A Probabilistic Approach to Organic Component Detection in Leishmania Infected Microscopy Images

Pedro A. Nogueira, Luís Filipe Teófilo

LIACC – Artificial Intelligence and Computer Science Lab., University of Porto, Portugal
FEUP – Faculty of Engineering, University of Porto – DEI, Portugal
{pedro.alves.nogueira, luis.teofilo}@fe.up.pt

Abstract. This paper proposes a fully automated method for annotating confocal microscopy images, through organic component detection and segmentation. The organic component detection is performed through adaptive segmentation using a two-level Otsu's Method. Two probabilistic classifiers then analyze the detected regions, as to how many components may constitute each one. The first of these employs rule-based reasoning centered on the decreasing harmonic patterns observed in the region area density functions. The second one consists of a Support Vector Machine trained with features derived from the log likelihood ratios of incrementally Gaussian mixture modeling detected regions. The final step pairs the identified cellular and parasitic components, computing the standard infection ratios on biomedical research. Results indicate the proposed method is able to perform the identification and annotation processes on par with expert human subjects, constituting a viable alternative to the traditional manual approach.

Keywords: Leishmania; Gaussian mixture models; Computer Vision, SVM.

1 Introduction

Leishmania is the parasite responsible for Leishmaniasis, a disease currently affecting over 12 million people throughout 88 countries [1]. Leishmaniasis is treatable by chemotherapeutics, which, nevertheless, suffer from poor administration regimens and high host toxicity [2]. Although the disease is not generally deadly, it severely damages the immune system, leaving the body exposed to other deadly pathogens, which often prove fatal [2]. The inadequate means to treat Leishmaniasis render the research for new treatments an urgent task.

Research in microscopy images produces large amount of data, which require anywhere from days to weeks to classify and annotate. In a single laboratory the number can easily ascend to thousands of images with merely a dozen different experiments. Not only does this detract the researchers from exploring new alternatives, as it also introduces inter-person variance, as many images are extremely cluttered and contain several hundreds of cells and parasites. This results in a time consuming and mentally straining process, which expresses itself as a decaying function over time as the sub-

© Springer-Verlag Berlin Heidelberg 2011

ject gets tired, frustrated or bored. These reasons justify the need for the development of automatic mechanisms to replace or aid researchers in the annotation task, for which and to the best of our knowledge no current solution exists. The proposed method provides a fully automatic pipeline for the identification of cells and parasites in Leishmania infected microscopy imaging, enabling more accurate annotations.

Pertaining this paper's organisation, it is structured as follows. Section 2 describes the main characteristics of fluorescence microscopy imaging, as well as the dataset used in this study. Section 3 discusses the state of the art in cell identification and segmentation in microscopy imaging. Section 4 briefly describes the proposed method, followed by the description of its steps. In section 5, the results for the segmentation and classifiers, as well as from the method's application to two real drug trials are presented. Finally, section 6, presents conclusions on the developed work, commenting on its performance and readiness for real-world applicability.

2 Fluorescence Microscopy Imaging

In contrast to the classical optical microscopy, the use of fluorescence microscopy allows simultaneous labeling of different cell components, which can be easily distinguished based on the fluorescence properties of their specific dyes [3]. The images collected for this study used three fluorophores [3], which emitted three distinct wavelengths. These corresponded to the cell nuclei DNA (in blue), cytoplasmic and nuclear DNA (in red) and the parasitic DNA (in green). This provided three separate sets of data per image (Figure 1), motivating the identification of cells, parasites and cytoplasm individually in the three channels as independent images.

Although very popular, fluorescence microscopy imaging (FMI) presents some well-known issues that also characterized our datasets. The most noticeable issues include: non-linear illumination (due to poor lighting conditions and sub-optimal experimental setup), photo bleaching, varying contrast, Gaussian noise, chromatic aberrations and overlapping cells and parasites (due to various focal planes).

In this study 794 fluorescence microscopy images from random drug trials with different experimental setups were collected and used. These images were collected through a light microscope and annotated manually by a Leishmania research team at the INEB/IBMC laboratory. Refer to section 5 for further details.

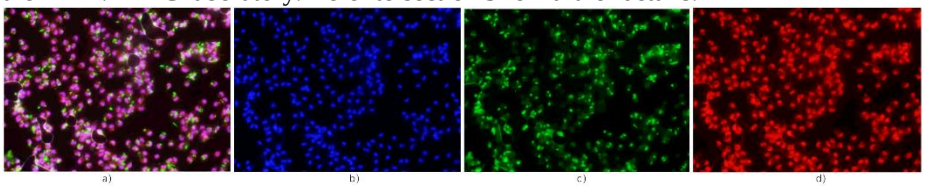

Figure 1. Details of a fluorescence microscopy image. a) Original image; b) Cell nuclei channel; c) Parasite nuclei channel; d) Cytoplasmic channel.

3 Related Work

Microscopy image analysis has been an active field for several decades. In related work, Liao et al. [4] used a simple thresholding method, coupled with mathematical morphology and contextual shape detection to detect white blood cells. However, their approach does not tolerate cells outside the defined conditions (e.g.: poorly segmented regions, forming a cell cluster region). An automated method for cellular membrane segmentation is described in [5]. This method also allows the reconstruction of unstained tracts through the nuclear membranes as a spatial reference. Jiang et al have also proposed white blood cell segmentation using scale-space filtering and watershed clustering in HSV color space [6].

Park has proposed bone marrow cell segmentation through an iteratively relaxed watershed algorithm [7]. However, this work is sensitive to illumination and noise conditions, since it overly relies on the fixed mean color values of each patch for the relaxation procedure. Begelman [8] performs cell nuclei segmentation using color, shape features and a fuzzy logic engine. This work is more robust than the aforementioned one because the extracted shape features, serve as an auxiliary classification input. However, it is still not able to account for non-circular geometries or abnormally colored cells due to the implemented rules' simplicity.

Yu proposes using an adaptive thresholding technique to detect cell nuclei, which are then expanded via level sets to determine cell boundaries [9]. Yan proposes a similar approach [11]. Yan improves on Yu by replacing the adaptive thresholding with a distance map of the initial adaptive histogram-based thresholding step. This distance map is then used to create a watershed transform, serving as a region list representing the level-set seed points. The only drawback to this approach is that it is not able to deal with highly cluttered images, as the distance map would not provide enough information to accurately parameterize the watershed transform, thus leading to an erroneous number/location of seed points for the level-set step.

From this review, it is clear that most of the literature does not attempt or is unable to deal with highly cluttered or overlapped image regions. This is a major concern in microscopy image analysis as the great majority of real-world data is heavily cluttered and saturated. The proposed method in this paper aims at addressing this issue.

4 Proposed Method

The proposed method focuses on developing robust methods for identifying and segmenting cellular and parasitic agglomerations in confocal microscopy images. The method starts by splitting the original image f in three channels: f_c (blue); f_p (green); f_{cyt} (red). Each channel is then normalized and segmented through Otsu's Method. This yields three region vectors, corresponding each one to cellular DNA, parasitic DNA and weak[1] DNA signatures (cytoplasm). Low-level features are computed for

[1] In order to register a weak DNA signature, this fluorophore must be highly sensitive. Thus, it also registers the cell nuclei's DNA. Since the cell nuclei can be trivially subtracted through

the first two aforementioned region vectors, and then used to train a rule-based classifier and a Support Vector Machine, both of which attempt to determine how many cells or parasites each region contains. To resolve disputes between these two classifiers, a voting system taking into account both of the classifiers' error margins is employed. Each region is then further segmented into the predicted number of subregions by Gaussian unmixing. Figure 2 can be inspected bellow for a more structured and clear understanding of the described pipeline.

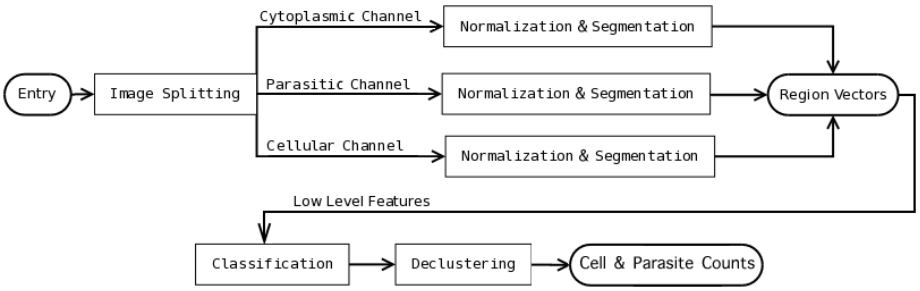

Figure 2. Developed method's architecture.

4.1 Pre-processing and Segmentation

The method first splits each of the target image's *f* color channels, as they are independently processed. Each image channel is then normalized and segmented into background and foreground components. An initial study on the general image characteristics was conducted in order to choose an appropriate segmentation technique. In this study, the intensity values of 120 randomly selected images presented clear bimodal distributions for all color channels (Figure 3).

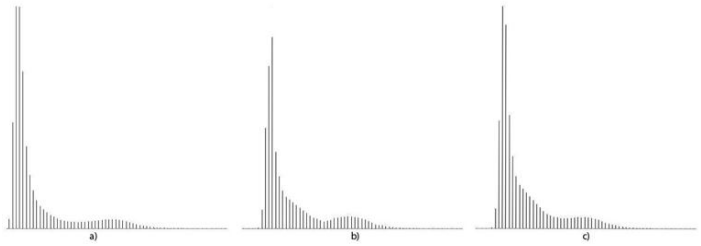

Figure 3. Bimodal distributions observed in the RGB components of 120 images (averaged). a) Red color component; b) Green color component; c) Blue color component.

Thus, Otsu's Method presented itself as a fitting approach due to its, low temporal and spatial complexity, non-parameterisable characteristics and segmentation principle. Otsu's Method's principle assumes a bi-modal distribution in the target dataset,

---

set operations involving the cell channel, we denominated this channel as the cytoplasmic channel.

for which it attempts to determine an optimal threshold value by minimizing intra-class variance [12]:

$$\sigma_b^2(t) = \sigma^2 - \sigma_\omega^2(t) = \omega_1(t)\,\omega_2(t)\,[\mu_1(t) - \mu_2(t)]^2 \quad (1)$$

Thus, each color channel was binarised using Otsu's Method and then proceeded to a connected component analysis, resulting in a region vector representing the cellular, parasitic and cytoplasmic regions present in the image. Note that the cytoplasmic regions are not used in this work, as they are intended for associating cell-parasite pairs. They were, however, computed and integrated into the method in hindsight of future work. Figure 4 depicts the result of the segmentation step for the cellular channel.

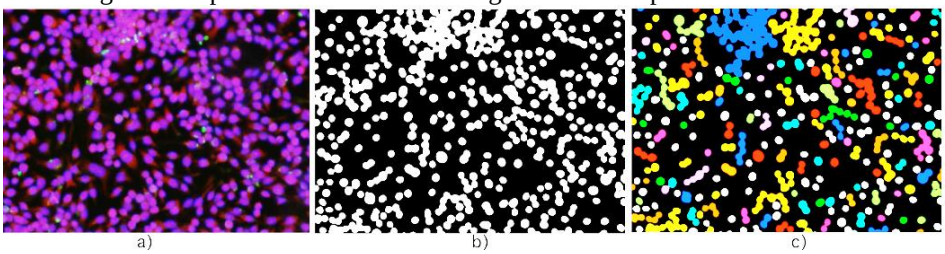

Figure 4. Segmentation output of a moderately cluttered image. a) Original image; b) Segmentation output; c) Visual representation of the cellular region vector obtained from the connected component analysis (randomly color-coded).

Following the region extraction, a low-level feature vector $F_i=[a,\ ll_{1..N},\ d^1_{1..N}(ll),\ d^2_{1..N}(ll)]$, is computed for each region. The features comprising the vector are: a, the area value (in pixels); ll, the log-likelihood ratios for modelling the region with 1 to N Gaussian mixtures; $d^1(ll)$, the first discrete derivate of ll and $d^2(ll)$, the second discrete derivate of ll. The area feature was used to define the rule-based classifier. The log-likelihood ratios and their derivates were used to train the SVM classifier.

4.2 Classification

The classification step is based on the assumption that the regions obtained from the segmentation process may not always correspond to a single cell or parasite. Based on this, it attempts to discern in how many sub-regions each region must be split. This is achieved by employing two separate classifiers: a rule-based classifier (RBC) and a support vector machine (SVM).

The RBC exploits the low area overlap percentage observed between cell and parasite pairs. This low overlap percentage is due to the high depth of field observed in the collected images, resulting in a near-perfect 2D cross-section of the 3D space within the tissue sample. Following this principle, and as single cells/parasites presented normal distributions, it was hypothesised that the area functions for cells and parasites could be approximated by a harmonic function. Since larger multi-nucleic regions are increasingly less frequent, the functions present a decreasing harmonic pattern. Although simple, this approach was found to be quite accurate (around 89.0% accuracy,

hitting a maximum value of 98,9% when considering an error of ±2 regions as acceptable). Figure 5 illustrates a detail for one of the described decreasing harmonic functions.

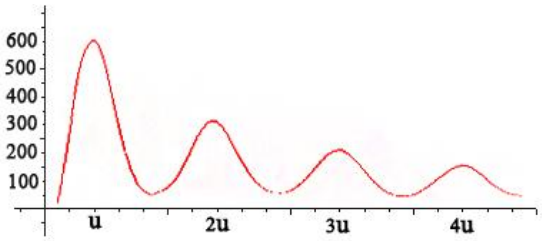

Figure 5. Detail of a decreasing harmonic function described by the normalized (cellular) area values of 120 randomly sampled images (~870 regions per image: 106.318 region samples total). Horizontal axis: area values; Vertical axis: area value occurrences.

The classifier was programmed with rules reflecting this concept and taught to ignore values outside its knowledge space, as there was no data to accurately model regions over 10 cells or parasites. Note that cells and parasites each have their own harmonic function, since their area distributions do not exhibit the same standard deviation.

The SVM classifier relies on the concept that circles and ellipses can be described as Gaussian distributions and, as macrophages and parasites partake such geometry, clusters of these objects can be formulated as a mixture problem. Our conjectured hypothesis was that: as mixtures are added to the modeling process of each region, the improvement rate is described in the log-likelihood ratio evolution sequence. Thus, if an initial annotated dataset with the correct number of mixtures *N* is available, it should be possible to model a function that is able to predict this *N* for new, non-annotated observations. Following this hypothesis, the classifier was trained with a subset of features from the main feature vector, consisting of the log-likelihood ratios, it's first and second order discrete derivates. The training set was obtained from roughly 150 regions, modeled with N = [1..15] Gaussian mixtures. Various machine-learning classifiers were tested. Ultimately, a SVM model was chosen as it achieved the highest (85,3%) sequential split and cross-validation classification rates.

In order to reconcile diverging predictions, a voting system was developed. As previously mentioned, the RBC exhibits an overall accuracy of nearly 98%, when considering an error margin of ±2. The SVM is generally much more accurate[2], but its error margins are, in counterpart, much wider. Making use of these characteristics, the voting system makes its choice based on the assumption that the correct decision does not deviate more than 2 from the RBC, so if the SVM's prediction is within this window, it is considered correct and incorrect if otherwise.

[2] Note that the 85% accuracy percentage referred for the SVM classifier refers to the classification of only multi-nucleic regions, as the accuracy ratings for the RBC refer to multi and uni-nucleic regions, giving it a considerable advantage.

4.3 Declustering

Having obtained a prediction for the number of nuclei in each region, these were unmixed using the Expectation-Maximization [13] method. The algorithm was parameterized with a minimum standard deviation of 10^{-6}, and a maximum of 200 iterations. To minimize runtime, the seeds for the centroids of each mixture were set by the averaged centroids of a 10 fold cross-validated K-Means. Figure 6, portrays the method's expert performance, even in the presence of large nucleic clusters.

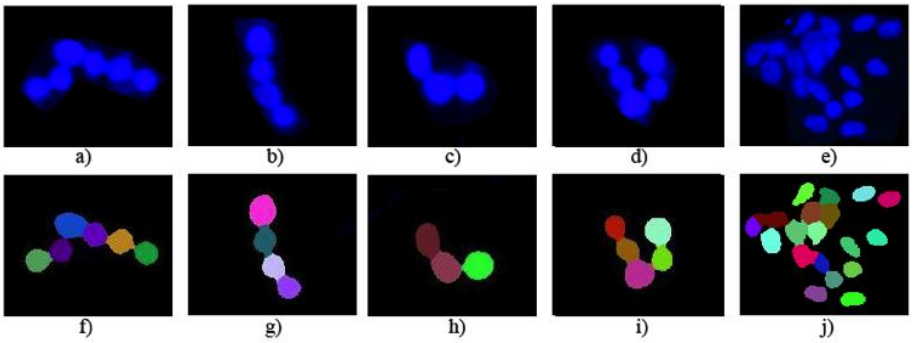

Figure 6. Several *declustering* examples for 6, 4, 3, 5 and 11 region clusters. a) - e) Original region patches. f) - j) Respective *declustered* patches.

5 Results

In order to evaluate the practical applicability of our method, we choose to compare the method's classification output with the final annotations made by biomedical researchers in real drug trials. Upon surveying the current drug trials undergoing in the IBMC lab two specific drug trails were chosen. These were labeled trial 1 (T_1) and trial 2 (T_2). Trial one was chosen for presenting a low number of regions, most of which were difficultly differentiated from the background, thus straining the method's segmentation step. Trial two was chosen due to the sheer number of oversaturated and overlapping regions, hence straining mostly the classification step. These two trials were held as complementary, therefore constituting a complete test of the foreseeable experimental conditions in real-world applications. Our ground truth was taken as the individual annotations of three biomedical researchers. The researchers were asked to carefully perform the annotations in separate days and double-check them, so as to minimize human error.

5.1 Individual Component Results

In order to understand the method's general behavior across experimental conditions we measured its segmentation and classification accuracy in both trials. The following section details the summary ratings obtained.

Regarding the classification step, since the method assumes that the cell/parasite identification process is not completely performed in the segmentation step, multi-nucleic regions were considered as well segmented results. A region was considered ill segmented if: a) it was not detected or b) its geometry was not correctly identified. Figure 7 exemplifies these criteria.

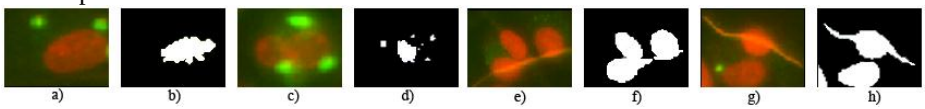

Figure 7. Examples of ill-segmented regions.

Table 1 presents a summary of the obtained segmentation accuracy on both trials.

Table 1. Accuracy ratings for the segmentation step of both datasets.

	Macrophages	Parasites
Segmentation total (T_1)	4873	6113
Ground truth (T_1)	5007	6437
Accuracy percentage (T_1)	97,32	94,96
Segmentation total (T_2)	7633	2571
Ground truth (T_2)	8014	2783
Accuracy percentage (T_2)	95,24	92,38

Regarding the rule-based classifier, it showed itself capable of identifying regions with 98.34% accuracy, when considering a ±2 error margin acceptable. This result was obtained by manually inspecting 1500 distinct regions from both datasets. No distinction was made between cellular or parasitic regions.

To train the SVM classifier, the log-likelihood ratios and their first and second order derivates were computed from mixture modeling 150 random clustered regions. The SVM classifier was trained using John Plat's sequential minimal optimization algorithm and an RBF kernel [14]. Validation was performed through a 66% sequential split (SS), for which the classifier obtained an 85.3% classification accuracy.

5.2 Stress Testing

In order to assert our method's real-world applicability, the computed infection ratios were compared with each of the three manual annotations. To account for inter-person variation, annotation values were modeled as a normal distribution and the method considered accurate if its output did not deviate more than 2 standard deviations from the mean value. Intra person variance was eliminated from this test, as each subject was instructed to carefully perform the annotations in a single pass. Table 2 presents the acceptable error margins for both trials.

Table 2. Stress-test error margins for both trials.

	Cells (total)	Parasites (total)	Infected Cells
Annotation Mean (T_1)	3020	4037	1873
Annotation Standard Deviation (T_1)	885	1110	546
Algorithm Error (T_1)	1353	1574	947
Annotation Mean (T_2)	5069	1967	1024
Annotation Standard Deviation (T_2)	294	34	38
Algorithm Error (T_2)	223	133	28

In trial one we observe a large standard deviation, both in the cellular and parasitic counts, which also has a bleeding effect into the infect cell count. Due to these large discrepancies, the method easily fits within the defined boundaries, actually being closer to one standard deviation in total parasites and infected cells. Hence, the method performs in a manner suitable for real-world application for images exhibiting similar experimental conditions as trial one.

Regarding trial two, the manual annotation discrepancies seen in trial one are no longer present, thus contributing to the low standard deviation verified. This further increased the error observed in the segmentation results. Although the algorithm error in total cells and parasites for trial two is not considerable in absolute numbers, it is in relative distance to the standard deviation, which translates to an error of over 3 standard deviations for the total cell count.

In sum, the method passed all tests, except the parasite detection category in trial two. Since this trial presented little to no multi-nucleic regions, no fault could be attributed to the classification step. Thus, the low parasitic detection (and ensuing error) falls upon the segmentation process, indicating future improvements should be directed at this step. The results indicate our method is robust to highly cluttered images, being able to expertly split region clusters and compute infection ratios.

6 Conclusions

In this work a robust, automatic analysis methodology for cell and parasite detection in fluorescence microscopy imaging was suggested. The proposed method has shown itself robust to poor lighting conditions and high cluttering indexes, falling well with-in the error margins observed in expert biomedical researcher annotations.

The obtained results demonstrate the method is capable of performing the image analysis task adequately and in less time than a human expert. Being a computer program, the method also boasts from being immune to traditional human errors related to distraction, fatigue or subjectivity. Since human errors are the major source of ambiguity in the traditional annotation process, we consider our alternative to be a more suitable choice. This claim is supported with the fact that the method has a fixed error margin; meaning, the error does not randomly vary through time, as human error does. The attained results in both stress tests further support our claim, proving the method's suitability for this specific task. As an added benefit, using the proposed method two or more drug trials can be safely compared as to their effectiveness, whereas if consider-

ing human error, the comparison would require validation via multiple annotations to attenuate the uncertainty generated by inter and intra-person variance.
Future work should focus on the methods used in the segmentation step, possibly employing mean shift or normalized cuts techniques, as well as increasing the training datasets of both classifiers. The built processing pipeline was made to be modular and applicable to other image types, thereby easily expandable to solve similar problems. This work has been successfully integrated with a pre-existing image annotation framework and is currently used in the INEB/IBMC laboratories in Portugal.

7 References

1. Ryan K. J., Ray C. G.: Sherris Med. Microbio. McGraw Hill. 749–754, (2004).
2. Myler P., Fasel N.: Leishmania: After The Genome. Caister Academic Press. (2008).
3. Spring K. R.: MicroscopyU: Introduction to Fluorescence Microscopy. (2010).
4. Liao Q., Deng Y.: An Accurate Segmentation Method For White Blood Cell Images. Proceedings IEEE International Symposium on Biomedical Imaging. 245-248, (2002).
5. Ficarra E., Cataldo S. D., Acquaviva A., Macii E.: Automated Segmentation of Cells With IHC Membrane Staining. IEEE Transactions on Biomedical Engineering, 58-5, 1421-1429, (2011).
6. Jiang K., Liao Q., Dai S.: A Novel White Blood Cell Segmentation Scheme Using Scale-Space Filtering And Watershed Clustering. Proceedings Second International Conference on Machine Learning and Cybernetics. (2003).
7. Park J. and Keller J. M.: Fuzzy Patch Label Relaxation in Bone Marrow Cell Segmentation. IEEE International Conference on Computational Cybernetics and Simulation, 1133–1138, (1997).
8. Begelman G., Gur E., Rivlin E., Rudzsky M., Zalevsky Z.: Cell Nuclei Segmentation Using Fuzzy Logic Engine. Proceedings IEEE International Conference on Image Processing. (2004).
9. Yu W., Lee H. K., Hariharan S., Bu W., Ahmed S.: Level Set Segmentation of Cellular Images Based on Topological Dependence. Proceedings of the 4th International Symposium on Advances in Visual Computing. (2008).
10. Yan P., Zhou X., Shah M., and Wong S. T. C.: Automatic Segmentation of High-Throughput RNAi Fluorescent Cellular Images. IEEE Transaction On Information Technology In Biomedicine, 12-1, (2008).
11. Morse B. S.: Brigham Young University, SH&B, Section 5. (2000).
12. Freeman H.: On the encoding of arbitrary geometric configurations. IRE Transactions on Electronic Computers, 260-268. (1961).
13. Platt J.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf, C. Burges, and A. Smola, eds., MIT Press. (1998).
14. Reynolds D.: Gaussian Mixture Models. MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA. (2007).