Automatic classification of human sperm head morphology

Violeta Chang, Laurent Heutte, Caroline Petitjean, Steffen Härtel, Nancy Hitschfeld

www.elsevier.com/locate/cbm

PII: S0010-4825(17)30083-5
DOI: http://dx.doi.org/10.1016/j.compbiomed.2017.03.029
Reference: CBM2634

To appear in: *Computers in Biology and Medicine*

Received date: 12 December 2016
Revised date: 28 March 2017
Accepted date: 29 March 2017

Cite this article as: Violeta Chang, Laurent Heutte, Caroline Petitjean, Steffen Härtel and Nancy Hitschfeld, Automatic classification of human sperm head morphology, *Computers in Biology and Medicine* http://dx.doi.org/10.1016/j.compbiomed.2017.03.029

Automatic classification of human sperm head morphology

Violeta Chang[a,b], Laurent Heutte[c], Caroline Petitjean[c], Steffen Härtel[b], Nancy Hitschfeld[a]

[a]*Department of Computer Science, University of Chile, Beauchef 851, Santiago, Chile.*
[b]*Laboratory for Scientific Image Analysis, (SCIAN-Lab), Centro de Espermiograma Digital Asistido por Internet (CEDAI SpA), Biomedical Neuroscience Institute (BNI), Program of Anatomy and Developmental Biology, Biomedical Science Institute (ICBM), National Center for Health Information Systems (CENS), Faculty of Medicine, University of Chile, Independencia 1027, Santiago, Chile.*
[c]*Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France.*

Abstract

Background and Objective: Infertility is a problem that affects up to 15% of couples worldwide with emotional and physiological implications and semen analysis is the first step in the evaluation of an infertile couple. Indeed the morphology of human sperm cells is considered to be a clinical tool dedicated to the fertility prognosis and serves, mainly, for making decisions regarding the options of assisted reproduction technologies. Therefore, a complete analysis of not only normal sperm but also abnormal sperm turns out to be critical in this context. This paper sets out to develop, implement and calibrate a novel methodology to characterize and classify sperm heads towards morphological sperm analysis. Our work is aimed at focusing on a depth analysis of abnormal sperm heads for fertility diagnosis, prognosis, reproductive toxicology, basic research or public health studies.

Methods: We introduce a morphological characterization for human sperm heads based on shape measures. We also present a pipeline for sperm head classification, according to the last Laboratory Manual for the Examination and Processing of Human Semen of the World Health Organization (WHO). In this sense, we propose a two-stage classification scheme that permits to classify sperm heads among five different classes (one class for normal sperm heads and four classes for abnormal sperm heads) combining an ensemble strategy for feature selection and a cascade approach with several support vector machines dedicated to the verification of each class. We use Fisher's exact test to demonstrate that there is no statistical significant differences between our results and those achieved by domain experts.

Results: Experimental evaluation shows that our two-stage classification scheme out-

Email addresses: vchang@dcc.uchile.cl (Violeta Chang), Laurent.Heutte@univ-rouen.fr (Laurent Heutte), Caroline.Petitjean@univ-rouen.fr (Caroline Petitjean), shartel@med.uchile.cl (Steffen Härtel), nancy@dcc.uchile.cl (Nancy Hitschfeld)

performs some state-of-the-art monolithic classifiers, exhibiting 58% of average accuracy. More interestingly, on the subset of data for which there is a total agreement between experts for the label of the samples, our system is able to provide 73% of average classification accuracy.

Conclusions: We show that our system behaves like a human expert, therefore it can be used as a supplementary source for labeling new unknown data. However, as sperm head classification is still a challenging issue that gets harder due to the uncertainty on the class label of each sperm head, with the consequent high degree of variability among domain experts, we conclude that there is room for improvement in designing a more accurate system by investigating other feature extraction methods and classification schemes.

Keywords: infertility, sperm head morphological descriptor, sperm head classification, two-stage classification

1. Introduction

Infertility affects up to 15% of couples worldwide with many implications [1]. The analysis of semen is usually made according to standard criteria [2] and is the first step in the evaluation of the male factor. In human semen samples, there are sperm cells with different kinds of anomalies which generally imply lower fertilizing potential and/or abnormal DNA [2]. The morphology of the sperm cells allows to classify each sperm cell as normal or abnormal [3], thus, it permits to clarify the potential fertility of a sample [4]. There is vast evidence that confirm that the male age, stress, nutrition, pathogens and inbreeding can influence the percentage of abnormal sperm in ejaculates from humans and non-humans [5, 6, 7, 8, 9]. The categories of defects that should be noted in a reliable morphological analysis include head, neck and mid-piece and tail defects, as well as excess residual cytoplasm (see Figure 1).

The close relationship between fertility and morphologically normal sperm has been demonstrated by many studies [10, 11, 12, 13, 14] showing that the morphology of human sperm serves, mainly, for making decisions regarding the options of assisted reproduction technologies [3]. Furthermore, emphasis on identifying the categories of abnormal sperm heads may have significant clinical utility when deciding on an infertility treatment. However, the spectrum of possible malformations is considerably wide and makes difficult the classification of abnormal sperm morphology [15].

There is evidence from previous decades that the aforementioned categorization is a challenging task. In 1966, [16] showed that the traditional method for performing the analysis was *personality oriented, subjective, qualitative, non repeatable and difficult to teach to students and technicians, when comparing protocols in 47 laboratories for human sperm morphological analysis. Although there was a simplification of the classification rules for morphological semen analysis* [2], many authors reveal a lack of standardization of the methods used in laboratories in many countries [17, 18] leading to inter and intra observer variability in labeling the data [19, 20, 21, 22]. Therefore, the visual analysis of sperm morphology still presents a substantial challenge concerning reproducibility and objectivity. Moreover, classifying defects according to normal

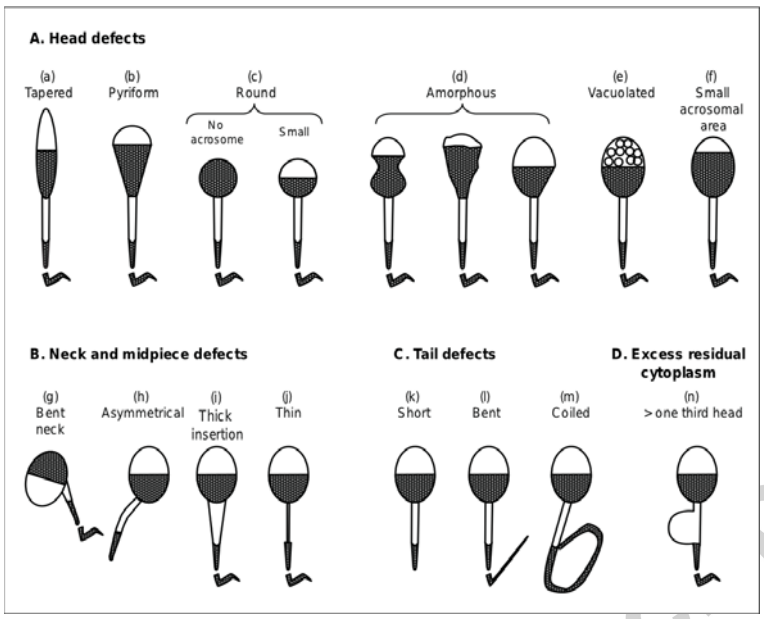


Figure 1: **Human sperm abnormalities.** Image reproduced exactly as appears in [2], showing schematic drawings of some abnormal forms of human sperm.

and abnormal sperm definitions in visual sperm classification under the microscope requires the evaluation of cellular and sub-cellular regions (size of the sperm head, tail length, residual cytoplasm area, etc.) and the detection/recognition of some characteristics (multiple heads or tails, absent tail, coiled tail, etc.) [3]. An alternative to replacing the poor visual ability to assess the size and shape of sperm is to analyze the sperm morphology with the help of a computer [3].

Currently, there are some Computer-Aided Sperm Analysis (CASA) systems. They were primarily developed to measure sperm concentration, the percentage of motile sperm and some details of sperm movement. To obtain useful information from the visual assessment of sperm morphology, it is crucial some kind of standardization of methods and variables be analyzed [23, 24, 25, 26, 3]. Although there are some commercial applications for sperm morphology assessment, none of these applications study abnormal sperm in depth, which has been proven to have a significant impact in research. In addition, these kinds of systems are sold as *black boxes* without any possibility of modification for research purposes, and, of course, there is no publication related to the algorithms implemented in these systems.

This paper sets out to develop, implement and calibrate a novel methodology to accurately characterize and classify sperm heads in the context of morphological sperm analysis. We expect that this automatic classification of sperm heads will behave at least as good as a human expert. Our work is aimed at focusing on a depth analysis of abnormal sperm heads for fertility diagnosis, prognosis, reproductive toxicology, basic research or public health studies. To this end, we propose a morphological characterization for human sperm heads, using single shape-based measures, for extracting features

from segmented sperm heads, and a pipeline for sperm head classification based on a two-stage classification scheme. Combining an ensemble strategy for feature selection and a SVM-based cascade classification, sperm heads are classified into one out of five different classes. Our experimental results show that our two-stage classification scheme outperforms monolithic classifiers and behaves as a human expert.

This paper is organized as follows. In Section 2 we review the research work in the area. Our proposed approach is justified in Section 3 and described in detail in Section 4. In Section 5 we present the dataset used, the experimental protocol and the experimental results achieved. Conclusion and future works are drawn in Section 6.

2. Related work

The analysis of sperm cells can be done regarding their vitality and/or their morphology. From a vitality point of view, a number of publications are concerned with the membrane integrity validation aimed at classifying sperm cells into live and dead cells [27, 28, 29, 30, 31]. In these works, textural features as well as moment-based descriptors were proposed for the characterization stage. For the classification stage, SVM, $k-$NN and multilayer perceptron neural networks were used, achieving up to 99% of correct classification [31]. Apart from the aforementioned works that classify sperm cells from their vitality point of view, the classification of sperm cells according to the morphology of their heads has received little attention. Although there is a plethora of papers focusing on human and non-human sperm head morphometry, most of these studies are limited to the use of some commercial computer-assisted morphometry software and the further statistical analysis of the resulting data. To the best of our knowledge, there are only four papers that deal methodologically with that issue: [32, 33, 34] only study the characterization of morphology for non-human sperm and [35] deals with classification of sperm images. We now review these works.

About morphological characterization of sperm heads, Beletti et al. [33] described the morphology of sperm heads using features such as: head area, perimeter, width, length, aspect ratio, ellipticity, shape factor, width of sperm basis, the three first Fourier values, side and anterior-posterior symmetry and hydrodynamic coefficient. In [32], the authors characterized the animal sperm head shape using a multiscale curvature estimation. They showed how a spectral approach to estimate the derivative property of the Fourier transform is useful for extracting relevant features for sperm head morphology assessment. The proposed method calculated three morphological features: width of sperm head, implantation symmetry and bending energy of the frontal portion of the head. Experiments were conducted on real data from several animal species. Severa et al. [34] developed a framework to characterize stallion sperm heads and evaluated the intrinsic shape variability. Sperm head shape characteristics including aspect ratio, position of the center of gravity, curvature and degree of roundness were assessed and analysed using Fourier descriptors and inverse Fourier transformation.

With respect to classification of sperm head morphology, the only published work seems to be that of Abbiramy and Tamilarasi who evaluated the accuracy of neural networks for classifying human sperm cells, discriminating between normal and abnormal sperm cells [35]. The feature vector proposed by the authors includes first order statistics, textural features (gray level co-occurrence matrix) and morphological

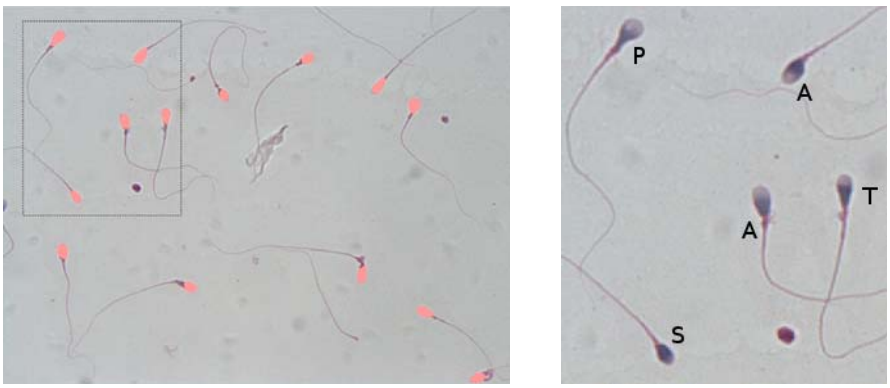
features (head area, perimeter width and length, excentricity and orientation, among others). For classification purposes, the authors evaluated three neural networks techniques: feed forward, radial basis and Elman back propagation. The experiments were done using images taken from WHO laboratory manual [2], and showed that the radial basis network produced the highest classification accuracy of 60%, 75% and 70% when trained with statistical features, combined features (statistical, textural and morphological) and morphological features, respectively. However, there is no evidence that the proposed approach works well with real images from clinical laboratories, captured with standard microscope and using rational resolution.

To the best of our knowledge, there is no published work that encompasses a full morphological analysis of human sperm cells with characterization and classification approaches. In this sense, we propose the first approach to characterize and classify human sperm heads in five different classes. Even though there is a lack of research on human sperm cell classification, it is important to realize that there are some good approaches for characterization of animal sperm heads. For example some single shape-based parameters such as head area, perimeter, ellipticity, aspect ratio, and curvature degree, among others, can be used in a more general human sperm head morphological characterization and combined with global shape-based measures, in order to conduct a multi-class classification towards morphological human sperm analysis. There is also a lack of published framework for identification of human sperm cell abnormalities. A depth analysis of abnormal sperm heads could however be useful for fertility diagnosis, prognosis, reproductive toxicology, basic research or public health study. Here lies the great impact and main contribution of our work.

3. Our strategy for sperm head classification

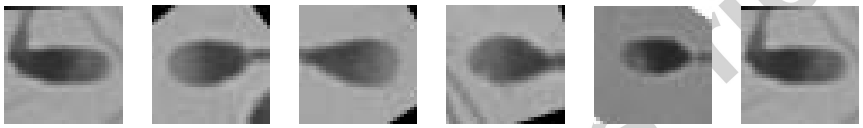
Our goal here is to perform the classification of human sperm heads into five classes: normal (N), tapered (T), pyriform (P), small (S) and amorphous (A). Looking at the sperm images and segmentation results (see Figure 2a), it is easy to figure out the difficulty for characterizing the shape of human sperm heads. For example, within the variety of sperm heads that can be observed in Figure 2b, we can find tapered, pyriform, small and amorphous sperm heads, with slight shape variations among them. The classification of human sperm heads is still a very hard task, even for a human expert. As a consequence, the human experts often disagree on the same image (see Figure 3) leading to a high degree of disagreement among experts. One factor that contributes greatly to this disagreement is the confusion that exists when trying to classify amorphous heads for example, due to the the close shapes of amorphous and normal, tapered, pyriform and small sperm heads.

To face this challenging task of classifying sperm heads, we thus need both discriminant features and an adapted classification scheme that will be able to cope with such strong inter and intra variability among sperm head classes. Yet it has been experimentally observed that creating a perfect monolithic classifier for a particular application is somewhat unfeasible for various reasons [37, 38]. One way to cope with this is to use a combination of classifiers where a set of single classification models is trained, and the output of the ensemble is obtained by aggregating the outputs of the single models, e.g., by majority voting. It has been shown that combined classifiers often outperform



a) b)

Figure 2: **Segmented sperm heads (a)** Original image in RGB color space containing a number of human sperm cells. Red color represents the results of head segmentation procedure as defined in [36]. Image size: 780×580 pixels $\approx 164 \times 122 \mu m$. **(b)** Zoom in to the marked zone in (a) with class label assigned by experts.



	a)	b)	c)	d)	e)	f)
Expert 1	Tapered	Normal	Tapered	Amorphous	Tapered	Normal
Expert 2	Tapered	Normal	Pyriform	Normal	Normal	Small
Expert 3	Tapered	Normal	Tapered	Normal	Amorphous	Amorphous

Figure 3: **Inter-expert disagreement.** Representative images with total agreement ((**a**) and (**b**)), partial agreement ((**c**) and (**d**)) and null agreement ((**e**) and (**f**)) among experts (Image size: 35×35 pixels $\approx 7 \times 7$ μm).

any single base learner [39]. Thus, our strategy aims at finding the best combination of couples descriptor/classifier to be combined in parallel in order to tackle the automatic classification of human sperm heads.

For that purpose, we have designed a two-stage classification scheme, aiming at, in a first stage, discriminating the amorphous sperm heads from the others, and in a second stage, classifying the four remaining classes: the first stage acts as a Class Filtering stage whose aim is to minimize the average confusion rate of the four classes (N, T, P and S) without taking into account class A; to verify if the result of the first stage is correct, the second stage is made of four verifiers, each one dedicated to the discrimination of class A from a specific class, thus acting as a Class Verification stage.

Due to the high variability of shapes, it is needed to use different shape-based descriptors that capture the subtle differences in the shapes of the heads of different classes. Rather than optimizing a monolithic classifier that takes as input all the features, we try in the first stage to find the best combination of features that minimizes the confusion rate of the four classes. For the second stage, as the problem is rather

different because we use four verifiers, we need to find, among the descriptors, that combination of features which maximizes the classification rate of each verifier. In this sense, we propose a feature selection process that takes advantage of the best shape-based descriptors for each class. In the feature selection process, either for the first stage or the second stage, we use a 1−NN classifier for its ability to assess the discriminative power of features. Once the features have been selected by 1−NN, in the first and the second stages, we use a combination of SVM classifiers.

More formally, our global strategy is summarized as follows. For the class filtering stage, we are interested in identifying the potential class of sperm heads that are not A. Thus, we look for the best combination of descriptors that minimizes the confusion rate between the four remaining classes discarding A (N, T, P and S). To do this, we design a descriptor selection strategy combining six different descriptors (see Section 4.2 for details) and using 1−NN as the base classifier for each descriptor, and any of different combination rules (e.g. majority voting) to evaluate our objective function. Let $DSC_1, DSC_2, \ldots, DSC_M$ be the best combination of descriptors selected in the previous step. For the classification process, we propose to use an independent SVM as base classifier for each one of these descriptors. Therefore, we will have one SVM that receives features from descriptor DSC_1, another SVM that receives features from descriptor DSC_2, and so on. Each one of these SVMs will be trained using only four out of five classes (N, T, P and S), but they will be tested using five classes (including A). To combine the outputs of all classifiers, we could use different combination rules designed for this purpose, such as unanimity, plurality, and majority voting, as well as one that considers the probability of the output from each classifier.

For the class verification stage, we are interested in verifying the potential class of sperm heads returned by the previous stage. Thus, the objective function of the process of descriptor selection in this stage is the maximization of mean precision rate between each of the four remaining classes discarding amorphous (N, T, P and S) versus A (in all cases). What we need to do is to assemble a descriptor selection strategy combining six different descriptors (see Section 4.2 for details) and using 1−NN as base classifier for each descriptor, and any of different combination rules to evaluate our objective function in four scenarios (N vs A, T vs A, P vs A, and S vs A). The best combination of descriptors will be used in each verifier of the second stage of the classification scheme. Let $FS_i = \{FS_{Ai}, FS_{Bi}, \ldots, FS_{Ni}\}$ be the best combination of descriptors selected for verifier V_i. For the verification stage, we use an independent SVM as the base classifier for each one of these descriptors. Therefore, we will have one SVM that receives features from descriptor FS_{Ai}, another SVM that receives features from descriptor FS_{Bi}, and so on. Each one of these SVMs will be trained using only two out of five classes (N, T, P or S, and A). They will be tested using the same two classes used for training. To combine the outputs of all SVMs, we could use different combination rules designed for this purpose, such as unanimity, majority voting, etc. Having the four verifiers $V_i, i = 1 \ldots 4$, we do not need any combination rule in this stage, because only one verifier should be used, and its output should be considered as the output of the whole stage.

In Figure 4, we show the architecture of the proposed classification scheme. For testing purposes, we evaluated sperm by sperm in a cascade approach. Let s_i be a given sperm head. We tested a *combined classifier* with s_i as input. If s_i is rejected, then the

testing process is finished and we considered sperm s_i classified as A. Otherwise, if s_i is accepted, we continued with the second stage considering the first stage output as following. If first stage output is N, then we tested $v1$ with s_i as input. Analogously, we tested $v2$ if first stage output is T, $v3$ if first stage output is P, and $v4$ if first stage output is S. In any case, we considered the output of the corresponding verifier ($v1$, $v2$, $v3$, $v4$) as the final output of the whole scheme.

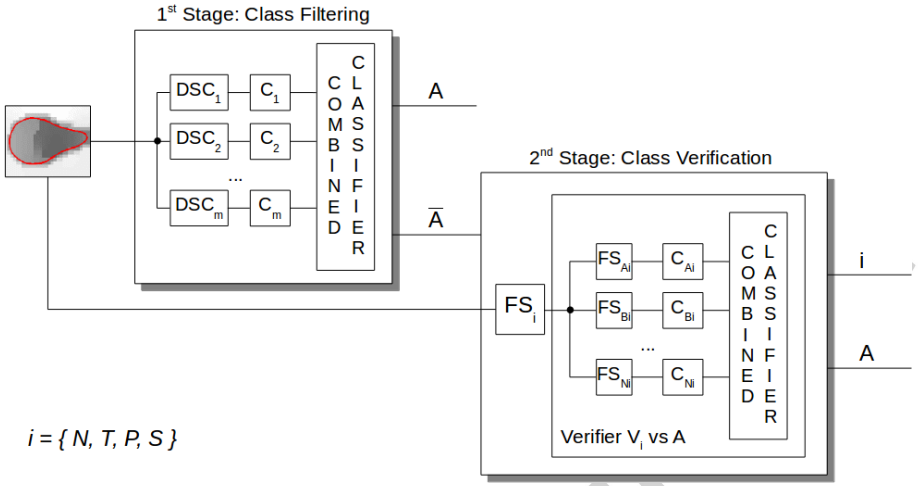


Figure 4: **Architecture of the proposed classification approach.** The proposed classification scheme has two stages: class filtering and class verification. The first stage discriminates between amorphous and other classes of sperm heads while acting as a prior classifier for the four remaining classes. The second stage is triggered just in case the first stage output is different from A. A stands for amorphous, N stands for normal, T stands for tapered, P stands for pyriform and S stands for small. DSC_j stands for the $j-$th descriptor selected in the best combination for the first stage. C_j stands for the $j-$th individual SVM whose result is combined in the first stage. FS_i stands for the combination of descriptors selected for verifier V_i. C_{ki} where $k = A, B, \ldots, N$, stands for the $k-$th individual SVM whose result is combined for the verifier V_i.

4. Feature extraction

Given a sperm sample image (see Figure 2a), we segmented the contained sperm heads following the two-stage procedure defined in [36]. The first stage detects regions of interest that define sperm heads using k-means, then candidate heads are refined using mathematical morphology. In the second stage, each region of interest is evaluated to segment accurately the sperm head using clustering and histogram statistical analysis techniques, working in three different color spaces (see Figure 2b). From the results of the segmentation procedure, isolated sperm head images are obtained (see Figure 5a). Further details can be found in our previous work on sperm head segmentation [36]. This kind of image is the input of our system and we use that image to classify the sperm head in only one class among normal (N), tapered (T), pyriform (P), small (S), and amorphous (A). To do this, we propose to divide the process in three main steps: preprocessing, characterization and two-stage classification. Regarding

that our strategy for classification was explained in the previous section, here we will describe in detail the two remaining steps of our proposed approach: preprocessing and characterization.

4.1. Sperm head preprocessing

In this work, special emphasis was given to a continuous representation of the curve defining the outline of a head rather than a discrete representation based on pixels. However, in order to have a reliable continuous representation, it is necessary to preprocess the image corresponding to each segmented head (regarded as a ROI) as the result of the method proposed in [36]. Therefore, the aim is to obtain a reliable representation of the closed curve constituting the contour of the head, given a 35×35 grey level image containing a sperm head (see Figure 5a).

To do this, first we need to remove from the image as much noise as possible. We propose using anisotropic diffusion [40] to preserve the border while the image is simplified, greatly reducing the noise in it (see Figure 5b). From an image (as presented in Figure 5b), we need now to generate a continuous representation of the curve that defines the sperm head. For this, we use active contours [41] regarding as initial curve the result of the segmentation approach presented in [36] (see Figure 5c). An example of how the outline for a given head would look after applying active contours is shown in Figure 5d. Finally, each sperm head is represented by a sequence B of n real coordinates (p_i, q_i) (and not in terms of pixels), where $(p_1, q_1) = (p_n, q_n)$. The number of points that form each contour is variable and depends exclusively on the perimeter of the curve. In this sense, a contour with a perimeter around 35 will be represented by 41 points approximately, while another with a perimeter around 64 will need approximately 73 points for its representation.

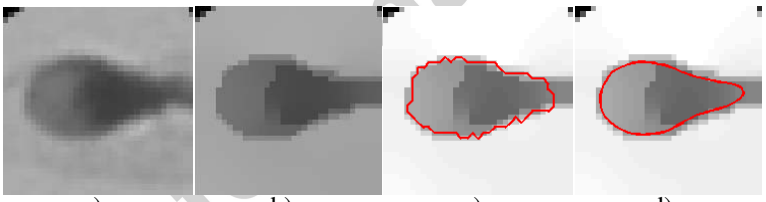


Figure 5: **Shape representation of sperm heads. (a)** Original grey level image (Image size: 35×35 pixels $\approx 7 \times 7$ μm). **(b)** Original image filtered using anisotropic diffusion. **(c)** Segmentation of sperm head (as returned according to [36]) used as the initial curve for active contours. **(d)** Shape contour after applying active contours.

4.2. Sperm head characterization

When designing a descriptor, considering how experts in the field would describe objects of study (ROIs in our case) can be of some help. Thus, rather than building a general theory of shape, a popular approach is to design shape-based descriptors sensitive to various aspects of shape. The importance of finding shape-based measures that are simple to compute, with intuitive meanings, has already been noted by Peura [42]. Common simple global shape-based measures are area, diameter, perimeter, and

eccentricity, among others. The majority of these measures not only have linear computational complexity in the number of (boundary or region) points, but also tend to be designed to be invariant to rotations, translations, and uniform scaling, and often have an intuitive meaning since they describe a single aspect of the object of study. If the object defined by a ROI is described by a combination of shape-based measures, this should be sufficient to provide discrimination between different classes of shapes. Deciding on the most appropiate measures depends on their suitability for a particular application.

In this section, we introduce a morphological descriptor designed as a combination of simple global shape-based measures. In addition, we also point out a number of state-of-the-art shape-based descriptors that will be used along with the proposed morphological descriptor to characterize human sperm heads.

4.2.1. Morphological descriptor

Global shape-based measures are a convenient way to describe objects defined by ROIs. They are generally simple and efficient to extract, and provide an easy means for high level tasks such as classification. Although global shape-based measures have direct intuitive meaning, they can only discriminate shapes with large differences. Therefore, they are usually used as filters to eliminate false hits or combined with other shape-based measures and/or descriptors to discriminate shapes. Although no single shape-based measure in the combination is descriptive enough to distinguish sperm heads from different classes, they contain enough information when combined with other shape-based descriptors to discriminate sperm head classes. All these measures are described for an object defined by a ROI S and summarized in Table 1. S is defined as the set of m points $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ that are enclosed by curve B, which are calculated from the boundary obtained after applying active contours.

It is important to point out specific comments for two measures. First, *quadrant fitness* measures how close each quadrant of the ROI defined by contour B is to the corresponding quadrant of an ellipse E with the same centroid, major and minor axis length, and orientation (see Figure 6). For each quadrant i, cf_i is calculated as the sum of the shortest distances from each point of B to ellipse E. We are interested not only in keeping the absolute fitness in each quadrant, but also in the relationships between them. In this sense, let cf_1 be the fitness value in the first quadrant, cf_2 the fitness value in the second quadrant and so on. We include the following feature values in the descriptor: cf_1, cf_2, cf_3, cf_4, $\frac{cf_1}{cf_2}$, $\frac{cf_2}{cf_4}$, $\frac{cf_1}{cf_3}$ and $\frac{cf_3}{cf_4}$. For instance, for normal sperm heads, $\frac{cf_1}{cf_2}$, $\frac{cf_2}{cf_4}$, $\frac{cf_1}{cf_3}$ and $\frac{cf_3}{cf_4}$ should have values close to 1, while for pyriform sperm heads only $\frac{cf_1}{cf_3}$ and $\frac{cf_2}{cf_4}$ should have values close to 1.

Second, *bilateral symmetry* measures the normalized area of overlap between the ROI defined by contour B and a reflected version of itself, in both directions, horizontal and vertical with respect to its centroid (see Figure 7 for a graphical explanation). Let R_V be a vertical reflected version of ROI S and R_H be a horizontal reflected version of ROI S, then, the feature values that we propose to include in the descriptor are: $\frac{area(S \bigcap R_H)}{area(S)}$ and $\frac{area(S \bigcap R_V)}{area(S)}$. For instance, for pyriform sperm heads, the value of bilateral symmetry in horizontal and vertical directions should be quite different, while in the case of tapered or normal sperm heads, both directions should yield similar values.

Table 1: **Summary of shape-based measures included in the proposed morphological descriptor.**

Shape Measure	Formulation	Size
area	$\frac{1}{2}\sum_{i=0}^{m-1}(x_i y_{i+1} - x_{i+1} y_i)$	1
perimeter	$\sum_{i=0}^{n-1} d(p_i q_i, p_{i+1} q_{i+1})$	1
eccentricity	$\frac{\sqrt{(\mu_{20}-\mu_{02})^2+4\mu_{11}^2}}{\mu_{20}+\mu_{02}} = \frac{majorAxis(S)}{minorAxis(S)}$	1
regularity	$\frac{\pi * majorAxis(S)^2 * minorAxis(S)^2}{4 * area(S)}$	1
circularity	$\frac{perimeter(S)^2}{area(S)}$	1
rectangularity	$\frac{area(S)}{area(MBR(S))}$	1
maximum curvature	$\max \frac{p_i' q_i'' - q_i' p_i''}{(p_i'^2 + q_i'^2)^{3/2}}$	1
minimum curvature	$\min \frac{p_i' q_i'' - q_i' p_i''}{(p_i'^2 + q_i'^2)^{3/2}}$	1
ellipticity	A fitness value given a contour against an ellipse is calculated as the shortest Manhattan distance between them [43]	1
quadrant fitness	*	8
bilateral symmetry	*	2

* See a detailed explanation in this section

4.2.2. Other shape-based feature descriptors

There are many shape-based descriptors in the literature, e.g. Fourier descriptors [44], geometric moments [45], and Zernike moments [46]. These descriptors have been proven to be effective in some applications, although a drawback is that their values are often not easily understandable. In many applications it is preferable that the measures from the object defined by a ROI can be analysed by the domain experts, as this aids validation of the whole proposed scheme. However, there is no single descriptor that is fully suitable for characterizing human sperm heads [47]. In this sense, we propose to combine five shape-based descriptors (as detailed in Table 2) with our proposed human sperm head morphological descriptor in order to accurately characterize human sperm heads towards sperm morphological analysis.

5. Experimental results and analysis

5.1. Dataset

For the experimental results that we show in this section, we have used the human sperm head classification gold-standard SCIAN-MorphoSpermGS introduced in [47]. Semen samples from volunteers, with age range of 28 − 35 years old, were obtained-expected at the Laboratory of Spermiogram, Program of Anatomy and Developmental Biology (ICBM), Faculty of Medicine, University of Chile, Santiago, Chile. After collection, the semen samples were stained with a modified Hematoxylin/Eosin procedure.

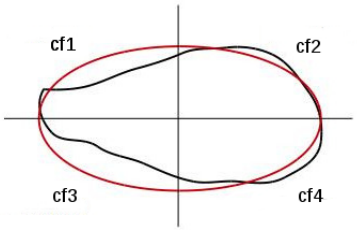
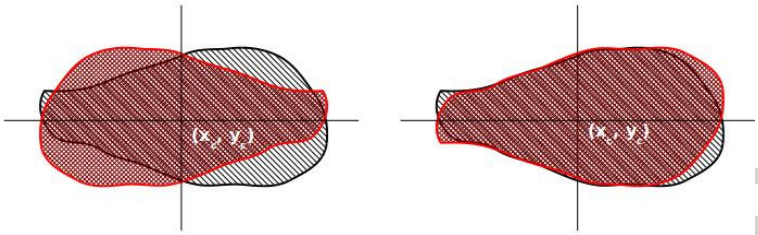


Figure 6: **Quadrant Fitness**. Overlay of the contour shape (black) with an ellipse with the same centroid, axis length and orientation (red). In each quadrant $i = 1 \ldots 4$, the fitness cf_i is calculated.



a) b)

Figure 7: **Bilateral Symmetry**. We show an overlay between a ROI defined by contour B (black) and a reflected version of itself (red), in both directions, vertical (a) and horizontal (b) with respect to its centroid.

In brief, semen smears were fixed with ethanol 70%, immersed in Harris' Hematoxylin for ten seconds, washed with tap water for ten minutes, immersed in 1% Eosin for two minutes, washed with distilled water for one minute and air-dried. Digital images were captured using optical, bright field microscopy (Axiostar Plus, Carl Zeiss Inc, Wetzlar, Germany), a 63x objective (oil, NA 1.4) with an adapter of 0.63x and a digital camera (scA780-54gc, Basler AG, Ahrensburg, Germany). SCIAN-MorphoSpermGS is a dataset of sperm head images, containing 1,854 observable and evaluable sperm cells with expert-classification labels in one of the following classes: N, T, P, S, or A. Figure 8 shows representative sperm cells from each class. The manual classification process was performed independently, per patient/smear, by three referent Chilean experts with vast experience in morphological sperm analysis.

As this gold-standard was built with the cooperation of three experts, there are three different agreement scenarios: one (basis set), two experts (partial agreement - PA), or three experts agree on the same label for a given sperm head (total agreement - TA). The first set contains 1,854 sperm head labels, but a sperm head can be classified into three different classes by the three different experts. The second set contains 1,132 sperm heads, meaning that there are 1,132 sperm heads with partial agreement and without overlapping. The third set contains only 384 sperm heads, with total agreement between the three experts. Table 3 shows the number of sperm cells per class for each agreement scenario. We choose to work with the partial agreement dataset (further referenced only as *dataset*) with 1,132 sperm heads with partial agreement among experts and without overlapping, distributed in five imbalanced classes (N, T, P, S, and A).

The dataset has been partitioned in three subsets, according to [47], named *Dataset*

Table 2: **Summary of used shape-based descriptors.**

	Shape Descriptor	Type	Dimensionality
1	Morphological (Section 4.2.1)	Contour and region based	19
2	Fourier [44]	Contour based	15
3	Geometric Hu moments [45]	Region based	7
4	Zernike moments [46]	Region based	36 [a]
5	Convexity measures [48]	Region based	5 [b]
6	Ellipticity measures [49]	Region based	10 [c]

[a] The first 36 Zernike moments up to order 10
[b] We vary $\alpha = 1/2^b$ where $b = 1 \dots 5$
[c] We vary λ from 0.5 to 5, with step size of 0.5

Agreement among experts	Normal (N)	Tapered (T)	Pyriform (P)	Small (S)	Amorphous (A)
At least one (Basis set)	175	420	188	152	919
Partial agreement (PA)	100	228	76	72	656
Total agreement (TA)	35	69	7	11	262

Table 3: Inter-expert agreement

1 (DS1), Dataset 2 (DS2) and Dataset 3 (DS3), aiming at having a training (60% of the whole dataset), validating (20%) and testing (20%) dataset, respectively. In addition, we have an extra testing dataset. Special Testing Dataset (DST) is a subset of DS3, but with a particular feature: all the sperm heads that are contained in DST have been manually classified within the same class by all experts (total agreement between experts). In Table 4, the size and distribution of classes in each partition and extra testing dataset are presented.

	Dataset	DS1	DS2	DS3	DST
Number of Normal sperm heads	100	60	20	20	9
Number of Tapered sperm heads	228	137	46	45	18
Number of Pyriform sperm heads	76	44	15	16	2
Number of Small sperm heads	72	45	14	14	2
Number of Amorphous sperm heads	656	394	131	131	56
Total number of sperm heads	1132	680	226	226	87

Table 4: **Dataset partition**. $Dataset = DS1 \cup DS2 \cup DS3$. DST is a special dataset for testing purposes with total agreement between experts (subset of DS3).

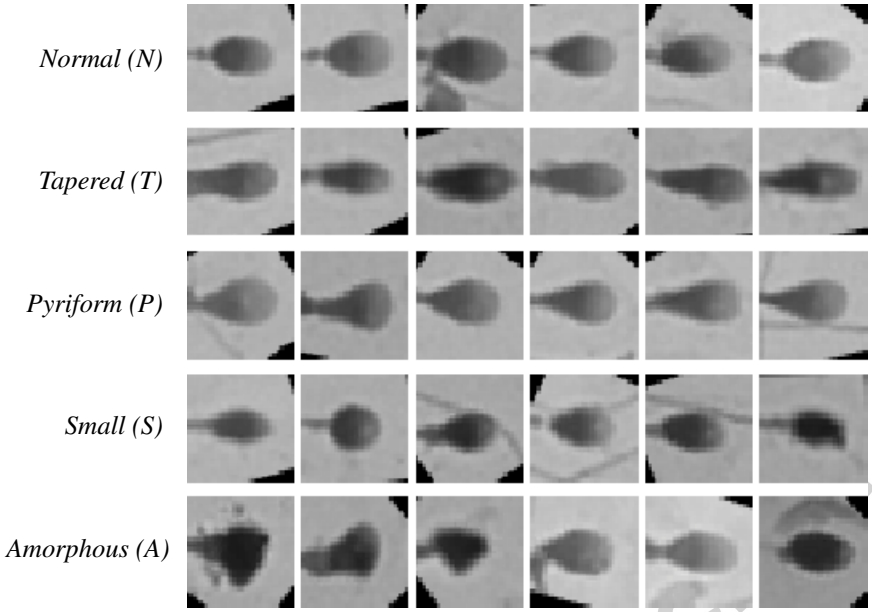


Figure 8: **Classification gold-standard.** Representative images of normal, tapered, pyriform, small and amorphous sperm cells that showed total agreement (TA) among experts (Image size: 35×35 pixels $\approx 7 \times 7$ μm).

5.2. Preliminary experiments

Trying to show how difficult the discrimination of five classes is, we have conducted some experiments to evaluate the accuracy rates per class in the classification of human sperm heads using monolithic classifiers. These experiments were conducted aiming at evaluating the difficulty of classification with a monolithic classifier. We worked with the whole set of features (without selection) by aggregating the six families of features, i.e. 92−size feature space, using DS1 for training and DS3 for testing purposes with five classes in both cases. We have run 100 iterations for each monolithic classifier: 1−NN, Naive Bayes, Decision Trees and SVM. In Table 5, we show the True Positive Rate for each class and the mean accuracy rate using different monolithic classifiers and without any dimensionality reduction technique.

In addition, we have performed some experiments to evaluate dimensionality reduction (DR) techniques using the same features and monolithic classifiers. These experiments were conducted following two goals: a) evaluate the impact of using different feature spaces (according to DR technique), and b) identify the most discriminant features (no feature families) according to MLE intrinsic dimension. We use DS1 for training and DS3 for testing purposes, with five classes in both cases. We have run 100 iterations for each monolithic classifier using different dimension reduction techniques: PCA (Principal Components Analysis), MDS (Multidimensional Scaling), Kernel PCA and Diffusion Maps. In Table 6 we report only the results of the SVM classifier because it was the classifier with the best accuracy reached with all DR techniques. We present the True Positive Rate for each class and the mean accuracy with the best number of

Classifier	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
1−NN	0.46	0.53	0.40	0.42	0.18	0.40
Bayes	0.31	0.65	0.45	0.80	0.17	0.48
Decision Trees	0.45	0.50	0.41	0.54	0.28	0.44
SVM	0.39	0.64	0.33	0.69	0.28	0.47

Table 5: **Results of using monolithic classifiers without dimensionality reduction techniques.** *tpr* stands for True Positive Rate, *N* means Normal, *T* means Tapered, *P* means Pyriform, *S* means Small, and *A* means Amorphous. *acc* stands for accuracy understood as mean of True Positive Rates of classes Normal, Tapered, Pyriform, Small and Amorphous.

dimensions, for each DR technique, which allows the best accuracy rate.

DR Technique	nDims	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
PCA	5	0.74	0.68	0.38	0.46	0.13	0.48
MDS	5	0.73	0.68	0.38	0.47	0.13	0.48
Kernel PCA	55	0.90	0.12	0.33	0.03	0.08	0.29
DiffusionMaps	20	0.80	0.10	0.24	0.18	0.03	0.27

Table 6: **Results of applying dimensionality reduction techniques using the best number of dimensions.** *nDims* means number of dimensions, *tpr* stands for True Positive Rate, *N* means Normal, *T* means Tapered, *P* means Pyriform, *S* means Small, and *A* means Amorphous. *acc* stands for accuracy understood as mean of True Positive Rates of classes Normal, Tapered, Pyriform, Small and Amorphous.

Observing Table 5, we can see that around 31 – 46% can be reached for classes Normal and Pyriform, around 50 – 65% for class Tapered, around 42 – 80% for class Small, and less than 30% for class Amorphous. From Table 6, we can observe that with DR techniques, up to 90% can be reached for class Normal, up to 38% for class Pyriform, up to 68% for class Tapered, up to 47% for class Small, and less than 15% for class Amorphous. In summary, the total accuracy rate without DR techniques reaches 48% in the best of cases, the same accuracy rate when using PCA or MDS as DR techniques.

From these preliminary results, we can conclude that the introduction of dimension reduction techniques has not yielded improved results. However these experiments allow us to confirm that even without DR technique, the problem is really difficult. This only confirms that the class Amorphous is the most difficult to discriminate (with and without TR techniques) and, therefore, the solution must be focused at an early separation of amorphous in a first stage to avoid confusing the discrimination of other classes in a second stage.

5.3. Feature selection experiments

In our strategy for human sperm head classification, we propose a descriptor selection step in each one of the two stages of the scheme. As mentioned earlier, we had

$d = 6$ different descriptors, and we needed to evaluate $2^d - 1 = 63$ different combinations of these descriptors for each stage. We used DS1 for selecting the best combination of descriptors. We perform this descriptor selection following the same procedure for each stage, but with different goals in each one. For each stage, we performed 10 runs of the experiment.

In this sense, for the first stage, we calculated the confusion rate among N, T, P and S for each one of the 63 combinations of descriptors, using 1−NN LOO for each descriptor and majority voting as combination rule. We look for minimizing this confusion rate in this stage. While in the second stage, we look to maximize the mean True Positive Rate of both classes of each verifier (A and one of {N, T, P, S}). Thus, for a given verifier, we calculated the mean TPR of both classes for each one of the 63 combinations of descriptors, using 1−NN LOO and majority voting as combination rule.

In Table 7, we show the five best combinations of descriptors with confusion rates using majority as combination rule.

Combination of descriptors	Confusion rate using majority
MorphoD,FourierD,ZernikeD	**0.44**
MorphoD,GeomD,ZernikeD	0.46
MorphoD,FourierD,GeomD	0.46
MorphoD,FourierD,GeomD,ZernikeD,ConvD	0.47
MorphoD,ZernikeD,ConvD	0.47

Table 7: **Ranking of descriptor combinations for Stage 1.** In column *Combination of descriptors*, *MorphoD* stands for morphological descriptor, *FourierD* stands for Fourier descriptor, *GeomD* stands for geometric moments descriptor, *ZernikeD* stands for Zernike moments descriptor, *ConvD* stands for convexity measures descriptor and *EllipD* stands for ellipticity measures descriptor.

From Table 7, we conclude that the best descriptor combination for this stage is the one that includes morphological descriptor, Fourier descriptor and Zernike moment descriptor as it achieves the lowest confusion rate (using majority voting).

For the second stage, we selected the most frequent descriptors that appear in the results of 10 run results (disregarding descriptor combinations as a whole). It is important to realize that in this stage, the best descriptor combination could differ from one verifier to another, because the main goal of this descriptor selection is to take advantage of the different features of classes versus A (See Table 8 for a summary). In Table 8, we show the selected descriptors for the four verifiers in stage 2.

Verifier	MorphoD	FourierD	GeomD	ZernikeD	ConvD	EllipD
Normal VS Amorphous		✓		✓	✓	✓
Tapered VS Amorphous	✓	✓	✓	✓		
Pyriform VS Amorphous		✓	✓			✓
Small VS Amorphous	✓			✓	✓	

Table 8: **Selected descriptors for four verifiers in stage 2**

5.4. Classification experiments

Once the descriptors are selected, we used a SVM for each selected descriptor to create one or more combined classifiers for each stage. We evaluated different combination rules: majority voting, unanimity voting and maximun probability using different threshold values. We used Dataset1 (DS1) for training and Dataset2 (DS2) for validating purposes.

For the first stage, the training procedure is as follows. We balanced the training data DS1 at first, creating a balanced training set (*trBag*) as $\{c0 \cup c1 \cup c2 \cup c3\}$ where $c0$, $c1$, $c2$ and $c3$ are balanced subsets of classes N, T, P and S. Next, for each selected descriptor, we apply cross-validation to choose the best parameters for individual SVM. We trained each SVM using selected descriptors of sperm heads contained in *trBag* and tested using DS2. We evaluated different combination rules. While for the second stage, the training procedure is similar to the one described above. In this case, we need to train four combined classifiers (verifiers), and the procedure to do this is the same as in the previous stage, changing only the composition of the training dataset in order to consider only the two relevant classes for each verifier.

As a result of training we obtain five combined classifiers. One called *svm1* as the combination of a number of SVMs, each one for one selected descriptor in the first stage. The one called *v1* consists of the combination of a number of SVMs, each for one selected descriptor looking to distinguish between sperm heads from classes N and A. Similarly, *v2* looking to distinguish between sperm heads from classes T and A, *v3* to distinguish between sperm heads from classes P and A, and *v4* to distinguish between sperm heads from classes S and A.

For testing purposes, we evaluated sperm by sperm in a cascade approach. Let s_i be a given sperm head. We tested *svm1* with s_i as input. If s_i is rejected, then the testing process is finished and we considered sperm s_i classified as an amorphous sperm head. Otherwise, if s_i is accepted, we continued with Stage 2 considering the output *label1* as following. If *label1* is class label *Normal*, then we tested *v1* with s_i as input. Analogously, we tested *v2* if *label1* is class label *Tapered*, *v3* if *label1* is *Pyriform*, and *v4* if *label1* is *Small*. In any case, we considered the output of the corresponding verifier (*v1, v2, v3, v4*) as the final output of the whole scheme. That is, suppose that *label1* is class label *Tapered*, thus, we tested only verifier *v2*. If s_i is accepted, then we considered sperm s_i classified as a tapered sperm head, otherwise s_i is classified as an amorphous one. The same reasoning is applied in all other cases.

In Table 9, we show the mean value of 10 runs of True Positive Rate (TPR) of four classes: N, T, P and S. We search for the best compromise between the accuracy rate (mean of TPR of the four classes) and the rejection rate (percentage of amorphous

sperm heads that are discarded) to achieve the main goals of this stage: to separate amorphous sperm heads from sperm heads from other classes, and to identify the potential class of sperm heads that are not amorphous. Using the maximum probability with a threshold of 0.4 as combination rule, we achieved this best compromise, thus we decided to use it as the combination rule of the outputs for our three different SVMs in this stage.

tpr(N)	tpr(T)	tpr(P)	tpr(S)	acc	rej
0.50	0.63	0.69	0.75	0.64	0.22

Table 9: **Results of the first stage of the classification scheme.** *tpr* stands for True Positive Rate, *N* stands for Normal, *T* stands for Tapered, *P* stands for Pyriform, and *S* stands for Small. *acc* stands for accuracy while *rej* stands for Amorphous rejection rate.

In Table 10, we show the mean value of 10 runs of True Positive Rate (tpr) of two classes: A and one of {N, T, P, S}. The best accuracy rate (mean of TPR of the two classes) was reached using majority voting to combine the outputs for the different SVMs of each verifier in this stage.

Verifier	Not amorphous	Amorphous	Accuracy
Normal VS Amorphous	0.81	0.60	0.70
Tapered VS Amorphous	0.63	0.81	0.72
Pyriform VS Amorphous	0.83	0.64	0.73
Small VS Amorphous	0.81	0.71	0.76

Table 10: **Results of the second stage of the classification scheme.** True positive rate for Amorphous and Not Amorphous sperm heads classification is shown, as well as the accuracy reached in each verifier.

5.5. Influence of the manual labeling on the performance of the automatic system

In order to evaluate the influence of the manual labeling of sperm heads on the performance of the automatic system, i.e. the impact of (dis)agreement between human experts on the classification performance, we performed two different experiments. For a fair comparison, in both experiments, we used the same configuration of the classification scheme, the same training and validating datasets (DS1 and DS2), the same selected features and the same combination rules, following the experimental protocol introduced in Section 5.4. We only varied the testing dataset. In the first case, we used Dataset3 (DS3) for testing purposes. In the second case, we used the special testing dataset (DST). It is important to realize that in the first case, the testing dataset has only partial agreement between experts, while in the last case, the testing dataset has total agreement between experts. We show in the following subsections that better classification performance are achieved, as expected, on easier-to-classifiy sperm heads

(DST dataset), and that in support of our automatic system, there is no difference with a human expert when dealing with partial agreement (DS3 dataset). Finally we put into perspective for clinical purpose the automatic classification of normal vs abnormal sperm heads on the two datasets.

5.5.1. Impact of expert (dis)agreement on classification performance

For comparing the performance of our classification system on DS3 and DST, we did 30 runs on each dataset and present the mean of the True Positive Rate for each class in Table 11. We show in Figure 9 a graphical representation of this comparison.

From the table, we can see that the correct classification for normal sperm heads reaches 62% if we use DS3 as the testing dataset; this increases to 74% if we use DST as the testing dataset. A similar situation occurs with tapered and small sperm heads, which are the easiest class to discriminate: in the case of tapered sperm heads, the correct classification ranges from 64% to 70%, using DS3 and DST, respectively; for small sperm heads, our proposed method achieves the best compromise between DS3 and DST testing datatsets, reaching 82% and 100%, respectively. The situation is kind of different with pyriform sperm heads, because achieving around 50% of correct classification using DS3 becomes 92% when using DST. Even if the difference between using DS3 and DST as testing datasets is not significant, this difference can be explained according to what really happens with pyriform sperm heads in manual classification of experts. When looking at the agreement between experts for identifying pyriform sperm heads [47], we observed that there were very few pyriform sperm heads in which all experts agree in manual classification, thus, it is supposed that for those sperm heads remaining in DST dataset, they are better characterized and easier to discriminate by our classification approach. The last case is about amorphous sperm heads, the most difficult class, where our proposal achieves around 30%, with no regard as to which testing dataset is used, showing the complexity of correctly identifying this type of sperm head.

Testing dataset	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
DS3	0.62	0.64	0.50	0.82	0.30	0.58
DST	0.74	0.70	0.92	1.00	0.30	0.73

Table 11: **Results of our proposed classification scheme.** *tpr* stands for True Positive Rate, *N* means normal, *T* means tapered, *P* means pyriform, *S* means small, and *A* means amorphous. *acc* stands for accuracy understood as mean of tpr of five classes.

In overall these results show clearly that when a total agreement can be reached on the manual labeling of sperm heads (DST dataset), i.e. on easier-to-classify sperm heads, our classification system exhibits a better accuracy than when human experts disagree (DS3 dataset). However, there is still room for improvement even for the easiest sperm heads: we only achieved 73% of accuracy when human experts fully agree on the class of the sperm heads.

In support for our classification system, we also compared our results with those obtained from referent domain experts. Figure 10 shows the inter-expert and automatic

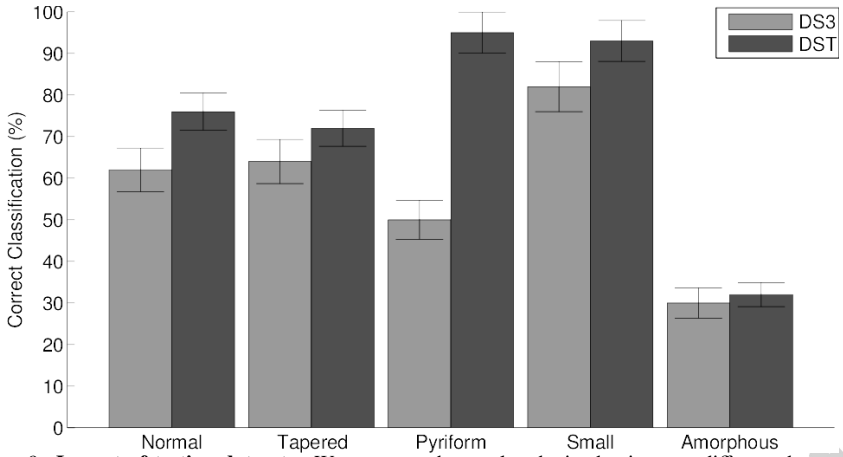


Figure 9: **Impact of testing datasets**. We compare the results obtained using two different datasets for testing: DS3 (Dataset 3 with partial agreement among experts) and DST (Special Testing Dataset with total agreement among experts). The percentage of correct classification ± SE for each class is shown.

classification variability in five-class classification: our proposed approach is confused with domain experts according to Fisher's exact test for some classes ($*p < 0.01$), using DS3 as testing dataset. The near results appear in pyriform and tapered classes, in which our proposal could be confused with a human expert. The confusing classes are normal, small and amorphous. These three classes have slight inter-class variations. For instance, the main difference between a small and a normal sperm head is often only the size, but the difference is poorly related to shape. Comparing normal against amorphous sperm heads, the slight shape variations are very notorious in a visual analysis. Furthermore, the intra-class variation in amorphous class is actually high, with many possibilities of confusion among experts [47].

5.5.2. Two-class classification for clinical purpose

For clinical purposes, the classification of human sperm heads can be reduced to a *two-class classification*, regarding only normal and abnormal (including tapered, pyriform, small and amorphous sperm heads), the results are summarized in Table 12. In this case, we can see that the correct classification as normal sperm heads reaches 62% while for abnormal sperm heads reaches 57% when using DS3 as the testing dataset. The results are much better if we use DST as the testing dataset, getting 74% for normal sperm heads and 73% for abnormal ones. In this case, we also did 30 runs for each experiment and presented the mean of the True Positive Rate for each class in the 30 runs. These results show that even the two-class classification problem is difficult, furthermore when trying to identify *not* normal sperm heads.

In the case of two-class classification, Figure 11 shows the inter-expert and automatic classification variability where our approach is confused with all domain experts according to Fisher's exact test for both classes ($p < 0.01$), using DS3 as testing dataset.

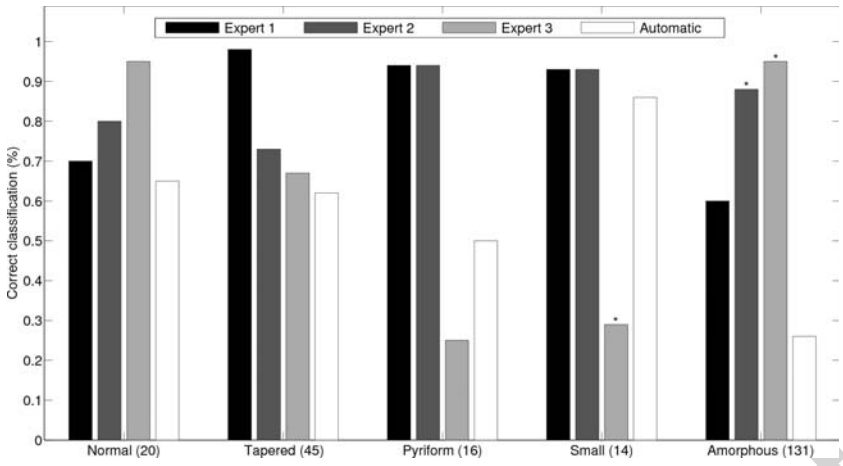


Figure 10: **Inter-expert and automatic classification variability in five-class classification.** For each class, the classification accuracy rate is shown according to each expert and our approach, using DS3 as testing dataset. Our proposed approach (denoted as Automatic) is confused with a number of domain experts in most of classes, however our approach presents statistical significant differences with Expert 3 (classes Small and Amorphous) and Expert 2 (class Amorphous) according to Fisher's exact test ($*p < 0.01$), using DS3 as testing dataset. For each class, the count of sperm cells is shown next to the name of the class.

Testing dataset	tpr(N)	tpr(Ab)	acc
DS3	0.62	0.57	0.60
DST	0.74	0.73	0.74

Table 12: **Results of the whole classification scheme for two-class classification.** *tpr* stands for True Positive Rate, *N* means normal, and *Ab* means abnormal. *acc* stands for accuracy understood as mean of tpr of two classes.

6. Summary and conclusions

In this paper, we have introduced a two-stage classification scheme for classifying human sperm head in five classes (normal, tapered, pyriform, small and amorphous), according to WHO criteria [2]. The approach of combining classifiers together with an ensembled feature selection technique, yields a method for characterizing and classifying sperm heads towards an accurate morphological sperm analysis. We have also presented a new characterization for human sperm heads, named *morphological descriptor*. This descriptor adopts and adapts a number of ROI shape-based measures focusing on ellipse fitness and symmetry. Special emphasis was given to a continuous representation of the curve defining the outline of a head rather than a discrete representation based on pixels. Our experimental evaluation shows that our proposed scheme outperforms a number of monolithic classifiers. Our results achieved more than 70% of classification accuracy on a dataset with total agreement among domain experts, showing that the results of our classification scheme could be easily confused

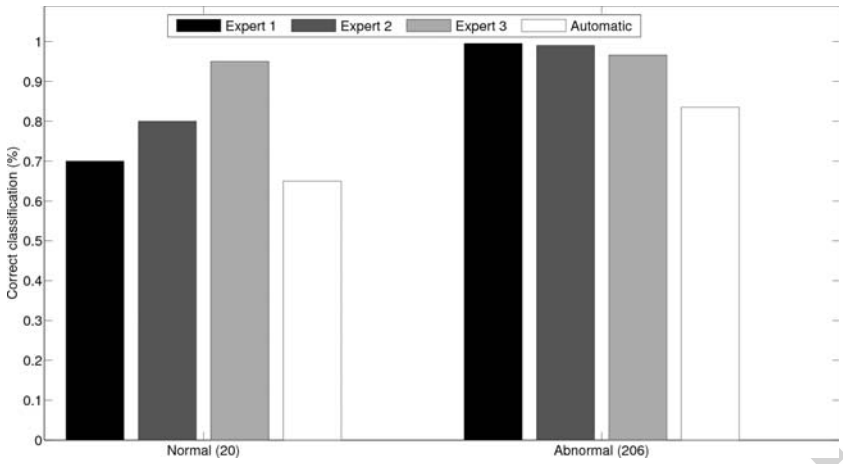


Figure 11: **Inter-expert and automatic classification variability in two-class classification.** For each class, the classification accuracy rate according to each expert and our approach is shown, using DS3 as testing dataset. Our proposed approach (denoted as Automatic) achieves results without statistical significant differences to those of referent experts for both classes, according to Fisher's exact test ($p < 0.01$), using DS3 as testing dataset. For each class, the count of sperm cells is shown next to the name of the class.

with those of a human expert.

We conclude that there is room for improvement even for the easiest sperm heads: we only achieved 73% if looking at the accuracy rates while using a testing dataset with full agreement among experts. This is a very difficult classification problem that gets harder due to the uncertainty of the class label of each sperm head, taking into account the high degree of variability among domain experts.

This paper suggests several directions for future research. First, we plan to explore and experiment with univariate feature selection, specifically feature ranking as a way to select the features for each single classifier in our proposed classification scheme, instead of ranking whole families of features. Second, as the integration of multiple classifiers to improve classification results is currently an active research area in the machine learning community, we plan to continue our research in this area. We are interested in finding out the impact in classification rates when using different base classifiers, such as naive-Bayes classifiers, decision trees and neural networks. Finally, future research focused on the applying of Convolutional Neural Networks (CNN) will be interesting. CNN has demonstrated its suitability in hard classification scenarios with slight intra-class variability, showing promising experimental results in different domain classification problems.

7. Acknowledgements

The authors thank J. Jara for support with the active contour function implementation. The authors thank J. Saavedra for support with the anisotropic diffusion and ellipse fitness function implementation. The authors thank J. Saavedra, L. Oliveira, A. Britto, A. Koerich, S. Bernard and V. Castañeda for invaluable discussion of the

classification scheme. This research (V. Chang, N. Hitschfeld, L. Heutte and C. Petitjean) was funded by CONICYT (STIC-AMSUD 14STIC-01). Violeta Chang was partially funded by CONICYT (NAC-DoctoradoLatin 57090057/ NAC-ApoyoTesis 24100118) and FONDECYT (Postdoctorado 3160559). Research in SCIAN-Lab (S. Härtel) is funded by FONDECYT 1151029, CONICYT PIA ACT 1402, ICM P09-015-F, CORFO 16CTTS-66390 (Chile) and DAAD 57220037 and 57168868 (Ger).

References

[1] WHO, Mother or nothing: the agony of infertility, World Health Organization Bulletin 88 (12) (2010) 881–882.

[2] WHO, World Health Organization - Laboratory Manual for the Examination and Processing of Human Semen, 5th Edition, Cambridge University Press, 2010.

[3] J. Auger, Assessing human sperm morphology: top models, underdogs or biometrics?, Assian Journal of Andrology 12 (1) (2010) 36–46.

[4] D. Katz, J. Overstreet, S. Samuels, P. Niswander, T. Bloom, E. Lewis, Morphometric analysis of spermatozoa in the assessment of human male fertility, Journal of Andrology 7 (4) (1986) 203–210.

[5] F. H. Bronson, Mammalian Reproductive Biology, 1st Edition, University of Chicago Press, Chicago, USA, 1991.

[6] M. Barone, M. Roelke, J. Howard, J. Brown, A. Anderson, D. Wildt, Reproductive characteristics of male florida panthers: Comparative studies from florida, texas, colorado, latin america, and north american zoos, Journal of Mammalogy 75 (1) (1994) 150.

[7] A. Jung, H.-C. Schuppe, W.-B. Schill, Comparison of semen quality in older and younger men attending an andrology clinic, Andrologia 34 (2) (2002) 116–122.

[8] S. A. Kidd, B. Eskenazi, A. J. Wyrobek, Effects of male age on semen quality and fertility: a review of the literature, Fertility and Sterility 75 (2) (2001) 237 – 248.

[9] A. J. Wyrobek, B. Eskenazi, S. Young, N. Arnheim, I. Tiemann-Boege, E. W. Jabs, R. L. Glaser, F. S. Pearson, D. Evenson, Advancing age has differential effects on dna damage, chromatin integrity, gene mutations, and aneuploidies in sperm, in: Proceedings of the National Academy of Sciences of the United States of America, Vol. 103, 2006, pp. 9601–9606.

[10] G. Moench, H. Holt, Sperm morphology in relation to fertility, American Journal of Obstetrics and Gynecology 22 (1) (1931) 199–210.

[11] J. MacLeod, R. Gold, The male factor in fertility and infertility – Sperm morphology in fertile and infertile marriage, Fertility and Sterility 2 (1) (1951) 394–414.

[12] T. Kruger, R. Menkveld, F. Stander, C. Lombard, J. Van der Merwe, J. van Zyl, K. Smith, Sperm morphologic features as a prognostic factor in in-vitro fertilization, Fertility and Sterility 46 (6) (1986) 1118–1123.

[13] M. Enginsu, J. Dumoulin, M. Pieters, M. Bras, J. Evers, J. Geraedts, Evaluation of human sperm morphology using strict criteria after diff-quik staining: correlation of morphology with fertilization in vitro, Human Reproduction 6 (6) (1991) 854–858.

[14] T. Kobayashi, M. Jinno, K. Sugimura, S. Nozawa, T. Sugiyama, E. Iida, Sperm morphological assessment based on strict criteria and in-vitro fertilization outcome, Human Reproduction 6 (7) (1991) 983–986.

[15] T. Kruger, T. du Toit, D. Franken, A. Acosta, S. Oehninger, R. Menkveld, C. Lombard, A new computerized method of reading sperm morphology (strict criteria) is as efficient as technician reading, Fertility and Sterility 59 (1) (1993) 202–209.

[16] C. Freund, Standards for the rating of human sperm morphology. A cooperative study, International Journal of Fertility 11 (1) (1966) 97–180.

[17] R. Walczak-Jedrzejowska, K. Marchlewska, E. Oszukowska, E. Filipiak, L. Bergier, J. Slowikowska-Hilczer, Semen analysis standardization: is there any problem in polish laboratories?, Asian Journal of Andrology 15 (2013) 616–621.

[18] A. Rivera-Montes, A. Rivera-Gallegos, E. Rodríguez-Villasana, A. Juárez-Bengoa, M. Díaz-Pérez, M. Hernández-Valencia, Estimate of the variability in the evaluation of semen analysis, Ginecologia y Obstetricia de Mexico 81 (11) (2013) 639–644.

[19] G. Barroso, R. Mercan, K. Ozgur, M. Morshedi, P. Kolm, K. Coetzee, T. Kruger, S. Oehninger, Intra- and inter-laboratory variability in the assessment of sperm morphology by strict criteria: impact of semen preparation, staining techniques and manual versus computerized analysis, Human Reproduction 14 (8) (1999) 2036–40.

[20] J. Auger, F. Eustache, B. Ducot, T. Blandin, M. Daudin, I. Diaz, S. Matribi, B. Gony, L. Keskes, M. Kolbezen, A. Lamarte, J. Lornage, N. Nomal, G. Pitaval, O. Simon, I. Virant-Klun, A. Spira, P. Jouannet, Intra- and inter-individual variability in human sperm concentration, motility and vitality assessment during a workshop involving ten laboratories, Human Reproduction 15 (11) (2000) 2360–2368.

[21] C. Soler, J. de Monserrat, R. Gutiérrez, J. Nuñez, M. Nuñez, M. Sancho, F. Pérez-Sánchez, T. Cooper, Use of the Sperm-Class Analyzer for objective assessment of human sperm morphology, International Journal of Andrology 26 (5) (2003) 262–270.

[22] A. Cipak, P. Stanić, K. Durić, T. Serdar, E. Suchanek, Sperm morphology assessment according to WHO and strict criteria: method comparison and intra-laboratory variability, Biochemia Medica 19 (1) (2009) 87–94.

[23] C. Wang, A. Leung, W. Tsoi, J. Leung, V. Ng, K. Lee, S. Chan, Computer-assisted assessment of human sperm morphology: comparison with visual assessment, Fertility and Sterility 55 (5) (1991) 983–988.

[24] R. Davis, C. Gravance, Standardization of specimen preparation, staining, and sampling methods improves automated sperm-head morphometry analysis, Fertility and Sterility 59 (2) (1993) 412–417.

[25] F. Lacquet, T. Kruger, T. du Toit, C. Lombard, C. Sánchez Sarmiento, A. de Villiers, K. Coetzee, Slide preparation and staining procedures for reliable results using computerized morphology, Archives of Andrology 36 (2) (1996) 133–138.

[26] K. Coetzee, T. Kruger, C. Lombard, Repeatability and variance analysis on multiple computer-assisted (IVOS) sperm morphology readings, Andrologia 31 (3) (1999) 163–168.

[27] L. Sánchez, N. Petkov, Similarity-Based Clustering, Springer-Verlag, 2009, Ch. Estimation of Boar Sperm Status Using Intracellular Density Distribution in Grey Level Images, pp. 169–184.

[28] E. Alegre, V. González-Castro, R. Alaiz-Rodríguez, M. García-Ordás, Texture and moments-based classification of the acrosome integrity of boar spermatozoa images, Computer Methods and Programs in Biomedicine 108 (2) (2012) 873–881.

[29] V. González-Castro, E. Alegre, O. García-Olalla, D. García-Ordás, M. García-Ordás, L. Fernández-Robles, Curvelet-based texture description to classify intact and damaged boar spermatozoa, in: Image Analysis and Recognition, Vol. 7325 of Lecture Notes in Computer Science, 2012, pp. 448–455.

[30] E. Alegre, M. Biehl, N. Petkov, L. Sánchez, Assessment of acrosome state in boar spermatozoa heads using n-contours descriptor and RLVQ, Computer Methods and Programs in Biomedicine 111 (3) (2013) 525–536.

[31] O. García-Olalla, E. Alegre, L. Fernández-Robles, P. Malm, E. Bengtsson, Acrosome integrity assessment of boar spermatozoa images using an early fusion of texture and contour descriptors, Computer Methods and Programs in Biomedicine 120 (1) (2015) 49–64.

[32] M. Beletti, L. Costa, M. Viana, A spectral framework for sperm shape characterization, Computers in Biology and Medicine 35 (6) (2005) 463 – 473.

[33] M. Beletti, L. Costa, M. Viana, A comparison of morphometric characteristics of sperm from fertile Bos taurus and Bos indicus bulls in Brazil, Animal Reproduction Science 85 (2005) 105 – 116.

[34] L. Severa, L. Máchal, L. Švábová, O. Mamica, Evaluation of shape variability of stallion sperm heads by means of image analysis and Fourier descriptors, Animal Reproduction Science 119 (12) (2010) 50 – 55.

[35] V. Abbiramy, A. Tamilarasi, A comparative study on human spermatozoa images classification with artificial neural network based on FOS, GLCM and morphological features, in: Advances in Digital Image Processing and Information Technology, Vol. 205 of Communications in Computer and Information Science, Springer Berlin Heidelberg, 2011, pp. 220–228.

[36] V. Chang, J. Saavedra, V. Castañeda, L. Sarabia, N. Hitschfeld, S. Härtel, Gold-standard and improved framework for sperm head segmentation, Computer Methods and Programs in Biomedicine 117 (2) (2014) 225 – 237.

[37] R. Ranawana, V. Palade, Multi-classifier systems: Review and a roadmap for developers, International Journal of Hybrid Intelligent Systems 3 (1) (2006) 35–61.

[38] A. Britto Jr., R. Sabourin, L. Oliveira, Dynamic selection of classifiersa comprehensive review, Pattern Recognition 47 (11) (2014) 3665 – 3680.

[39] T. Dietterich, Ensemble methods in machine learning, in: Multiple Classifier Systems, Vol. 1857 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2000, pp. 1–15.

[40] P. Perona, J. Malik, Scale-space and edge detection using anisotropic diffusion, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (7) (1990) 629–639.

[41] L. Cohen, I. Cohen, Finite-element methods for active contour models and balloons for 2D and 3D images, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (11) (1993) 1131–1147.

[42] M. Peura, J. Iivarinen, Efficiency of simple shape descriptors, in: Proceedings of the 3rd International Workshop on Visual Form, 1997, pp. 443–451.

[43] J. Saavedra, B. Bustos, Sketch-based image retrieval using keyshapes, Multimedia Tools and Applications 73 (3) (2014) 2033–2062.

[44] C. Zahn, R. Roskies, Fourier descriptors for plane closed curves, IEEE Transactions on Computers C-21 (3) (1972) 269–281.

[45] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Transactions on Information Theory 8 (2) (1962) 179–187.

[46] M. Teague, Image analysis via the general theory of moments, Journal of the Optical Society of America 70 (8) (1980) 920–930.

[47] V. Chang, A. Garcia, N. Hitschfeld, S. Härtel, Gold-standard for computer-assisted morphological sperm analysis, Computers in Biology and Medicine 83 (2017) 143–150.

[48] E. Rahtu, M. Salo, J. Heikkila, A new convexity measure based on a probabilistic interpretation of images, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (9) (2006) 1501–1512.

[49] M. Aktas, J. Žunić, Sensitivity/robustness flexible ellipticity measures, in: Pattern Recognition, Vol. 7476 of Lecture Notes in Computer Science, 2012, pp. 307–316.