined as the variation in run-length distribution within the block; This is obtained by calculating the ratio between the run-length distribution of the input pattern and that of the template of the c-th category. The stroke run-length $l_{m,i}$ for the i-th direction on the m-th block is calculated from the input pattern

using the approach mentioned above. By using $l_{m,i}$ and $\bar{l}^c_{m,i}$, the degree of degradation on the m-th block against the c-th category, $p^c_{m,i}$, is given by

$$p^c_{m,i} = \begin{cases} (l_{m,i} - \bar{l}^c_{m,i})/l_{m,i} & (l_{m,i} > \bar{l}^c_{m,i}) \\ (\bar{l}^c_{m,i} - l_{m,i})/\bar{l}^c_{m,i} & \text{otherwise.} \end{cases} \quad (4)$$

$p^c_{m,i}$ approaches 1 as the input pattern become degraded or dissimilar. $p^c_{m,i}$ becomes 0 for the comparison of same patterns. On the other hand, the degree of degradation over the complete pattern is defined as the average of the degrees on all blocks. The global degree of degradation against the c-th category, g^c, is given by

$$g^c = \frac{\sum_{m=1}^{N^2} \sum_{i=1}^{4} p^c_{m,i}}{4 \cdot N^2} \quad (0 < g^c < 1). \quad (5)$$

g^c expresses the dissimilarity for the complete pattern between the input pattern and the template of the c-th category.

A feature value varied by degradation is compensated with $p^c_{i,x,y}$ and g^c in each block. A new feature value against the c-th category, $d^c_{m,i}$, is obtained by compensating $d_{m,i}$ as follows:

$$d^c_{m,i} = (1 - (1 - g^c) \cdot p^c_{m,i}) \cdot d_{m,i}. \quad (6)$$

Through the use of g^c, the degree of compensation is adjusted so as to avoid over compensating the feature values to the point that they resemble those of an incorrect/dissimilar category. Substituting $d^c_{m,i}$ for $d_{m,i}$ yields the components of the feature vector against the c-th category. Finally, C feature vectors are obtained by repeating the above procedure for every category.

Figure 4 visualizes the horizontal and vertical feature values in the direction contributivity obtained using compensated technique and that based on the original feature for the character with background noise. The compensated feature yielded by the proposed method still retains stroke direction and suppresses the influence of background noise.

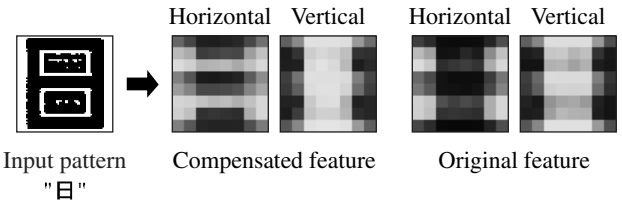


Figure 4. Feature values of the compensated and original features.

4. Recognition experiments

4.1. Data and experimental conditions

We used the following data in the recognition experiments. As the training data set, we used 67 fonts of machine printed Kanji characters from 3,190 categories. As the test data set for evaluating robustness against noise and blur, 9,918 samples were selected from the data set mentioned in section **2.2**; They contained 7,779 clean/slightly noisy characters and 2,139 noisy and blurred ones.

Each character was normalized into 64×64 pixels. Each feature vector consisted of 256 dimensional components (8 $\times$ 8 blocks $\times$ 4 directions). The dictionary was constructed by averaging features from the training samples for each category. Euclidean distance was used as the classifier.

4.2. Experimental results and discussion

We compared the compensated feature yielded by our proposed method to the original feature and the image-based method[3]. Figure 5 shows classification rates of the three methods for (a) the data containing only background noise and blur, and (b) all test data. The compensated feature achieved about 10% better classification rates than the original one for the data (a). In particular, for the top ten candidates, the compensated feature yielded 34% fewer errors than the original feature. Moreover, the compensated feature yields about 2% better classification rates than the other methods for data set (b); this represents a 30% fewer errors for the top ten candidates. These results prove that our method can effectively offset the variation in features caused by degradation without lowering the recognition accuracy for clean data. On the other hand, the image-based method offered the lowest accuracy among the three methods due to its sensitivity and weakness to the distortion created by the variety of fonts and styles.

Figures 6 shows examples recognized correctly by the proposed method which were recognized erroneously by the conventional alternatives (correct result by the compensated feature $\leftarrow$ erroneous result by the original one). With regard to these examples, the proposed method effectively compensates the fluctuation in directional information caused by background noise and blur, and so avoids erroneousness.

One advantage of our method is that the classifier can be freely selected. By comparison, conventional and category-independent methods [14, 11] set the compensation procedure in the classification stage, so it is difficult to change the classifiers specified.

[3]Each pixel value of a character (64×64 pixels) was used as feature values.

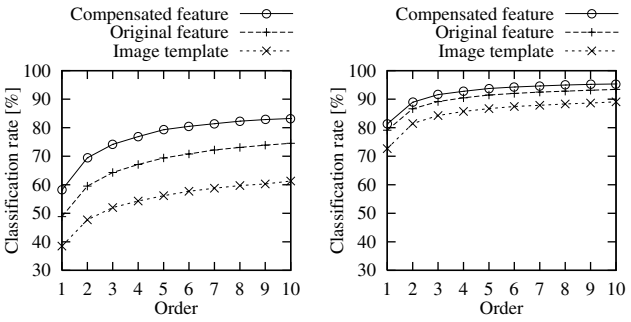


Figure 5. Classification rates for test data set.

Figure 6. Characters recognized by the proposed method.

5. Conclusion

We have proposed a feature compensation technique that is based on a category-dependent method, and have applied it to the recognition of video text exhibiting varying levels of background noise and blur. Our method estimates the degree of degradation of the input pattern by exploiting a template of each category. This estimation enables the adaptive compensation of fluctuated feature values extracted from degraded characters. It was applied to the directional feature based on stroke run-length for extracting approximate feature values. Recognition experiments using characters extracted from videos showed that our proposed method achieves higher classification rates than the original alternatives of the original feature and the image-based method.

Future works are to examine an iterative compensation procedure, and to expand our method to achieve robustness against subtractive noise.

Acknowledgements

I am grateful to Dr. Noboru Sugamura for support. I am indebted to Dr. Norihiro Hagita, Mrs. Minako Sawaki, and Prof. Hiroshi Murase for fruitful discussions. I also wish to thank Prof. George Nagy and Dr. Daniel Lopresti for suggestions.

References

[1] Content-based image retrieval systems. *IEEE Computer*, 28(9):18–62, Sept. 1995.

[2] T. Akiyama and N. Hagita. Automated entry system for printed documents. *Pattern Recognit.*, 23(11):1141–1154, Nov. 1990.

[3] T. K. Ho. Bootstrapping text recognition from stop words. In *Proc. of 14th ICPR*, volume 1, pages 605–609, 1998.

[4] A. K. Jain and B. Yu. Automatic text location in images and video frames. *Pattern Recognit.*, 31(12):2055–2076, Dec. 1998.

[5] K. Kashino et al. Time-series active search for quick retrieval for audio and video. In *Proc. of ICASSP'99*, volume 6, pages 2993–2996, 1999.

[6] G. E. Kopec. Supervised template estimation for document image decoding. *IEEE Trans. PAMI*, 19(12):1313–1324, Dec. 1997.

[7] H. Kuwano et al. Telop character extraction from video data. In *Proc. of DIA'97*, pages 82–88, 1997.

[8] H. Li et al. Automatic text detection and tracking in digital video. *IEEE Trans. IP*, 9(1):147–156, Jan. 2000.

[9] R. Lienhart. Automatic text recognition for video indexing. In *Proc. of ACM Multimedia'96*, pages 11–20, 1996.

[10] T. Mita and O. Hori. Improvement of video text recognition by character selection. In *Proc. of 6th ICDAR*, pages 1089–1093, 2001.

[11] M. Mori et al. Robust telop character recognition in video for content-based retrieval. In *Proc. of 5th ICDAR*, pages 13–16, 1999.

[12] M. Mori et al. Robust feature extraction based on run-length compensation for degraded handwritten character recognition. In *Proc. of 6th ICDAR*, pages 650–654, 2001.

[13] M. Mori et al. Category-dependent feature extraction for recognition of degraded handwritten characters. In *Proc. of 16th ICPR*, volume 3, pages 155–159, 2002.

[14] S. Omachi et al. A noise-adaptive discriminant function and its application to blurred machine-printed kanji recognition. *IEEE Trans. PAMI*, 22(3):314–319, Mar. 2000.

[15] A. Sato. A learning method for definite canonicalization based on minimum classification error. In *Proc. of 15th ICPR*, volume 2, pages 199–202, 2000.

[16] M. Sawaki and N. Hagita. Text-line extraction and character recognition of document headlines with graphical designs using complementary similarity measure. *IEEE Trans. PAMI*, 20(10):1103–1109, Dec. 1998.

[17] S. Srihari et al. Machine-printed japanese document recognition. *Pattern Recognit.*, 30(8):1301–1313, Aug. 1997.

[18] T. Wakahara and K. Odaka. Adaptive normalization of handwritten characters using global/local affine transformation. *IEEE Trans. PAMI*, 20(12):1332–1341, Dec. 1998.

[19] Y. Xu and G. Nagy. Prototype extraction and adaptive OCR. *IEEE Trans. PAMI*, 21(12):1280–1296, Dec. 1999.

[20] J. Zhu et al. Image-based keyword recognition in oriental language document images. *Pattern Recognit.*, 30(8):1293–1300, Aug. 1997.