Mosaicing-by-recognition for recognizing texts captured in multiple video frames

Seiichi Uchida, Hiromitsu Miyazaki, and Hiroaki Sakoe
Graduate School of Information Science and Electrical Engineering,
Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka-shi, 812-8581 Japan

Abstract

Text recognition in video frames is promising because of its following superiorities over text recognition in a still camera image: (1) it is possible to recognize longer texts by concatenating the frames, and (2) it is also possible to improve the quality of the text image by integrating the frames. In this paper, a mosaicing-by-recognition technique is proposed where video mosaicing and text recognition are simultaneously and collaboratively performed in a one-step manner by a dynamic programming-based optimization algorithm. In this optimization algorithm, rotation, scaling, vertical shift, and speed fluctuation of camera motion are efficiently compensated. The results of experiments to evaluate not only the accuracy of text recognition but also that of video mosaicing indicates that the proposed technique is practical and can provide reasonable results in most cases.

1. Introduction

Text recognition for a single image captured by a camera, i.e., a still image, becomes a practical technique and is often equipped in commercial cellular phones for recognizing e-mail addresses, URLs, single words, and so on. In spite of its practical property, it has several limitations. For example, (1) long texts often cannot be recognized, and (2) it is generally dif cult to improve the quality (e.g., resolution and noise level) of a still image.

Text recognition for multiple video frames (Fig. 1) has been investigated [1] as an alternative to text recognition in a still image, because it has a potential to overcome the above limitations. That is, it is possible to recognize longer texts by mosaicing consecutive frames, i.e., by matching and concatenating the frames. In addition, it is also possible to improve the quality of the text image (e.g., super-resolution, noise removal) by utilizing overlapped ares between consecutive frames.

In this paper, a *mosaicing-by-recognition* technique is proposed. Previous attempts to recognize texts in video sequences generally assumes a two-step manner that video frames are rstly concatenated into one large image by mosaicing techniques (e.g., [2]) and then the texts in the large mosaic image is recognized. In contrast, the proposed technique is organized in a one-step manner that video mosaicing and text recognition are simultaneously and collaboratively performed. Speci cally , multiple frames capturing a long text line are optimally matched and concatenated with a guide of the text recognition framework. The optimization is performed by a dynamic programming (DP)-based algorithm while compensating various distortions of the frames.

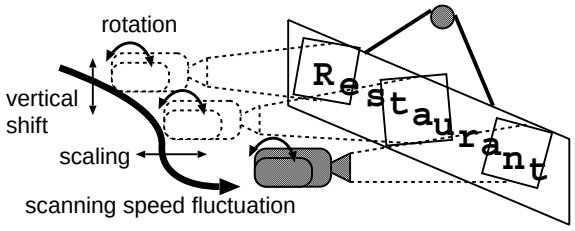

Figure 1. Recognition of the text captured in multiple video frames. When using a hand-held camera, several distortions will appear in the frames due to rotation, vertical shift, scaling, and scanning speed fluctuation.

2. Mosaicing-by-recognition

2.1. Problem formulation

Assume that a long text line is continuously and fragmentarily captured in video frames by a hand-held camera which moves from left to right along the text. Major distortions appeared in the frames are: rotation, scaling, vertical shift, and speed uctuatio n of the camera motion. Our task is the recognition of the captured texts while mosaicing the frames and removing the distortions.

In the remaining part of this section, we will rstly discuss a *simple case* that video frames undergo only speed

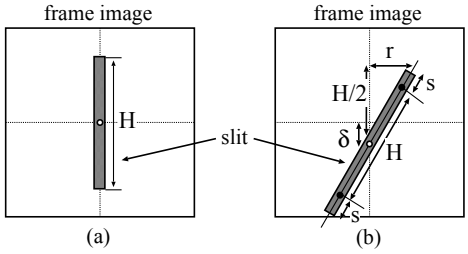

Figure 2. (a) One-pixel slit ($r = s = \delta = 0$) and (b) its controlled version.

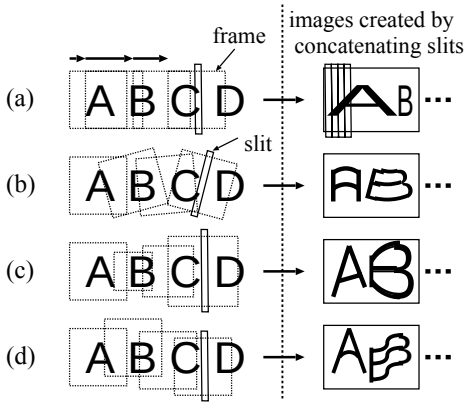

Figure 3. Major distortions in video sequence obtained by a hand-held camera. (a) Scanning speed fluctuation. (b) Rotation. (c) Scaling. (d) Vertical shift. The rightmost images of (b)–(d) indicate the necessity of controlling slit shape.

uctuatio n. This simpli cation is quite useful to grasp the basic principle of the proposed technique. In fact, by this simpli cation our mosaicing-by-recognition problem is reduced to a well-known *segmentation-by-recognition* problem for continuous speeches [3] and texts [4]. Secondly, we will discuss a *general case* that video frames undergo not only speed uctuation but also rotation, scaling, and vertical shift. The mosaicing-by-recognition problem on the general case is derived as an extension of the simple case.

Our discussion is further simpli ed by the use of a *one-pixel slit* (shown in Fig. 2(a)), which is a central part of the frame and has 1 pixel width and H pixel height. Although this simpli cation is useful to understand the principle of the proposed technique, most of information contained in frames is disregarded. Thus, the use of *wider slits* whose width is two or more pixels is discussed in Section 2.4.

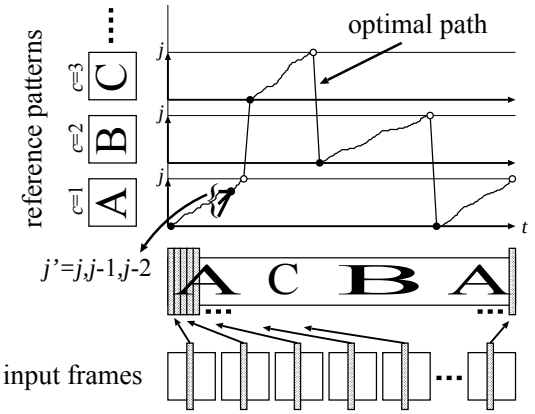

Figure 4. Mosaicing-by-recognition for the simple case that video frames undergo only speed fluctuation.

/* Initialization */
for $c := 1$ **to** C **do begin**
  $g_1(c,1) := d_1(c,1)$
  for $j := 2$ **to** J_c **do**
    $g_1(c,j) := \infty$
end
$D_1 := \infty$
/* DP Recursion */
for $t := 2$ **to** T **do begin**
  for $c := 1$ **to** C **do begin**
    $g_t(c,1) := d_t(c,1) + \min\{g_{t-1}(c,1), D_{t-1}\}$
    for $j := 2$ **to** J_c **do**
      $g_t(c,j) := d_t(c,j) + \min_{j' \in \{j,j-1,j-2\}} g_{t-1}(c,j')$
  end
  $D_t := \min_{c' \in C} g_t(c', J_{c'})$
end

Figure 5. The DP algorithm for mosaicing-by-recognition for the simple case. Several steps for backtracking operation is omitted.

2.2 DP algorithm for simple case

In this section, a mosaicing-by-recognition algorithm for the simple case is provided, where only the uctuation of scanning speed is assumed. The one-pixel slit is also assumed here. Other distortions and wider slits will be considered in later sections.

On the simple case, the problem is reduced to the well-known optimization problem, called segmentation-by-

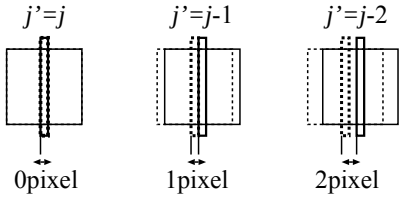

Figure 6. The relation between the selection of j' and scanning speed.

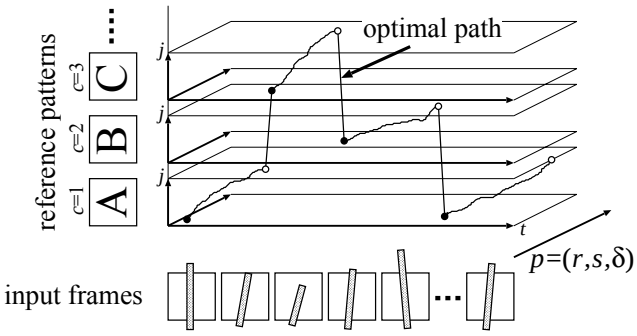

Figure 7. Mosaicing-by-recognition for the general case that video frames undergo not only speed fluctuation but also rotation, scaling, and vertical shift.

recognition problem, of a character sequence. The text contained in the frames can be treated as a deformed character sequence in the image created by concatenating the one-pixel slits of all T frames (shown in the right side of Fig. 3(a)). Thus, the text in the image can be recognized and partitioned into its component characters by solving an optimal path problem on the search space indexed by t and (c, j), where $c \in \{1, \ldots, C\}$ is the character category and $j \in \{1, \ldots, J_c\}$ is the index for the row of the reference pattern image of the category c (Fig. 4).

It is also well-known that this optimal path problem can be solved effectively by DP. Figure 5 shows a DP algorithm for the simple case, where $d_t(c, j)$ is the matching cost between the one-pixel slit of the tth frame and the jth column of the reference pattern of category c. The value $g_t(c, j)$ is the minimum cost accumulated along with the optimal path to the point (so-called the "state") indexed by t, c and j.

The speed uctuation can be compensated by controlling j' in the DP recursion of Step 11. Speci cally , as shown in Fig. 6, $j' = j - 2$ is selected when the scanning speed is 2 pixel/frame and $j' = j$ is selected when it is 0pixel/frame.

The result of character recognition is obtained by backtracking the optimal (c, j)-sequences (illustrated as the optimal path in Fig. 4) after performing the DP algorithm. An optimal mosaic image is also obtained by backtracking as will be shown in the Section 2.5. Thus, the mosaicing of video frames is optimized simultaneously with the text recognition, and therefore we call the above procedure *mosaicing-by-recognition*.

2.3 DP algorithm for general case

In this section, we derive a DP algorithm for the general case, where not only the speed uctuation but also the other distortions are considered. The DP algorithm for the general case is an extension of the foregoing DP algorithm for the simple case. The main idea of the extension is to control (i.e., rotate, scale, and vertical shift) the slit according to the distortions. Figure 2(b) shows a slit controlled by three parameters r, s, and δ which represents rotation, scaling, and vertical shift, respectively. When $r = s = \delta = 0$, the controlled slit is reduced to the original slit of Fig. 2(a) and means that no distortion appears.

The optimal parameters are searched for in the DP framework. Speci cally , as shown in Fig. 7, the problem becomes an optimal path problem in the search space in-

/* Initialization */
1 **for all** $p \in \{(r, s, \delta)\}$ **do begin**
2 **for** $c := 1$ **to** C **do begin**
3 $g_1(p, c, 1) := d_1(p, c, 1)$
4 **for** $j := 2$ **to** J_c **do**
5 $g_1(p, c, j) := \infty$
6 **end**
7 $D_1(p) := \infty$
8 **end**
/* DP Recursion */
9 **for** $t := 2$ **to** T **do begin**
10 **for all** $p \in \{(r, s, \delta)\}$ **do begin**
11 **for** $c := 1$ **to** C **do begin**
12 $g_t(p, c, 1) := d_t(p, c, 1) + \min_{p' \in \mathrm{pre}(p)} \{g_{t-1}(p', c, 1), D_{t-1}(p')\}$
13 **for** $j := 2$ **to** J_c **do**
14 $g_t(p, c, j) := d_t(p, c, j) + \min_{\substack{p' \in \mathrm{pre}(p) \\ j' \in \{j, j-1, j-2\}}} g_{t-1}(p', c, j')$
15 **end**
16 $D_t(p) := \min_{c' \in C} g_t(p, c', J_{c'})$
17 **end**
18 **end**

Figure 8. The DP algorithm for the general case.

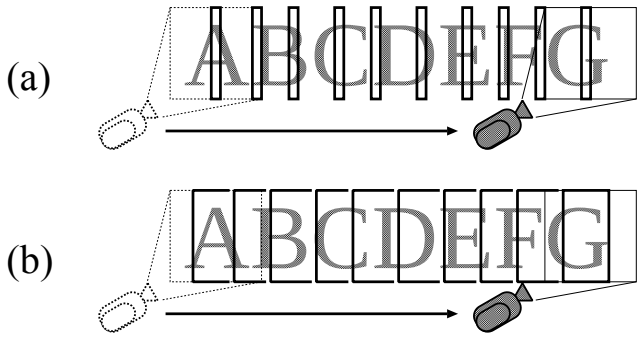

Figure 9. The relation between slit width and scanning speed.

dexed by t and (p, c, j), where $p = (r, s, \delta)$ is a parameter vector. Figure 8 shows a DP algorithm the general case, where $d_t(p, c, j)$ is the matching cost between the one-pixel slit whose shape is controlled by the parameter p and the jth column of the reference pattern of category c. In the DP recursion of Step 14, the smoothness of the distortion is assumed by constraining the parameter vectors of consecutive frames (p for frame t and p' for frame $t-1$) by

$$\mathrm{pre}(p) = \{p' = (r', s', \delta') \mid r - 1 \le r' \le r + 1, \\ s - 1 \le s' \le s + 1,\ \delta - 1 \le \delta' \le \delta + 1\}$$

The computational complexity of the algorithm is $O(TCJRS\Delta)$, where R, S and Δ are the ranges of r, s, and δ, respectively. Similar to the simple case, the result of character recognition is obtained by backtracking the optimal path after performing the DP algorithm.

2.4. Expansion of slit width

In the above discussion, the width of the slit is x ed at one pixel for simplifying the problem. This means, however, that most of information contained in each frame is disregarded.

The modi catio n of the above algorithms for using a wider slit is very straightforward. Speci cally , the modi cation can be done by simply changing the de nition of the matching distance $d_t(c, j)$ to be a distance between the wider slit (a rectangular area on a input frame) and a rectangular area of a reference pattern[1]. Using a wider slit, the recognition accuracy can be improved because "over- tting" and "over-segmentation" can be suppressed as discussed in Section 3.

[1]Strictly speaking, the projective distortion within a wider slit should be considered for recognizing texts captured from a non-frontal video camera.

A wider slit produces another promising effect; a wider slit allows higher scanning speeds. Figure 9 shows the relation between slit width and acceptable scanning speed. As shown in Fig. 9 (a), when the one-pixel slit is used, non-negligible gaps will appear in captured frames as scanning speed becomes higher. Thus, most information will be lost and the accuracy of recognition/mosaicing results will be seriously decreased. On the other hand, as shown in Fig. 9 (b), when a wider slit is used, the gaps will disappear because some overlaps between consecutive frames can be expected. For allowing scanning speeds of K pixel/frame, the DP recursions of the above algorithms (i.e., Step 11 of Fig. 5 and Step 14 of Fig. 8) also should be modi ed so that j' can be chosen not only from $\{j, j-1, j-2\}$ but also from $\{j-3, \ldots, j-K\}$.

2.5 Mosaic image

Although conventional video mosaicing techniques require several corresponding points among consecutive frames, the proposed technique does not. In the simple case, the mosaic image can be obtained by placing the tth frame with a 0$\sim$ K pixel horizontal shift according to the relation between j' and j, which can be obtained by the backtracking operation for the optimal path. (See Fig. 6 for the case $K = 2$.)

Even in the general case, the mosaic image can be obtained by a similar procedure. The only difference is a dewarping operation of the controlled slit of each frame is necessary in advance to placing it with a 0$\sim$ K pixel horizontal shift. The dewarping can be done by using the optimal parameter p at frame t, which can be obtained by the backtracking operation.

On creating a mosaic image by the above procedures, we should manage the overlapped area between two consecutive frames. In the following experiment, a simple averaging was performed to determine a pixel value of the overlapped area. In future, this overlapped area will be utilized to improve the quality of the mosaic image by super-resolution or other image restoration techniques [5, 6, 7].

3. Experimental results

3.1. Data preparation

As test samples for performance evaluation, 20 text lines printed on white A4-sized papers were prepared. Each text line contains about 50 characters (of capital/small English alphabets and digits) and thus about 1000 characters were prepared in total. Each character was printed in the same Times-Roman font. The character height ($\sim H$) in the frame was about 40 pixels.

Each text line was then captured in multiple frames by moving a video camera. A special equipment with a variable speed motor was used for moving the camera horizontally. Thus, we could uctuate the speed of camera movement, while excluding rotation, scaling, and vertical shift. According to this manner, the video frames of the simple case were prepared. Note that naive gray-level was used as the pixel feature for calculating the matching cost $d_t(c, j)$ or $d_t(p, c, j)$

For preparing the video frames of the general case, the above video frames of the simple case were arti cially rotated, scaled and vertically shifted. That is, the video frames for the general case was synthesized from those of the simple case. On the synthesis, the maximum amplitude of distortions were limited so that the distortions can be theoretically compensated by $p = \{(r, s, \delta) \mid |r| \leq k, |s| \leq k, |\delta| \leq k\}$, where k was xed at 1, 2, 3, or 4 (pixels).

3.2. Qualitative analysis

Figure 10 shows a result of the simple case. The one-pixel slits were used here to observe the minimum performance of the algorithm. The camera scanning speed was uctuated between 0$\sim$ 2 pixel/frame. Figure 10 (a) shows several input frames and (b) shows the image created by concatenating the one-pixel slits. This gure (b) indicates that scanning speed became very low around "t" of the word "Character". Figures 10 (c) and (d) show the mosaic image and the recognition result. While most part of the mosaic image was successfully created, several misrecognitions can be observed. The misrecognitions were mainly due to segmentation errors, called *over-segmentation*, such that "m" is misrecognized as "r" and "n". The misrecognitions of this type are often found in the results of segmentation-by-recognition techniques. A well-known remedy for this problem is the use of a word lexicon. The use of a wider slit is also effective to suppress such misrecognitions as will be shown later.

Figure 11 shows a result of the general case where the one-pixel slits were used. As noted in Section 3.1, the video frames were synthesized from the video frames used in the above simple case experiment. (That is, the scanning speed uctuation between 0$\sim$ 2 pixel/frame was appeared together with rotation, scaling, and vertical-shift.) While most part of the mosaic image (c) is well created, the part around misrecognitions shows degradation. For example, the last character "o" is deformed to be close to "v" by abusing the e xibility on controlling slits. Thus, this misrecognition ("o"$\rightarrow$ "v") is caused by so-called *over-fitting*, which often degrades the performance of elastic matching-based character recognition (e.g., [8]). The use of some sophisticated pixel feature (e.g., directional feature, background feature, crossing feature, localized moment feature, etc.), a

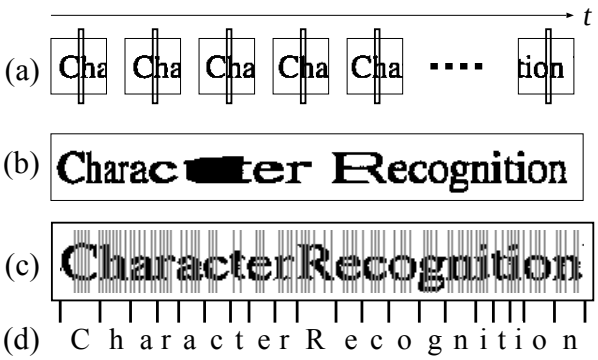

Figure 10. Result of the simple case. The original text is "Character Recognition". (a) Input video frames with one-pixel slits. (b) Image created by simply concatenating their one-pixel slits. (c) Mosaic image and (d) recognition result provided by the simple case algorithm.

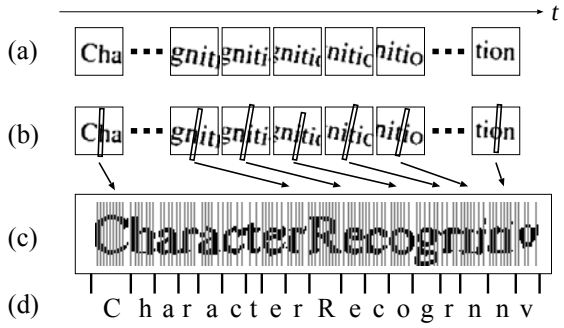

Figure 11. Result of the general case. The original text is "Character Recognition". (a) Input video frames. (b) The optimally controlled one-pixel, (c) mosaic image, and (d) recognition result, provided by the general case algorithm.

word lexicon, and a wider slit will be still helpful to reduce such misrecognitions due to over- tting.

3.3. Quantitative analysis

Figure 12 shows the recognition rates for the general case. A wider slit with 20 pixel width was used here. The camera scanning speed was uctuated between 0$\sim$ 2 pixel/frame. This result shows that the proposed technique could provide recognition rates over 95% even when the video frames undergo scaling and vertical shift. Considering that we only use a naive gray-level feature to obtain

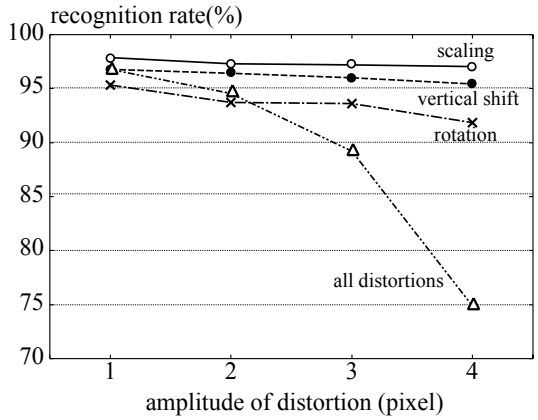

Figure 12. Recognition rate for the general case. The horizontal axis represents the amplitude of distortions, k (pixels). The slit width W was fixed at 20.

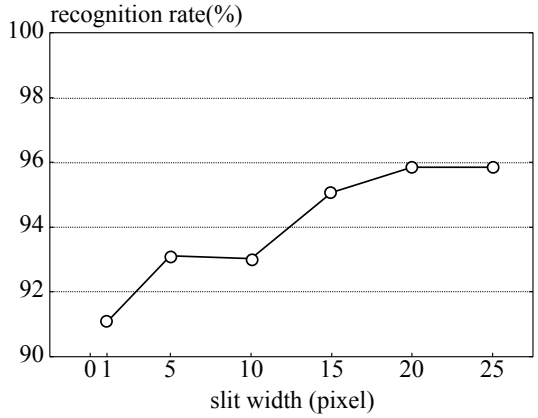

Figure 13. Recognition rates as a function of slit widths W. Here, the simple case with scanning speed fluctuation between 0$\sim$ 2 pixel/frame was assumed.

matching score $d_t(c, j)$, those rates are acceptable one. The recognition rates were degraded by rotation. The reason of this degradation was quantization errors on dewarping to compensate the rotation. Thus, this degradation can be minimized by using blurring operation, local perturbation matching, invariant feature, and so on.

Figure 13 shows the effect of slit width W on recognition accuracy. This result was of the simple case; that is, only camera scanning speed uctuation (0$\sim$ 2 pixel/frame) was imposed. The constant K which de nes the acceptable scanning speed was xed at 2. The result shows that recognition accuracy is improved by increasing W to 20 pixels, i.e., about half of average character width. When

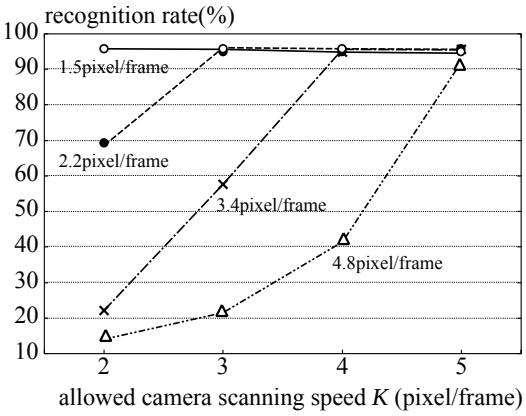

Figure 14. Recognition rates at higher camera scanning speeds. ($W = 20$)

using wider slits, the misrecognitions between similar characters (e.g., "I" to "1") and the misrecognitions due to oversegmentation (e.g., "m" to "r"+"n") were successfully reduced.

Figure 14 shows the recognition rates when higher camera scanning speeds were allowed by using larger K. In this experiment, the video frames whose scanning speed was x ed at about 1.5, 2.2, 3.4, or 4.8 pixel/frame were used. No geometric distortion (rotation, scaling, and vertical-shift) was imposed on those frames. The result of Fig. 14 clari es that K should be x ed at larger values for compensating higher scanning speeds. For example, when the scanning speed is 4.8 pixel/frame, K should be set at 5 or more. Conversely, when K was smaller than the scanning speed, the recognition rate was seriously degraded.

4. Conclusion and future work

A mosaicing-by-recognition technique was proposed for recognizing texts in multiple video frames and mosaicing those frames. Those two procedures, i.e., recognition and mosaicing, are simultaneously and collaboratively performed in a one-step manner by a DP-based optimization algorithm. Experimental results showed that the proposed technique can attain about 90% character recognition rate even when rotation, scaling, vertical shift, and speed uctuation appear in the frames.

Future work will focus on the following points:

- *Lexicon*: The proposed technique often produces misrecognitions by over-segmentations (e.g., "m"$\rightarrow$ "r" and "n") and over- tting (e.g., "o" $\rightarrow$ "v"). Like the other text recognizer based on segmentation-by-recognition framework, the use of lexicon will be help-

ful to exclude such misrecognitions.

- *Sophisticated pixel feature*: In the experiment conducted in this paper, only naive pixel feature, i.e., gray-level feature, was used. Since this feature is very weak to geometrical distortions, sophisticated pixel features, such as directional feature, background feature, crossing feature, and localized moment feature, should be used for improving the matching between a slit and a reference pattern.

- *Reduction of computational complexity*: Beam search techniques (cost-based pruning and lexicon-based pruning) will be effective to reduce the computational complexity.

Acknowledgment: This work was supported in part by the Research Grant of The Okawa Foundation and the Research Grant (No.17700198) of The Ministry of Education, Culture, Sports, Science and Technology in Japan.

References

[1] D. Doermann, J. Liang and H. Li: "Progress in Camera-Based Document Image Analysis," Proc. ICDAR, pp. 606–616, 2003.

[2] A. Zappala, A. Gee, M. Taylor: "Document mosaicing," Image and Vision Computing, vol. 17, no. 8, pp. 585–595, 1999.

[3] H. Sakoe, H. Fujii, K. Yoshida, and M. Watari: "A high-speed DP-matching algorithm based on frame synchronization, beam search and vector quantization," Systems and Computers in Japan, vol. 20, no. 11, pp. 33-45, 1989.

[4] R. Plamondon and S. N. Srihari: "On-Line and Off-Line Handwriting Recognition : A Comprehensive Survey," IEEE Trans. Pat. Anal. Mach. Intell., vol. 22, no. 1, pp. 63–84, Jan. 2000.

[5] T. Sato, S. Ikeda, M. Kanbara, A. Iketani, N. Nakajima, N. Yokoya, and K. Yamada: "High-resolution video mosaicing for documents and photos by estimating camera motion," Proc. SPIE Electronic Imaging, vol. 5299, 2004.

[6] H. Li and D. Doermann: "Text Enhancement in Digital Video Using Multiple Frame Integration," Proc. ACM Multimedia, pp. 19–22, 1999.

[7] J. Kosai, K. Kato, and K. Yamamoto: "Recognition of low resolution character by a moving camera," Proc. 5th Int. Conf. Quality Control by Arti cial Vision (QACV'99), pp. 203-208, 1999.

[8] S. Uchida and H. Sakoe : "Eigen-deformations for elastic matching based handwritten character recognition," Pattern Recognition, vol. 36, no. 9, pp. 2031–2040, 2003.