GASepo – a software solution for quantitative analysis of digital images in Epo doping control

I. Bajla\textsuperscript{a}, I. Holländer\textsuperscript{a}, M. Minichmayr\textsuperscript{a}, G. Gmeiner\textsuperscript{b}, and Ch. Reichel\textsuperscript{b}

\textsuperscript{a} Department of High Performance Image Processing,

\textsuperscript{b} Doping Control laboratory,

ARC Seibersdorf research GmbH,

A-2444 Austria, tel: +43 50550 2903, e-mail: ivan.bajla@arcs.ac.at

Abstract

A software has been developed that is aimed at quantitative analysis of images acquired by isoelectric focusing and double blotting procedures used for recombinant erythropoietin doping control. It represents a unified and easy-to-use tool for Epo doping experts in WADA accredited laboratories. It is based on image segmentation philosophy that enables identification of individual bands whose characteristics are needed for evaluation of the Epo doping positivity criteria. Several modules implemented in the GASepo software include an original know how, in particular, the method of robust calculation of the cut-off line, band segmentation and classification algorithms. GASepo is being used in several doping control laboratories worldwide.

Keywords: erythropoietin doping control, quantitative image analysis, geometry-driven diffusion, image segmentation, machine learning classification
1 Introduction

Human erythropoietin (Epo) is a glycoprotein responsible for the proliferation of erythrocytes in the human body. Approximately 90% of human Epo production takes place in the kidneys whenever a tissue oxygen sensor detects oxygen depletion. The recent advances in molecular engineering led to a possibility of production of recombinant human erythropoietin (rEpo) by chinese hamster ovary cells. Subsequently, rEpo has been approved for the treatment of anemia attributable to renal failure. Since the late eighties, rEpo is available on the market in various forms: epoietin α, β (rEpo) and darbepoietin (also known as NESP–novel erythropoiesis stimulating protein). Unfortunately, rEpo is also being used by athletes in endurance sports, like cross country skiing or cycling, as doping. rEpo boosts athletic performance by up to 10% as a result of increasing the number of erythrocytes ([1]).

The International Olympic Committee added rEpo to its “List of prohibited substances” in 1990, though no method existed at that time to detect it in body fluids. The possibility to produce rEpo comparably cheaply and the difficulty of its detection has catalyzed its use in recent years. Doping with erythropoietic proteins is one of the most complex and serious issues sport authorities are facing today.

2 Background

Previous research has shown that recombinant Epo differs from human Epo in post-translational modifications [2–4]. This difference manifests itself in different charge ratios of sugar moieties. It was found that for detection of such small differences, electrophoretic approaches to molecule separation are suitable [5–8]. Especially, isoelectric focusing (IEF), [9–10], proved to be a method of choice. However, due to very low concentrations of endogenous and recombinant Epo among
all other proteins, present in human urine, the detection of rEpo present in a urine sample is very
difficult. Lasne proposed to solve this problem by double-blotting (DB) [11].

Such techniques use antibodies for detection and identification of proteins. The process involves
the separation of Epo isoforms on a polyacrylamide gel followed by the transfer of the proteins
onto a thin membrane (blot). An Epo-specific antibody is incubated on the membrane resulting in a
mirror image of the first membrane. Isoforms are detected by a chemiluminescence reaction after
incubation of a second antibody, an enzyme catalyst and luminescence reagent. After imaging
(analog or digital), a typical pattern of lanes (vertical stripes) is finally generated. As can be seen
in Fig. 1, the lanes comprise spots (bands) of individual isoforms, which have been separated by
\(pH\) gradient. When a sample containing rEpo is subjected to IEF, a shift to more basic isoforms
(upper parts of the lanes denoted as standard in Fig. 1 with 4–6 bands at different \(pH\) positions) is
observed compared to endogenous Epo. When urine with natural uEpo (urine Epo) is subjected to
the same process, the bands observed (7–15 bands in the lanes 2-6, 8-11 in Fig. 1) are separated.
These bands partially overlap the region of those ones belonging to rEpo. The detection of a
positive doping case (i.e. detection of rEpo in presence of endogenous uEpo) is based on setting
the reference cut-off line (\(col\)) and on comparison of characteristics of the bands located above the
\(col\) in sample and standard lanes.

At doping control laboratories Epo images were originally evaluated by a combination of mea-
urement on the image and using various multipurpose software packages. However, these are
not tailored to this specific task. The use of different software tools for data analysis interferes the
search for a common basis of Epo image interpretation. To provide a systematical basis for quanti-
tative analysis of Epo images and to contribute to the process of standardization and harmonization
in this area, an interdisciplinary team of information technology and doping control (DC) experts
was created in ARC Seibersdorf research GmbH in 2003. At the beginning of cooperation a pilot
software GASepo1 has been developed in MATLAB\textsuperscript{TM}. The software was tested on Epo images
provided by several DC laboratories and discussed at international forum of DC experts. Based on the first experience and positive response of DC community, a new project was formulated. The goal of this project was to develop a reliable and easy-to-use software package, specially designed for quantification of recombinant erythropoietin in urine samples as a standardized and unified tool for the use across DC laboratories worldwide thus providing for international harmonization of Epo analysis. In 2004 the international project was accepted and supported by the WADA grant. Eight other DC laboratories are involved in this project as permanent testers of the software which is being developed and updated in ARC Seibersdorf research GmbH.

Figure 1: Typical Epo image acquired by a CCD camera. The lanes denoted as standard represent the samples in which only rEpo with known concentration has been included.

2.1 User requirements

The main goal of the interdisciplinary research project is to design a software system which would make possible to carry out Epo doping tests in WADA accredited DC laboratories on a systematic, unified and user-friendly basis. The requirements to such a software are divided into two groups: 1. core functionality requirements (derived from WADA criteria), and 2. efficient work-
Core functionality requirements

- using the information from standard lanes, the software has to calculate the reference cut-off lines (cols) (one for epoietin $\alpha$, $\beta$, and one for darbepoietin),
- it has to detect individual bands in all lanes as unique objects whose individual intensity characteristics have to be evaluated according to the doping positivity criterion in regard to the col,
- it has to make possible superimposing a reference photography of the gel (with some methyl red added to the catholyte) with the corresponding Epo image of this gel; the deformation of the methyl red line reflects the local deformation of the gel caused by inhomogeneities of pH gradient in the gel,
- it has to incorporate a rigorous reporting of all visual and numerical data generated during the quantitative analysis of the given Epo image,
- image noise has to be suppressed before subsequent analysis of the given Epo image,
- inhomogeneity of the image background, which is frequent in Epo images, has to be corrected.

Efficient work-flow requirements

- possibility to load images in various formats,
- simple adjusting of brightness and contrast of the Epo image,
- user-friendly selection of geometrical region of interest (ROI) in the image,
- automatic estimation of lane margins with a possibility of their manual correction,
• possibility of definition of the type of standards (epoietin $\alpha$, $\beta$, darbepoietin, or a combination),

• variability of tools for easy work with an image, if interaction is necessary (zooming, panning, etc.),

• the proportion between automatic and interactive analysis should follow the rule: the higher the quality of the gel image, the more analysis tasks are solved automatically,

• instructive presentation of all relevant numerical characteristics of the bands in selected lanes,

• documenting of the doping control results in a suitable form and format.

3 Design considerations

3.1 General aspects

The above mentioned requirements to the software system GASepo (Gel Analysis System for epo) and our previous experience with the development of the GelMaster system ([12]), aimed at analysis of electrophoretic gel images of DNA fragments, constituted a basis of our design concept. We decided to orientate the system core towards the methodology of band segmentation in Epo images. This methodology enables us to operate with isolated objects whose characteristics (intensities and volumes) are needed for evaluation of Epo doping positivity criteria.

It is known that gel images, in particular Epo images, suffer from Gaussian and impulse noise. Since results of image segmentation operations are sensitive to noise, it is necessary to suppress the noise of Epo images before their segmentation. In [13] we explored modifications of the nonlinear image noise suppression method that is based on the principle of geometry-driven diffusion (GDD)
This method was successfully applied to our system GelMaster ([12]). Therefore we have chosen it also for preprocessing of Epo images in GASepo.

Based on analysis of several hundreds of Epo images originated from all DC laboratories included in the WADA project, a conclusion was made that the image background is significantly inhomogeneous even within individual lanes of Epo images. It is difficult to model the random behaviour of local defects occurring in the background by a physical (stochastic) model which would make possible to correct them by a parametrized function of two variables. Therefore, for solving this problem we have chosen a combination of heuristic and numerical approaches, that is described in Appendix 2.

The prerequisite of the use of quantitative characteristics of individual bands for evaluation of the Epo positivity criteria is knowledge of reference cut-off line. We have analysed ([17]) the conventional 1D method of cut-off line calculation and found out its weakness. We decided to explore alternative approaches to the cut-off line calculation. The method we proposed is described in Appendix 3.

Computer tests showed, that in spite of the correction of most degradation effects in the preprocessing stage, the application of edge detectors or adaptive thresholding to band segmentation in Epo images does not yield satisfactory results. Research into modification and combination of various methods was needed. In particular the problems of merged and disrupted bands occurring in Epo images had to be solved within the segmentation algorithm. The know how we developed is addressed in Appendix 4.

There is no guarantee that the segmentation procedure segments only true bands, and false positives (artifacts) can be detected too. To separate the artifacts we suggested a post-segmentation classification procedure based on machine learning approach (see Appendix 5).
Besides the mathematical and algorithmic problems, implementation tasks related to each of functional module had to be considered. The main software feature, we decided to implement, is automatic sequence of individual operations. This basic mode is supported by a number of graphical tools which enable to control automatic operations and to correct them in relevant cases.

The overall philosophy of software use is concentrated in a built-in sequence of operations to be performed with Epo images. These are invoked individually and in chronological order by the user.

### 3.2 Flow of image processing operations in GASepo

The flow of operations, which served as a basis for the design of the GASepo structure, is depicted in Fig. 2.

The GASepo system is structured into a sequence of five subsequent basic stages: *Load, Select ROI* (Region Of Interest), *Rectify, Analyze and Report* (Fig. 2). The *Load* stage enables loading of images from files in various formats and visualizing them. The *Select ROI* stage is aimed at selecting a rectangular fragment of the input image in which only useful information is comprised. The *Rectify* stage is used to correct geometrical distortions (bended bands and slanted lanes) occurring usually in Epo gel images. GASepo supports rectification by a combination of user intervention and automatic calculation. The user introduces graphically a source mesh (solid lines in Fig. 9) as an approximation of the shape of geometrical distortion, while the regular target mesh (dashed lines in Fig. 9) is superimposed on the input image. The warping coordinates are calculated by fitting each quadrangle of the source mesh into the corresponding rectangle of the target mesh, via bilinear interpolation. The software calculates and displays the rectified image in real time with the user interaction.

The core part of GASepo is constituted by the *Analyze* stage, which is dedicated to interactive
Figure 2: Main flow of operations in the GASepo software.
analysis of Epo images. The lane separations, estimated by the software, can further be adjusted by the user who also selects the standard lanes (epoietin α, β and darbepoietin). The software calculates the image background, the position of the reference cut-off line, and the doping substance ratio for each individual lane. It performs the band segmentation operation which is based on a combination of three LoG (Laplacian of Gaussian) filters, using different values of the parameters, with subsequent morphological operations (see details in the following chapters and Appendix 4). Once the bands are represented as individual objects, the summation of intensities, needed for evaluation of the doping positivity criteria, is applied only to band pixels. Thereby the inaccuracy introduced by background intensities can be considerably reduced. For visualizing segmented objects a flexible transparency setting has been incorporated into the software. The segmentation operation is followed automatically by a classification procedure that classifies the segmented objects into two classes: bands and artifacts. A simple interactive tool is added for making all the artifacts invisible. The novelties of the methods proposed and algorithms implemented into the program system GASepo are concentrated in this operation stage.

The fifth stage is the Report stage which serves for exporting the analysed data and the results to external files. A multipage PDF document file is generated which contains full description of the analysis results, including Epo image itself with graphical overlay showing lanes, cols, segmented bands, etc.

GASepo provides a comprehensive view at the Analyzed gel by using variety of visualization techniques. These include 1D, 2D, 3D visualization, and pseudocolour representation.
4 System description

GASepo has been developed in the C# programming language [18] in combination with the Microsoft™ .NET Framework.

The GASepo software has been designed for use on computers with the following minimum hardware requirements: Processor: Intel-based architecture, minimum specification: P4, Intel Pentium III Processor, 800 MHz, RAM: 256 MB, Free hard disk space: 100 MB, Graphics adapter resolution: 1024x768.

The minimum software requirements are as follows: Microsoft Windows 2000 or Windows XP, Direct X 9.0, Internet Explorer, Version 5.1 or higher, Adobe Acrobat reader. Direct X 9.0 and Adobe Acrobat Reader are distributed with the GASepo software.

4.1 The system overview

GASepo is a window application. Immediately after start, GASepo opens its main window with the logo as displayed in Fig. 3. Generally, all interaction takes place within this window. Exceptions include file selectors, warning, error and similar messages. Standard way of working with GASepo is to load an Epo image from an image file (buttons “Input” and “Load”), to make necessary settings for better visual appearance, to select a ROI, to perform geometrical corrections via rectification tools, if necessary, to analyze the given Epo image, including all necessary steps for evaluating doping positivity criterion, and finally, to generate a final report in PDF format and to save the session file for archiving. The basic steps of the session (as depicted in Fig. 3) – stages are initiated by the buttons located in the left upper corner of the main GASepo window (Fig. 3). There is an optional utility incorporated into GASepo that is devoted to superimposing a reference red line photo on the given Epo image.
4.2 The input stage

Contrast, inversion and pseudocolours

Epo images, generated in various doping control laboratories, vary considerably in brightness and contrast. To make sure the user will be able to spot the tiny variances in the intensities and not miss any of the true bands, one can adjust the way the software displays images. The gray-levels of Epo images are restricted to certain values within 16 bit wide originally acquired data. To take advantage of this range, two sliders have been incorporated into the software for restricting the
Figure 4: The complete layout of GASepo with buttons for interaction: an example of the ANA-LYZE stage, in which the image view contains all relevant lanes with bands and reference cut-off lines, and in the lane view quantitative results of the analysis of the given lane are displayed.

The black and white range of 16-bit Epo data files (tools marked as group 5 in Fig. 4). To set the correct contrast of a visualized image, the left slider (as depicted in Fig. 5) defines the maximum value that is mapped as black, whereas the right slider (as depicted in Fig. 5) defines the minimum value that is mapped as white. The range between these two values is linearly stretched into the visible gray scale.

Sometimes it is useful to see inverse of the image, i.e. an image with bright bands on dark background will show as dark bands on bright background and vice versa (a button in the area 5 in Fig. 4). Other buttons are dedicated to pseudocolour mapping of the Epo image and contrast stretching independently in each lane.
The analysis begins by loading the Epo image from an image file. Usually, this file is obtained by scanning the film in a flat-bed scanner, or by scanning the IEF membrane directly by a digital camera. This software is able to process 8-bit or 16-bit image files. The recommended spatial resolution for optimum analysis of a typical Epo image is about 30 to 40 (width) by 300 to 500 (length) pixels per lane.

The GASepo software is able to read the following image formats: TIFF (8 or 16 bit Grayscale, and 24bit RGB LZW compressed), BMP, PNG, PGM (Grayscale or colour). Colour images which are loaded as Epo images are automatically converted into grayscale. Extra-large image files with vertical size over 1024 are automatically (with a confirmation request) downscaled to the vertical size 1024 (aspect ratio is preserved). This feature allows to process extra large Epo image files without significant performance loss.

**Annotation of the Epo image of bands**

The annotation of the session (such as identification, date, operator and general comment) can be entered during the input stage. The annotation does not affect the processing in any way. It appears in the report document. The session annotation field contains important information about
4.3 Selection of the ROI

A typical Epo image usually covers a larger area than the region with the relevant lanes and bands. The processing has to be restricted to the ROI which covers all lanes and all bands but not more. Such a region is typically a rectangle which may be slightly rotated with respect to the input image rectangular frame. In GASepo, the selection of a rectangular ROI with arbitrary rotation can be performed in the following interactive way. After entering the ROI stage, a ROI definition marker (yellow in the original display) is overlaid on the image in the default position. It consists of a rectangle and four control points in the corners: TL-top left, TR-top right, BL-bottom left, and BR-bottom right. In the original display, the TL control point is marked blue, whereas the remaining three points are green. Active state of any of the control points, depressed, is indicated by red colour of the points. The interior of the rectangle represents the ROI. The position, size and slope of the ROI can be adjusted using all four control points (in Fig. 6 a state after such an adjustment is displayed). Return from any position of the ROI rectangle is possible.

Double view

With the ROI, RECTIFY and ANALYZE stages, a double view can be used. In the double view, the size proportion between the upper and lower window can be adapted by simply dragging the separation line up and down (Fig. 6, Image preview).

Magnification, reduction, panning and reset

There are four mutually exclusive tools associated with image magnification operation (area 1 in Fig. 4). After interactive definition of a rectangle, magnification (zooming in) can be applied to this rectangle. The actual visual magnification is also dependent on the size of the windows. Reducing
(zooming out) the image is an operation opposite to magnification. Pan operation can be used for moving the image to a convenient position within the basic (upper) image window. The last optional operation (area 1 in Fig. 4) is return of the image into the initial size and position. All these operations are available in all stages where this makes sense. The current magnification/reducing of the image does not affect the processing in any way and serves solely for convenient viewing. All these functions can be independently related to the active lane image displayed in the lane view (Fig. 4).

![Image of image selection](image.png)

Figure 6: Selection of the rectangular region of interest (ROI) on an original input Epo image: its default position and sizes are transformed by the user (via moving the control points TL, TR, BL, BR with various functions); in the image preview the result of the selection is displayed online.

**3D mode**

There is a possibility to view the Epo image in a 3D view with intensities presented as elevations. By using the left or the right mouse button the 3-D image can be turned round the x- and y- axis. With the middle mouse button zoom in and zoom out is possible.
4.4 The overlay stage

Following the idea of Lasne ([19]) we have incorporated into the software an option to combine the Epo image with a photography of the gel with some methyl red added to the catholyte (pH reference image, Fig. 7).

Figure 7: The gel image window (left), the pH reference image window (right), and the window for description of the image analysis by the user (bottom).

After activating the overlay stage, the Epo image and the pH-reference image are automatically overlayed. Both images have to be registered to each other by using the operation controls for setting the ROI. The software is able to filter the selected colour by using transparency slider (marked by an ellipse in Fig. 8).
Figure 8: Overlay of the image with the red reference line onto the original input Epo image; transparency of the red line reference image is controlled by the slider marked by the ellipse.

### 4.5 Rectification stage

The main purpose of the rectification is to compensate for nonlinear geometrical distortions of the Epo image, mostly caused by inhomogeneous electrophoretic process. The rectification stage is initiated by pressing the button *Rectify* in the stage selection menu. The rectification is based on the following paradigm: in the distorted image, we define discrete source point positions, which should be transformed to target positions within a regular mesh in the rectified image. The underlying image gets transformed by means of interpolation, so that it “follows” the geometrical transform of the mesh (Fig. 9). In comparison to the implementation of the rectification module in our GelMaster system ([12], the interaction and graphical tools of the rectification in GASepo have been extended.

The user’s interaction is organized as follows. Any number of vertical or horizontal straight lines can be introduced by means of navigating the arrow cursor and opening a specific window. The
position of the line being inserted within the image is given by the position of the arrow cursor at the moment of clicking the mouse button. In Fig. 9 the case of vertically distorted Epo image is displayed. As the slope of distortion in individual lanes may differ, for correcting the image the user should adjust each vertical line separately. The goal is to shift the white lines of the rectification mesh into positions which follow the slope of the lane distortion (the stage Before in Fig. 9). Red lines at places of initial yellow lines appear. For each white line (yellow in original display) of the mesh the corresponding red line represents the points of the desirable output mesh towards which the input points are transformed.

Figure 9: Illustration of the rectification principle used in GASepo: a curvilinear mesh (“Before”) is superimposed on a distorted Epo image following geometrical distortions, the rectified (geometrically corrected) image is illustrated in the part “After”.

4.6 The analyze stage

For the ANALYZE stage two windows are specific: image view and lane view (Fig. 4). Its layout comprises a number of specific buttons and objects (areas marked by black rectangles 2, 3, 4 in Fig. 4).
4.6.1 Image settings

First thing which the user should check is related to the type of the gel image, since the mistake in this stage of the analysis may lead to processing of incorrectly represented data. Epo images can come in two variants: 1. those, which comprise bright bands on dark background, and 2. vice versa those, which are constituted by dark bands on bright background. In most cases the algorithm incorporated into GASepo solves this task and when the second type of the gel image is present, it is automatically converted to the type one because the data which are to be processed in the subsequent steps need to be represented in this form. This operation is performed transparently for the user. Some gel images have very unregular background and it can happen that algorithm does not perform needed inversion of intensities. To manage such cases in a simplest way, the software provides a specific window in which the indication of the gel image type is displayed. In this window the representation of the data which will be processed is indicated. As can be seen in Fig. 4 (right upper part), the indication shows that the data has not been inverted, therefore the user must press the radio button for bright bands. The user can orientate himself in the situation by using also the plot of the profile displayed within the lane view (Fig. 4 left bottom part). As the profile values increase from left to right, the given plot indicates dark bands on bright background. Inverted image intensities are confirmed in the profile window by changing the plot into the correct shape.

The lane profile pane (lane view in Fig. 4) has a button which invokes the stretching slider. The user can move the slider, and the slider disappears after releasing the mouse button. It is calculated as a vector of median intensity values for each row of this lane. Using the slider, the plot of the lane profile may be stretched by logarithmic transformations. The goal is to enable the user to observe the differences of the profile peaks in more details. The thin curve displayed together with the profile represents a profile (a baseline curve) of the background correction surface for the
given lane. In the calculation of the final epoetin $\alpha, \beta$ or darbepoetin doping ratios the values of the base curve are subtracted from the lane profile.

Some images manifest overexposed intensities which are cut to the visible domain of gray scale values (0 to 255). For the images with such saturated areas, it is questionable, if the analysis can lead to acceptable results. Therefore it is on the decision of user, whether to proceed the analysis or not. For making this decision reliably, saturated regions of the given image are artificially highlighted.

### 4.6.2 Lane identification

The algorithm of lane identification calculates an estimate of the number of lanes. It is based on the approach described in [12]. Individual lanes are marked by blue vertical rectangles (in original display). The accuracy of this estimate depends on the accuracy of the physical division of the tracks of bands in the gel and on the appropriate position of the ROI rectangle set by the user. If the estimated number of the lanes does not match, the increment/decrement buttons (right upper part of Fig. 4) can be used for additional adjustment. The accuracy of the lane boundary positions depends similarly on the regularity of the physical tracks of bands. For adjusting the boundary of any individual lane a specific cursor image is invoked by clicking on this boundary. Clicking on the area within the lane makes this lane active (that is indicated in original display by yellow colour of its boundaries). The lower part of the layout (lane view in Fig. 4) serves for displaying objects and characteristics related only to this active lane. The bands within the neutral (endogeneous) region are marked by greek letters. The bands in the basic region (alone epoetin $\alpha, \beta$) are marked by integers, the bands in the acidic region (alone darbepoietin) are marked by capital letters. The user can select a band by either using the band properties field or by clicking on the band itself. The selected band is marked by a yellow rectangle and with a yellow line within the band properties.
4.6.3 Lane properties

The user can assign each lane a different lane identification name and sample ID. He/she can set the concentration, and leave an individual comment for the selected lane which will be showed in the analysis report. The numbers of the individual lanes and their type are listed and indicated by colours at the lane headers. Each lane can be assigned an own property. By clicking on the areas “Prop” (lane headers), a separate lane properties window is opened. There are three different lane property options available: 1. Type: four lane types are possible: sample, epoietin standard, darbepoietin standard, and combined standard. 2. Behaviour: allows cut-off line repositioning. 3. Background correction: the user can select the background correction surface from a lane different to the given lane.

After pressing the button from the main operation menu all lanes have the default type Sample. When all necessary adjustments of individual lane boundaries are finished the interactive assigning the attributes of standard lanes follows.

4.6.4 Cut-off line detection

For development of a software that would be tailored particularly to the analysis of digitized Epo images, it was necessary to implement an automatic and robust method of searching for a reference \( col \). Based on computer tests with calculation of the \( col \) using conventional 1D profile of the lane, we have decided to change the 1D paradigm to 2D basis. The method ([17]) we proposed is described in Appendix 3.

After defining the type of a standard lane, the \( col \) (one or two) are automatically calculated for this
lane. If more lanes in the given Epo image are specified as standards, the *cols* of the whole image are calculated as linear interpolation between the *cols* found in the specified standard lanes.

Sometimes the gel images are heavily corrupted, however they are still acceptable for doping control. The defects in homogeneity of the background obviously violates assumptions on which the method of the *col* calculation is based. It implies a wrong position of the *col*. For these cases the GASepo software is equipped with an interactive tool that can be used for excluding an inhomogeneous region from a standard lane (area 2 in Fig. 4). For cancelling this intervention an additional button is included.

### 4.6.5 Band segmentation

The band segmentation operation in the whole image is performed simultaneously with any change of the parameters of the lanes (boundaries adjusted or regions removed). The results of the segmentation are visualized by pseudocolours whose level of transparency is controlled by the slider in the layout area 1 (Fig. 10). The segments are displayed as a transparent layer over the original Epo image. The level of transparency can be controlled by the user. The colour of the bands varies from blue to red and represents the relative height of the band peak (red being the most prominent band peak).

The segmentation of the bands is a complex computational process governed by specific parameters (details are given in Appendix 4). The most important of them are triples of thresholds applied to the result of *LoG* operation. To provide the user with a possibility to select various values of these triples a special slider (“Segmentation sensitivity”) has been designed (area 4 in Fig. 10). The way these values affect the result of segmentation can be described as follows: when the slider is set to “high” values of the thresholds (the leftmost lane in Fig. 11), the segmentation produces smaller number of objects, i.e. minimum of artifacts are present. However, it can happen
Figure 10: In the ANALYZE stage all segmented objects are visualized by colour transparency. The lanes contain a number of artifacts which are to be discarded by a classification procedure. that by this setting some bands are missing. In Fig. 11 the results of segmentation for gradually decreased values of the thresholds are illustrated. If the position of the segmentation parameter slider is set to the minimum threshold values (the rightmost lane in Fig. 11), the maximum number of objects is generated. Besides the bands, which were missing before, additional artifacts can appear. Therefore adjusting an optimum position of the segmentation parameter slider is a question of trade-off that is to be decided by the user.

**Band quantification**

Once the bands are delineated by segmentation, they can be quantified. Quantification is based on two parameters: The maximum height of a band and the volume of a band.

**Band height.** This is calculated as the maximum intensity value within the area of the band.
Figure 11: Interactive adjusting of segmentation parameters by a slider: the individual positions of the slider represent segmentation thresholds resulting in different number of segmented objects. The intensities are taken from a filtered background-corrected image, so that random peaks caused by impulse noise are eliminated and do not affect the result.

**Band volume.** This is calculated as the sum of filtered background-corrected intensities of all pixels within the area of the band segment. It is clear that the two quantitative parameters correlate but they might slightly differ. Both band height and band volume parameters come in two versions:

- **Band height absolute** and **band volume absolute** (values of the parameters). These values are direct results of the internal calculations which are related to pixel intensities. They make little sense as such (since pixel intensities are themselves relative values) but can be used for comparisons of bands in different lanes.

- **Band height normalized max** and **Band volume normalized max.** Parameters are normalized with respect to the maximum value within the lane. Thus, the highest (or most voluminous) band in the lane has the normalized max value 100, and the other bands within the same lane have normalized max values between 0 and 100.
**Band height normalized sum** and **Band volume normalized sum**. Parameters are normalized with respect to the total sum within the lane. Thus, the sum of all band heights (or volumes) in the lane is 100, and the individual bands within the same lane have normalized sum values between 0 and 100, reflecting the proportion of this band’s value to the total. The parameters with attribute absolute, normalized max, and normalized sum are linearly dependent.

**Band centroid x**. The pixel position of the centroid of the band area. The x coordinate is perpendicular to the direction of the electrophoresis and increases from left to right.

**Band centroid y**. The pixel position of the centroid of the band area. The y coordinate is identical to the direction of the electrophoresis and increases from top to bottom.

The values of all the characteristics for an active lane are displayed in Table which is located in the lane area together with lane image and profile to which it is related. The detailed view of the lane area is displayed in Fig. 12.

<table>
<thead>
<tr>
<th>Name</th>
<th>Comment</th>
<th>Centroid X</th>
<th>Centroid Y</th>
<th>Volume Absolute</th>
<th>Volume Mean</th>
<th>Height Absolute</th>
<th>Height Mean</th>
<th>Height Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>43.4</td>
<td>75.1</td>
<td>1983.7</td>
<td>14.0</td>
<td>2.0</td>
<td>0.5</td>
<td>16.7</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>48.0</td>
<td>74.1</td>
<td>3804.5</td>
<td>26.6</td>
<td>3.6</td>
<td>14.4</td>
<td>26.5</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>45.6</td>
<td>106.3</td>
<td>5265.5</td>
<td>37.3</td>
<td>5.4</td>
<td>20.6</td>
<td>37.3</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>45.3</td>
<td>134.1</td>
<td>7901.8</td>
<td>95.2</td>
<td>8.0</td>
<td>27.7</td>
<td>59.0</td>
</tr>
</tbody>
</table>

![Image of GASepo layout in the ANALYZE stage](image)

Figure 12: The lane area of the GASepo layout in the ANALYZE stage: the diminished version of the selected lane with segmented bands is displayed at the left side of the window, the mean lane intensity profile is depicted in the next window (the slider enables various mapping of the profile), and finally, the table with characteristics of individual segmented bands in the given lane is depicted.
4.6.6 Band classification

The band classification operation in the whole image proceeds after segmentation operation. It is transparent for the user and it appears to him/her as simultaneous with any change of the parameters of the lanes. A special button has been designed (area 2 in Fig. 10), the initiation of which results in disappearing of all artifacts found by classification. If the user needs to define an artifact object as a band, he/she can do it by clicking on the relevant object in the image while this button is on. The resulting image in which the artifacts are suppressed is displayed in Fig. 13.

Based on experience of GASepo testers another interactive tool has been designed, namely, the stepwise dilation of the bands by one pixel (extension of the band boundaries by one pixel). This operation is controlled by the user using the incremental counter (area 3 in Fig. 10). The result of dilation of the bands by 3 pixels is illustrated in Fig. 14.

![Figure 13: Segmented bands displayed without disturbing artifacts in original shape, i.e. the objects classified as artifacts are suppressed.](image)

**Special interactive tools – cutter and gluer**

27
Figure 14: All the bands are displayed after their dilation by 3 pixels; the size of structuring element can be selected by the user using the button marked by the rectangle.

Due to occurrence of specific degradation effects in Epo images, a segmented object is sometimes constituted by a band merged with an artifact. For separating these two objects, a tool “cutter” has been designed (area 4 in Fig. 4). Assume, an object is constituted by a true band glued together with an artifact. After pressing the cutter button, all segmented objects in the image considered in this analysis step as bands appear boxed in white rectangles. By holding the left mouse button a cut line across the band can be dragged. The segment will be divided into two bands which will get their own bounding boxes and new labels.

Only denoted objects (bands) are enabled for cutting. If the object, the user wants to cut, is not classified as a band, it has to be reclassified first. The other way round, all new segments of a cut band will be classified by default as bands. If some of the segments are artifacts, they have to be reclassified manually. The colour of the band will change after a cut. This is obvious since, basically, the old band disappeared and two new bands have been created. The cutter is an
intelligent tool. It tries to predict the right way of cutting. Often, it is enough to draw a short line in the middle of the segment, and the tool will complete the cut.

An interactive tool band “gluer” has also been incorporated into the ANALYZE stage. This tool is not exactly reverse to the cutter, since it does not create new glued objects. Its purpose is just to integrate characteristics of the individual objects which are being glued. This function is necessary for some cases of image analysis. The main interaction is concentrated on the relevant rows in the Table (lane view) which are marked by yellow colour. Then a small menu window is invoked in which the functions “glue” or “unglue” can be chosen. Characteristics of the new object are displayed in a new row where the names of all objects glued together are listed using the symbol “+”.

4.7 Reporting stage

In the REPORT stage, a multipage PDF (Portable Document Format) document file can be generated which contains full description of the analysis results, including the Epo image itself with graphic overlay showing lane, cut-off lines, etc. PDF is a widely used document format which can be read and manipulated by the Acrobat Reader software. This format is suited for archiving, as well as for document exchange.

On the first page (Fig.15) of the GASepo Analysis Report, the whole Epo image selected as ROI is displayed in gray level domain. The headers of the individual lanes are equipped with necessary description including the ratios of band intensities above and below cols for epoietin α, β and darbepoietin. These data serve only for reference purpose and they are not used for Epo doping positivity evaluation any more. All bands are denoted by symbols according to the convention explained before.

On the second page (Fig.16) of the GASepo Analysis Report, the analysed Epo image with its
lane specific settings and evaluations is displayed. The following pages of the GASepo Analysis Report show the results of analysis of each lane and tables containing detailed information. The last page of the report contains a short summary of the Epo image analysis.

Figure 15: An example of the first page in a report file: the selected region of an Epo image with the results of the ANALYZE stage is displayed.

**Session save and load**

The GASepo software allows to interrupt the user’s work at any time during the analysis and to save the work to a disk file. Such files are called *session files* and they include all data and settings used during the session. After saving the session file, it is possible to correctly exit the program. When the program is started again, the session file can be loaded by the “load” button and it is possible to continue working at exactly the same stage as was achieved before session saving. The session files are also suitable for archiving, since they represent the complete documentation of the entire Epo image analysis process.

**Undo and redo**
The GASepo software also features a virtually unlimited undo/redo stack, i.e. it is allowed to undo virtually as many analysis steps (actions) as is needed. The capacity of this stack is only limited by the physical memory available in the system. Simple general display settings like inverted display, pseudocolour display, magnification operations, and others are not covered by the undo/redo mechanism, as they do not directly affect the analysis process. The undo/redo stack is obviously not saved in the session files.

5 Status report

After the pilot research project of the GASepo1 software had been evaluated by experts from the Doping Control laboratory in ARC Seibersdorf research GmbH, we proposed a project of the new GASepo software. The goal of this project was to develop a reliable and easy-to-use software package, specially designed for quantification of recombinant erythropoietin in urine sample as a
standardized and unified tool for the use across doping control laboratories worldwide. The interna-
tional project was accepted and supported by the WADA grant. Besides the developer, ARC Seibersdorf research GmbH, ten other DC laboratories have been involved into the project with the aim to test intermediate releases of GASepo. For this purpose a web site (www.antidoping.at/epo) for the GASepo project has been officially released in June 2004. This site serves as a basic platform of user feedback and communication between users and developers. A number of functions of the GASepo software have been motivated by the users from the mentioned DC laboratories.

The evaluation of the Epo images was originally based on simple integration of lane profiles. GASepo introduces the concept of segmented bands, i.e. bands identified as 2D objects in the lane using the mathematical procedure of image segmentation. The evaluation in GASepo is based on true volume and/or height computation of these objects. The advantage over the conventional approach is that the objects which are overlapping in 1D profile can still be totally separated as segmented objects. Moreover, having the full numerical description of bands opens possibility of extending Epo doping positivity criteria in the future.

In [20] we validated the col calculation procedure. The validation was carried out on a standard lane phantom particularly designed for this purpose. It consists of a sequence of bands with various mean intensity and smooth shape which are located in different positions within a lane. The bands are superimposed on a background image with various profiles. In our validation tests the position of the col has been investigated in dependence of various width of lane margins set by the user. On the basis of a number of tests, this operation was found crucial for proper calculation of the col position. We simulated five cases of Gaussian noise ($\sigma = 2300, 2600, 3100, 3600, 4500$ in 16 bit Epo image data). For the cases of noise with standard deviations characteristic for current Epo images ($\sigma = 3000$), the col position deviations do not exceed two pixels over the whole range of the lane width values.

32
Testing the method for the \textit{col} calculation on real 50 Epo images from different DC laboratories showed its robust behaviour in Epo images of medium quality (measured by the level of background inhomogeneity and number of artifacts present in gaps between individual lanes).

As a part of the WADA Educational Program 2004, two urine samples were distributed to 18 WADA accredited DC laboratories in 2004. For the detection of rEpo, performed by these laboratories, the method of Lasne et al. ([9]) was used, as the current reference method. All gel images were evaluated by the GASepo software for Epo positivity using the criteria of the WADA Technical Document TD2004EPO in its recent version (October 2004). The use of the GASepo software gave consistent results for 11 laboratories the samples of which were provided in acceptable form for evaluation. In every case, in which the sample had to be declared positive for containing rEpo according to the declaration of the provider of the sample, the software produced identical result.

\textbf{Epo doping positivity}

To demonstrate the capability of the GASepo software to evaluate final Epo doping positivity criterion according to the valid WADA criteria in a unique, accurate and clear way, we have documented in Fig. 17 two representative cases of Epo doping, positive and negative results of the doping control. First, the segmented bands of a fragment of the Epo image containing standard lanes and lanes with positive and negative samples are displayed. The profiles of the positive and negative lanes provide auxiliary information enabling comparison to the previous approaches of doping positivity evaluation. The values of the normalized volumes of the denoted bands in the tables (only their relevant parts are displayed) are ordered in decreasing order that makes possible to check the WADA criteria immediately.
Figure 17: Examples of positive and negative samples including the standard lanes; fragments of the tables with characteristics of the positive and negative lanes are displayed.
6 Lessons learned

Based on the theoretical and simulation research and using the feedback to testers from WADA accredited doping control laboratories, the GASepo software has been developed. The know-how behind this software comprises several types of novelties: (i) philosophy of analysis, (ii) mathematical solutions, (iii) image data processing algorithms, and (iv) graphical tools. We summarize the most important of them:

- original algorithm of background correction,
- replacement of 1D profile processing by 2D image processing for calculation of the position of the reference cut-off line that is based on a novel know how (proposals for two Austrian patents [21], [22] are pending),
- a novel methodology of coupled image object segmentation and classification,
- an algorithm of band segmentation which solves the problems of merged and disrupted band,
- original application of machine learning approach to supervised classification of segmented objects, in particular a fuzzy decision tree has been derived on which a simple band classifier has been based,
- a number of graphical tools designed for user interaction suitable for common image analysis or for solving problematic cases of analysis of heavily degraded Epo images.

7 Future plans

The first release of the GASepo software package was used during the Summer Olympic Games held in Athens in 2004. Another release was applied during the 10th IAAF World Championships
in Athletics in Helsinki in August 2005. The next deployment of the GASepo software is planned during the Winter Olympic Games in Torino in 2006. During the remaining time of the WADA project (till December 2006) update releases are planned twice a year.

In ARC Seibersdorf research GmbH another project associated with the GASepo development is being carried out. Its goal is to develop a compact workstation *epoCAM* for acquisition of chemiluminescent and fluorescent images. The system is characterized by very high sensitivity at low price. It is equipped with user-friendly calibration software for everyday practice. The GASepo software together with *epoCAM* system represent a unique system solution suitable for Epo doping control laboratories worldwide.

8 Acknowledgement

The project of the GASepo software has been carried out with the support of World Anti-Doping Agency.
Appendix 1 (Noise filtering using geometry-driven diffusion)

The first problem with the Epo images (both scanned from a film or acquired by a CCD camera) is the noise superimposed on the useful signal. The noise can significantly affect the evaluation, so it has to be removed. In GASepo, similarly as in the GelMaster software ([12], a special noise suppression filter based on geometry-driven diffusion (GDD) algorithm is used. In ([13]) we explored modifications of this nonlinear image filter, studied, e.g. in ([14], [15]). This filter is given by the equation

$$\frac{\partial I}{\partial t} = \text{div} \left[ c \left( |\nabla I(x, y, t)| \right) \nabla I(x, y, t) \right],$$

(1)

where $\nabla I(x, y, t)$ is the gradient of an image intensity function $I(x, y, t)$, the conductance $c(\cdot)$ is a function of spatial coordinates $x, y$ and $t$ is time which corresponds in discrete implementation to the index of the iteration step. The GDD filter outperforms other linear and nonlinear image noise filters, as proved in papers [13, 14, 15, 16]. Additive noise and impulse noise is removed equally well whilst the boundaries of the bands are not smoothed out.
Appendix 2 (Inhomogeneous background correction)

We based our background correction method on vertical margins of each lane. Let us define the rectangular margins (left and right) of a lane in such a way that they do not touch bands and contain a part of the empty space between the neighboring lanes (Fig. 18). Then, as depicted in Fig. 18 (the horizontal line shows a position of a plane cutting the lane in the i-th row), for each lane row we consider its intensity profile. For the parts of the profile corresponding to the left and right margins we calculate the mean intensity values. Thus, two points are obtained. The first point with x-coordinate identical to the x-coordinate of the starting profile point, and y-coordinate equal to the left margin mean. The second point is given by x-coordinate of the last profile point and y-coordinate equal to the right margin mean. By joining these two points we obtain intensity line for the i-th row that approximates background intensities in the i-th row. The lines for all the rows in the lane constitutes a correction surface for the background in this lane. For illustration of the shape of the background approximation, in Fig. 18 (right part), the intensity mean profile of the lane and the central column of the background correction surface are displayed.

Figure 18: Left: i-th row intensity profile with lane margins for background correction, Middle: the given lane with margins, Right: lane intensity profile and the background correction curve for the central lane column.
Appendix 3 (Cut-off line calculation)

In an rEpo standard lane we can assume that the set of bands is located in upper acidic part of the gel. Then we can consider two specific image blocks in the lane image: (i) the upper image block which bounds all bands as tightly as possible, and (ii) the complementary image block in the lane which contains only background. These two image blocks differ considerably in variability of intensities. In ideal case, the former block comprises a harmonic-like pattern of band intensities varying along vertical axis, while the latter one represents just homogeneous background. Therefore the neighboring line separating these two blocks is a good estimate of the col. For finding this separating line by an automatic procedure, we propose the following methodology:

1. a set \( \{A_k, B_k\} \) of all possible partitions of the lane into neighboring rectangular blocks \( A_k, B_k, k = 1, 2, \cdots, m \) (Fig. 19) is constructed,

2. we characterize each image block by a measure of intensity variability,

3. for every position of the separating line of two adjacent blocks we define two functions:
   (i) a function of vertical indices \( k \) given by the values of the intensity variability measure \( m_t(A_k) \) of the blocks \( A_1, A_2, \cdots, A_m \), and (ii) a function given by the values of the intensity variability measure \( m_b(B_k) \) of the blocks \( B_1, B_2, \cdots, B_m \),

4. finally, we define a difference function of these two functions and search for a maximum of this function which represents a pair of adjacent image blocks for which the difference between their intensity variability is the greatest; the position of the maximum determines the cut-off line.

We have evolved this methodology into the design of a particular algorithm that takes into account specificity of Epo images. We describe the steps of this algorithm.
To suppress the image noise a median profile $P_{med}(x)$ of the intensities of the lane is first calculated by finding medians of intensity values in each row $x$ of the lane image $I(x,y)$:

$$P_{med}(x) = \text{median}\{I(x,y) : y \in [c,d]\}, \quad (2)$$

(for denotation of the individual parameters see Fig. 19). To smooth irrelevant maxima we filter the 1D median profile by convolution with a mean filter kernel $\phi$:

$$P_{fil}(x) = P_{med}(x) \otimes \phi(x). \quad (3)$$

The choice of a convenient measure of the intensity variability represents the crucial point of the method. We propose to use the following differential measures characterizing individual image blocks, $m_t$ for the upper block and $m_b$ for the lower block:

$$m_t(A_k) = \frac{1}{a_k - a} \int_{a}^{a_k} \left| \frac{dP_{fil}(x)}{dx} \right| dx,$$

$$m_b(B_k) = \frac{1}{b - a_k} \int_{a_k}^{b} \left| \frac{dP_{fil}(x)}{dx} \right| dx. \quad (4)$$

Finally, the cut-off line position is defined as the index $K$ of the maximum difference

$$K = \arg\{ \max_k \{\text{Dif}(k)\}\} = \arg\{\max_k |m_t(A_k) - m_b(B_k)|\}. \quad (6)$$

Figure 19: The novel method of calculation of the cut-off line.
Appendix 4 (Band segmentation)

A compound segmentation method we developed ([23]) can be described as a sequence of operations specifically designed to solve the following crucial cases of band degradation in Epo images: (i) blurred bands, (ii) bands which are merged into one blob object, and (iii) disrupted bands which are represented by separate individual objects. The developed method is structured into several steps:

- application of the sequence of LoG filters with various window sizes and parameter $\sigma$ values,
- thresholding of the obtained images,
- subsequent operation of logical conjunction of the resulted binary images,
- region growing that includes the morphological operation of dilation,
- merging of the regions representing one band by the operation of projection.

**Threefold LoG filtering**

The rotationally symmetric filter Laplacian of Gaussian (LoG) is defined as:

$$\text{LoG}(x, y) = \left( 1 - \frac{x^2 + y^2}{2\sigma^2} \right) e^{-\frac{x^2 + y^2}{2\sigma^2}}. \quad (7)$$

Variable sizes of bands, different levels of their blurring (smearing) and presence of variable local artifacts in Epo images prevent from obtaining a satisfactory band segmentation by application of one single LoG operation with fixed window size and fixed value of the parameter $\sigma$. Based on computer experiments we proposed a triple of the LoG filters with the following window sizes: $w_1 = 15 \times 5, w_2 = 15 \times 15, w_3 = 25 \times 25$. For the segmentation of a lane image $I(x, y)$, this image is subsequently convolved by three LoG filters $L(x, y) = I(x, y) \otimes \text{LoG}(x, y)$.

**Thresholding of filtered data**

41
The second step of the proposed method consists in binarization of the filtered results. We define a binary mask $O$ as the thresholding

$$O(x, y) = \begin{cases} 
1 & \text{if } L(x, y) \leq \beta \\
0 & \text{otherwise}
\end{cases} .$$

(8)

The choice of the threshold value $\beta$ which could be fixed for a sufficient large set of Epo images is not a trivial task. The band boundary is defined as zero crossing positions of Laplacian of Gaussian. However, application of this conventional condition to our case results in a number of false bands on the background. Based on computer experiments, the very small, but nonzero values of the threshold $\beta$ have been found which yield better results of band segmentation.

**Combination of the binary images obtained after thresholding the LoG filter results**

Application of LoG filters with various values of parameter $\sigma$ and window sizes results in segmented objects (masks $O_i$) of different area. The overall tendency observed is towards getting the objects with larger area than is the area of the corresponding bands. To reflect the differences in band shape and in their intensity characteristics, we proposed to combine the results of segmentations by the operation of logical conjunction of the distinct masks $O_i$:

$$O(x, y) = \bigwedge_i O_i(x, y) .$$

(9)

Reducing the undersegmentation via morphological region growing

As usually, the choice of proper threshold is a question of trade-off between undersegmentation and oversegmentation. In our case, the problem of oversegmentation is a priori reduced, since by increasing the threshold values the increase of number of resulting segments is limited. Only the sizes of individual segments (bands) are decreasing. Therefore we propose to find an appropriate undersegmentation, i.e. smaller number of segmented objects (masks) with area greater than area
of the true bands. Then a segmentation into more objects with areas smaller than true bands is found. We proposed to combine these two operations using region growing via morphological operation. We applied dilation with the structure element given by $3 \times 3$ matrix of ones. The two-stage algorithm can be characterized as follows. First, some starting small regions of bands are found, which are subsequently grown into larger final masks. All regions of the final masks which do not contain any starting region can optionally be included into the final segmentation result.

*Merging of objects which represent one band*

In some cases the degradation of the gel image may cause that bands are represented in the segmented image by several separated objects. To resolve this particular problem, i.e. to ensure that the identical index is assigned to all objects belonging to one band, a special algorithm has been developed which we call *band projection*. We denote $P_i$ horizontal projection of every region mask $R_i$ in the lane. This mask is equal to 1 if the pixel $(x, y)$ belongs to the $i$–th region, otherwise it is equal to zero. Then the projection is defined as the sum

$$P_i(y) = \sum_x R_i(x, y).$$

Let $U_{i,j}$ denote the intersection (pixels) of two projections:

$$U_{i,j} = \{y|P_i(y) > 0 \land P_j(y) > 0\}.$$

Then a new combined mask $comb$ is defined for two segmented regions as a set of all pixels satisfying the following conditions for the overlapped projections:

$$comb =$$

$$\left\{(i, j) \mid \sum_{y \in U_{i,j}} P_i(y) > 0.5 \sum_{y} P_i(y) \land \sum_{y \in U_{i,j}} P_j(y) > 0.5 \sum_{y} P_j(y)\right\}.$$

43
Fuzzy decision tree generated by Machine Learning Framework

Since expert information on bands for typical Epo images had been available, we decided to use the concept of decision trees that is suitable for supervised classification tasks. It was popularized by Quinlan [24] in ID3 algorithm, which stands for *iterative dichotomizer 3*. It is characterized by using an information theoretic measure of entropy for assessing the discriminatory power of each attribute. ID3 is a popular decision making approach to classification of symbolic data. Since most real life problems deal with nonsymbolic, i.e. numerical data, they have to be discretized prior to selection of appropriate attributes. The fusion of fuzzy sets with decision trees enables one to combine the uncertainty handling and approximate reasoning capabilities of fuzzy sets with comprehensibility and ease of application of the ID3 decision tree generation. The geometrical and shape properties of segmented objects in Epo images can be described numerically, so that fuzzy ID3 decision tree is a suitable approach to be applied to our supervised classification task.

Regardless of specific measures used in the process of generation of fuzzy decision trees [25, 26, 27], this process involves two basic elements: (i) representation of the input parameter space by fuzzy sets, and (ii) generation of fuzzy predicates. In our case, the following five natural language expressions of quantity are associated with fuzzy sets: *very low (VL)*, *low (L)*, *medium (M)*, *high (H)*, *very high (VH)*. To find a segmentation of the input space according to the distribution of the sample data, the centers of the sets are first found by dividing the input range into five equally sized intervals. Then the centers of these sets are iteratively updated by using a modified $k$-means algorithm [28]. The fuzzy sets are then generated as trapezoidal membership functions around the centers computed previously. The set of all possible predicates describing a variable (quantity) is rather big. Therefore, it is necessary to reduce the number of descriptions that are analyzed to a manageable amount.
A fuzzy set $A$ induces a fuzzy predicate

$$t(x \text{ is } A) = \mu_A(x), \quad (14)$$

where $t$ is the function which assigns a truth value to the assertion “$x$ is $A$” (e.g. “$x$ is medium”). Other relevant fuzzy predicates associated with the fuzzy set $A$ are [26]:

$$t(x \text{ is not } A) = 1 - \mu_A(x), \quad (15)$$

$$t(x \text{ is at least } A) = \sup\{\mu_A(y) | y \leq x\}, \quad (16)$$

$$t(x \text{ is at most } A) = \sup\{\mu_A(y) | y \geq x\}. \quad (17)$$

The fuzzy ID3 method, theoretically described by Mitra et al [25] and Drobics et al [26], has been implemented in the specific software $MLF^{TM}$ (Machine Learning Framework: http://www.unisofwareplus.com/products/mlf/) developed as an add-on to $MATHEMATICA^{TM}$. This software package was selected to generate a fuzzy decision tree (FDT) on the basis of which a FDT-classifier has been designed.

**Application of FDT to segmented objects in Epo images**

The methodology of the development of a band classifier for Epo images ([29]) and its implementation into the GASepo system consists in the following stages:

1. Expert segmentation of Epo images and its use for generation of a training set (a subset of all input expert data).

2. Design of quantitative measures of properties of segmented objects in Epo images.

3. Input of the training set (as a text file) to MLF template program and generation of a fuzzy decision tree (FDT) with fuzzy predicates and accompanying values of fuzzy set parameters.


6. Testing the FDT classifier on a testing set (a subset of all input expert data).

Quantitative measures of properties of segmented objects

For characterization of geometrical and shape properties of the segmented objects, which are represented as labeled regions in each lane, we have chosen five common and three Epo image-specific measures:

*Relative Band Area, BandBox Ratio, Band Eccentricity, Band Orientation, Band Solidity, Band Centroid Eccentricity, Band Boundary Complexity, BandBox Fullness.*

These measures are related to binary objects, we did not apply any grey level measure of a band.

Fuzzy predicates

In Fig. 20 a fuzzy decision tree is depicted which has been derived by the MLF module. As an input of this module the training data was used which had been produced by an expert segmentation (classes of bands and artifacts were found) of 99 lanes selected from actual Epo images. Fuzzy predicates are positioned at individual nodes. The leaf predicates Classmemb_Is_0 and Classmemb_Is_1 stands for the statement that the given object belongs to the class of artifacts or bands, respectively. The branches denoted by T represent true value and those denoted by F represent false value. The numbers listed in the brackets denote the number of objects resolved at the given level. As can be seen in the graph, only four of the initial eight measures have been selected by the MLF procedure of fuzzy ID3 into the decision tree.

Based on the derived fuzzy decision tree six membership functions have been defined. By using the parameters of fuzzy sets provided in the output of the MLF template, the membership functions $mf_1, mf_2, ..., mf_6$ have been characterized numerically. The second and the fourth function are just
negations of the corresponding functions. As an example in Fig. 21 we display the membership function of the fuzzy predicate BBRatio\_IsAtLeast\_L together with the histogram of the values of the measure \textit{BBRatio} for all objects of the training set.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{bb_ratio_histogram}
\caption{Membership function for the fuzzy predicate: \textit{BBRatio\_IsAtLeast\_L}.}
\end{figure}

\textbf{Rules for FDT classification}

The rules given implicitly by the fuzzy decision tree can be expressed in the following form:

\begin{align*}
\text{\textit{rule1}} & \quad = \quad \text{If (BBRatio\_IsAtLeast\_L) then Classmemb\_Is\_0} \\
\text{\textit{rule2}} & \quad = \quad \text{If (non(BBRatio\_IsAtLeast\_L) and Eccentr\_Is\_VH) then Classmemb\_Is\_1}
\end{align*}
rule3 = If (non(BBRatio_IsAtLeast_L) and non(Eccentr_Is_VH) \\
    and CentEcc_IsAtLeast_M and non(Solidit_Is_VH)) \\
    then Classmemb_Is_0.

Using the operations of minimum for logical “and” and maximum for “or”, the rules can be ex-
pressed in another convenient form:

\[\text{bandmf} = \min\{mf2, mf3\}\]

\[\text{artemf} = \max\{mf1, \min\{mf2, mf4, mf5, mf6\}\}\]

The final decision on assigning the given object to the class of bands or artifacts is implemented
in the form:

If \(\text{bandmf} \geq \text{artemf}\) then object is BAND else object is ARTIFACT.

9 List of abbreviations

DP         doping control
WADA       World Anti-Doping Agency
GASepo     gel analysis system for epo
IEF        isoelectric focusing
DB         double blotting
ROI        region of interest in an image
GDD        geometry driven diffusion
References


[27] I. Hayashi, T. Maeda, A. Bastian, L. C. Jain, Generation of Fuzzy Decision Trees by fuzzy ID3 with adjusting mechanism of and/or
