Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 38–55
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Automatic segmentation of metaphase cells based on global context and variant analysis☆

Gunter Ritter*, Le Gao

Fakultät für Mathematik und Informatik, Universität Passau,D-94 030 Passau, Germany

Received 1 August 2006; received in revised form 6 February 2007; accepted 18 May 2007

Abstract

We treat the problem of chromosome segmentation with the aid of shape analysis and classification. Our approach consists of a combination of two phases, a purely rule-based phase and a phase driven by constrained discriminant analysis. In the first phase, obvious prototypical shape elements related to touchings and overlaps are recursively identified, in the second, remaining complex and ambiguous cases are treated. The latter phase exploits global context by using variant analysis, a statistical theory of ambiguity recently established. The method turns out to be quite accurate. The system works on whole clinical cells and to a certain degree when band patterns are not or not well visible.

Keywords: Automatic chromosome segmentation; Karyotyping; Shape analysis; Context; Variant analysis; Diagnostic classification

1. Introduction

1.1. Automation of karyotyping

The period of reproduction during the cell cycle of a eukaryotic organism is called mitosis, that between mitoses the interphase. While chromosomes are in an extended state at interphase, they form oblong, detached objects at two phases of mitosis, the late prophase and the metaphase. It is at these stages that chromosomal structures can be made visible under a light microscope after suitable staining, see Fig. 1. This enables the cytogeneticist to detect gross aberrations in the chromosomal structure caused by pathological processes, genetic degeneration or environmental factors such as radiation. The number of chromosomes in the cell is specific to the species, in the case of humans 46 subdivided in 24 types. Each chromosome carries a constriction called the *centromere*. A chromosome is called *acrocentric* if the centromere is off center and *metacentric*, otherwise.

As a first procedure during the analysis of a pro- or metaphase cell, the cytogeneticist usually produces a karyogram. This is an arrangement of all its chromosomes displaying their biological classes [1] together with their polarities, see Fig. 2.

Karyotyping is a routine cytogenetic task, nowadays usually performed with the support of an interactive system. It lends itself to automation. In fact, its automation has a long history beginning in the 1960s, [2,3]. Fully automatic systems usually follow a number of consecutive steps.

(i) *Cleaning* of the image from stains and interphase nuclei;
(ii) *segmentation* of the cleaned metaphase cell in its different chromosomes;
(iii) extraction of *features* from all chromosomes;
(iv) *classification* of the feature sets into the biological classes.

The first two steps employ methods from image processing, the last two from pattern recognition and statistics. Since stains are in general smaller than the smallest chromosome and nuclei are big, round, and dark objects, both can be easily recognized and step (i) is not much of a problem. Opposed to (i), step (ii) remains a challenge, at least if it is to be performed with an accurateness close to that of the expert cytogeneticist and for all kinds of cells and preparations. The reasons are clusters of

☆ Work supported by Deutsche Forschungsgemeinschaft, Ri477/4.
* Corresponding author. Tel.: +49 851 509 3110.
E-mail address: ritter@fim.uni-passau.de (G. Ritter).

0031-3203/$30.00
doi:10.1016/j.patcog.2007.05.013

Fig. 1. A human metaphase cell.

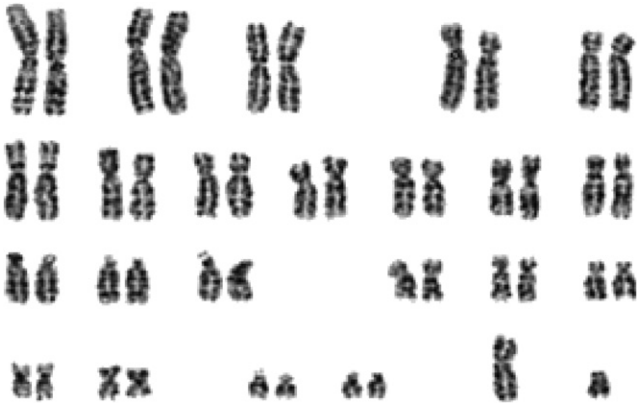

Fig. 2. The karyogram associated with the cell of Fig. 1.

touching and overlapping chromosomes that regularly appear in the two-dimensional images of modern preparations. Steps (iii) and (iv) are not easy either but have been satisfactorily solved in the past. In step (iii), it is favorable to extract about 30 features from each chromosome. They are mainly the size, the mean density, and a large number of features computed from the density and shape profiles, see [4].

The pattern recognition stages (iii) and (iv) receive from a successful segmentation process all chromosomes of the cell as isolated objects. However, feature extraction needs in addition the shapes of the chromosomes in the form of their longitudinal axes and their polarities. Both are *ambiguous* at this stage, the former due to possibly bent shapes. As a remedy, we proposed the application of a general concept for resolving ambiguities in pattern recognition based on statistical decision theory: variants [5–7]. A variant of an object is a feature set extracted from an object under a certain interpretation. Since the correct interpretation may not be known at a certain stage in a feature extraction process, several variants corresponding to different interpretations are extracted. The variant corresponding to the correct interpretation is the *regular* variant, the others are *irregular*. The problem is to find the regular variant. Variant analysis has found applications to the recognition of polarities [8] and shape [9] and to motif discovery in regulatory genomics [7].

Of each chromosome, at least two feature sets (variants) are extracted, one for each polarity [8], in the case of bent chromosomes or otherwise unclear shapes even up to 12 [9]. Each one accounts for possible interpretations of shape and polarity. Constrained classifier–selectors were designed that use all variants at a time [10]. For each chromosome, they estimate the regular variant associated with the true polarity and shape interpretation classifying all feature sets simultaneously. The most accurate *classifier* published to date for the present purpose is a robust maximum – a posteriori estimator derived from a statistical model of the (random) karyotype. It postulates independent chromosomes after normalization of the features across the cell. The model of (the features of) a chromosome consists of a mixture of a normal distribution and a quadratically asymmetric distribution [11,12] based on elliptical symmetry. The latter accounts for outliers and is responsible for the robustness of the classifier, see also [13,14].

At first sight, estimating the biological classes of chromosomes and, at the same time, selecting the correct variants may appear to be computationally infeasible. However, it turns out that the classifier–selector reduces to a transportation problem well known in operations research; there are efficient algorithms for its solution. To our knowledge, the discovery of the connection between a constrained ML-classifier as above and the transportation problem is due to Tso and Graham [15]. It was applied to karyotyping by Tso et al. [16] and to classification in the presence of variants by Ritter and Gallegos [10].

In this paper, we take a look at the *segmentation* process (ii) showing among other things that variant analysis may be successfully applied to this problem, too. We use it to take into account *global pictorial context*, a notion recognized as important also in other fields of image analysis, see, e.g., Torralba [17].

1.2. State of the art in chromosome segmentation

Automation of segmentation of metaphase images has a long history, two of the earliest sources being Ledley et al. [3] and Hilditch and Rutovitz [18]. However, in the early years, cells at late metaphase were used for analysis. At this stage, chromosomes are in a contracted state so that touchings and overlaps do not occur frequently and segmentation reduces mainly to finding the connected components in the image. The situation changed when cytogeneticists began to exploit the advantages of the early metaphase and late prophase for their analyses. Such preparations display many more bands and more detail. Modern preparations of amnion and blood cells consist of chromosomes whose shapes resemble short pieces of rope. Due to their greater lengths, chromosomes tend to touch and overlap to a significant degree. Clusters of 10 or more chromosomes are not rare. Therefore, a fully automated analysis today cannot dispense with a sophisticated component for disentangling clusters.