25 – Text Localization and Recognition in Images and Video

Seiichi Uchida

Short introductory text

Nowadays, most textual information is captured by camera rather than scanner. Camera-based OCR is the technology to recognize characters captured by camera. Since camera is far handier than scanner and possible to capture various textual information in scene (such as characters on a signboard), realization of camera-based OCR will develop new applications of OCR. On the other hand, characters captured by camera are different from those in scanned documents and their recognition is very difficult due to various reasons. For example, their appearance is often distorted by perspective, complex background, blur, low-resolution, non-uniform lighting, and special decoration. In addition, their localization is also difficult because there is no prior knowledge of their location and there are many character-like non-character patterns. For the solution of those difficulties, we have to not only evolve the traditional image processing methods but also incorporate the state-of-the-art pattern recognition technologies. In addition, camera-based OCR deals with a "real-world" problem, general computer vision technologies are more and more important for OCR. This chapter introduces those problems and solution technologies to the realization of camera-based OCR.

1 Recognition of texts in images

1.1 Background

In past researches, the target of OCR is mainly document images acquired by scanner; this is because scanner can provide images with sufficiently high resolution for OCR. Assume that each character should have 32 pixels in its height for an OCR. In this case, for an A4-sized document (8.3×11.7 inches) with 40 text lines and line spacing of 18-pixels, the height of the entire document image must be larger than 2,000 pixels. This indicates that resolution of the scanner should be higher than 170dpi. For dealing with more complex characters, such as Chinese and Japanese characters, a higher resolution, say 300dpi, is necessary. Fortunately, even old scanners can have enough resolution and thus they have been used for OCR.

Recently, digital camera also becomes an important device to acquire document images for OCR. Even a popular digital camera has a huge number of CCD elements. For example, my small digital camera (whose price was less than 200USD in 2011) can provide images with 4,608×3,456 pixels. Clearly, my camera has a potential to capture an A4 document with 50 text lines with a sufficient resolution for OCR. In addition, camera-equipped mobile phones (and smartphones) become very popular. Accordingly, not only fans of photography but also many people carry cameras and have chances to acquire various kinds of text information through their cameras.

This is good news for OCR to discover its new targets. Since digital camera is far handier than scanner, we can consider various new OCR targets. Scanners generally deal with texts on paper sheet (while they have their own improvement [1]). In contrast, digital camera has no limitation on the target. In other words, digital camera can capture any texts around us. For example, digital camera captures big signboards in a town, digital texts on a computer display, precious historical documents, pages of a thick and heavy bound book, a timetable on a station, notes on a blackboard, slides projected on a screen, vehicle license plates in an omnidirectional image, etc.

Consequently, OCR for digital camera images (often called "camera-based OCR") becomes one of important directions for researches on character recognition and document image analysis. In fact, as introduced in survey papers [2][3] and this chapter, there are many papers which tackle various problems around camera-based OCR. Moreover, camera-based OCR has been incorporated even in commercial services and thus becomes more popular.

1.2 Applications

There are many applications of camera-based OCR. The first and most straightforward application is to recognize various characters and documents without putting them on any scanner.

- Car license plates are popular and important targets (e.g., [4]). The systems recognizing the plates are called automatic number plate recognition (ANPR) and have been used around early 1980s for police systems, traffic control, toll collection, and parking systems. In some scene browsing service, such as Google Street View, car license plates should be detected and hidden for privacy preservation [5].
- Several smart desktop systems have been proposed, where a camera will capture documents or texts on the desk and understand their location and contents (e.g., [6]).
- Business cards are popular targets of camera-based OCR on mobile phone.
- Thick books and precious historical books are also a target. In fact, book scanners capture page images by a single digital camera (or a stereo camera pair) (e.g., [7]). In [8], a complete digitization system based on a stereo camera-based book scanner has been proposed. See also dewarping methods listed in Chapter 2.1

It will be easy to find other targets, such as presentation slides on screen, messages on digital signage, signboards, and posters and plates on wall. Even popular paper documents will be recognized by camera-based OCR because image acquisition by camera is handier than that by scanner.

Image search is also a good application of camera-based OCR. In this application, a personal or large public image database is searched for images including a keyword specified. For example, we can find scene images capturing the word "park" by using the keyword "park". In 2011, several commercial web services, such as Evernote, Google Goggle, and Microsoft OneNote, are already available. An important thing is that these services can assign multiple recognition results to each text region for better retrieval accuracy (or, better recall rate). Consequently, they may have many excessive and erroneous recognition results. Fortunately,

those misrecognition results are not a serious problem for the services, because the misrecognition results are not fatal for their search results.

A similar but different image search application is camera-based document image retrieval. Document image retrieval is a technology of finding the similar or identical document image stored in a database, given a query document image. Its essential difference from the keyword-based image search is that it is not always necessary for document image retrieval to recognize individual characters and words. For example, it is possible to retrieve the identical document image by matching entire document images. Recently, cameras become popular for capturing document images (e.g., [9][10]). For more discussion related to document retrieval, refer Chapters 3.3 (Page Similarity and Classification) and 6.6 (Image Based Retrieval and Keyword Spotting in Documents).

Another application is mobile dictionary or translator (e.g., [11][12][13][14]). Assume that now you are traveling around Japan and find a nice restaurant for your dinner. You want to have some special dishes and find the document of Figure 1 on your table. With camera-based OCR and translator on your mobile phone, you will understand that the first "column" of the document (Japanese texts are often written vertically) means "Today's menu", even though you cannot understand those Japanese characters at all.

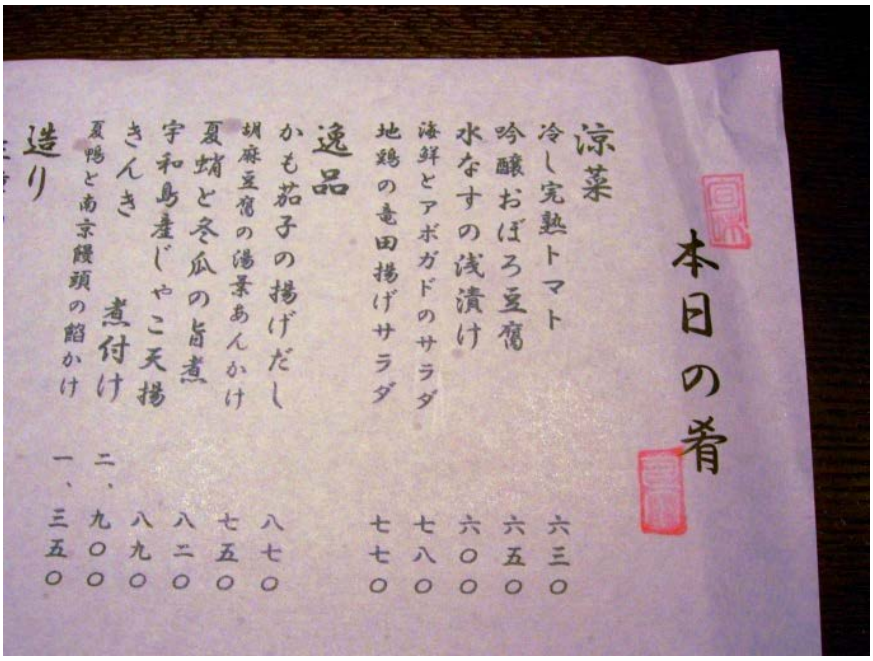

Figure 1 A food menu written in Japanese. (The photograph was taken by Gustavo Veríssimo and uploaded at Flickr Creative Commons.)

For navigation, scene texts are important clues. In our daily life, we extract text information from scene and then navigate ourselves. In fact, we will lose our way in an unfamiliar place without text. In a shop, we cannot find what we want without text. Sign recognition [15] will

be useful for navigation. A similar trial was done to navigate blind persons (e.g., [16][17]). Another navigation application has been done by cameras equipped on a mobile robot (e.g., [18]).

Handwritten characters are also target of camera-based OCR. For example, in [19], handwritten characters on whiteboard are to be recognized by a video camera system. In [20] , an early attempt of a video camera-based system has been proposed, where handwriting patterns are captured by pen-tip tracking. As camera has smaller size and higher resolution, the combination of camera and pen will be examined continuously, like [21].

Technologies used in camera-based OCR are helpful to realize various paper-based user-interfaces, as reviewed in [22]. This kind of user-interfaces can be enriched by printing or embedding watermarks, small dots, and barcodes. Anoto digital pen technology utilizes small dots printed on paper for not only precisely determining pen-tip position but also identifying the paper itself. These dots are captured by a small camera equipped inside a pen device. Consequently, the Anoto pen can acquire its movement by using the image of dots from the camera. Without embedment, it is possible to utilize "paper finger print", which is a micro-fiber structure of paper and also can be used for paper identification.

2 Recognition of texts in video

2.1 Background

A similar research topic to camera-based OCR is OCR for texts in *videos*. This research topic can be further divided into two types [23]. The first type is to recognize *scene text*s, which are the texts in the scene captured in video frames. This is almost the same as the above camera-based OCR except for the fact that we can utilize multiple frames to recognize a text. Specifically, we can find the multiple images of a word in 30 video frames, if the video camera (30 fps) captured the word for 1 second. We can utilize them for better performance of camera-based OCR.

The second type is to recognize *caption texts*, which are the texts superimposed on video frames. Since caption texts are attached intentionally to explain or summarize the contents of the video frames, they are also useful as an accurate index of the video. Caption texts are also captured in multiple video frames like scene texts, and have their own characteristics. For example, the location, color, and size of caption texts are almost constant within a single video sequence. This constancy can simplify the recognition task. Some caption texts are scrolling from bottom to top with a constant speed regardless of background movement.

Since caption texts are generally machine-printed characters superimposed on frame, and their poses and positions are often fixed within each video. Thus, caption texts are more constrained than scene texts and thus more tractable. There have been many trials on extraction and recognition of caption texts, such as [24][25]. A review is found in [23].

2.2 Applications

Applications of recognition of texts in video are summarized as follows. Since the applications of scene text recognition are already described in 1.2, the applications of caption text recognition are focused here.

The most typical application of caption text recognition is content-based video retrieval and annotation. Since the length of each video often exceeds several hours, manual inspection of large video database, such as broadcast video database, is almost impossible. Thus, automatic retrieval by text query is necessary and many trials have been made.

TRECVID [28] is a famous competition for improving content-based video retrieval and annotation and other video understanding technologies. Its active and long history from 2001 indicates the difficulty of video understanding. In fact, this task includes recognition of general visual objects, which is a well-known difficult problem. For example, if we want to retrieve videos capturing a coffee cup, we need to recognize cups in the videos. This is difficult because of the ambiguity of the object class “coffee cup”. Coffee cups have various appearances (by different colors, shapes, materials, and poses). In addition, we cannot define the explicit and general difference between bowl and cup.

Captions (i.e., characters) are far less ambiguous than visual object and thus useful for content-based video retrieval and annotation. This is because any character has been developed and used for accurate human communication during thousands of years. In addition, captions are often strongly related to the contents of the video frame. For example, the caption of a news program will be the headline of the current news topic. Consequently, if captions are recognized, video sequences related a query keyword can be retrieved automatically and accurately.

There are other applications than video retrieval. In [24], caption texts are used for video skimming, which is a technique to extract key frames for video compaction. Zhang and Chang [29] have developed an event analysis method for baseball game videos based on caption text recognition. Bertini et al. [30] have tried to recognize not only caption text but also jersey number for identifying soccer players in video. In their system, face is also used a cue for the identification and they showed those textual cues are helpful to improve reliability of the identification result. TV logo detection and tracking methods have been developed [31][32] for detecting and skipping commercial block, removing the TV logo, and finding unauthorized distribution. Time stamp on photography, which is not a video text but similar to a caption text, is also a target [33].

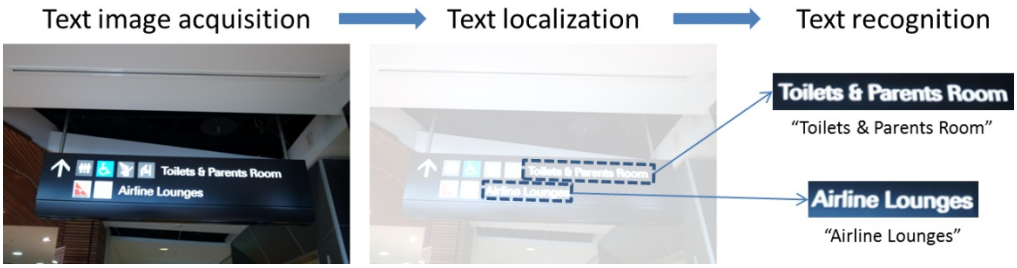

Figure 2 Three tasks of camera-based OCR

3 Three tasks and their difficulties

Generally speaking, recognition of texts in scene images and video is comprised of three tasks; *text image acquisition*, *text localization* and *text recognition*. Figure 2 illustrates the three tasks. The text image acquisition task is image capturing and preprocessing for improving quality of the captured image. This preprocessing will make the succeeding text localization and recognition tasks easier. Rectification and deblurring are examples of the preprocessing. The text localization task is to find texts in the image. The text recognition task is to identify the class of each character in the text.

We can find the same three tasks in ordinary OCR; we scan the document and remove its skew, find text lines, and recognize characters and words on the text lines. However, texts in scene images acquired by camera are far more difficult to be localized and recognized than texts in ordinary document images acquired by scanner. Major difficulties of camera-based OCR are listed in Figure 3 and they will be detailed in this section.

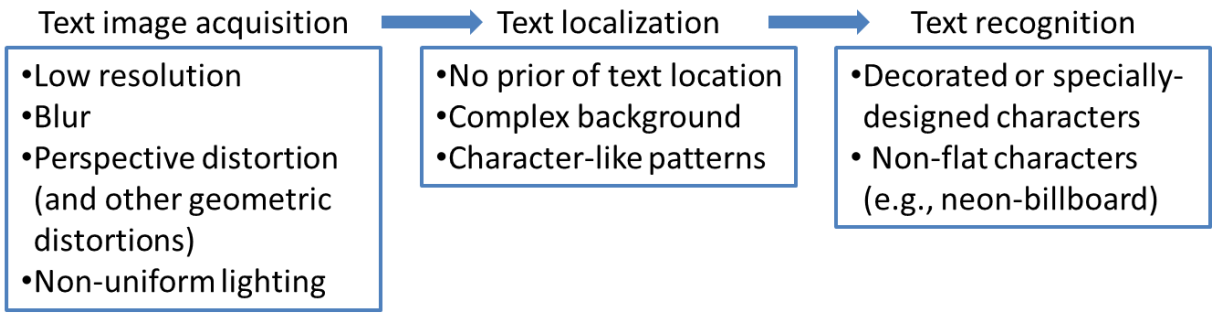

Figure 3 Major difficulties for camera-based OCR.

In [2], Jung et al. have defined that recognition of scene texts is comprised of five tasks; detection, localization, tracking, extraction and enhancement, and recognition. In their definition, the detection task is to find text regions and the localization task is to group the detected regions. The tracking task is to find the same text regions captured in multiple contiguous frames. The extraction and enhancement task is to convert the localized text regions into a binary image. Although the following sections assume the foregoing three tasks (acquisition, localization, and recognition), these five tasks are included in them.

3.1 Text image acquisition

The first step of camera-based OCR is to acquire text images through various kinds of still cameras or video cameras. Many difficulties of camera-based OCR are caused by this acquisition step; that is, low resolution, motion blurring, perspective distortion, non-uniform lighting, specular by flash, and occlusion. Those difficulties are particular for the camera-based OCR. Accordingly, if we want to use an OCR engine of the traditional systems, we must apply image processing techniques to camera-captured images for making them as much similar as scanner-captured images in advance. As noted above, rectification and deblurring are examples of the preprocessing. In the following section, those preprocessing techniques for camera-based OCR will be reviewed.

3.2 Text localization

The second step of camera-based OCR is text localization, which is the task to detect texts in a scene image or a video frame. If we capture an ordinary document by a camera, the localization task is similar to text line extraction process for scanned documents. In contrast, if we capture a scene, the localization task becomes more complicated. In the following, the difficulties of text localization in scene images are discussed.

The first difficulty of scene text localization is that texts and words are scattered in the scene image. What is worse, we do not have any prior information of text location, text size, text orientation, and the number of texts. Since there is no formatting rule in scene texts, we cannot localize them by text segmentation methods for document images. Text size is not constant. For example, the text on a signboard close to the camera may be very larger than the text on another signboard far from the camera. The orientations of text lines are not constant. In addition, text lines, such as a label text on a bottle, are often curved.

The second difficulty of scene text localization is complex background. If texts are printed on a solid-colored background, the separation of the texts may be rather easy like ordinary scanned documents. However, in scene images, characters and their background often have a very low contrast and thus the separation is difficult even for the solid-colored background. If texts are printed on a textured background, or a background with color gradation, or a picture, the separation becomes a difficult problem. Captions are often superimposed on video frames directly; that is, they do not have any solid background. This case is one of the most difficult cases of text localization.

The third difficulty is that there are many patterns which seem like characters. In other words, there are many ambiguous and confusing shapes in scene. On the corner of a room, we will find “Y”-shaped edges. Around leaves of trees, we also find dense and fine (i.e., high-frequency) edge structures which look like dense text lines. Conversely, we know that some decorated characters seem like a branch of a tree. This difficulty invites a very important question; *what are character patterns?* More precisely, what is the difference between character patterns and non-character patterns? In fact, we never read the corner as “Y”.

The results of localization will be given as a set of rectangles (or, more generally, arbitrarily-shaped connected regions) each of which contains a single character, or a single word. For

single character rectangles, neighboring character rectangles will be grouped (i.e., concatenated) to form a word. As noted above, texts in a scene image are often rotated or curved originally or by non-frontal camera. Accordingly, this grouping process is not trivial.

Figure 4 Various characters in scene texts. (The photographs were taken by Steve Snodgrass and uploaded at Flickr Creative Commons.) The characters in the lower three images are three-dimensional characters.

3.3 Text recognition

After the localization, the third task, that is, recognition of detected texts is performed. This step is less severe than the above two tasks, if the target of camera-based OCR is ordinary business document printed on paper. This is because that such a document is printed with regular fonts, such as Time-Roman, and thus we can use some traditional character/word recognition engine (if we can expect that the above two tasks are solved sufficiently).

Text recognition, however, becomes very difficult when the target is not ordinary paper document. In fact, general scene texts often contain decorated or specially-designed characters. Figure 4 shows several examples of those characters. Although some of them seem machine printed characters, they have particular appearance and thus difficult to be recognized by traditional OCR engines. In captured scene images, we will also encounter more difficult characters to be recognized, such as calligraphic characters, pictorial characters, multi-colored characters, textured characters, transparent characters, faded characters, touching characters, broken characters, very-large (or small) characters. In addition, we encounter handwritten characters on notebook and whiteboard, and “handwritten fonts” on signboard and poster.

There are further variations in character appearance by the three-dimensional nature of scene. The lower three images of Figure 4 include non-flat characters. Some of them are engraved

characters and some are three-dimensionally formed characters (for a neon-billboard). In addition, partial occlusion will damage the character appearance very severely.

Table 1 Problems on image acquition and their typical solutions.

Problem	Solution	Short description	Remarks
Low resolution	Super-resolution by multiple-frame processing	Formulated as an optimal estimation problem of the original high-resolution image from multiple low-resolution inputs.	Cons: Not applicable to still image
	Instance-based super-resolution	Reference to the instances each of which is a pair of a high-resolution image and its low-resolution version.	Pros: Applicable to still image Cons: A sufficient number of instances
Blur	PSF estimation	Solving inverse problem with document/character-specific prior knowledge, such as existence of edge, binary color, font style, etc.	
Perspective distortion	Boundary-based method	Parameter estimation using document boundary shape.	Pros: Simple and easy implementation Cons: Only for texts on rectangular paper with visible boundary
	Text line-based method	Parameter estimation using the direction of regularly aligned text lines.	Pros: Accurate in the direction of text lines Cons: Less accurate in the direction perpendicular to the text lines; not applicable to irregular text
	Character shape-based method	Parameter estimation using the direction of character strokes.	Pros: Robust to irregular text. Cons: Language-dependent
	Instance-based method	Parameter estimation referring stored instances.	Pros: Applicable even to a single character. Cons: Large memory and computation.
Non-uniform lighting	Local binarization methods	Non-uniform background elimination by using a threshold value optimized locally.	See Chapter 2.1 for various binarization methods.
	Image composition method	To remove a strong specular reflection, two (or more) images with different lighting conditions are composed into one image.	

4 Methods to solve problems in image acquisition

In this section, the methods to solve various problems caused in image acquisition step are reviewed. Table 1 lists the problems discussed in this section. Note that several topics are related to scanner-based OCR and thus will not be described in this chapter. For example, the method to tackle with document images on non-flat surface will not discussed here but in Chapter 2.1. Various binarization methods for document images are also described in Chapter 2.1.

4.1 Low resolution

The image processing to convert a low-resolution image or a set of low-resolution images into a more visible image is a kind of image enhancement processes. This is called *super-resolution* because this image processing often provides higher-resolution images for better visibility. There are two approaches. The first approach is multiple-frame processing and the second is the instance-based approach. The first approach is applicable for video texts and the second is applicable even for scene texts, i.e., texts in a single still image.

4.1.1 Multiple-frame processing

The first approach based on multiple-frame processing utilizes the fact that the same text appears in multiple frames in a video. Simply speaking, low resolution images of a text in multiple frames are redundant but erroneous representation of the original text. Thus, by applying some error correction technique (like error correcting code), the original high resolution image can be recovered.

A simplest version of super-resolution is simple averaging of multiple frames [34][35]. Since caption texts are often fixed at a certain position, the averaging operation will enhance the caption text part and cancel its background clutter. If a caption text moves by, for example, scroll-up/down, the text should be tracked precisely before averaging. If a text undergoes other deformations, such as affine deformation, the deformation should be estimated and removed before averaging. Mancas-Thillou and Mirmehdi [36] have improved this averaging approach by adding a high-frequency component. This high-frequency component is extracted from the low-resolution images by the Teager filter, which is a kind of high-pass filters.

Capel and Zisserman [37] have examined performance of a maximum likelihood (ML) based approach, where observed low resolution images are assumed via a Gaussian blurring process and the original image which most likely produces those low resolution images are derived. In their paper, starting from this simple ML based approach, its iterative version, called back-projection [38], is also examined. They also proposed two maximum a posterior (MAP) based approaches by introducing a prior into the ML approach. One incorporates a prior to evaluate the total image gradient, and the other incorporates a prior to evaluate the total variation.

Donaldson and Myers [39] have proposed an improved version of [37] by introducing a new prior to the MAP framework. This prior is based on the bimodal (i.e., black and white) property of document images. They also introduced a smoothness term not to lose the step discontinuity around the boundary of characters.

Like the bimodal prior of [39], it is possible to introduce other priors suitable for MAP-based document image super-resolution. Banerjee and Jawahar [40] have proposed a MAP-based super-resolution method where the bimodal prior is employed along with an edge-preserving smoothness term. Bayarsaikhan et al. [41] have proposed a prior evaluating the total variance. Their prior is different from the simple total variance prior on the point of using a weight to the local direction of edges for avoiding over-smoothing around character edge.

An important point for multiple-frame processing methods is that they require precise geometrical alignment of multiple frames in advance to averaging procedure or MAP estimation. Consequently, some tracking method in sub-pixel accuracy is often employed for caption texts. More generally, some nonlinear image registration is necessary for two or more consecutive frames. For example, RANSAC is utilized in [36]. In [42][43], document mosaicing methods are proposed for perspective registration of multiple images.

4.1.2 Instance-based approach

The instance-based approach does not require multiple video frames as its input. Instead, this approach prepares many instances in advance. Each instance is a pair of an original high-resolution image and its low-resolution version. This instance is an example representing how a high-resolution image becomes in its low-resolution version. Conversely, the instance also represents how a low-resolution image becomes in its high-resolution version. Consequently, this instance provides a well-grounded evidence for super-resolution.

An important point of this approach is that a sufficient number of instances are necessary for better super-resolution results. Clearly from the above principle, if we cannot find a low-resolution image instance close to the input low-resolution image, it is impossible to guess the original high-resolution image of the input image. In other words, if we want to apply the method for a specific class of documents, it is better to collect a sufficient number of document images of the class. Another important point is that this approach is free from the problem of tuning priors, such as smoothness.

Baker and Kanade [44] have proposed a super-resolution method based on this approach. In their method, called *hallucination*, high-resolution image instances are firstly collected into a database. Then, each of them is represented in multiple resolutions and stored in the database together. When a super-resolution image of an input low-resolution image is necessary, the database is searched for the nearest-neighbor of each local part of the input image. Since the image of the nearest-neighbor part in its original high resolution is known, it is possible to reconstruct a super-resolution image. Although this method is mainly applied for face images, several results for document images are also shown in [44]. Dalley et al. [45] also have developed a similar method independently based on their past trial [46] on instance-based super-resolution for general images.

Like the multiple-frame approach, the instance-based approach also can be considered to be based on a prior model. In fact, in [44], the instance is considered as a “recognition-based prior”. In this sense, the instance-based approach also can incorporate other priors, such as a bimodal prior. Park et al. [47] have proposed a hybrid method where the recognition-based prior by instances is combined with a bimodal prior in a Bayesian framework.

A more direct approach for dealing with low-resolution text images has been developed by Jacobs et al. [48]. In their method, low-resolution text images are directly processed by a word recognizer which is trained by low-resolution images. Specifically, a low-resolution word image is first decomposed into fixed-size sub-images by a sliding window and each sub-image is then processed by a character recognizer based on a convolutional neural network. The character recognizer is trained by a sufficient number of camera-captured, i.e., degraded character examples. The final word recognition result is obtained by a dynamic programming search, like general analytical word recognition methods.

4.2 Blurring

Camera-captured document images are often blurred. Since no blur is happened in scanned document image, blurry image is a difficult target for conventional OCR. Several trials have been done for blur removal, or deblurring. Note that deblurring is a kind of image enhancement processes and thus related to the other image enhancement process, i.e., super-resolution.

There are two reasons of blurring; imperfect focus and motion. For both cases, a blurred image is considered as the result of convolution between an original image and point spread function (PSF). PSF represents how a single pixel in the original image is spread over in the blurred image.

Consequently, for deblurring, we need to estimate two signals, that is, and the original image, simultaneously from a single input image. This simultaneous estimation problem is so-called *blind deconvolution*, which is a well-known ill-posed inverse problem. A very intuitive explanation of this problem is as follows: a number “24” is given and it is known that this number is the result of multiplication of two numbers. What are those two numbers? Of course, there is no unique answer. They may be 1 and 24, or 4 and 6, or -4 and -6, etc. So, we need some additional information, i.e., a prior, to choose the best answer from those three candidates.

In the above naïve example, if we know one of those two numbers, the other number is immediately determined. This fact suggests that if we can estimate the PSF in any way, the original image is directly recovered through the inverse filtering based on the estimated PSF. Unfortunately, for the deblurring problem, even if the PSF is given, the determination of the original image is still not straightforward. For example, in [49], a regularization-based method is proposed for determining the original image with a given PSF.

A key idea for deblurring camera-captured document images is to utilize specific properties of document images as prior knowledge for regularizing the PSF estimation problem. For example, Tian and Ming [50] have proposed a method to estimate two-dimensional PSF by observing character edges. Specifically, the intensity changes across an edge in horizontal and vertical directions are directly considered as the horizontal and vertical projections of PSF, respectively. The second idea is that PSF is estimated locally. This is because the degree of blur depends on the depth (i.e., distance from the lens) of target scene and thus multiple objects located in different depths have different blurs. In [50], the entire image is simply

divided into several sub-images and for each sub-image its local PSF is estimated in the above way.

In addition to the existence of edges, other kinds of prior knowledge on the target document image can further improve the estimation condition. Kanungo and Zheng [51] have proved that the knowledge of the font type of the target document is helpful to estimate the parameters of document image degradations, such as blurring. Cheng et al. [52] used the "bimodal" prior which indicates document images are generally binary images. In [49], a traditional smoothness prior has also introduced for deblurring with special consideration to suppress its side-effect around character edges.

Different from isotropic PSF of imperfect focus, PSF of motion blur has a direction which reflects motion direction. Accordingly, in order to remove motion blur, it is necessary to estimate not only blur size but also blur direction [53]. In [53], the motion blur is estimated in the two-dimensional Fourier spectrum domain because the motion blur diminishes the high frequency components of the motion direction.

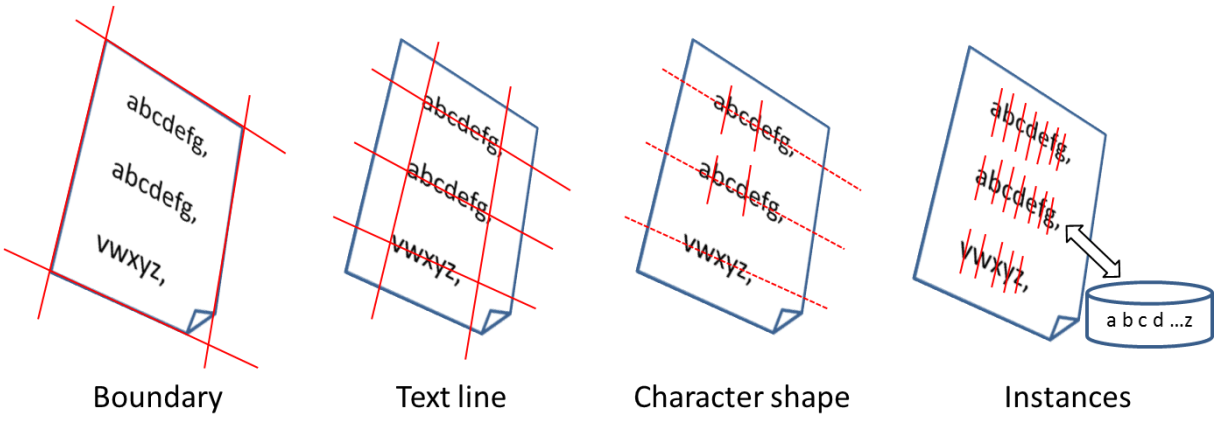

Figure 5 Four approaches for estimating perspective distortion.

4.3 Perspective distortion

When document image is captured by camera, the target image often undergoes *perspective distortion* due to non-frontal camera. Since perspective distortion will degrade performance of text line extraction, word/character segmentation, and character recognition, it should be removed. For image capturing by scanner, the target image undergoes just skew distortion, i.e., rotation distortion. Skew is represented by a single parameter (rotation angle) and thus rather easier to be estimated and removed. In contrast, perspective distortion is represented by more parameters because camera posture (position and direction) relative to the target document has much more freedom than scanner. Consequently, the estimation and removal of perspective distortion, or rectification, becomes more difficult.

Figure 5 illustrates four typical approaches to estimate perspective distortion. Each of these approaches is detailed in the following. Note that skew removal (called deskew) is described in Chapter 2.1. Removal of nonlinear distortion of curled documents and crumpled documents is called dewarping and described also in Chapter 2.1.

4.3.1 Estimation of Perspective Distortion by Document Boundary

The simplest approach to estimate perspective distortion is to utilize the four boundary edges of document sheet. Clark and Mirmehdi [54] have proposed a method on this approach. In their method, line segments are found by Hough transform and then four line segments forming a quadrilateral are selected as a candidate of a document sheet region. If the inside of the candidate quadrilateral is evaluated to contain characters, it is determined as a document region. The evaluation is done by checking the sum of edge vectors. The shape of the quadrilateral shows the perspective distortion and the distortion can be removed. A boundary based removal of affine transformation is proposed in [15].

4.3.2 Estimation of Perspective Distortion by Text Lines

When the document boundary is difficult to be found, the direction of linear text lines can be utilized to estimate perspective distortion. There are two types of the methods which are utilizing text lines for estimating perspective distortion. In the first type, text lines are not extracted explicitly. Instead, a projection profile along a certain direction is derived to evaluate how the direction is valid as the direction of text lines. In the second type, text lines are explicitly extracted by some line tracking method.

The methods of the first type are as follows. Clark and Mirmehdi [55] have utilized this fact to determine two vanishing points. Specifically, the horizontal vanishing point is determined by using projection histogram. For each candidate point, a projection profile is created by counting the number of black pixels on each of lines radially-emerged from the point. The point where the large peaks are found is selected as the optimal horizontal vanishing point. (More precisely, the point with the maximum squared sum of histogram values is selected.) The vertical vanishing point is then determined by using the optimal horizontal vanishing point. The task of determining the vertical vanishing point is more difficult because no line structure is expected in the vertical direction. In [55], two end points (the left end and the right end) and the center point are detected along each of the above lines radially-emerged from the horizontal vanishing point and their arrangement is used for the task. Dance [56] has proposed a similar method of estimating two vanishing points.

The methods of the second type are as follows. Myers et al. [57] have proposed a method based on text line extraction. They extracted each text line by a simple blob-linking process. Then the bounding box of each individual text line is determined and rectified. Since the characters in the resulting image still undergo slant deformation, it is estimated by observing vertical projection profile. An important point is that those processes are performed on each text line independently. In other words, this method does not assume that the target image contains several text lines. Consequently, this method can be used even for isolated short text line images, such as a signboard showing a restaurant name.

A similar text line extraction process is utilized in [7]. Different from [57], this method tries to utilize directions of multiple text lines for estimating perspective distortion. Specifically, RANSAC is used to estimate the optimal horizontal component of perspective distortion for the entire document. Then the remaining vertical component is estimated by using the arrangement of ending points of the text lines like [55].

In Yamaguchi et al. [67], the perspective distortion of each individual character is estimated. Their method deals with a single text line of digits. After extracting digit candidate regions by using several heuristics, global skew is first estimated by Hough transform. Then the slant of each candidate is estimated by fitting a tilted rectangle. This two-step approach is similar to [57].

4.3.3 Estimation of Perspective Distortion by Character Shape

In addition to document boundary and text line, character shape has also been utilized for estimating perspective distortion. Characters of many languages have vertical and horizontal strokes. For example, some Latin characters, such as “E”, “I”, “d”, and “h”, have a vertical stroke and some characters, such as “H”, “T”, “e”, and “t”, have a horizontal stroke. Among those strokes, the vertical strokes are helpful to estimate the vertical vanishing point. In fact, the estimation of the vertical vanishing point is generally more difficult than that of the horizontal one. As we see the above, the horizontal vanishing point can be estimated by using the directions of text lines but the vertical one cannot be estimated by them. Consequently, the above methods have used the arrangement of the end points or the center points of the text lines, although they are often unstable.

In Lu et al. [59], stroke boundary and top and bottom tip points are first detected at each character. Then, vertical strokes are detected from the stroke boundary. The vertical strokes are used for estimating the vertical component of perspective distortion and the tip points are used for the horizontal component. Liang et al. [60] also have developed a similar method where vertical strokes are used for the vertical component. The difference from [59] is that their method estimate horizontal text line direction in a local manner and thus is applicable for non-flat surface deformation.

Since the estimation of perspective distortion using character shape generally requires a larger amount of computation than the document boundary-based methods and the text line-based methods. Yin et al. [61] have proposed a multi-stage method where first a boundary-based estimation is performed. If its result is not reliable, a text line-based method is applied. The reliability of this result is also evaluated, and if it is not reliable, a character shape-based method is finally applied.

4.3.3 Estimation of Perspective Distortion by Instances

The above methods generally assume the linearity (i.e., one-dimensionality) of text line as an important clue for estimating perspective distortion. Camera-captured texts, however, sometimes do not have clear linearity. In a case, characters are laid out freely. In another case, texts are captured only partially and thus not enough to have a linear structure.

Uchida et al. [62] have proposed an instance-based approach for estimating text skew (i.e., rotation), which is considered as a constrained case of perspective distortion. In their method, the value of a skew-variant, such as the size of the upright bounding box, is calculated for each connected component. The skew of the connected component is determined by referring the instance which describes the relationship between the skew-variant and the skew angle. For example, the character “l” has a small bounding box size without skew and has a larger size with a skew and thus we can guess the skew by the size. Since at the skew estimation step,

the character class (such as “l”) is not known, they used skew-invariants to refer a proper instance. This method is extended as a part-based method [63] where local skew is estimated at each local part detected by a keypoint detector, such as SIFT and SURF.

The skew estimation method by Lu and Tan [64] is a more radical and promising instance-based approach. From each connected component, vertical component cut is first calculated. The vertical component cut counts how many times the center vertical line crosses strokes and thus will change according to skew. The normalized histogram of the vertical component cuts for all connected components is considered as a feature vector of the document. The nearest neighbor is searched for the database containing the feature vectors of various skewed documents and the skew angle of the nearest neighbor is determined as the estimation result. An interesting point is that this method detects languages simultaneously by preparing enough document instances printed in those languages.

4.4 Non-uniform lighting

Since camera-based OCR cannot expect a controlled uniform lighting environment, it should deal with non-uniform lighting, such as shade by a non-frontal light and specular reflection by a flashlight. Many camera-based OCR researches have tried to remove shade in their binarization process. The binarization process generally employs some local thresholding technique, which is detailed in Chapter 2.1. The effect of specular reflection can be reduced by the local thresholding technique; however, under strong reflection, it is impossible to restore the text image around the reflecting area. In [7], a dual-flash system is introduced, where two pictures are taken under different flashlights and then composed as a single picture without reflecting area. Koo et al. [65] have proposed a similar image composition method.

5 Methods to solve problems in text localization

As noted in Section 3.2, text localization is a very difficult task due to three main reasons:

1. No prior information of text location, size, orientation, etc.
2. Complex background.
3. Character-like non-characters.

For dealing with this difficult task, many researches have been conducted. In fact, text localization is, currently, the hottest topic of camera-based OCR and fascinating not only OCR specialists but also computer vision and machine learning researchers. A huge variety of approaches are developed and thus difficult to cover all of them. A couple of survey papers are found as [2][23].

It is noteworthy that competitions on text localization, called "ICDAR Robust Reading Competitions", have been organized [66][67][68] and played an important role to make this research topic very active. The result of the latest competition in 2011 [68] reported that the best method have achieved detection accuracy of around 70% in f-value on a dataset of scene images including texts. This means that we now (2012) can extract 70% true text areas with 30% erroneous non-text areas.

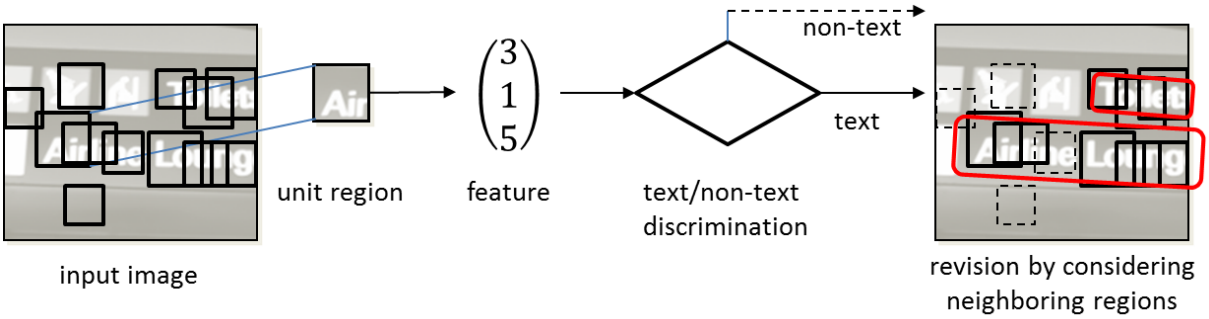

Figure 6 Typical process of text localization.

5.1 Typical process of text localization

Figure 6 illustrates a typical process of text localization. Its first step is feature extraction from a region. Then, the region is classified as a text region or non-text region. As an optional step, this classification result is revised by considering the classification results of its neighboring regions. If this revision tries to form a cluster of text regions, they are combined as a larger text region, such as a word region.

Of course, there may be text localization processes which do not follow the process of Figure 6 exactly. For example, we can consider a process where the first text/non-text discrimination step is performed at neighboring regions simultaneously. In addition we can consider a process where different features are extracted from different regions and then combined for discrimination.

5.2 Four aspects characterizing text localization methods

As noted above, there are a huge number of text localization methods and thus it is impossible to detail individual methods. This section, therefore, tries to classify them by four aspects according to Figure 6:

- Feature
- Method of text/non-text discrimination
- Region type
- Method of revision by considering neighboring regions

Each localization method is characterized by their choice on these aspects. For example, a method will extract a color feature ("feature") from a fix-sized block ("region type"), then perform an SVM classification ("method of text/non-text discrimination"), and finally revise the discrimination result by an MRF ("method of revision by considering neighboring regions").

5.3 Features for text localization

Table 2 lists typical features for text localization. Those features have been used because they are considered to be good for text/non-text discrimination. In other words, they are expected to capture some "character-ness". For example, a contrast feature is expected to have a large value around text regions because the characters and their background often have a large difference in their intensity values and/or colors characters. Note that most features, especially low-level features such as color, contrast, and gradient, have been designed by researchers' intuitive expectation. On the other hand, we can use a feature selection method, which is a kind of machine learning methods, for selecting good features. For example, we can use AdaBoost or random forest to select good features from a bank of features.

Among the listed features, context features have been used by a different purpose. Context features try not to evaluate "character-ness" directly but "contextual suitability for text existence". In fact, we have many empirical priors for text existence. In the sky, there may be no character. On leaves, there may also be no character. On a planar area, a text, i.e., a sequence of characters can exist. Gandhi et al. [103] have tried to detect texts by finding planar areas using a shape-from-motion technique. Park et al. [104] have tried to estimate the position of caption texts by the existence of a speaker. Kunishige et al. [105] have utilized environment recognition results (where possible classes of the environment were, ground, sky, green, building, etc.) for text/non-text recognition.

Note that we can consider that the recognized texts as context information for recognizing other objects in the scene. In other words, a text in a scene will useful to recognize its surrounding object and the scene itself. A clear example is that a word "cola" on a cylindrical object indicates that the object is a can or glass. A word "menu" indicates restaurant scenery. In [30], the numbers detected and recognized in a soccer game video are used to recognize player's face.

User interaction is also a promising approach to ease the difficult text localization problem. Nowadays, we have various pointing interfaces, such as touch pads, and thus it is easy to tell

the location of texts to machine. If the machine knows the rough text position by user's interaction, it can extract the corresponding text region precisely by using, for example, *grabcut* [108]. Focusing operation on a digital camera is also a promising user interaction which can be used as a prior for character detection [107].

Table 2 Features for text localization.

Features	Short description	Remarks	Example
Color / intensity	A text region is expected to have a higher contrast level or a clear bimodality in color histogram.	Shivakumara et al.[70] recommended using different discrimination criteria according to the contrast level.	Ezaki et al. (2005) [71](Bimodality by Fisher's Discriminant Rate)
Edge	A text region expected to have more edges than non-text region.		Kuwano et al.(2000)[72] (Edge pair on scanline), Sin et al.(2002)[73] (Spectrum of edge image), Liu and Samarabandu (2006)[74] (Multiscale edge)
Corner	A text region is expected to have corner points more densely than non-text regions.	Corners are detected as *keypoints* and utilized in visual object recogntion.	Bertini et al. (2001) [75](Harris corner), Huang and Ma (2010)[76] (Dense corner point area), Zhao et al.(2011) [77] (Harris corner)
One-dimensional gradient	A scanline across a text region is expected to have a specific gradient pattern.	Need of combination of results on adjacent scanlines for two-dimensional regions.	Wolf et al.(2002) [78], Wong and Chen (2003) [79], Kim and Kim (2009) [80] (Precise gradient aruond character edge), Phan et al.(2009)[81] (Maximum gradient difference),
Two-dimensional gradient	A text region is expected to have a specifc gradient pattern.	So-called « local features », such as SIFT and SURF, evaluate the 2D gradient of a region around a keypoint. HOG (Histogram of oriented gradients) also evaluates 2D gradient.	Wang and Belongie (2010) [82] (HOG), Uchida et al.(2011)[83] (SURF) Mishra et al.(2012)[84] (HOG)
Local texture	A text region is expected to have a specific local texture.	Pros : For example, local binary pattern (LBP) is a texture features invariant to intensity change.	Anthimopoulos et al. (2010) [85](LBP)
DCT coefficients	Spectrum features by DCT. As a kind of texture, a text	Pros : JPEG and MPEG also employ DCT.	Chaddha et al. (1994)[86], Zhong et al.(2000)[87],

	region is expected to have a specific frequency band.		Goto (2008) [88]
Wavelet / Gabor		Pros : Multiresolution is suitable to scale-invariant character detection.	Jain and Bhattacharjee (1992) [89] (Gabor) Haritaoglu (2001) [13](Edgeness by DWT), Saoi et al.(2005)[90], Kumar et al. (2007)[91] (Matched DWT), Weinman et al., (2009) [92] (Gabor)
Hybrid	Combination of the above (rather simple) features	Pros : Robustness by complementary feature usage. Note: Many features are introduced in [95].	Kim et al.(2004)[93], Chen et al.(2004)[94] (Edge+color) Pan et al.(2008)[96] (HOG+LBP), Peng et al. (2011)[97] (Edge+corner+Gabor), Yi et al. (2011)[17] (Gradient+edge)
Linearity of CC alignment	CCs (or other elements) aligned in a line are considered as characters.	This feature is generally considered in the postprocessing step to combine neighboring regions.	See Table 5.
Character elements	Representation of the target region by a set of typical character elements, or substrokes.	Machine learning methods are utilized to prepare the typical character elements.	Pan et al.(2008)[98] Coates et al.(2011)[99] Yi and Tian(2011)[100]
Visual saliency	A scene text region shuold have enough visual saliency to attract potential readers.	Since visual saliency is defined by combining low-level image features, such as color contrast and edge orientation, it can be considered as a hybrid feature.	Shahab et al. (2011)[101], Shahab et al.(2012)[102]
Context	Context, i.e., the area surrounding a target region, will be a strong prior. For example, a region surrounded by sky, it has a small probability to be a text region.	Pros : Good for reducing false detection. Cons: A further effort is necessary to understand the context.	Gandhi et al.(2000) [103], Park et al.(2008) [104], Kunishige et al.(2011) [105]
Language-dependent feature	Feature unique to the target language are utilized as detection clue.		Bhattacharya et al. (2009)[106] (Horizontal line of Indian script)
User interaction	User's action is utilized as a strong clue of indicating a text region.		Kim et al.(2009)[107] (Focus indicated by user is used as the region for first color clustering.)

5.4 Methods of text/non-text discrimination

Table 3 lists typical methods of text/non-text discrimination. If the discrimination result by the feature extracted at a specific region becomes "text", the region is considered to be a text region; that is, a text region is detected. The possible features have already overviewed in Table 2.

5.4.1 Discrimination by machine learning

Nowadays, the most popular discrimination strategy is to employ a machine learning method for training a text/non-text classifier. Multi-layer perceptron (MLP) and Support Vector Machine (SVM) are traditional choices. Several methods, called classifier ensemble methods, integrate results by multiple classifiers to have a final result. The classifier ensemble methods, such as AdaBoost and Random Forest, can provide a reliable result by their weighted-voting strategy. Note that AdaBoost and Random Forest have a promising function of feature selection. They automatically select discriminative features and thus can relieve the problem of "the curse of dimensionality".

Machine learning methods are useful because they do not need "manual optimizations" of classification rules; however, they need a sufficient number of training samples. In fact, the performance of the trained classifier heavily depends on the quantity and quality of training samples.

- For the quantity, we should not forget "the curse of dimensionality" and "overfitting"; if we use a higher dimensional feature and/or a flexible classifier with a larger number of parameters, we need to prepare (exponentially) more training samples. Although this problem is relieved by feature selection as noted above, feature selection also needs an enough number of training samples.
- For the quality of training samples, we should be careful of large variations of not only l character shapes but also non-character shapes. Especially, it is almost impossible to cover all non-character shapes by a limited number of training samples. One possible remedy on this problem is to use a 1-class classifier, where a classifier is trained by using only positive samples (i.e., samples from text regions).

5.4.2 Recognition-based localization

Recognition-based localization utilizes the discrimination power of OCR for text localization. This approach is similar to so-called recognition-by-segmentation (or recognition-based segmentation) for text recognition. The conventional recognition-by-segmentation techniques are applicable to the one-dimensional segmentation problem, where an input text image is partitioned into individual characters from left to right. In contrast, the text localization problem of scene images becomes a two-dimensional segmentation problem and thus the conventional techniques cannot applicable.

Two possible styles of recognition-based localization are as follows:

- OCR as a verifier: A pre-detection operation of a text region is performed and then the region is verified by OCR (ex. [116]). If OCR provides any valid character string, the region is verified as a text region.
- Exhaustive application of OCR: With a sliding window technique, OCR is performed at every window (ex. [117]).

Although recognition-based localization is promising, we have to recall that scene characters are often different from document characters and difficult to be recognized. Researches for recognizing various (often decorated) scene characters are now active, but their recognition accuracy is less than OCR for ordinary documents. Consequently, there is still an inevitable risk of rejecting text regions by false rejection by OCR.

Table 3 Methods of text/non-text discrimination.

Discrimination approches		Short description	Remarks	Example
Heuristics		Discrimination by a set of if-then rules, each of which is manually designed.	Pros : Intuitive Cons : Weak theoretical validity ;verification is often necessary	Messelodi and Modena (1999)[109], Chen et al. (2004) [25], Gatos et al. (2005)[110], Ezaki et al. (2004) [16]
Machine learning		Text/non-text discrimination using a classifer by some machine learning technique.	Pros : Automatic and robust. Cons : A sufficient number of training samples are necessary. For text/non-text discrimination, it is often difficult not only to balance the number of training samples, but also to cover all possible variations.	Chen et al. (2004) [25] (MLP+SVM for verification) Chen and Yuille (2004)[111] (AdaBoost) Hu and Chen (2005) [112] (SVM), Pan et al. (2008)[96] (AdaBoost), Xu et al. (2008)[114] (AdaBoost), Hanif and Prevost (2009)[115] (AdaBoost), Peng et al. (2011)[97] (SVM+CRFwGC), Kunishige et al. (2011) [105] (Random forest), Ma et al. (2012) [95] (Random Forest)
Recogniton -based localization	OCR as a verifier	If the target region is recognized as a chracter or a text, the region is determined as a text region.	Pros : Accurate discrimination by the power of OCR. Cons : Sensitive to the ability of OCR.	Ohya et al.(1994) [116]
	Exhaustive application of OCR	While sliding the location of a window, character	Pros :Resonable realizaton of localization-by-	Kusachi et al. (2004) [117]

		recognition is performed at individual locations. If the window can provide any recognition result, the window is determined as a character.	recogntion. Cons : Prohibitive computations ; a large number of training samples.	
	Aggregation of multiple OCR results	After generating multiple (and tentative) recognition results around the target region, their coherence is checked.	Like clasiffier ensemble, complementary effect is expected.	Li et al.(2001)[130], Rong et al.(2012) [131]
Muliframe analysis		A temporally stable region is deermined as a caption text.	For moving captions, text tracking [31][32] is often employed.	Bertini et al. (2001) [75](Temporal stability of Harris corner), Antani et al. (2000)[118] (Temporal stability of extracted CCs)

5.5 Region types for text localization

For extracting a feature for text/non-text discrimination, we have to define the region where the feature is extracted. This is not a trivial problem because the size of the target character is unknown in advance to text localization.

Table 4 lists possible region types. One reasonable choice to avoid the above "unknown size" problem is to use connected components after a binarization process. If a character is printed on a uniform background with a sufficient contrast value, the character will be extracted as a connected component by the binarization process. Color clustering and stroke width transform (SWT) [124] are also possible choices for extracting the appropriate connected component.

However, the extraction of the appropriate connected component is also difficult due to various factors. Figure 7 shows binarization results under different global threshold values. It is possible to observe that the shape of each connected component depends on the threshold. Especially, the shape of "M" changes drastically because a shadow is casted on it. Recently, MSER [125] has often been chosen as a more stable method for connected component analysis; however, even MSER may not be perfect on, for example, the case of this "M".

Extraction of multiple connected components under different parameters and their combinatorial use is a possible remedy. This will be followed by some recognition-based localization approach; the individual connected components are recognized and their recognition results are aggregated to have the final result.

Table 4 Region types for text localization.

Region type		Short description	Remarks	Example
Single pixel, keyponit		Feature extraction and discrimination at individual pixels.	Pros : Precise evaluation Cons : Large computation ; need of gap filling, less stability	Uchida et al. (2011)[83] , Shahab et al. (2012)[102]
Superpixel		A group of adjacent and coherent pixels [69].		Cho et al. (2011)[119]
Line		1-D gradient-based discrimination methods extract gradient features on a line.	Need of combination of results on adjacent scanlines for two-dimensional regions.	Wolf et al.(2002) [78], Wong and Chen (2003) [79], Kim and Kim (2009) [80], Phan et al. (2009) [81],
Block		Generally, a fix-sized square region.	Pros : Simple Cons : Rough. A single character may be devided into multiple blocks and thus postprocessing is often necessary like [84].	Hu and Chen (2005) [112], Hanif and Prevost (2009)[115], Mishra et al. (2012) [84], DCT-based methods
Connected component	Binalization	CC containing pixels with similar intensity values. For binarization methods, see Chapter 2.1.	Pros : Simple ; many binarization methods. Cons : Weak against non-uniform light and gradation	Gatos et al. (2005)[110] (Binarization result)
	Color clustering	CC containing pixels with similar colors..	Pros : Simple ; many clustering methods. Cons : Weak against multi-colored characters	Wang et al. (2001) [121], Ezaki et al. (2004) [16], Wang and Kangas (2005)[122], Mancas-Thillou and Gosselin (2007)[123]
	Stroke filter	For example, stroke width transform (SWT)[124] enhances the region surrounded by a pair of parallel edges	Pros : Accurate extraction of stroke-shaped area.	Kuwano et al.(2000) [72], Liu et al.(2006)[120], Epshtein et al. (2010) [124], Ma et al. (2012) [95](SWT)
	Maximally stable extremal region (MSER)	MSER[125] is a region with sufficient difference from its surroundings.	Pros : Accurate extraction of high-contrast area.	Donoser et al.(2007) [126], Neumann et al. (2010)[127], Merino-Gracia et al.

				(2011)[128], Nuumann and Matas (2012) [129]
	Multiple CCs	Multiple CCs are extracted by different criteria and therefore overlapped.	Pros : Robust to variations Cons : Aggregation step is necessary	Li et al.(2001)[130], Chen et al. (2004) [25], Rong et al. (2012) [131]

Figure 7 Binarization results under different (global) threshold values. It is possible to observe that the shape of connected components varies by the threshold. Green annotations are OCR results of connected components.

5.6 Method of revision by considering neighboring regions

Although the above three aspects (features, discrimination, and region type) can characterize a text localization method, that is, can determine localization results, a revision process is often employed as an optional process. Revision of the discrimination result is necessary because the text/non-text discrimination is done on individual regions independently and thus unstable. On the other hand, a scene image usually contains one or more words and sentences and each of them are linearly aligned characters. Hence, the neighboring regions are often mutually dependent and this dependence can be used for the revision. For example, if three neighboring regions are discriminated independently as text, non-text, and text, the middle region may be a text region.

Table 5 lists the revision methods by considering neighboring regions. The most typical revision method is the optimization-based revision, where a certain graphical model, such as MRF, is used. In MRF, each region is treated as a node and nodes of neighboring regions are connected by an edge. Each node has some cost, such as a discrimination cost, and each edge

also has some cost, such as coherence of the neighboring nodes. The latent variable at each node takes a text or non-text label. The value of the latent variable is optimized in various strategies, such as dynamic programming, belief propagation, and graph cut.

Table 5 Methods of revision by considering neighboring regions

Revision methods		Short description	Remarks	Example
Forced connection		Connect all neihboring text-candidate regions.	Need of further text/nontext-discsimination on the connected result.	Hanif and Prevost (2009)[115] (MLP was used for the final discrimination)
Heuristics		Connect neighboring candidate regions if they satisfy handmade rules.	Pros : Simple Cons :Manual parameter setting is necessary.	Goto and Aso (2001)[133], Wang and Kangas (2005)[122], Epshtein et al. (2010)[124]
Morphology		Neihboring text candidate regions are connected by a morphological operation.	Pros : Simple Cons : No function to exclude non-character regions.	Huang and Ma (2010)[75] (Filling space among corner points)
Optimization	Graph-based representation and optimization ; Markov random field (MRF), Conditional random field (CRF)	After creating a graph connecting text candidate regions, their final and optimal text/non-text classification is performed under a criterion.	Pros : Erroneous candidates can be excluded by considering neighbor regions. Cons : Evaluation functions (e.g., edge cost) shoud be designed on the graph. Note: Many evalutation functions are introduced in [95].	Zhang and Chang (2002) [113](Higher-order MRF), Kumar et al. (2007)[91] (MRF), Pan et al. (2008) [96] (MRF), Xu et al. (2008) [114](MRF), Wang and Belongie (2010) [82], Cho et al. (2011) [119] (CRF), Peng et al. (2011) [97] (CRF)
	Stochastic relaxation			Hase et al. (2001)[132]
Word lexicon		Neighboring		Mishra et al.

	regions are combined if they form a word.		(2012)[84]

6 Methods to solve problems in text recognition

6.1 Scene character recognition using standard OCR techniques

Since characters in images and video have wider variations in their appearance than those in scanned documents, their recognition module should be robust enough to deal with the variations. As listed in Table 6, there are two major approaches to make standard OCR techniques robust.

- The first approach is to use sufficient preprocessing methods for normalizing the target text region like a scanned document before applying the standard OCR. Any preprocessing method listed in Table 1 can be used for this purpose. For example, if we want to recognize low-resolution characters, we first apply a certain super-resolution technique to the input image and then the standard OCR.
- The second approach is to train the character recognizer by using a set of training samples covering the variations of scenery characters. For example, if we want to recognize low-resolution characters, we will train our recognizer with low-resolution training samples. Note that training samples can be generated (i.e., synthesized) by applying "degradation models" to an original training sample set. For example, we can generate low-resolution character images by applying a down-sampling operation to the original high-resolution character images. Similarly, we can generate character images with blur and perspective by convolution with a PSF and transformation with a projection matrix.

Both of the above approaches are related to text localization. In fact, as we discussed in Section 5 of this chapter, there are methods of "recognition-based localization", where the location with a high score by a recognizer is detected as a text region. Clearly, these methods provide the recognition result of a text when the text is localized.

6.2 Specific features for scene character recognition

Table 6 also lists more elaborated approaches with features which are not typical for ordinary character recognition. The "Bag-of-features" approach uses a histogram-based feature vector. This approach decomposes a target character into parts, or local features, and makes their histogram through a quantization process. Its important property is that it will not represent global structure of the target character. This property realizes a recognizer which is robust to global and severe deformations in scene characters. (This property has been utilized in general object recognition.)

Invariant features are also a promising approach. Since their feature value does not change under a family of deformations (such as affine deformation and perspective), they have been employed in the recognizers. Theoretically, if we use an invariant feature, say, a perspective invariant feature, we can recognize characters perfectly regardless of perspective distortions.

Furthermore, it is also beneficial that we need not to remove the perspective distortion by a preprocessing step. One point to be considered is that invariant features sometimes become less effective for character discrimination. For example, if a rotation invariant feature is used without any limitation, two different symbols "+" and "×" become identical.

6.3 Recognition of decorative characters

In scenery images, we find various decorative characters. Figure 8 shows a very limited number of "A"s. Some of them are rather popular fonts and the others are highly decorative fonts. It is rather easy to deal with the popular fonts because we can add them to the training samples. Highly decorative fonts are still difficult to be recognized because it is impossible to cover huge variations of all possible decorative fonts. In fact, highly decorative fonts are often unreadable even by human without context. Figure 8 (c) shows an extreme example of "A". It is difficult not only to recognize as "A" but also to detect as a character.

One possible strategy to deal with highly decorative fonts is to extract a topological structure, which is (nearly) invariant to decorations. Despite of the importance of this strategy, there have been only a few research trials, such as [144]. Feature representation by a set of character elements (introduced in Table 2), as well as bag-of-features approach, is also promising because the shape of character elements, i.e., local parts of a character is often preserved even under decorations. Similarly, deformable template (also called image warping, nonlinear registration, and elastic matching) [142] is a possible choice. Since it fits two images by nonlinear mapping, it can minimize the difference between a standard font image and a decorated font image. In any case, recognition of decorated characters is still an open problem and further researches are expected.

Table 6 Methods for recognizing texts in images and video.

Recognition approach		Short description	Remark	Example
Use of standard OCR techniques	Direct use of standard OCR after sufficient preprocessing steps	By normalizing the input image by a set of preprocessing steps, some standard OCR technique is applied.	In [24] [84], word lexicon is employed like a traditional analytical word recognizer. Similarly, other recognition methods are found as the "recognition-based localization" entries of Table 3 , as well as Table 5	Sato et al. (1998) [24], Okada et al. (1998) [11], Newman et al. (1999) [6], Chen et al. (2004)[15], Myers et al. (2004) [57], Mishra et al. (2012) [84], Neumann and Matas (2012) [127]
	Training OCR by scene character	Scene characters are directly used for training a		Sawaki et al. (2000) [134], Jacobs et al.

	samples	standard OCR technique.		(2005)[48], Saidane and Garcia (2007) [135], Weinman et al., (2009) [92], Netzer et al. (2011) [136]
Recognition by specific features or similarity evaluation	Bag-of-features	Usually, local features are extracted and then quantized to form a histogram.	Pros: Robust to global deformation since each character is decomposed into a set of parts (local features).	Campos et al. (2009) [137]
	Invariant features	Use of deformation (e.g., perspective) invariant feature.	In [140] [141], the matching process is accelerated by a hash technique.	Lu and Tan (2006) [138], Lin and Tan (2010) [139], Iwamura et al. (2010)[140], Pan et al. (2011) [141]
	Deformable template	Similarity evaluation after fitting a template nonlinearly to the input image.	Theoretically, it provides a deformation-invariant similarity. A review of deformable templates is found in [142].	Yokobayashi, Wakahara (2006) [143]
Recognition of special characters	Recognition of highly decorative characters		Increase of font templates is a naïve solution. Another idea is to understand the global character structure [144]. The above deformable template is also a possible choice.	Omachi et al. (2001)[144], Yokobayashi, Wakahara (2006) [143]
	Recognition of 3D characters	Recognition of characters with self-shadow.	In [145], engraved characters are recognized.	Mancas-Thillou and Mancas (2007) [145]
	Recognition of caption text	Since caption texts are degraded in various ways, special consideration is necessary.		See the papers cited in Section 2 of this chapter.
	Recognition of a text captured by a hand-held	Combination of video mosaicing and text recognition	A simultaneous optimization of video mosaicing and text recognition is found in	Uchida et al. (2008) [146]

	video camera		[146].	
	Recognition of license plate characters		In [4], there is a comprehensive list of trials to recognize license plate characters.	Anagnostopoulos et al. (2006) [4]
	Recognition of characters designed for camera-based OCR	Development of so-called “OCR fonts” for camera-based OCR.	The literature [147] tried to embed class information into the character shape by invariants.	Uchida et al. (2007)[147]
Post correction by a trained model		Post correction to various errors due to the problems particular to scene characters.	Typical misrecognitions (such as “o” to “n” due to the cut of lower part) are learned and modeled by a graphical model.	Beaufort and Mancas-Thillou (2007) [148]

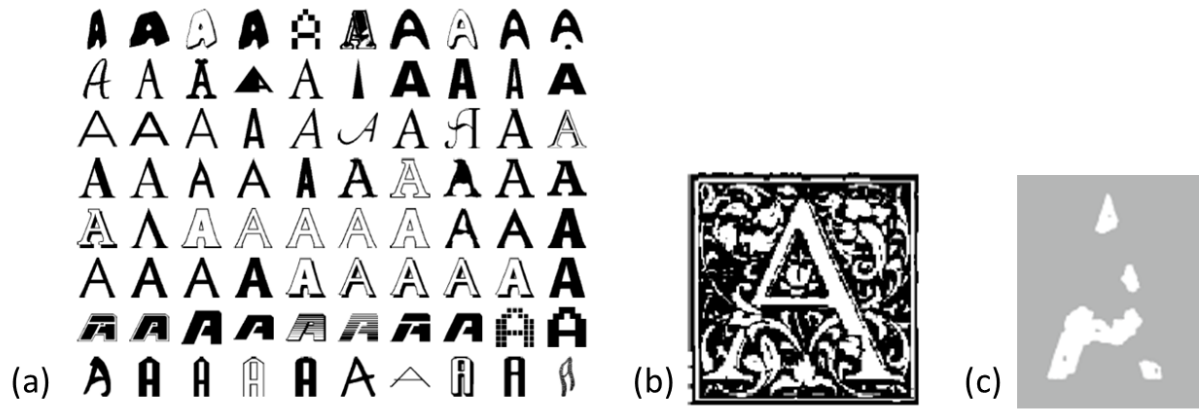

Figure 8 Various appearances of “A”. (a) Popular “A”s and decorative “A”s. (b) “A” as an initial capital. (c) A highly decorated “A” from the font set named “Asphalt road surface” (by Yokokaku, Japan).

7 Dataset for evaluation

As we saw in this chapter, camera-based OCR and video-text OCR are comprised of various modules, and therefore we need various datasets to evaluate those modules. The most widely used datasets are provided at ICDAR Robust Reading Competitions [66][67][68]. Their ICDAR2011 version contains 485 scene images where 1564 words exist. There are other datasets for camera-based OCR, that is, KAIST scene character dataset (3000 images), Google Street View dataset (249 images with 647 words and 3796 letters), The Street View House Numbers dataset (600,000 digit images), IUPR dataset of camera-captured document images (100 images of warped document pages), NEOCR dataset (659 images with 5238 text boxes), The Chars74K dataset (7705 scene character images together with handwritten and machine-printed characters). See also the list in Chapter 8.1.

Unfortunately, the sizes of the above datasets are not sufficient for covering huge variations in character shapes, appearances, lighting conditions, etc. In addition, if we want to train a classifier by any machine learning method, we need to prepare training samples as many as possible. Consequently, we need to increase its size, although it is highly costly to create a large dataset. A good news is that nowadays a huge number of scene images are freely downloadable from the Web (e.g., Flickr). Another good news is that it is possible to use a crowd-sourcing service (e.g., Amazon MechanicalTurk) for labeling images.

Conclusion

This chapter described the tasks and its solutions of text localization and recognition in camera-captured images and videos. In some cases, such as recognition of camera-captured document images, the recognition task can be reduced to the ordinary OCR task by applying sufficient preprocessing steps, such as binarization, deblurring, estimation and removal of perspective distortion. In more general case, such as recognition of the shop name on a signboard, the task is far more difficult than the ordinary OCR task because it is necessary to localize the target text region (often without any prior) and then recognize the characters. This text localization step is a difficult problem despite of a large number of trials, which were introduced in this chapter in detail. Furthermore, recognition of scene characters is not a simple task; those characters are often different from those of ordinary documents and sometimes highly decorative. To overcome those difficulties, we need to integrate not only traditional image processing and character recognition techniques but also recent machine learning and computer vision techniques.

Cross-references

Chapter 2.1 (Imaging Techniques in Document Analysis Processes) will provide several techniques for image binarization and image dewarping, which can be applicable to camera-captured document images. Chapter 8.1 (Datasets and Annotations for Document Analysis and Recognition) will provide information on datasets for camera-based OCR.

Further readings

Since scene text recognition and caption text recognition are one of the hottest topics for document analysis researchers, papers introducing a new technique are easily found in various document-related conferences, such as ICDAR (International conference on document analysis recognition) and DAS (International workshop on document analysis systems). CBDAR (International workshop on camera-based document analysis and recognition) is a rather new workshop specialized for this topic. Note that several competitions (called "Robust reading competition") on this topic have been conducted at ICDAR and provided very good reference information to understand the performance of the state-of-the-art techniques. An interesting thing is that this topic has attracted researchers with various interests; many papers, therefore, also can be found in more general conferences, such as CVPR (IEEE conference on computer vision and pattern recognition).

Journal papers on the topic are found in IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Pattern Recognition (PR), Pattern Recognition Letters (PRL), and International Journal of Document Analysis and Recognition (IJDAR), like the papers on other topics of document analysis and character recognition.

Since this is an emerging topic, there has been no complete book specialized for this topic (to my best knowledge). A couple of review papers have been published as introduced in this chapter. As noted above, for realizing practical scene text localization and recognition methods, we need to integrate various machine learning techniques and/or computer vision techniques. Consequently, introductory books on those techniques are useful for developing new techniques, as well as for understanding past trials.

References

[1] Sebastiano Impedovo, Raffaele Modugno, Anna Ferrante, Erasmo Stasolla, New Trends in Digital Scanning Processes, International Conference on Document Analysis and Recognition (ICDAR2009), pp.1071-1075, 2009.

[2] Keechul Jung, Kwang In Kim, Anil K. Jain, Text Information Extraction in Images and Video: A Survey, Pattern Recognition, 37, pp.977 – 997, 2004.

[3] Jian Liang, David Doermann, Huiping Li, Camera-Based Analysis of Text and Documents: A Survey, International Journal on Document Analysis and Recognition, 7(2-3), pp.84-104, 2005.

[4] C.N.E. Anagnostopoulos, I.E. Anagnostopoulos, V. Loumos, E. Kayafas, A license plate-recognition algorithm for intelligent transportation system applications, IEEE Transactions on Intelligent Transportation Systems 7 (3), pp. 377-391, 2006.

[5] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, Luc Vincent, Large-scale Privacy Protection in Google Street View, IEEE International Conference on Computer Vision, pp. 2373-2380, 2009.

[6] William Newman, Chris Dance, Alex Taylor, Stuart Taylor, Michael Taylor, Tony Aldhous, CamWorks: A Video-Based Tool for Efficient Capture from Paper Source Documents, International Conference on Multimedia Computing and Systems (ICMCS1999), pp.647–653, 1999.

[7] Stephen Pollard, Maurizio Pilu, Building Cameras for Capturing Documents, International Journal on Document Analysis and Recognition, 7(2-3), pp.123-137, 2005.

[8] Faisal Shafait, Michael Patrick Cutter, Joost van Beusekom, Syed Saqib Bukhari, Thomas M. Breuel, Decapod: A Flexible, Low Cost Digitization Solution for Small and Medium Archives, International Workshop on Camera-Based Document Analysis and Recognition (CBDAR2011), pp.41-46, 2011.

[9] Tomohiro Nakai, Koichi Kise, Masakazu Iwamura, Hashing with Local Combinations of Featur Points and Its Application to Camera-Based Document Image Retrieval - Retrieval in 0.14 Second from 10,000 Pages-, International Workshop on Camera-Based Document Analysis and Recognition (CBDAR2005), pp.87-94, 2005.

[10] Xu Liu, David Doermann, Mobile Retriever: Access to Digital Documents from Their Physical Source, International Journal on Document Analysis and Recognition, 11(1), pp.19-27, 2008.

[11] Yoshihiro Okada, Tetsuya Takeda, Yeun-Bae Kim, Yasuhiko Watanabe, Translation Camera, International Conference on Pattern Recognition (ICPR1998), pp. 613-617, 1998.

[12] Jiang Gao Jie Yang, An Adaptive Algorithm for Text Detection from Natural Scenes, IEEE Computer Society Conference on Vision and Pattern Recognition (CVPR2001), pp.2:84 -89, 2001

[13] Ismail Haritaoglu, Scene Text Extraction and Translation for Handheld Devices, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), pp.2:408-413, 2001.

[14] K. Yokomizo, K. Sono, Y. Watanabe, Y. Okada, Translation Camera on Mobile Phone, International Conference on Multimedia and Expo (ICME2003), pp. 177-180, 2003.

[15] Xilin Chen, Jie Yang, Jing Zhang, Alex Waibel, Automatic Detection and Recognition of Signs From Natural Scenes, IEEE Transactions on Image Processing, 13(1), pp.87-99, 2004.

[16] Nobuo Ezaki, Marius Bulacu, Lambert Schomaker, Text Detection from Natural Scene Images: Towards a System for Visually Impaired Persons, International Conference on Pattern Recognition (ICPR2004), pp.2:683-686, 2004.

[17] Chucai Yi, Yingli Tian, Assistive Text Reading from Complex Background for Blind Persons, International Workshop on Camera-Based Document Analysis and Recognition (CBDAR2011), pp.21-26, 2011.

[18] Yongmei Liu, Tsuyoshi Yamamura, Toshimitsu Tanaka, Noboru Ohnishi, Character-Based Mobile Robot Navigation, International Conference on Intelligent Robots and Systems (IROS1999), pp. 2:610-616, 1999.

[19] Markus Wienecke, Gernot A. Fink, Gerhard Sagerer, Toward Automatic Video-based Whiteboard Reading, International Journal on Document Analysis and Recognition, 7(2–3), pp.188–200, 2005.

[20] Mario E. Munich, Pietro Peronaa, Visual Input for Pen-Based Computers, International Conference on Pattern Recognition (ICPR1996), pp.33-37, 1996.

[21] Kazumasa Iwata, Koichi Kise, Masakazu Iwamura, Seiichi Uchida and Shinichiro Omachi, Tracking and Retrieval of Pen Tip Positions for an Intelligent Camera Pen, International Conference on Frontiers in Handwriting Recognition (ICFHR2010), pp.277-283, 2010.

[22] Qiong Liu and Chunyuan Liao, PaperUI, International Workshop on Camera-Based Document Analysis and Recognition (CBDAR2011), pp.3-11, 2011.

[23] Jing Zhang, Rangachar Kasturi, Extraction of Text Objects in Video Documents: Recent Progress, International Workshop on Document Analysis Systems (DAS2008), pp. 5 – 17, 2008.

[24] Toshio Sato, Takeo Kanade, Ellen K. Hughes, Michael A. Smith, Video OCR for Digital News Archives, IEEE International Workshop on Content-Based Access of Image and Video Database, pp. 52-60, 1998

[25] Datong Chen, Jean-Marc Odobez and Hervé Bourlard, Text detection and Recognition in Images and Video Frames, Pattern Recognition, 37(3), pp.595-608, 2004.
[26] Wonjun Kim and Changick Kim, A New Approach for Overlay Text Detection and Extraction From Complex Video Scene, IEEE Transactions on Image Processing, 18 (2), pp. 401-411, 2009.
[27] Palaiahnakote Shivakumara, Weihua Huang, Trung Quy Phan, Chew Lim Tan, Accurate Video Text Detection through Classification of Low and High Contrast Images, Pattern Recognition, 43(6), pp. 2165-2185, 2010.
[28] Alan F. Smeaton and Paul Over and Wessel Kraaij, Evaluation Campaigns and TRECVid, ACM International Workshop on Multimedia Information Retrieval (MIR2006), pp. 321—330, 2006.
[29] Dongqing Zhang, Shih-Fu Chang, Event detection in baseball video using superimposed caption recognition, ACM International Conference on Multimedia (MULTIMEDIA '02), pp.315-318, 2002.
[30] Marco Bertini, Alberto Del Bimbo, Walter Nunziati, Automatic Detection of Player's Identity in Soccer Videos Using Faces and Text Cues, ACM International Conference on Multimedia (MULTIMEDIA '06), pp. 663-666, 2006.
[31] Jinqiao Wang, Lingyu Duan, Zhenglong Li, Jing Liu, Hanqing Lu, Jesse S. Jin, Robust Method for TV Logo Tracking in Video Streams, IEEE International Conference on Multimedia and Expo (ICME2006), pp. 1041-1044, 2006.
[32] Nedret Özay, Bulent Sankur, Automatic TV Logo Detection and Classification in Broadcast Videos, European Signal Processing Conference (EUSIPCO2009), pp.839—843, 2009.
[33] Asif Shahab, Faisal Shafait, Andreas Dengel, Bayesian Approach to Photo Time-Stamp Recognition, International Conference on Document Analysis and Recognition (ICDAR2011), pp.1039-1043, 2011.
[34] Huiping Li and David Doermann, Text enhancement in Digital Video Using Multiple Frame Integration, ACM International Conference on Multimedia (Part 1) (MULTIMEDIA '99), pp.19-22, 1999.
[35] Huiping Li and David Doermann, Superresolution-Based Enhancement of Text in Digital Video, International Conference on Pattern Recognition (ICPR2000), pp.1:847-850, 2000.
[36] Celine Mancas-Thillou, Majid Mirmehdi, Super-Resolution Text using the Teager Filter, International Workshop on Camera-Based Document Analysis and Recognition (CBDAR2005), pp.10–16, 2005.
[37] David Capel, Andrew Zisserman, A Super-resolution enhancement of text image sequences, International Conference on Pattern Recognition (ICPR2000), pp.1 :600-605, 2000.
[38] M Irani, S Peleg, Improving Resolution by Image Registration, Graphical Models and Image Processing, 53(3) , pp.231-239, 1991.
[39] Katherine Donaldson, Gregory K. Myers, Bayesian super-resolution of text in video with a text-specific bimodal prior, International Journal on Document Analysis and Recognition, 7(2-3), pp.159–167, 2005.
[40] Jyotirmoy Banerjee, C.V. Jawahar, Super-resolution of Text Images Using Edge-Directed Tangent Field, International Workshop on Document Analysis Systems (DAS2008), pp.76-83, 2008.
[41] Battulga Bayarsaikhan, Younghee Kwon, Jin Hyung Kim, Anisotropic Total Variation Method for Text Image Super-Resolution, International Workshop on Document Analysis Systems (DAS2008), pp. 473-479, 2008.
[42] Anthony Zappalá, Andrew Gee, Michael Taylor, Document mosaicking, Image and Vision Computing, 17, pp.589–595, 1999.

[43] T. Sato, S. Ikeda, M. Kanbara, A. Iketani, N. Nakajima, N. Yokoya, K. Yamada, High-resolution video mosaicing for documents and photos by estimating camera motion, Proc. SPIE Electronic Imaging, 5299, 2004.
[44] Simon Baker, Takeo Kanade, Limits on Super-Resolution and How to Break Them, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(9), pp.1167-1183, 2002.
[45] Gerald Dalley, Bill Freeman, Joe Marks, Single-frame text super-resolution: a Bayesian approach, International Conference on Image Processing (ICIP2004), pp.5:3295-3298, 2004.
[46] William T. Freeman, Thouis R. Jones, Egon C. Pasztor, Example-Based Super-Resolution, IEEE Computer Graphics and Applications, 22(2), pp.56-65, 2002.
[47] Jangkyun Park, Younghee Kwon, Jin Hyung Kim, An Example-based Prior Model for Text Image Super-resolution, International Conference on Document Analysis and Recognition (ICDAR2005), pp.374-378, 2005.
[48] Charles Jacobs, Patrice Y. Simard, Paul Viola, James Rinker, Text Recognition of Low-resolution Document Images, International Conference on Document Analysis and Recognition (ICDAR2005), pp.2 :695 – 699, 2005.
[49] M. J. Taylor, C. R. Dance, Enhancement of document images from cameras, Proceedings of SPIE, 3305, pp. 230-241 1998.
[50] Yibin Tian, Wei Ming, Adaptive Deblurring for Camera-Based Document Image Processing, Lecture Notes in Computer Science, 5876, pp.767-777, 2009.
[51] T. Kanungo and Q. Zheng, Estimating degradation model parameters using neighborhood pattern distributions: an optimization approach, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(4), pp.520 - 524, 2004.
[52] Xiaogang Chen, Xiangjian He, Jie Yang, Qiang Wu, An effective document image deblurring algorithm, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2011), pp. 369-376, 2011.
[53] Xing Yu Qi, Li Zhang, Chew Lim Tan, Motion Deblurring for Optical Character Recognition, International Conference on Document Analysis and Recognition (ICDAR2005), pp. 389-393, 2005.
[54] P. Clark, M. Mirmehdi, Location and recovery of text on oriented surfaces, *SPIE Conference on Document Recognition and Retrieval VII*, 3967, pp.267-277, 2000.
[55] P. Clark, M. Mirmehdi, Estimating the Orientation and Recovery of Text Planes in a Single Image, British Machine Vision Conference (BMVC2001), pp.421-430, 2001.
[56] Christopher R. Dance, Perspective Estimation for Document Images, SPIE Document Recognition IX, pp.20-25, 2002.
[57] Gregory K. Myers, Robert C. Bolles, Quang-Tuan Luong, James A. Herson, Hrishikesh B. Aradhye, Rectification and recognition of text in 3-D scenes, International Journal on Document Analysis and Recognition, 7(2-3), pp.147–158, 2004.
[58] Takuma Yamaguchi, Minoru Maruyama, Hidetoshi Miyao and Yasuaki Nakano, Digit recognition in a natural scene with skew and slant normalization, International Journal on Document Analysis and Recognition, 7(2-3), pp. 168-177, 2004.
[59] Shijian Lu, Ben M. Chen, C.C. Ko, Perspective rectification of document images using fuzzy set and morphological operations, Image and Vision Computing, 23(5), pp.541-553, 2005.
[60] Jian Liang, Daniel DeMenthon, David Doermann, Geometric Rectification of Camera-Captured Document Images, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(4), pp.591-605, 2008.

[61] Xu-Cheng Yin Jun Sun Satoshi Naoi, Katsuhito Fujimoto, Hiroaki Takebe, Yusaku Fujii, Koji Kurokawa, A Multi-Stage Strategy to Perspective Rectification for Mobile Phone Camera-Based Document Images, International Conference on Document Analysis and Recognition (ICDAR2007), pp. 574-578, 2007.
[62] Seiichi Uchida, Megumi Sakai, Masakazu Iwamura, Shinichiro Omachi, Koichi Kise, Skew Estimation by Instances, International Workshop on Document Analysis Systems (DAS2008), pp.201-208, 2008.
[63] Soma Shiraishi, Yaokai Feng, Seiichi Uchida, A Part-Based Skew Estimation Method, International Workshop on Document Analysis Systems (DAS2012), pp.185-189, 2012
[64] Shijian Lu, Chew Lim Tan, Automatic Detection of Document Script and Orientation, International Conference on Document Analysis and Recognition (ICDAR2007), pp.237 – 241, 2007.
[65] Hyung Il Koo, Jinho Kim, Nam Ik Cho, Composition of a Dewarped and Enhanced Document Image from Two View Images, IEEE Transactions on Image Processing, 18(7) , pp.1551-1562, 2009.
[66] S. M. Lucas et al., ICDAR 2003 Robust Reading Competitions: Entries, Results and Future Directions, International Journal on Document Analysis and Recognition, 7, pp. 105-122, 2005.
[67] S. M. Lucas, ICDAR 2005 Text Locating Competition Results, International Conference on Document Analysis and Recognition (ICDAR2005), pp. 80-84, 2005.
[68] Asif Shahab, Faisal Shafait, Andreas Dengel, ICDAR 2011 Robust Reading Competition Challenge 2: Reading Text in Scene Images, International Conference on Document Analysis and Recognition (ICDAR2011), pp.1491-1496, 2011.
[69] X. Ren, J. Malik, Learning a Classification Model for Segmentation, International Conference on Computer Vision (ICCV2003), 1, pp.10-17, 2003.
[70] Palaiahnakote Shivakumara, Weihua Huang, Trung Quy Phan, Chew Lim Tan, Accurate video text detection through classification of low and high contrast images, Pattern Recognition, 43(6), pp.2165-2185, 2010.
[71] Nobuo Ezaki, Kimiyasu Kiyota, Bui Truong Minh, Marius Bulacu, Lambert Schomaker, Improved Text-Detection Methods for a Camera-based Text Reading System for Blind Persons, International Conference on Document Analysis and Recognition (ICDAR2005), pp. 257 – 261, 2005.
[72] H. Kuwano, Y. Taniguchi, H. Arai, M. Mori, S. Kurakake, H. Kojima, Telop-on-demand: video structuring and retrieval based on text recognition, IEEE International Conference on Multimedia and Expo (ICME 2000), pp.2: 759 – 762, 2000.
[73] Bong-Kee Sin, Seon-Kyu Kim, Beom-Joon Cho, Locating Characters in Scene Images Using Frequency Features, International Conference on Pattern Recognition (ICPR2002), 3, 2002.
[74] Xiaoqing Liu, Jagath Samarabandu, Multiscale Edge-Based Text Extraction from Complex Images IEEE International Conference on Multimedia and Expo (ICME2006), pp.1721-1724, 2006.
[75] M. Bertini, C. Colombo, A. Del Bimbo, Automatic Caption Localization in Videos Using Salient Points, IEEE International Conference on Multimedia and Expo (ICME'01), pp.69-72, 2001.
[76] Xiaodong Huang, Huadong Ma, Automatic Detection and Localization of Natural Scene Text in Video, International Conference on Pattern Recognition (ICPR2010), pp.3216 – 3219, 2010.
[77] Xu Zhao, Kai-Hsiang Lin, Yun Fu, Yuxiao Hu, Yuncai Liu, Thomas S. Huang, Text From Corners: A Novel Approach to Detect Text and Caption in Videos, IEEE Transactions on Image Processing, 20(3), pp.790 – 799, 2011.

[78] Christian Wolf, Jean-Michel Jolion, Francoise Chassaing, Text Localization, Enhancement and Binarization in Multimedia Documents, International Conference on Pattern Recognition (ICPR2002), pp.2:1037 – 1040, 2002.
[79] Edward K. Wong, Minya Chen, A new robust algorithm for video text extraction, Pattern Recognition, 36(6), pp.1397-1406, 2003.
[80] Wonjun Kim, Changick Kim, A new approach for overlay text detection and extraction from complex video scene, IEEE Transactions on Image Processing, 18(2) , pp.401-411, 2009.
[81] Trung Quy Phan, Palaiahnakote Shivakumara, and Chew Lim Tan, A Laplacian Method for Video Text Detection, International Conference on Document Analysis and Recognition (ICDAR 2009), pp.66-70, 2009.
[82] Kai Wang, Serge Belongie, Word spotting in the wild, European conference on Computer vision (ECCV2010), pp.591-604, 2010.
[83] Seiichi Uchida, Yuki Shigeyoshi, Yasuhiro Kunishige, Yaokai Feng, A Keypoint-Based Approach Toward Scenery Character Detection, International Conference on Document Analysis and Recognition (ICDAR 2011), pp.819-823, 2011.
[84] Anand Mishra, Karteek Alahari, C. V. Jawahar , Top-down and bottom-up cues for scene text recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR2012), pp.2687-2694, 2012.
[85] Marios Anthimopoulos, Basilis Gatos, Ioannis Pratikakis, A two-stage scheme for text detection in video images. Image and Vision Computing 28(9), pp.1413-1426, 2010.
[86] Navin Chaddha, Rosen Sharma, Avneesh Agrawal, Anoop Gupta, Text Segmentation in Mixed-Mode Images, Asilomar Conference on Signals, Systems and Computers, pp.2:1356-1361, 1994.
[87] Yu Zhong, Hongjiang Zhang, Anil K. Jain, Automatic Caption Localization in Compressed Video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(4), pp.385-392, 2000.
[88] Hideaki Goto, Redefining the DCT-based feature for scene text detection Analysis and comparison of spatial frequency-based features, International Journal on Document Analysis and Recognition, 11(1), pp.1-8, 2008.
[89] Anil K. Jain and Sushil Bhattacharjee, Text segmentation using Gabor filters for automatic document processing, Machine Vision and Applications, 5(3), pp. 169-184, 1992.
[90] Tomoyuki Saoi, Hideaki Goto, Hiroaki Kobayashi, Text Detection in Color Scene Images based on Unsupervised Clustering of Multi-channel Wavelet Features, International Conference on Document Analysis and Recognition (ICDAR 2005), pp.690-694. 2005.
[91] Sunil Kumar, Rajat Gupta, Nitin Khanna, Santanu Chaudhury, Shiv Dutt Joshi, Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model, IEEE Transactions on Image Processing, 16(8) , pp.2117-2128, 2007.
[92] Jerod J. Weinman, Erik Learned-Miller, Allen R. Hanson, Scene Text Recognition Using Similarity and a Lexicon with Sparse Belief Propagation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(10) , pp.1733-1746, 2009.
[93] K. C. Kim, H. R. Byun, Y. J. Song, Y. W. Choi, S. Y. Chi, K. K. Kim, Y. K. Chung, Scene Text Extraction in Natural Scene Images using Hierarchical Feature Combining and Verification. International Conference on Pattern Recognition (ICPR2004) , pp.2:679-682, 2004.
[94] Xilin Chen, Jie Yang, Jing Zhang, Alex Waibel, Automatic Detection and Recognition of Signs From Natural Scenes, IEEE Transactions on Image Processing, 13(1), pp.87 – 99, 2004.

[95] Yi Ma, Wenyu Liu, Xiang Bai, Cong Yao, Zhuowen Tu, Detecting texts of arbitrary orientations in natural images, IEEE Conference on Computer Vision and Pattern Recognition (CVPR2012), pp. 1083-1090, 2012.
[96] Yi-Feng Pan, Xinwen Hou, Cheng-Lin Liu , A Robust System to Detect and Localize Texts in Natural Scene Images, International Workshop on Document Analysis Systems (DAS2008), pp. 35-42, 2008.
[97] Xujun Peng, Huaigu Cao, Rohit Prasad, Premkumar Natarajan, Text Extraction from Video Using Conditional Random Fields, International Conference on Document Analysis and Recognition (ICDAR2011), pp.1029-1033, 2011.
[98] Wumo Pan, T. D. Bui, C. Y. Suen, Text detection from scene images using sparse representation, International Conference on Pattern Recognition (ICPR2008), pp. 1-5, 2008.
[99] Adam Coates, Blake Carpenter, Carl Case, Sanjeev Satheesh, Bipin Suresh, Tao Wang, David J. Wu, Andrew Y. Ng, Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning, International Conference on Document Analysis and Recognition (ICDAR2011), pp.440-445, 2011.
[100] Chucai Yi, Yingli Tian, Text Detection in Natural Scene Images by Stroke Gabor Words, International Conference on Document Analysis and Recognition (ICDAR2011), pp.177-181,2011.
[101] Asif Shahab, Faisal Shafait, Andreas Dengel, Bayesian Approach to Photo Time-Stamp Recognition, International Conference on Document Analysis and Recognition (ICDAR2011), pp.1039-1043, 2011.
[102] Asif Shahab, Faisal Shafait, Andreas Dengel, Seiichi Uchida, How Salient is Scene Text, International Workshop on Document Analysis Systems (DAS2012), pp.317-321, 2012.
[103] Tarak Gandhi, Rangachar Kasturi, Sameer Antani, Application of Planar Motion Segmentation for Scene Text Extraction, International Conference on Pattern Recognition, 1, pp. 1445-1449, 2000.
[104] Seung-Bo Park, Kyung-Jin Oh, Heung-Nam Kim, Geun-Sik Jo, Automatic Subtitles Localization through Speaker Identification in Multimedia System, IEEE International Workshop on Semantic Computing and Applications (IWSCA2008), pp.166-172, 2008.
[105] Yasuhiro Kunishige, Yaokai Feng and Seiichi Uchida, Scenery Character Detection with Environmental Context, International Conference on Document Analysis and Recognition (ICDAR2011), 2011.
[106] Ujjwal Bhattacharya, Swapan Kumar Parui, Srikanta Mondal, Devanagari and Bangla Text Extraction from Natural Scene Images, International Conference on Document Analysis and Recognition (ICDAR2009), pp.171-175, 2009
[107] Egyul Kim, Seong-Hun Lee, and Jin-Hyung Kim, Scene Text Extraction Using Focus of Mobile Camera, International Conference on Document Analysis and Recognition (ICDAR 2009), pp.166-170, 2009.
[108] C. Rother, V. Kolmogorov, A. Blake, GrabCut: Interactive foreground extraction using iterated graph cuts, Proc. SIGGRAPH, pp.309–314, 2004.
[109] S. Messelodi, C.M. Modena, Automatic identification and skew estimation of text lines in real scene images, Pattern Recognition, 32, pp.791-810, 1999.
[110] B. Gatos, I. Pratikakis, S.J. Perantonis, Text detection in indoor/outdoor scene images, International Workshop on Camera-based Document Analysis and Recognition (CBDAR2005), pp. 127-132, 2005.

[111] X. Chen, A.L.Yuille, Detecting and reading text in natural scenes, IEEE Conference on Computer Vision and Pattern Recognition (CVPR2004), pp.2:366-373, 2004.
[112] Shiyan Hu, Minya Chen, Adaptive Frechet Kernel Based Support Vector Machine for Text Detection, International Conference on Acoustics, Speech, and Signal Processing (ICASSP2005), pp.5: 365-368, 2005.
[113] Dong-Qing Zhang, Shih-Fu Chang, Learning to Detect Scene Text Using a Higher-Order MRF with Belief Propagation, Conference on Computer Vision and Pattern Recognition Workshop (CVPRW2004), pp.101-108, 2000.
[114] Lianli Xu, Hiroto Nagayoshi, Hiroshi Sako, Kanji Character Detection from Complex Real Scene Images based on Character Properties, International Workshop on Document Analysis Systems (DAS2008), pp. 278-285, 2008.
[115] Shehzad Muhammad Hanif, Lionel Prevost, Text Detection and Localization in Complex Scene Images using Constrained AdaBoost Algorithm, International Conference on Document Analysis and Recognition (ICDAR2009), ,pp1 -5, 2009.
[116] J. Ohya, A. Shio, S. Akamatsu, Recognizing Characters in Scene Images, IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(2), pp. 214-220, 1994.
[117] Yoshinori Kusachi, Akira Suzuki, Naoki Ito, Kenichi Arakawa, Kanji Recognition in Scene Images without Detection of Text Fields - Robust Against Variation of Viewpoint, Contrast, and Background Texture -, International Conference on Pattern Recognition (ICPR2004) , pp.1:457-460, 2004.
[118] S. Antani, D. Crandall, R. Kasturi, Robust Extraction of Text in Video, International Conference on Pattern Recognition (ICPR2000), pp.3:831–834, 2000.
[119] Min Su Cho, Jae-Hyun Seok, Seonghun Lee, Jin-Hyung Kim, Scene Text Extraction by Superpixel CRFs Combining Multiple Character Features, International Conference on Document Analysis and Recognition (ICDAR2011), pp.1034-1038, 2011.
[120] Qifeng Liu, Cheolkon Jung, Youngsu Moon, Text segmentation based on stroke filter. Annual ACM international conference on Multimedia (MULTIMEDIA 2006), pp.129-132, 2006.
[121] Xuewen Wang, Xiaoqing Ding, Changsong Liu, Character Extraction and Recognition in Natural Scene Images, International Conference on Document Analysis and Recognition (ICDAR2001), pp. 1084-1088, 2001.
[122] Kongqiao Wang, Jari A. Kangas, Character location in scene images from digital camera, Pattern Recognition, 36(10), pp.2287-2299, 2003.
[123] Céline Mancas-Thillou, Bernard Gosselin, Color text extraction with selective metric-based clustering, Computer Vision and Image Understanding, 107(1-2), pp.97-107, 2007.
[124] Boris Epshtein, Eyal Ofek, Yonatan Wexler, Detecting Text in Natural Scenes with Stroke Width Transform, IEEE Conference on Computer Vision and Pattern Recognition (CVPR2010), pp. 2963-2970, 2010.
[125] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust Wide-Baseline Stereo from Maximally Stable Extremal Regions, Image and Vision Computing, 22, pp.761-767, 2004.
[126] Michael Donoser, Clemens Arth, Horst Bischof, Detecting, tracking and recognizing license plates, Asian Conference on Computer Vision (ACCV2007), pp. II:447-456, 2007.
[127] Lukas Neumann, Jiri Matas, A method for text localization and recognition in real-world images, Asian Conference on Computer Vision (ACCV2010), pp.III:770-783, 2010.

[128] Carlos Merino-Gracia, Karel Lenc, Majid Mirmehdi , A Head-mounted Device for Recognizing Text in Natural Scenes, International Workshop on Camera-Based Document Analysis and Recognition (CBDAR2011), pp.27-32, 2011.
[129] Lukas Neumann, Jiri Matas, Real-time scene text localization and recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR2012), pp. 3538-3545, 2012.
[130] Chuang Li, Xiaoqing Ding, Youshou Wu, Automatic Text Location in Natural Scene Images, International Conference on Document Analysis and Recognition (ICDAR 2001), pp.1069-1073, 2001.
[131] Rong Huang, Shinpei Oba, Shivakumara Palaiahnakote, Seiichi Uchida, Scene Character Detection and Recognition Based on Multiple Hypotheses Framework, International Conference on Pattern Recognition (ICPR2012), pp.--, 2012.
[132] Hiroyuki Hase, Toshiyuki Shinokawa, Masaaki Yoneda, Ching Y. Suen, Character string extraction from color documents, Pattern Recognition, 34 (7), pp.1349-1365, 2001
[133] Hideaki Goto, Hirotomo Aso, Character pattern extraction from documents with complex backgrounds, International Journal on Document Analysis and Recognition, 4(4), pp.258-268, 2001.
[134] M.Sawaki, H.Murase, N.Hagita, Automatic acquisition of context-based image templates for degraded character recognition in scene images, International Conference on Pattern Recognition (ICPR2000), pp.4:15-18, 2000
[135] Zohra Saidane, Christophe Garcia, Automatic Scene Text Recognition using a Convolutional Neural Network, International Workshop on Camera-Based Document Analysis and Recognition (CBDAR 2007), pp.100–106, 2007.
[136] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, Andrew Y. Ng, Reading Digits in Natural Images with Unsupervised Feature Learning, NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.
[137] Teofilo E. de Camposm, Bodla Rakesh Babu, Manik Varma, Character Recognition in Natural Images, International Conference on Computer Vision Theory and Applications (VISAPP2009), pp. 273-280, 2009.
[138] Shijian Lu, Chew Lim Tan, Camera Text Recognition based on Perspective Invariants, International Conference on Pattern Recognition (ICPR 2006), pp.2:1042-1045, 2006.
[139] Linlin Li, Chew Lim Tan, Recognizing Planar Symbols with Severe Perspective Deformation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(4) , pp.755-762, 2010.
[140] Masakazu Iwamura, Tomohiko Tsuji, and Koichi Kise, Memory-based recognition of camera-captured characters, International Workshop on Document Analysis Systems (DAS2010), pp. 89-96, 2010.
[141] Pan Pan, Yuanping Zhu, Jun Sun, Satoshi Naoi, Recognizing Characters with Severe Perspective Distortion Using Hash Tables and Perspective Invariants, International Conference on Document Analysis and Recognition (ICDAR 2011), pp.548-552, 2011.
[142] Seiichi Uchida and Hiroaki Sakoe, A survey of elastic matching techniques for handwritten character recognition, IEICE Transactions on Information & Systems, E88-D(8), pp.1781-1790, 2005.
[143] Minoru Yokobayashi, Toru Wakahara, Binarization and Recognition of Degraded Characters Using a Maximum Separability Axis in Color Space and GAT Correlation, International Conference on Pattern Recognition (ICPR 2006), pp.2:885-888, 2006.

[144] Shin'ichiro Omachi, Masaki Inoue, Hirotomo Aso, Structure Extraction from Decorated Characters Using Multiscale Images, IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3), pp. 315-322, 2001.

[145] Celine Mancas-Thillou, Matei Mancas, Comparison between Pen-scanner and Digital Camera Acquisition for Engraved Character Recognition, International Workshop on Camera-based Document Analysis and Recognition (CBDAR 2007), pp.130-137, 2007.

[146] Seiichi Uchida, Hiromitsu Miyazaki, Hiroaki Sakoe, Mosaicing-by-recognition for video-based text recognition, Pattern Recognition, 41(4), pp.1230-1240, 2008.

[147] Seiichi Uchida, Megumi Sakai, Masakazu Iwamura, Shinichiro Omachi, Koichi Kise, Extraction of Embedded Class Information from Universal Character Pattern, International Conference on Document Analysis and Recognition (ICDAR 2007), pp.1:437-441, 2007.

[148] Richard Beaufort, Celine Mancas-Thillou, A Weighted Finite-State Framework for Correcting Errors in Natural Scene OCR, International Conference on Document Analysis and Recognition, (ICDAR 2007), pp.2:889-893, 2007.