Quantifying the proportion of damaged sperm cells based on image analysis and neural networks

Conference Paper · September 2008

CITATIONS
8

READS
323

4 authors:

Rocío Alaiz
Universidad de León
53 PUBLICATIONS **1,111** CITATIONS
SEE PROFILE

Enrique Alegre
Universidad de León
208 PUBLICATIONS **1,577** CITATIONS
SEE PROFILE

Víctor González-Castro
Universidad de León
83 PUBLICATIONS **852** CITATIONS
SEE PROFILE

Lidia Sánchez
Universidad de León
92 PUBLICATIONS **407** CITATIONS
SEE PROFILE

Quantifying the Proportion of Damaged Sperm Cells based on Image Analysis and Neural Networks

R. Alaiz-Rodríguez, E. Alegre-Gutiérrez, V. González-Castro and L. Sánchez
University of León
School of Industrial, Computing and Aeronautic Engineering
Campus de Vegazana s/n, 24071 León.
SPAIN

Abstract: Insemination techniques in the veterinary field demand more objective methods to control the quality of sperm samples. In particular, different factors may damage a number of sperm cells that is difficult to predict in advance. This paper addresses the problem of quantifying the proportion of damaged/intact sperm cells in a given sample based on computer vision techniques and supervised learning. Unlike common supervised classification approaches, neither the individual example classification is important nor the class distribution assumed in learning can be considered stationary. To deal with this, an estimation process based on Posterior Probability estimates (PP), and known to increase the classifier accuracy under changes in class distributions, is assessed here for quantification purposes. It is compared with two approaches based on the classifier confusion matrix (Adjusted Count and Median Sweep) and the naive approach based on classifying and counting. Experimental results with boar sperm samples and back propagation neural networks show that the PP based quantification outperforms any of the previously considered approaches in terms of the Mean Absolute Error, Kullback Leibler divergence and Mean Relative Error. Moreover, in spite of an imperfect classification, the PP approach guarantees a uniform Mean Absolute Error (around 3%) for whatever combination of training and test class distributions, what is very promising in this practical field.

Key–Words: Class distribution estimation, Quantification methods, Neural Networks, Semen image analysis.

1 Introduction

The evaluation of the semen viability is an important challenge in fertility medical research. Some computer assisted approaches have been proposed to assess the quality of semen samples [13, 8] in terms of spermatozoon motility, morphology, hyperactivation or concentration. Another important aspect to consider is the sperm membrane integrity. It is known that a high percentage of spermatozoa with an intact membrane (not capacitated) becomes crucial for fertilizing purposes. The reason is that sperm cells that are already capacitated prior to the insemination are likely to be of reduced longevity and therefore, of questionable use for fertilization in vivo.[11]. However, determining the cell class proportions in terms of its membrane (intact and damaged acrosome) is usually carried out manually using stains because the commercial computer tools do not provide this feature. The stains, though, have several drawbacks like its high cost in terms of time and required equipments and the need of specialized veterinarian staff. Therefore, developing a method to automate this task would be an asset in this field.

Several recent studies have followed this research line [9, 7], focusing on improving the classification accuracy on individual cells by extracting better features from the images and optimizing the classifier. It is assumed that once the classifier is designed, it is applied as-is to a sperm sample with unknown class distribution of damaged/intact sperm cells. Then, the proportion of damaged spermatozoa can be estimated as the fraction of the instances classified as damaged. This has been referred as *Classify & Count* (CC) by Forman [5, 6] in the context of news classification.

However, a key aspect that is somehow overlooked in previous studies is the fundamental assumption made in supervised learning: training and test data are expected to follow the same, although unknown, distribution [4]. In particular, prior class probabilities estimated from the training data set are considered to truly reflect the target class distribution. This assumption, though, does not always hold in

[1]This work has been partially supported by the research project DPI2006-02550 from the Spanish Ministry of Education and Science.

practice. Unfortunately, time or space class stationarity can not be assumed in this practical field: factors like the animal/farm variability or the manipulation and conservation conditions make the class distribution imprecise. Indeed, the ultimate goal of the application is to determine the class distribution of an unlabeled test set what is referred as *quantification* in [6]. Therefore, if the class distribution in the present sample differs from the assumed in training, the classifier will be suboptimal. It is worth to highlight that according to the veterinary experts only samples with a proportion of damaged cells lower than 20% have practical interest and that defines the uncertainty region we will focus on.

Recently, some approaches have been proposed to address the mismatch between training and future (real) class distribution in order to prevent a significant drop in classifier performance. Some works rely on an eventual perfect knowledge of the new conditions by the end user [3]. On the other hand, Saerens et al. [10] propose a re-estimation process of the new conditions as long as the classifier provides estimates of the posterior probabilities of class membership and an unlabeled data set is available. Finally, other proposals deal with this uncertainty from a robust minimax approach [1]. However, all of them have been originally proposed in order to minimize classifier error rate or risk (by means of adjusting the classification threshold, the classifier outputs or the learning process) but with no practical interest in the estimation itself of the new class distribution.

The goal in seminal quality control applications is not directly to maximize the overall cell classification accuracy but reliably estimate the class proportion of damaged/intact cells with no concerns about the individual classification. Up to our knowledge, only Forman [5, 6] has tackled this issue in the context of the news trend and text classification. In his studies, methods based on the classification confusion matrix (*Adjusted Count*(AC) and *Median Sweep*(MS)) are found to outperform other approaches based on modeling distributions.

On the other hand, Saerens et al. [10] found that in order to improve the classifier accuracy by readjusting the classifier outputs for the new conditions, methods based on the classifier posterior probability estimates outperform those that rely on the confusion matrix. For this reason, this paper addresses the quantification problem in the veterinary insemination field by exploring this latter method (based on posterior probability estimates). We will compare it with the CC, AC and MS methods, and show that it outperforms the techniques based on the confusion matrix and gives a reasonable quantification in our practical problem with boar sperm images and using a back-propagation neural network as classifier. The rest of this paper is organized as follows: Sect. 2 covers the class distribution methods assessed in this work. Experimental results are shown in Sect. 3 and Sect. 4 summarizes the main conclusions.

2 Class Distribution Estimation

Consider a binary classification problem with a training labeled data set $S_t = \{(\mathbf{x}^k, d^k), k = 1, \ldots, K\}$ where $\mathbf{x}^k$ is a feature vector, d^k is the class label with $d \in \{0, 1\}$[1] and $\widehat{Q}_i = P\{d = i\}$ is the class prior probability estimated from S_t. Let us also consider a classifier that makes decisions in two steps: it first computes a soft output $\widehat{y}^k$ and based on it, makes the final hard decision $\widehat{d}^k \in \{0, 1\}$.

Once the classifier is trained, consider now an unlabeled data set $S = \{(\mathbf{x}^l), l = 1, \ldots, N\}$ with unknown class distribution P_i and the decisions y^l and $\widehat{d}^l$ provided by the classifier for each instance in S.

The naive approach to estimate the actual class distribution is based on counting the labels assigned by the classifier (CC method). It is considered here as a baseline for comparison purposes. Next, some techniques based on the classifier confusion matrix and another one based on the posterior probability estimations provided by the classifier are briefly described.

2.1 Estimation of priors based on the Confusion Matrices

Classifier performance can be summarized by its confusion matrix, from which the following rates can be computed for a binary problem:

- True Negative rate: $TN = P\{\widehat{d} = 0 | d = 0\}$
- False Negative rate: $FN = P\{\widehat{d} = 0 | d = 1\}$
- True Positive rate: $TP = P\{\widehat{d} = 1 | d = 1\}$
- False Positive rate: $FP = P\{\widehat{d} = 1 | d = 0\}$

Methods relying on the confusion matrix (estimated from training data and cross-validation procedures) have been proposed in [10, 5]. Basically, they are employed in order to estimate the new class prior probabilities by solving the following system of two (in a binary case) linear equations with respect to $\widehat{P}_i$.

$$\widehat{P}\{\widehat{d} = i\} = \sum_{j=0}^{1} \widehat{P}\{\widehat{d} = i | d = j\} \widehat{P}_j, \quad i = 0, 1 \tag{1}$$

[1] Damaged class will be denoted as class-1 or positive class. Intact class will also be called class-0 or negative class.

where $\widehat{P}_i$ is an estimation of P_i and $\widehat{P}\{\widehat{d}=i\}$ is the observed class probability by looking at the classifier labels $\widehat{d}$. In order to keep probability values in the range $[0,1]$, the solution of system (1) must be clipped as suggested in [5]. Following [5], we will refer to this method as Adjusted Count (AC).

Based on AC, Forman [6] also proposes the Median Sweep (MS) method. Briefly, it can be described as follows: first, several confusion matrices are computed for different classification thresholds; Then, the AC method is applied for each matrix and finally, the class distribution estimation is computed as the median of the estimations derived from each confusion matrix.

2.2 Estimation of priors based on posterior probabilities

Given a model whose outputs y_i provide estimates of posterior probabilities, Saerens et al. [10] proposes an iterative procedure based on the EM algorithm in order to adjust the classifier outputs for the new deployment conditions without re-training the classifier. This is carried out by indirectly computing the new class prior probabilities, what is the goal in our work. Here, we briefly present the algorithm and we refer the interested reader to [10] for more details. The new class prior and a posteriori probabilities are initialized with the values assumed in training and the outputs given by the model, respectively, as

$$\widehat{P}_i^{(0)} = \widehat{Q}_i \tag{2}$$

$$\widehat{P}^{(0)}\{d=i|\mathbf{x}^l\} = y_i^l \tag{3}$$

Estimation of the class prior probabilities at iteration k is given by (4) and the adjusted classifier outputs by (5). Note that N is the number of instances in the new unlabeled data set.

$$\widehat{P}_i^{(k)} = \frac{1}{N}\sum_{l=1}^{N}\widehat{P}^{(k-1)}\{d=i|\mathbf{x}^l\} \tag{4}$$

$$\widehat{P}^{(k)}\{d=i|\mathbf{x}^l\} = \frac{\frac{\widehat{P}_i^{(k)}}{\widehat{P}_i^{(k-1)}}\widehat{P}^{(k-1)}\{d=i|\mathbf{x}^l\}}{\sum_{j=0}^{1}\frac{\widehat{P}_j^{(k)}}{\widehat{P}_j^{(k-1)}}\widehat{P}^{(k-1)}\{d=j|\mathbf{x}^l\}} \; . \tag{5}$$

We will refer to this iterative approach as the Posterior Probability method (PP).

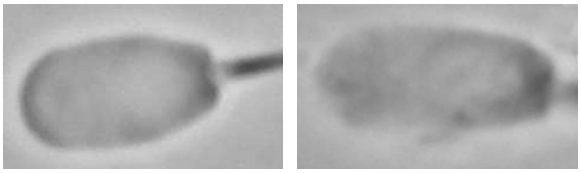


Figure 1: Examples of grey level images of acrosome-intact (left) and acrosome-damaged (right) boar spermatozoa.

3 Experimental Results

In this section we assess the performance of two quantifying approaches that rely on the confusion matrix (AC, MS), another based on the posterior probability estimates (PP) and the baseline approach CC.

3.1 Sperm cell data set

Experiments are carried out with boar sperm samples. We use an image data set with 393 instances (166 damaged and 227 intact spermatozoon heads) provided by the Faculty of Veterinary Sciences of the University of Leon. An example of these two acrosome states are shown in Fig. 1.

Texture features are extracted by means of the Discrete Wavelet Transform (DWT). In particular, 20 features for each head image are derived from the co-occurrence matrix using the Wavelet Co-occurrence Features (WCF) [12]. For further details of the data set itself and the image processing and feature extraction process we refer the interested reader to [7].

3.2 Neural network Classifier

Classification stage is carried out by means of a back-propagation Neural Network with one hidden layer and a logistic sigmoid transfer function for the hidden and the output layer. Learning was carried out with a momentum and adaptive learning rate algorithm. It is well known that classifier outputs provide estimates of class posterior probabilities when training is carried out minimizing some loss functions such as the mean square error used in this work [2].

Data were normalized with zero mean and standard deviation equal to one. The neural network architecture as well as the number of training cycles were determined by 10-fold cross validation. A two node hidden layer network with 200 training cycles lead to the optimal configuration evaluated in terms of the overall misclassification rate (5.6% error rate estimated by 10 fold-cross validation on the whole set).

3.3 Performance metrics

The mismatch between the real class distribution and the estimation provided by the different approaches assessed in this work is measured by means of the Mean Absolute Error (MAE), the Mean Relative Error (MRE) and the Kullback-Leibler divergence (KL).

The MAE metric focuses on the class of interest (class-1, positive or damaged cell class) and is defined as the absolute value of the difference between its actual prior probability and the estimated one:

$$MAE(\mathbf{P}, \widehat{\mathbf{P}}) = |P_1 - \widehat{P_1}| \quad (6)$$

The drawback of this metric is that does not allow us to evaluate the importance of the error. That means this metric makes impossible to distinguish between a 3% error produced for an estimated distribution of 20% when the actual distribution is 23% and the same error obtained for an estimated distribution of 0% when the real distribution is 3%. The MRE metric includes such information and is defined as follows:

$$MRE(\mathbf{P}, \widehat{\mathbf{P}}) = \frac{|P_1 - \widehat{P_1}|}{P_1} \quad (7)$$

The KL metric [2] also covers the error relative importance and it is given by

$$D_{KL}(\mathbf{P}||\widehat{\mathbf{P}}) = P_0 \log \frac{P_0}{\widehat{P_0}} + P_1 \log \frac{P_1}{\widehat{P_1}} \quad (8)$$

where $\mathbf{P}$ and $\widehat{\mathbf{P}}$ are the real and estimated probabilistic vectors, respectively. Note that this metric takes the value zero when there is no mismatch between the predicted and the actual class distribution.

3.4 Quantification of the damaged sperm cells

In order to evaluate the four quantification methods (CC, AC, MS, PP) with the neural-based learning machine described in Section 3.2, the following experiment was designed.

The aim of this experiment is to evaluate the classifier for different combinations of training and test distributions, whereas the same set size (160 instances for training and 60 for test) is fixed across the experiments. The training set is formed by 160 examples, varying the percentage of the images from the class-1 from 50% to 5% with steps of 5%. The class-1 distribution in the test set focuses on the region of interest and goes from 5% to 20% by 5% steps (higher proportions have no practical interest). This makes a total of 40 different combinations of training and test scenarios. For each particular training/test combination, results are the average of 30 randomly training sets extracted from the whole data set and for each training set, 20 test sets were randomly selected among the examples that were not considered in training. The confusion matrices required for the AC and MS methods were estimated exclusively from the training data set by k-fold cross-validation (with k=50 as in [6]).

Performance metrics averaged over the 40 scenarios considered in the experiments are gathered in Table 1. We can see that PP method achieves on average the lowest MAE (3.15% against 7.92% for MS or 4.01% for AC) keeping a low divergence (0.009 against 0.015 for CC and 0.015 for both AC and MS). In terms of MRE, the baseline CC method leads to the highest score (0.42), whereas the AC/MS methods reduce it to 0.38 and the PP method achieves the lowest error (0.32). Next, a detailed analysis is carried out.

	Method			
	CC	**AC**	**MS**	**PP**
MAE	4.00	4.01	7.92	**3.15**
MRE	0.42	0.38	0.38	**0.32**
KL	0.014	0.015	0.015	**0.009**

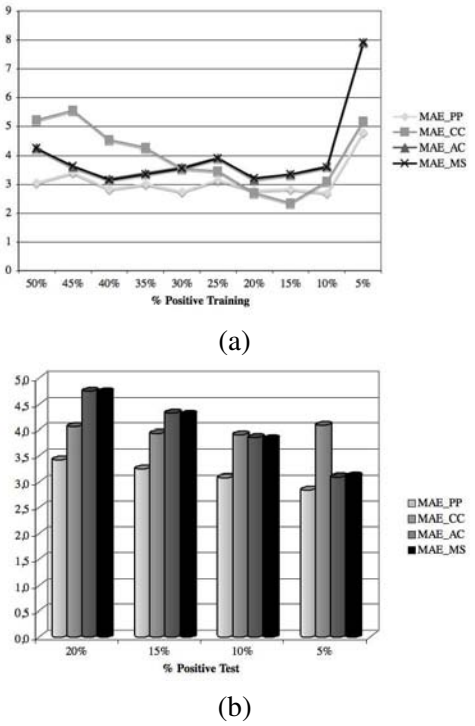
Table 1: Performance metrics averaged over 40 scenarios with different training and test class distributions.

Fig.2 depicts the MAE across the different training and test conditions. Unlike [5], note that no significant differences are observed between AC and MS. This may be due to an estimation of the confusion matrices for the threshold used in AC as good as the median value used in MS. In the following we will only refer to AC since MS and AC show similar performance.

Fig.2(a) shows the MAE for different training scenarios averaged for the four testing conditions evaluated (5%, 10%, 15%, 20%). It can be seen that the quantification performance for all methods degrades when training is carried out with the most imbalanced data set (5%). This suggests that for quantification tasks, even though we expect the test class distribution to be imbalanced, we should not look for a match between training and test class proportions (a more balanced distribution will provide better estimations). Note also that the PP method is quite robust against the class distribution used in training (MAE around 3%), whereas the CC method and AC are more dependent on the training class distribution and in general, show higher MAE than the PP method.

When the analysis is carried out for a fixed test distribution, Fig.2(b) depicts the MAE (averaged for

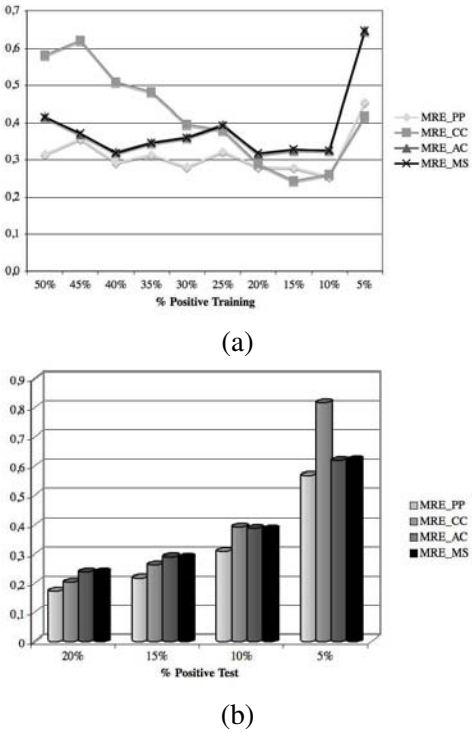
[2] This metric is referred as normalized Cross Entropy in [5]



(a)

(b)

Figure 2: Mean absolute errors (MAE) for the estimation methods CC, AC, MS and PP. (a) MAE for ten different training scenarios scenarios. (b) MAE for four test scenarios.



(a)

(b)

Figure 3: Mean Relative Errors (MRE) for the estimation methods CC, AC, MS and PP. (a) MRE for ten different training scenarios scenarios. (b) MRE for four test scenarios.

the ten different training conditions). Although the relative performance between CC and AC vary for different test scenarios (CC is better than AC when the test class-1 distribution is 20%, while AC outperforms CC for the scenario with 5%), the MAE for the PP method is always the lowest (around an absolute deviation in the estimation of 3%) and quite stable across different scenarios.

As the MAE does not provide information about the relative deviation in the estimation, we analyze the MRE. We can see in Fig. 3(b) that relative errors get higher as the positive percentage in the test set decreases. It can be seen that the CC method shows a great variability: from 0.20 to 0.82 depending on the scenario, what makes it not suitable for these applications. The AC and PP quantification techniques have the same tendency, but the PP method always outperforms the other one (0.17 against 0.24; 0.22 against 0.29; 0.31 against 0.39, 0.57 against 0.62).

Regarding the KL divergence, the AC and PP techniques show similar relative performance (see Fig. 4(b)). The PP method always outperforms the techniques based on the confusion matrix (in the 20% test scenario, for instance, with a KL that is three times lower than the AC method).

Although the PP method shows good performance for a wide variety of training and test conditions, it is also important for this application to select (at least an interval) of training class distributions that guarantee a good estimation for whatever test class distribution. In Fig. 4(a), we see that the PP technique is almost the lower envelop of the plot, and shows a tendency to decrease as the percentage of positive examples decreases. Keeping in mind that in terms of MRE and MAE (see Fig.2(a) and Fig.3(a)), apart from the 5% scenario, the performance is quite stable, empirical work suggests a training class distribution between 10% and 30% of positive examples.

4 Conclusions and Future Work

The veterinary field demands tools able to determine the proportion of damaged cells in a given sperm sample without any concern about the individual cell classification. This quantification task is tackled in this work following an approach based on *Posterior Probability* (PP) estimates.

Experimental results with a boar sperm data set were carried out for quantification purposes within the uncertainty region of interest (class distribution in the test stage lower than 20% of damaged cells). In general terms, evaluation as a function of Mean

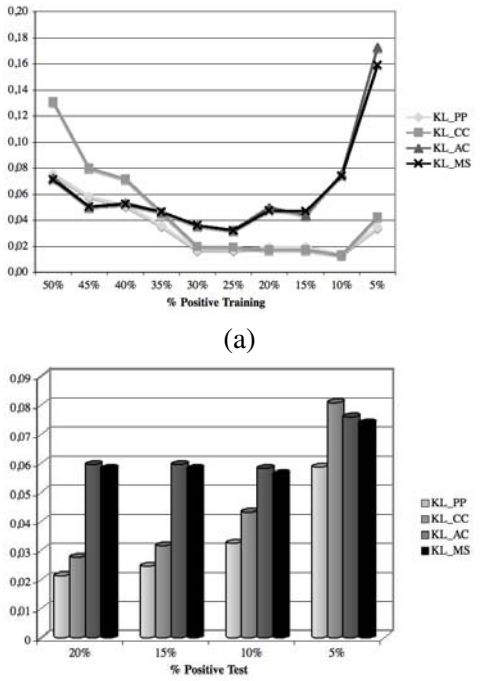


Figure 4: Kullback-Leibler divergences (KL) for the estimation methods CC, AC, MS and PP. (a) KL for ten different training scenarios scenarios. (b) KL for four test scenarios.

Absolute Error , Mean Relative Error and Kullback-Leibler divergence shows that: (1) there are no significant differences between the *Adjusted Count*(AC) and *Median Sweep*(MS) techniques in this veterinary field, (2) in general, the AC, MS and CC methods yield poorer estimations than those provided by the approach based on the PP estimates.

Results show that PP guarantees an stable MAE (around 3%) for whatever combination of training and test conditions. On the contrary, techniques based on the confusion matrix show higher MAE on average and maximum MAE values twice higher than the MAE for the PP technique.

In future work we will address the design of switching and/or combining modules to select among different classifiers that may have been trained with different class distributions with the goal to reduce the relative deviation in the quantification.

References:

[1] R. Alaiz-Rodríguez, A. Guerrero-Curieses, and J. Cid-Sueiro. Minimax regret classifier for imprecise class distributions. *Journal of Machine Learning Research*, 2007.

[2] C. M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1996.

[3] Chris Drummond and Robert C. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2001.

[5] G. Forman. Counting positives accurately despite inaccurate classification. In *European Conference on Machine Learning*, pages 564–575, 2005.

[6] G. Forman. Quantifying trends accurately despite classifier error and class imbalance. In *Principles and Practice of Knowledge Discovery in Databases*, pages 157–166, 2006.

[7] M. González, E. Alegre, R. Alaiz, and L. Sánchez. Acrosome integrity classification of boar spermatozoon images using dwt and texture techniques. In *VipIMAGE -Computational Vision and Medical Image Processing*. Taylor and Francis, 2007.

[8] C. Linneberg, P. Salamon, C. Svarer, and L.K. Hansen. Towards semen quality assessment using neural networks. In *Proc. IEEE Neural Networks for Signal Processing IV*, pages 509–517, 1994.

[9] M. Biehl N. Petkov, E. Alegre and L. Sánchez. LVQ acrosome integrity assessment of boar sperm cells. In *Computational Modelling of Objects Represented in Images-Fundamentals, Methods and Applications*. Taylor and Francis, 2007.

[10] M. Saerens, P. Latinne, and C. Decaestecker. Adjusting a classifier for new a priori probabilities: A simple procedure. *Neural Computation*, 14:21–41, January 2002.

[11] J.C. Samper. *Equine Breeding and Management and Artificial Insemination*. W. B. Saunders Company, Philadelphia, Pennsylvania, 2000.

[12] S.Arivazhagan and L.Ganesan. Texture classification using wavelet transform. *Pattern Recognition Letters*, 24(9-10):1513–1521, June 2003.

[13] J. Verstegen, M. Iguer-Ouada, and K. Onclin. Computer assisted semen analyzers in andrology research and veterinary practice. *Theriogenology*, 57:149–179, 2002.