Chromosome Region Recognition Based on Local Band Patterns

Toru Abe Chieko Hamada Tetsuo Kinoshita

***Abstract*— To make the visual examination of a chromosome image for various chromosome abnormalities, individual chromosome regions have to be extracted from the image and classified into the distinct chromosome types in advance. To improve the accuracy and flexibility in this process, a subregion (local band pattern) based method has been proposed for recognizing individual chromosome regions in the image. This method regards each chromosome region as a series of subregions, and iterates a search for subregions in the image consecutively. As a result, chromosome region classification is performed simultaneously with its extraction for each subregion. Since the dimensions and intensities of chromosome regions vary with every image, effective subregion searches require templates whose dimensions and intensities correspond with those of chromosome regions in the subject image. In this paper, to develop an efficient subregion search, we present a method for adjusting the dimensions of templates to those of chromosome regions in the subject image and adapting the intensities in the subject image to those of the templates.**

I. Introduction

The examination of chromosome images for various chromosome abnormalities plays an important role in many clinical practices, including treatment and prevention of genetic disorders, radiation dosimetry, toxicology, etc [1]. To make the visual examination of a chromosome image, individual chromosome regions have to be extracted from the image and classified into the distinct chromosome types in advance.

To improve the accuracy and flexibility in this process, we have proposed a subregion (local band pattern) based method for recognizing individual chromosome regions in an image. This method regards each chromosome region as a series of subregions, and iterates a search for subregions in the image consecutively. As a result, chromosome region classification is performed simultaneously with its extraction for each subregion. Since the dimensions and intensities of chromosome regions vary with every image, to achieve effective subregion searches, the dimensions and intensities of templates for subregion searches are required to correspond with those of chromosome regions in the subject image.

In this paper, to develop an efficient chromosome subregion search, we present a method for adjusting the dimensions (widths and lengths) of templates to those of chromosome regions in the subject image and adapting the intensities in the subject image to those of the templates. Furthermore, to show the effectiveness of the proposed method, we also present the results of subregion search experiments on chromosome images.

T. Abe and T. Kinoshita are with Cyberscience Center, Tohoku University, Sendai, 980-8577 Japan.

C. Hamada is with Graduate School of Information Sciences, Tohoku University, Sendai, 980-8577 Japan.

II. Extraction and Classification of Chromosome Regions

This section explains the general procedures for examining chromosome images and the difficulties in those procedures.

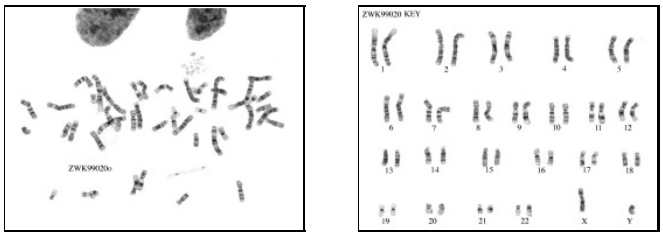
A. Examining Chromosome Images

Every cell nucleus in a normal human being contains 46 chromosomes consisting of 44 autosomes and two sex chromosomes. The autosomes are composed of 22 homologous pairs of chromosomes, and by convention, numbered from 1 to 22. The sex chromosomes are referred to as X and Y. A normal human female has two X chromosomes, while a normal human male has an X and a Y chromosome. Each chromosome has a narrow part, which is called a centromere, and it divides the entire region into two parts. The shorter part is called a short arm and the longer part is called a long arm. With proper staining methods, such as Giemsa staining (G-staining) method, a characteristic series of light and dark bands appears along the longitudinal axis of a chromosome (Fig. 1 (a)). The band appearance on a chromosome is called a band pattern, and it is unique to each type of chromosome.

Usually the examination of a chromosome image requires the following procedures [2]:

1) Staining a set of chromosomes and capturing its image.
2) Extracting individual chromosome regions from the image.
3) Classifying the chromosome regions into the 24 types (1, 2, ..., 22, X, and Y).
4) Inspecting the region appearances for chromosome abnormalities.

To make the visual examination of a chromosome image, individual chromosome regions are extracted from the image, and the extracted regions are classified into the 24 distinct chromosome types (Fig. 1 (b)). The dimensions of a chromosome change with the stage in a cell division, and the intensities of it change with staining conditions, therefore the dimensions and intensities of a chromosome region vary with every image. Meanwhile, the relative length,



(a) (b)

Fig. 1. (a) chromosome image, (b) classification result [3].

the relative centromere position, and the band pattern of each chromosome type vary little with every image. For this reason, the latter features are used for the classification [4].

According to the classification result, abnormalities of number, where there are one or more entire chromosomes additional to or missing from the normal complement, can be detected. From the region appearances (the band pattern on each chromosome region), abnormalities of structure, where part of the bands are lost (deletion), repeated (duplication), or shifted (translocation), can be examined visually.

B. Difficulties in Extracting and Classifying Chromosome Regions

The existing methods perform chromosome region extractions apart from chromosome region classifications, and their classification procedures suppose that individual chromosome regions are extracted accurately from a subject image beforehand [5, 6]. However, chromosome regions in the image frequently touch or overlap each other, and have some parts difficult to distinguish them from the background. Consequently, the accurate extraction of individual chromosome regions from the image is not an easy procedure.

Although extracted regions can be classified into several chromosome groups according to the relative lengths and the relative centromere positions of them, to discriminate between all 24 chromosome types, the use of band patterns is required in the classification. The classification methods using band patterns are generally categorized into two approaches: one is a global approach, and the other is a local approach [1, 2, 6]. In the global approach, the band pattern on an entire region (the longitudinal profile of intensity in an extracted region) is determined, and a chromosome type is assigned to the region by comparing its band pattern with reference band patterns [6, 7]. Therefore, when aberrant bands appear partly on a region because of various reasons (region extraction failure, region overlap, chromosome abnormalities, etc.), it is difficult to assign a chromosome type correctly. In the local approach, local features such as particular bands are determined in a region, and they are used for the classification. This approach can partially reduce the aberrant band influence on the classification accuracy [2, 5, 8]. However, it is reported that the local approaches are inferior to the global approaches in the classification accuracy [6]. The conceivable reasons for that are as follows:

- It is difficult to determine local features accurately in a chromosome region.
- Compared to the global approaches, the local approaches use fewer features for the classification.

III. Chromosome Region Recognition Based on Local Band Patterns

To overcome the problems in the existing methods, a subregion based method has been proposed for recognizing individual chromosome regions. This method regards each chromosome region as a series of subregions, and iterates a search for subregions in the subject image consecutively.

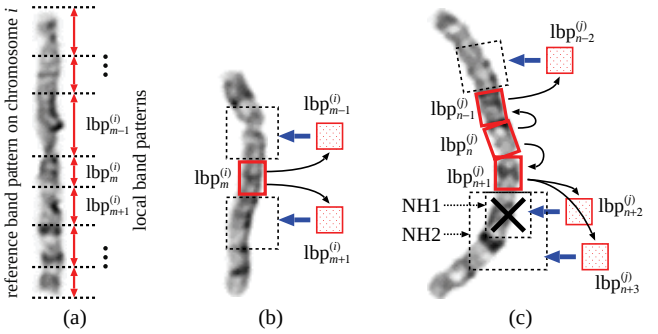


Fig. 2. (a) local band patterns, (b) extraction and classification with local band patterns, (c) control of template and search area.

In this method, a reference band pattern is prepared for every type of chromosome. Each reference band pattern for an entire chromosome region is divided into several parts, and they are used as the templates for extracting and classifying the chromosome region. In the following, the divided parts are referred to as local band patters, and as shown in Fig. 2 (a), the m th local band pattern on the chromosome type i is denoted by $\mathrm{lbp}^{(i)}_{m}$.

Firstly, the subject image is searched for the subregion corresponding to a local band pattern (template). As shown in Fig. 2 (b), if a subregion corresponding to $\mathrm{lbp}^{(i)}_{m}$ is detected, secondly, the neighborhood of the detected subregion is searched for the next subregion corresponding to the adjacent $\mathrm{lbp}^{(i)}_{m-1}$ (or $\mathrm{lbp}^{(i)}_{m+1}$). By iterating the search for subregion consecutively, with the first detected subregion as the starting point, one subregion after another is detected, and the entire region of a chromosome is determined in the image.

As shown in Fig. 2 (c), when a subregion corresponding to $\mathrm{lbp}^{(j)}_{n+1}$ is detected and the adjacent $\mathrm{lbp}^{(j)}_{n+2}$ cannot be found in the neighborhood NH1, it is surmised that the aberrations (chromosome region overlaps, chromosome abnormalities, etc.) occur in NH1. To deal with this difficulty and complete the search for the entire chromosome region, if the adjacent local band pattern cannot be found in the neighborhood, the template and search area are changed. For example, in Fig. 2 (c), the template is changed from $\mathrm{lbp}^{(j)}_{n+2}$ to $\mathrm{lbp}^{(j)}_{n+3}$, and the search area is extended from NH1 to NH2.

By taking these approaches, the following advantages are expected in this method:

- As the consecutive searches for subregions, simultaneously with the extraction, the classification is performed on part of a chromosome region, and the results of preceding searches are utilized for the following searches.
- By controlling the template and search area, the consecutive searches integrate features in the subregions while reducing the aberrant band influence.

IV. Adjusting Template Dimensions and Adapting Subject Image Intensities

Assume that a subregion search is made by scanning a subject image with a template and seeking in the image for subregions where the mean-squared-error (MSE) to the template are sufficiently small. To achieve effective subregion

searches, the dimensions and intensities of templates are required to correspond with those of chromosome regions in the subject image. This section presents a method for adjusting the dimensions of templates to those of chromosome regions in the image and adapting the intensities in the image to those of the templates.

A. Adjusting Template Dimensions

While the dimensions of chromosome regions vary with every image, the relative length of each chromosome type varies little from one image to another and the widths of chromosome regions are similar in each image. The proposed method binarizes an image by the intensities of pixels, and then determines the width W of chromosome regions and the sum of chromosome region lengths (total length L) in the binarized image. The determined W and L are used for adjusting the dimensions of templates.

Let p_c and p_b represent pixels corresponding to the chromosome regions and the background in the binarized image, respectively. To determine W, as shown in Fig. 3 (a), from every p_c bordered on the background, the Euclidean distances d_t to the other boundary are measured in eight directions $(t = 0, 1, \ldots, 7)$. With d_t and d_{t+2}, the estimated width w_t and its direction θ_t are calculated at p_c by

$$w_t = d_t \times d_{t+2} \Big/ \sqrt{d_t^2 + d_{t+2}^2}\,, \quad (1)$$

$$\theta_t = \tan^{-1}(d_t/d_{t+2}) + t \times \pi/4, \quad (2)$$

where $d_8 = d_0$ and $d_9 = d_1$. For calculating stability, d_t and d_{t+2} greater than a threshold Td are used for computing (1) and (2). In addition to w_t, from p_c to the other boundary in the direction θ_t, the actual measurement w'_t is taken (w'_t is the number of chromosome region pixels concatenate in the direction θ_t). As shown in Fig. 3 (b), where both boundaries are straight and parallel to each other, w_t and w'_t are both equal to the true width of the chromosome region. However, as shown in Fig. 3 (c), where boundaries curve or they aren't parallel to each other, w_t and w'_t are different and they may differ from the true width. Therefore, the proposed method accepts w'_t as the reliable width at p_c only when $e_t = |w_t - w'_t|$ is less than a threshold Te. If more than one reliable w'_t is accepted (i.e., for more than one t, the error e_t is less than Te) at a pixel p_c, the width w'_t with the smallest e_t is chosen as the region width at p_c. By choosing w'_t at each p_r bordered on the background and counting the occurrence frequency for every value of chosen w'_t, the most frequently occurred value is determined as the width W of chromosome regions in the image. Thus, by choosing reliable w'_t and using them for counting the occurrence frequency, the proposed method can determine W stably.

The sum of chromosome region areas in the image can be estimated as the total number S of pixels p_c in its binarized image, and it is approximated by the product of the region width W and the total region length L. Therefore, L can be determined by $L = S/W$.

Suppose that templates for subregion searches are made from an reference image I_R, and let I_S represent the subject image. In the proposed method, to adjust the dimensions of templates with those of chromosome regions in I_S:

- The width and total length of chromosome regions are determined in both I_R and I_S (those determined values are denoted by W_R, L_R, W_S, and L_S, respectively).
- The width of the templates are set to W_S, and the length of each template is multiplied by L_S/L_R.

B. Adapting Subject Image Intensities

Let $I_S(x, y)$ and $T(u, v)$ represent the intensities at (x, y) in the subject image and at (u, v) in the template, respectively. As shown in Fig. 4 (a), when the template is set at (x, y) and rotated by θ in the subject image, the MSE e^2 at (x, y) is computed by

$$e^2(x, y) = \frac{1}{UV} \sum_{u=0}^{U-1} \sum_{v=0}^{V-1} \left(I_S(x', y') - T(u, v)\right)^2, \quad (3)$$

$$x' = x + u\cos\theta - v\sin\theta, \quad (4)$$

$$y' = y + u\sin\theta + v\cos\theta, \quad (5)$$

where U and V are the width and length of the template, respectively. The rotation angle θ is set to minimize $e^2(x, y)$. The intensities of chromosome regions change with every image, and they vary locally in an image according to staining conditions. To achieve effective subregion searches, the proposed method adapts the intensities in the subject image so that the MSE $e^2(x, y)$ is reduced, and then uses the adapted MSE $\tilde{e}^2(x, y)$ for the subregion search.

As shown in Fig. 4 (b), the region of a template set in the subject image consists of two parts: one part O_b overlaps with the background, and the other part O_c overlaps with the chromosome regions in the image. O_b and O_c can be determined from the pixels corresponding to the background

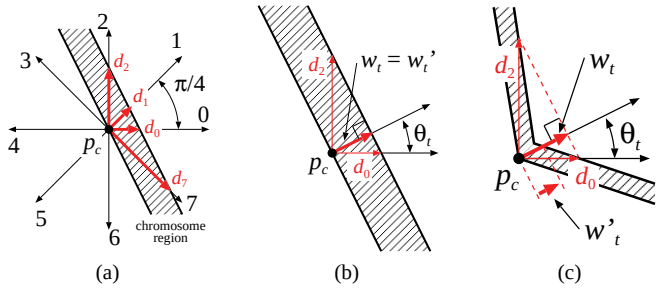


Fig. 3. Region width at boundary p_c.

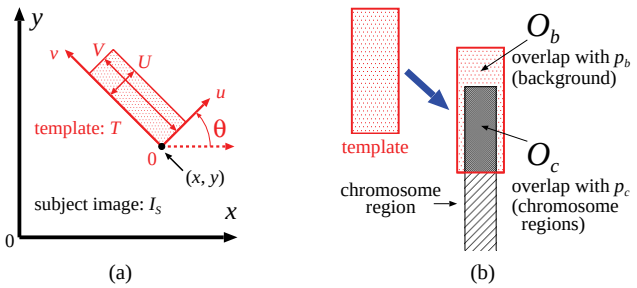


Fig. 4. Subregion search with a template.

p_b and chromosome regions p_c in the binarized subject image. The adapted MSE $\tilde{e}^2$ at (x, y) is determined by

$$\tilde{e}^2(x,y) = \frac{1}{UV}\left(\tilde{E}_b^2(x,y) + \tilde{E}_c^2(x,y)\right), \quad (6)$$

$$\tilde{E}_b^2(x,y) = \sum_{(u,v)\in O_b} \left(I_S(x',y') - T(u,v)\right)^2, \quad (7)$$

$$\tilde{E}_c^2(x,y) = \sum_{(u,v)\in O_c} \left(\alpha I_S(x',y') + \beta - T(u,v)\right)^2, \quad (8)$$

where the sums of squared-error $\tilde{E}_b^2(x,y)$ and $\tilde{E}_c^2(x,y)$ are computed in O_b and O_c, respectively. The intensities are similar almost everywhere in the background, and it is necessary that $\tilde{E}_b^2(x,y)$ is supplied to $\tilde{e}^2(x,y)$ as a penalty. Accordingly, $\tilde{E}_b^2(x,y)$ is computed from raw intensities $I_S(x',y')$ in the subject image, although $\tilde{E}_c^2(x,y)$ is computed from adapted intensities $\alpha I_S(x',y') + \beta$. For every subregion, constants α and β are set to minimize $\tilde{E}_c^2(x,y)$, and they are determined by

$$\alpha = \frac{|O_c| \sum_{(u,v)\in O_c} I_S(x',y')T(u,v) - \sum_{(u,v)\in O_c} I_S(x',y') \sum_{(u,v)\in O_c} T(u,v)}{|O_c| \sum_{(u,v)\in O_c} I_S^2(x',y') - \left(\sum_{(u,v)\in O_c} I_S(x',y')\right)^2}, \quad (9)$$

$$\beta = \frac{1}{|O_c|}\left(\sum_{(u,v)\in O_c} T(u,v) - a \sum_{(u,v)\in O_c} I_S(x',y')\right), \quad (10)$$

where $|O_c|$ is the number of pixels in O_c. If α not exceeding 0 is determined for any subregion, such subregion is excluded from the subregion search because the band pattern of it is reverse to that of the template.

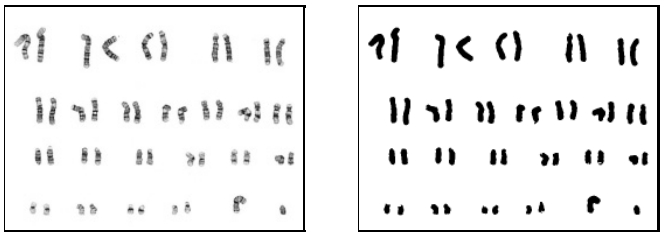
V. Subregion Search Experiments

To demonstrate the effectiveness of the proposed method, we carried out chromosome subregion search experiments.

A. Chromosome Images

Experiments were carried out on the chromosome images that are opened to public by the website of the Wisconsin State Laboratory of Hygiene and ZooWeb [3]. This site provides not only original chromosome images but also their classification results. An example of the original chromosome image and its classification result are shown in Fig. 1 (a) and (b), respectively. Although the proposed method can be applied to the original chromosome images, it is difficult to evaluate the subregion search results in them. Therefore, the subregion searches were conducted on the classification result images, where every chromosome region was extracted, classified, and arranged in standard order.

Thirty-one chromosome images (classification results) were used in the experiments. They consist of 19 female and 12 male chromosome images. This set includes 9 normal chromosome images (46 chromosomes in each image) and 22 numerical abnormal chromosome images (2 images with 45 chromosomes and 20 images with 47 chromosomes). Each image is 768×576 pixels in size, and symbols in it are removed beforehand. To conduct cross-validation, the images were divided into two sets A (16 images) and B (15 images). When one set was used as subject images, the other set was used as reference images and employed for making templates. Fig. 5 (a) and (b) show an example of the image and its binarized image. The binarized images were made by the method in [9], and they were used for determining chromosome regions, region width W, and total length L.



(a) (b)

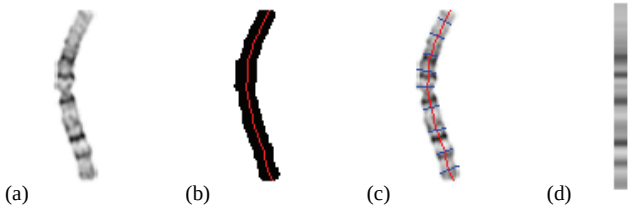
Fig. 5. (a) an example of the chromosome image, (b) binarized image.

B. Templates

To make templates for subregion searches, firstly, chromosome regions were extracted from the reference images in a chosen set, and the intensity profile was acquired in each extracted region. Secondary, for each chromosome type, the average intensity profile was made from the acquired profiles, and it was used as the reference band pattern of the chromosome type. Finally, templates were made by dividing the reference band patterns.

As shown in Fig. 6 (a) and (b), to acquire the intensity profile in a chromosome region, the medial axis is determined in each extracted region. On the determined medial axis, average intensities are taken perpendicularly to the medial axis (Fig. 6 (c)), and they are used as an intensity profile of the chromosome region (Fig. 6 (d)).

For each chromosome type i, intensity profiles $P_k^{(i)}$ $(k = 1, 2, \ldots, K^{(i)})$ are made. Since region extraction and medial axis determination may fail for some chromosome regions, the number $K^{(i)}$ of intensity profiles differs depending on the chromosome type i. In each i, the longest profile $P^{(i)*} = P_l^{(i)}$ is determined, the lengths of other profiles $P_k^{(i)}$ $(k \neq l)$ are extended to that of $P^{(i)*}$, and the average profile is made



(a) (b) (c) (d)

Fig. 6. (a) chromosome region, (b) extracting region and acquiring its medial axis, (c) taking average intensities perpendicularly to the medial axis, (d) intensity profile of the chromosome region.

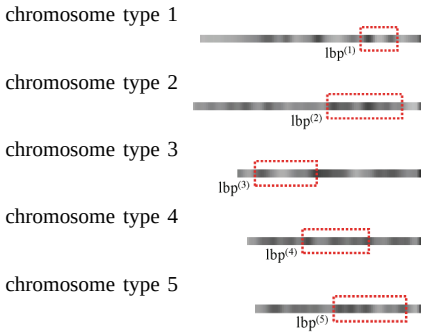


Fig. 7. Examples of the templates (lbp) used in the experiments.

from all $P_k^{(i)}$. The average profile is used as a reference band pattern of chromosome type i.

By dividing the reference band patterns into local band patterns (lbp), templates for subregion searches are made. To adjust the dimensions of a template for the chromosome type i, the width W_R and the total length L_R of chromosome regions are determined in the reference image where the longest intensity profile $P^{(i)*}$ is acquired. In the experiments:

- For the chromosome types 1, 2, ..., 5, a single template was made each in every set.
- Thresholds were set as $Td = 3$pixels and $Te = 2$pixels in estimating W_R.
- The mean and variance of intensity in each template were set to 100 and 50^2, respectively.

Fig. 7 shows examples of the templates made from set A and used in the experiments.

C. Experimental Results

The following four type methods of subregion search were applied to the subject images:

SRS1 without adjusting template dimensions and without adapting subject image intensities.
SRS2 without adjusting template dimensions and with adapting subject image intensities.
SRS3 with adjusting template dimensions and without adapting subject image intensities.
SRS4 (the proposed method) with adjusting template dimensions and with adapting subject image intensities.

Although several methods can be employed for improving the efficiency of scanning a subject image with a template, naive exhaustive searches were used in the experiments for the image scanning. The origin of a template ($(u, v) = (0, 0)$ in Fig. 4 (a)) was displaced to each pixel in the subject image except the background (the background was determined according to the binarized image), and the template was rotated on every displaced origin.

To evaluate their results, precision R and recall P were used. They are defined by

$$R = |C \cap D|/|C|, \quad (11)$$
$$P = |C \cap D|/|D|, \quad (12)$$

where C is a set of correct subregions (subregions corresponding to a template) and D is a set of detected subregions in a subject image. $|C|$ and $|D|$ denote the number of subregions in C and D, respectively. As shown in Fig. 8, for each template, the correct subregions C were set manually in every subject image. The detected subregions D for a template were defined as follows:

- Subregions in a subject image are sorted by their MSEs to the template in ascending order.
- If the MSE or the order of a subregion is less than or within a specified threshold, it is decided as 'detected.'

Fig. 9 (a) and (b) show examples of the detected subregions with SRS1 and SRS4. The numbers on the figures denote the order of each subregion (top 10 subregions are illustrated). These results were obtained by the same template whose correct subregions correspond to those on Fig. 8.

P and R were computed for each subject image by varying the specified thresholds, and the averages of them were calculated for each method. Fig. 10 shows the average R for all type methods (SRS1, 2, 3, 4) at different thresholds of order, and Fig. 11 shows the average R at different thresholds

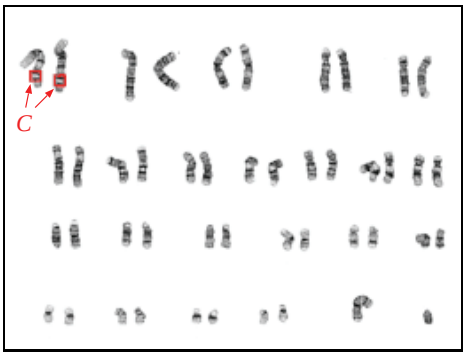


Fig. 8. Examples of the correct subregions in a subject image.

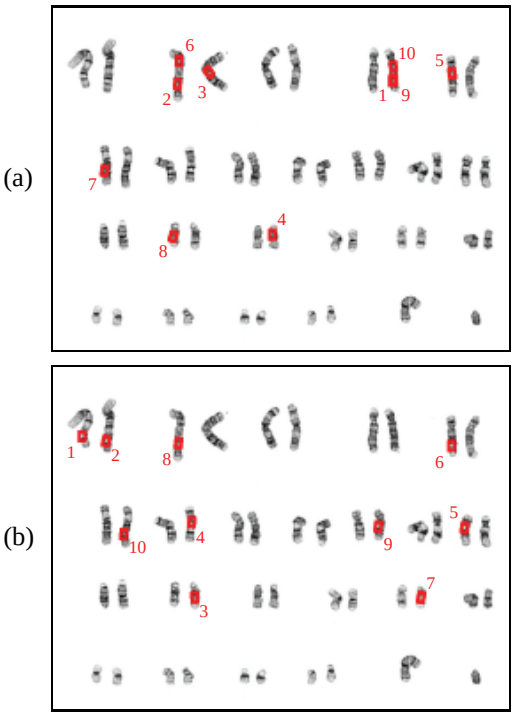


Fig. 9. Examples of the detected subregions: (a) with SRS1, (b) with SRS4.

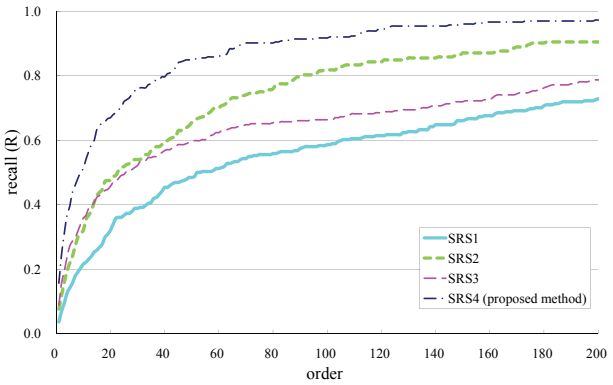


Fig. 10. The average R at different thresholds of order.

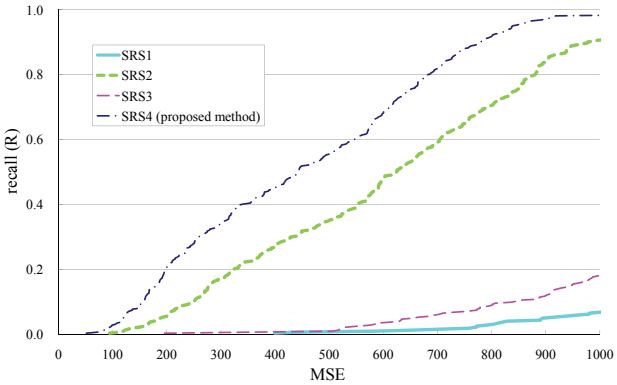


Fig. 11. The average R at different thresholds of MSE.

of MSE. Fig. 12 shows P at different R, where R was changed by varying the threshold of MSE.

These results show that adjusting the dimensions of templates and adapting the intensities in a subject image improve the accuracy in subregion searches, especially the proposed method, which uses both the approaches, improves the accuracy considerably. Consequently, it is expected that the proposed method can achieve subregion searches effectively.

VI. Conclusions and Future Works

In this paper, to develop an efficient chromosome subregion search, we have proposed the method for making the dimensions and intensities of templates correspond with those of chromosome regions in a subject image. By adjusting the dimensions of the templates to those of chromosome regions in the subject image and adapting the intensities in the subject image to those of the templates, the proposed method can improve the accuracy in subregion searches.

Several template matching methods have been proposed for image registration [10]. Those existing methods can accurately detect target regions in the subject image, however:

- Most of them require initial estimations of the target region dimensions.
- They cannot distinguish between the background and the chromosome regions, and they adapt the intensities in the background as well as in the chromosome regions.

Therefore, by adjusting templates with our proposed method and using them as initial templates for the existing methods,

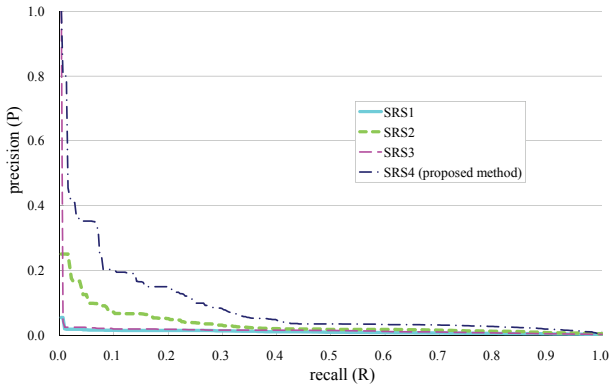


Fig. 12. The average P at different R.

it is expected that more effective subregion searches are achieved. In addition to this, by adapting the intensities with our proposed method, the accuracy in subregion searches can be improved further.

To achieve an effective recognition of individual chromosome regions in the subject image, we plan to develop following methods:

- A method for determining effective subregion templates in each reference band pattern.
- A method for extracting and classifying a whole chromosome region in the subject image efficiently.
- A method for recognizing a complement of chromosome regions in the subject image effectively.

VII. Acknowledgments

The authors thank H. Shimada, K. Asano, E. Sakai (Nihon Gene Research Laboratories Inc.), N. Ohba, and T. Satoh (IBM Japan, Ltd.), who have gave them valuable comments and suggestions for this work.

References

[1] A. Carothers and J. Piper, "Computer-aided classification of human chromosomes: a review," *Statistics Computing*, vol. 4, pp. 161–171, 1994.

[2] J. Graham and J. Piper, "Automatic karyotype analysis," *Meth. Mol. Biol.*, vol. 29, pp. 141–185, 1994.

[3] "ZooWeb — Karyotypes home page," http://worms.zoology.wisc.edu/zooweb/Phelps/karyotype.html.

[4] D. G. Harnden and H. P. Klinger, Eds., *ISCN1985: An international system for human cytogenetic nomenclature (1985)*. S. Karger AG., 1985.

[5] F. C. A. Groen, T. K. ten Kate, A. W. M. Smeulders, and I. T. Young, "Human chromosome classification based on local band descriptors," *Pattern Recognit. Lett.*, vol. 9, pp. 211–222, 1989.

[6] Q. Wu, Z. Liu, T. Chen, Z. Xiong, and K. R. Castleman, "Subspace-based prototyping and classification of chromosome images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 1277–1287, 2005.

[7] J. Piper and E. Granum, "On fully automatic feature measurement for banded chromosome classification," *Cytometry*, vol. 10, pp. 242–255, 1989.

[8] M. Moradi and S. K. Setarehdan, "New features for automatic classification of human chromosomes: a feasibility study," *Pattern Recognit. Lett.*, vol. 27, pp. 19–28, 2006.

[9] N. Otsu, "A threshold selection method from gray-level histogram," *IEEE Trans. Syst., Man. & Cybern.*, vol. SMC-9, no. 1, pp. 62–66, 1979.

[10] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image & Vision Computing*, vol. 21, no. 11, pp. 977–1000, 2003.