FEDERAL TECHNOLOGICAL UNIVERSITY OF PARANÁ
POSTGRADUATE PROGRAM IN COMPUTATIONAL TECHNOLOGIES
FOR AGRIBUSINESS

MARCELA MARQUES BARBOSA

# IDENTIFICATION OF FISH SPECIES: A computational approach using techniques of digital image processing and artificial intelligence

DISSERTATION

**MEDIANEIRA**

**2017**

MARCELA MARQUES BARBOSA

# IDENTIFICATION OF FISH SPECIES: A computational approach using techniques of digital image processing and artificial intelligence

Dissertation presented as a partial requirement to obtain the degree of Master of Postgraduate Program in Computational Technologies for Agribusiness, Federal Technological University of Paraná. Area of Concentration: Computational Technologies Applied to Agribusiness

Advisor Drª Saraspathy N T G De Mendonca

Co-Advisor Prof. Dr. Pedro Luiz de Paula Filho

**MEDIANEIRA**

**2017**

Ministry of Education
**Federal Technological University of Paraná**
Board of Medianeira *Campus*
Teaching and Research Management
**Postgraduate Program in Computational Technologies for Agribusiness**

## TERM OF APPROVAL

**IDENTIFICATION OF FISH SPECIES: A computational approach using techniques of digital image processing and artificial intelligence**

Per

**Marcela Marques Barbosa**

This Dissertation was presented at 14:00 on November 7, 2016 as a requirement partial in order to obtain a Master's Degree in the Master no Postgraduate Program in Computational Technologies for Agribusiness, da Federal Technological University of Paraná, Câmpus Medianeira. The candidate was accused by the examining bank composed of the professors below. After deliberation, the examining board considered the work approved.

Prof. Invited 1
Institution

Prof. Invited 2
Institution

Prof. Dr. Pedro Luiz de Paula Filho
UTFPR - Câmpus Medianeira
Co-Advisor

Drª Saraspathy N T G De Mendonca
UTFPR - Câmpus Medianeira
Advisor

Coordination Visa:

Prof. Dr. Cláudio Leones Bazzi
UTFPR - Câmpus Medianeira
(Coordenador of the PPGTCA)

Dedication of work (optional).

## ACKNOWLEDGMENT

For this moment to materialize, I owe thanks to many people, so many that a sheet here would be little, in a summarized way I will try to thank everyone.

For first I will be eternally grateful to UTFPR itself, which opened the doors giving me the opportunity to study in one of the best public institutions in Brazil. I thank you not only for the financial aid, which, incidentally, was fundamental for this moment to happen, but above all I thank you immensely for the warmth, the human warmth to me dispensed in the great majority by your teachers and servants and workers in general of the UTFPR. I am quite convinced that it would not have happened the other way.

By the second, I want to thank all my teachers, and those who in one way or another, even though not being my teacher, have always been on my side and taught me everything that was possible. Here the list is great, beginning with the teachers of mathematics, I owe a lot of thanks to Professor Fausto, they were years of coexistence in design, Professor Diego, Professor Cleverson, Professors Neusa and Rafaela, Professor Lucas. Many thanks also to my beloved computer teachers, Claudio Bazzi, Cezar Angonese, Nelson Betzek, Paulo Job, Alessandra Hoffmann, Fernando Schutz, Allan Gavioli, Paulo Lopes, Márcio Matté, Evandro Pessini, Hamilton Pereira, Jorge Aikes, Juliano Lamb, Arnaldo Candido, Neylor Michel, Patrícia Lopez, Ricardo Sobjack, Everton Coimbra, José Airton, Silvana Mendonça, Lairton Moacir, Elias Lira. I am and will forever be grateful to you all.

Besides my dear teachers, I would like to thank my classmates, who helped me a lot during the course, among all, which are not many, I would like to thank Gabriela Michellon and Samanta de Sousa very much. Last, companion not only of the academic works but also companion for the uncertain hours. Thank you very much to all of you.

Thirdly, I would like to thank those responsible for this project, Drª. Tania, Professor Dr Pedro and Professor Drª. Deisy, who have entrusted me with this project, giving me the opportunity to contribute to a significant project. Thank you very much, it was a unique and rewarding experience to have had the honor of working with you.

In particular, I would like to thank Dr. Arnaldo Candido for his patience always dedicated to me, not only regarding doubts regarding the subjects, but also incredible help in the articles, at the conclusion of this work and going beyond the academic subjects, giving me strength in the hours that But I needed it. Here is all my gratitude and admiration for Professor Dr. Arnaldo.

Epigraph of work (optional).

## ABSTRACT

BARBOSA, Marcela Marques. IDENTIFICATION OF FISH SPECIES: A COMPUTATIONAL APPROACH USING TECHNIQUES OF DIGITAL IMAGE PROCESSING AND ARTIFICIAL INTELLIGENCE. 19 f. Dissertation – Postgraduate Program in Computational Technologies for Agribusiness, Federal Technological University of Paraná. Medianeira, 2017.

The study and understanding of the genetic material of the species enables humans to better understand the characteristics of the species so that they can be stored as a source of species preservation, as well as to propose genetic improvements that can eradicate diseases. In the fish population, there is a great variety in its number of chromosomes, due to the great variety of fish known today. The objectives are to identify and separate the highlighted segments in the coloring, and through them to extract characteristics in order to classify it. For the development of the software will be used the programming language C ++, using the open source library Open Computer Vision. It will be used the similarity and dissimilarities method, which appropriates the concept of distances, Euclidian, Manhattan, Mahalanobis, Simple Wedding Coefficient and Jaccard Coefficient.



**Keywords:** petri, microorganisms, food, counting, image

# RESUMO

BARBOSA, Marcela Marques. IDENTIFICAÇÃO DE ESPÉCIES DE PEIXES: Uma abordagem computacional utilizando técnicas de processamento digital de imagens e inteligência artificial. 19 f. Dissertation – Postgraduate Program in Computational Technologies for Agribusiness, Federal Technological University of Paraná. Medianeira, 2017.

O estudo e a compreensão do material genético das espécies propicia aos seres humanos compreender melhor as características das espécies para que possam ser armazenado como fonte de preservação das espécies, bem com propor melhorias genéticas que possam erradicar doenças. Na população de peixes, existe uma grande variedade em seu número de cromossomos, devido a grande variedade de peixes conhecida atualmente. Os objetivos são de identificar e separar os segmentos destacados na coloração, e através dos mesmos extrair características afim de classifica-lo. Para o desenvolvimento do software será utilizada a linguagem de programação C++, utilizando a biblioteca opensource Open Computer Vision. Será utilizado o método de similaridade e dissimilaridades, que se apropriam do conceito de distancias, Euclidiana, Manhattan, Mahalanobis, Coeficiente de Casamento Simples e Coeficiente de Jaccard.

**Palavras-chave:** petri, micro-organismos, alimentos, contagem, imagens

# LIST OF FIGURES

# LIST OF TABLES

## LIST OF ACRONYMS

| | |
|---|---|
| Opencv | Open Computer Vision |
| (RAM) | Random Access Memory |

**SUMMARY**

# 1 INTRODUCTION

The study of the karyotype is an important source of information capable of distinguishing individuals from several species, since they are differentiated by their chromosome quantification as well as their morphology. This analysis allows that alteration can be suggested with the intention of manipulating the chromosomes, providing genetic improvements, contributed to the conservation of the species (AGOSTINHO et al., 2006).

Unlike the human karyotype, where there are a constant number of pairs of chromosomes, 23 pairs, 22 of which are autosomes (non-sexual) and a pair of allosomes (sexual). It is worth regreting that this number of 23 homologous pairs is not always satisfied, in case of down syndrome and tunner syndrome. In the fish population, there is a great variety in its number of chromosomes, due to the great variety of fish known today. Even karyotypes with different numbers of pairs of chromosomes were found. Another important feature of the karyotype of the fish distinct from the human karyotype is that in fish the pair of chromosomes are not always contradict, or it may be found in different forms (POVH et al., 2005).

For the study of the karyotype to be possible, it is necessary to obtain an image of the chromosomes, when they are in metaphase (when they are more condensed), at this stage they are better visualized. To obtain this image, they can be performed in two ways, the first and most common is the so-called “conventional coloring”, which give a uniform color to the chromosomes. The second is called “differential staining”, in which the technique of ”chromosome banding” is used, in which only or mainly chromatin region. Differential staining is only possible because some of the chromosomes are composed of segments, called “bands”. This feature allows the identification of homologous pairs and an analysis of their evolution. It is possible to perform the direct coloring in several ways. Something in common between them and that good results are not always achieved (GUERRA; SOUZA, 2002).

In this sense, observing that such images present color differentiation, it becomes possible to develop a software that uses digital image processing techniques, as it is possible to differentiate such highlighted segments, to verify its context in relation to the chromosome (as to its classification ), And perform extraction of characteristics: morphological; Statistical and texture for each segment. Such characteristics may provide sufficient information to

characterize each segment individually and thus make the karyotype distinguishable between species by their degree of dissimilarity.

## 1.1 PROBLEM

Is it possible to identify the species to which the fish belong in an automated way through the image of a karyotype?

## 1.2 OBJECTIVES

The objectives are to identify and separate the highlighted segments in the coloring, and through them to extract characteristics in order to classify it. To this end, it should be one of the objectives of this study, to carry out a rigorous and precise study on such characteristics that make them different from the other segments.

### 1.2.1 General objective

The general objective is to develop a software with low cost technologies that allow the identification of fish species through the image of their karyotype (idiogram)

### 1.2.2 Specific objectives

- Extrair característica morforlógicas dos cromossomos e do cariótipo;
- Implement and train a convolutional neural network from the extracted features;
- Validate the neural network.

## 1.3 JUSTIFICATION

Em laboratórios de microbiologia comumente necessitam-se fazer a avaliação da qualidade e da segurança dos alimentos, para isso realizam a contagem visual de micro-

organismos, em meio de cultura previamente conhecido. Essas contagens também são realizadas quando há a necessidade de adição de micro-organismos em produtos alimentícios, por exemplo a adição de bactérias ácido lática em produtos lácteos fermentados, com a finalidade de estimar a concentração bacteriana adicionada ao produto. Atualmente essas contagens são realizadas a olho nu pelo pesquisador, por meio de um contador de colônias munido de uma lupa de aumento, um marcador de colônias e quadrantes que facilitam a visualização e posterior contagem e estimativa do número de colônias na placa.

A estimativa da concentração de bactérias é determinada através da construção da curva de concentração de bactérias por meio da correlação dos resultados obtidos pela absorvância e o logaritmo da concentração de bactérias obtida por contagem das placas após a inoculação das culturas em meio específico e a incubação por termpo e temperatura específicos. Todos os experimentos são realizados em duplicata ou triplicata.

A contagem a olho nu também limita a estimativa de crescimento de colônias após um determinado tempo de incubação, em que, devido ao aumento do crescimento microbiano e consequente formação de unidades formadoras de colônias torna-se impossível a contagem, havendo a necessidade de diluição do meio, aumentando desta forma o tempo de análise e a imprecisão na contagem final das bactérias e consequentes erros quanto à contaminação do alimento ou construção da curva de crescimento bacteriano.

## 1.4 HYPOTHESIS

Utilizando tecnologias de segmentação de imagem é possível realizar a contagem e a classificação das bactérias em placas de petri.

# 2 MATERIALS AND METHODS

The materials related to the research will be given initially by the acquisition of the images, which will be acquired through the differential coloring method. For software development, the C ++ programming language will be used, using the Open Computer Vision (Opencv) version 3.1 library, which has implementations of algorithms for digital image processing and computational vision, together with the Qt development environment in Its opensource version, which facilitates the development of software with a graphical interface, facilitating its use by the user.

It will be used the similarity and dissimilarities method, which appropriates the concept of distances, Euclidian, Manhattan, Mahalanobis, Simple Wedding Coefficient and Jaccard Coefficient. Thus, measures of dissimilarity similarity allow the classification of the karyotype by their proximity or not of each other (BIOINFORMáTICA EXAMES, 2015).

# 3 EXPECTED RESULTS

It is expected that at the end of the whole process, the developed software will be able to identify which fish species belongs to this karyotype in the image. However, given the characteristic of the problem, it is known from the outset that it is not possible to obtain 100% recognition of the test base, however, efforts will be devoted to identifying positively the closest of the totality. In this way, it is expected that such research will produce significant results, increasing interest in the topic addressed by this study, and encouraging further work, this time more comprehensive, to be developed

# 4 FINAL CONSIDERATIONS

Foi possível definir a padronização na aquisição das imagens que formaram a base. Esta padronização se mostrou suficiente no que diz respeito a posição, distância, iluminação e posicionamento da câmera fotográfica. Essas características permitiram a definição de um algorítimo que se mostrou eficiente para localização da placa de petri em questão, obtendo uma pequena margem de erro em poucas placas, sem comprometer nenhuma placa em sua totalidade, inclusive se mostrando flexível quanto ao deslocamento da mesma dentro da imagem. Apesar de apresentar pequenas falhas que não comprometeram a identificação da placa e nem a contagem das UFCs de bactérias, foi possível observar que alguns cuidados no ato da aquisição da imagem podem evitar alguns problemas tais como: retirar a tampa da placa de petri, a presença da tampa ocasiona brilho excessivo nas bordas da placa de petri, fazendo com que a borda tenha valores de saturação semelhantes as UFCs de bactérias, dificultando assim a identificação das mesma nessa região; Outro cuidado refere-se ao fundo utilizado, que deve conter o mínimo possível de riscos ou sujeira, evitando assim que ruídos sejam adicionados as imagens. Estes ruídos podem interferir na contagem das UFCs de bactérias pequenas, causando falsos negativos em placas com UFCs de bactérias pequenas e falsos positivos em placas com pouca quantidade de UFCs de bactérias.

Para a contagem das UFCs de bactérias, foram estudadas técnicas de processamento de digital como, algoritmos de pré-processamento das imagens, para realçar pontos de interesse e minimizar os ruídos e técnicas de segmentação que propiciassem a separação de fundo e a identificação das UFCs de bactérias. Para o devido realce das áreas de interesse foram estudadas técnicas como *blur*, *gaussian blur*, filtro bilateral e filtro de mediana, tendo o último se mostrado mais útil devido ao caráter variável tanto das placas de petri bem como as próprias UFCs de bactérias, por eliminar uma boa quantidade de ruídos sem comprometer as UFCs de bactérias de tamanho reduzido em placas com alta densidade de UFCs de bactérias. Para a segmentação foram estudados espaços de cores, sendo identificado que os canais de cores da família H apresentam melhores resultados, se destacando o canal HLS. Além do espaço de cores também foram utilizadas técnicas de morfologia matemática tais como: erosão; dilatação; abertura; fechamento; *top hat* e *black hat*. Todas essas técnicas se mostraram úteis tanto na

identificação da placa de petri como a própria identificação das UFCs de bactérias. E por fim foram estudas técnicas de limiarização, aqui vale destacar seu valor ambíguo, sendo igualmente usada na fase de realce quanto na de segmentação. Para a abordagem utilizada nesse trabalho tais técnicas se mostram suficiente, pois permitiram de modo satisfatório uma segmentação possível baseando-se na saturação e luminosidade das imagens. Porém é importante ressaltar que esse conjunto de técnicas não abrangem as diversas características das imagens, por tanto, a adição de outras técnicas pode complementar este trabalho.

Dentre as possíveis abordagem para resolução do problema proposto em questão, a abordagem baseada apenas em técnicas de processamento de imagens, centrando-se na saturação e brilho das imagem foi a que se mostrou mais viável, não sendo definitiva. Tomando a contagem manual como referência, foi possível observar que a contagem automática obteve uma forte correlação em comparação com a contagem manual, segundo o cálculo de *pearson* apresentado o valor de 0.9486 e erro absoluto médio de 0.2243, levando em consideração apenas os tempos por horas (média dos logs). Nesse ponto é importante destacar que esse valor não está levando em consideração a margem de erro e também é necessário considerar a baixa quantidade amostral (apenas 16 tempos distintos). Essa ressalva fica ainda mais evidenciada quando é feita uma análise mais detalhada, comparando as contagens individualmente. Para esses casos os resultados demonstram uma queda significativa no cálculo de *pearson*, evidenciando uma correlação bem inferior e o erro médio absoluto apresenta valores relativamente altos.

Com tudo, é importante destacar o desempenho do *software* no que se refere ao tempo. Para que se obtenha a curva de crescimento do micro-organismo são necessárias inocular e incubar 45 placas em duplicatas, somando 90 placas a serem contadas. Desta forma um técnico analista de laboratório de microbiologia qualificado e experiente pode levar mais de 8 horas de trabalho, ou seja, mais de um dia, para realizar toda a contagem e todas as 90 placas. Para que pudesse ser executado pelo *software* foi obtida duas fotos de cada placa, totalizando 180 imagens e utilizou-se um *notebook* com processador core i3, 12 *gigabytes* de memória *Random Access Memory* (RAM) de configurações, no qual realizou a contagem das 180 imagens com 15 *megapixeis* em 14 min e 54 segundos. A diferença entre os tempos é realmente significativa, mostrando que a utilização do *software* pode agregar benefícios ao laboratório de microbiologia, uma vez que além do ganho no tempo, a utilização do *software* não exige conhecimento prévio, podendo ser operado por qualquer pessoa.

Considerando os resultados apresentados neste trabalho concluí-se que a abordagem baseada no brilho e saturação das imagens apresentou-se em um primeiro momento suficiente para inicio de estudo, porém é de fundamental importância enfatizar que tais resultados foram obtidos a partir de uma base de imagens pequena, e que se faz necessário estudo posteriores,

com base de imagens maiores afim de verificar tais resultados aqui apresentados se fazem igualmente satisfatórios. Este trabalho de alguma forma contribui para o inicio de um trabalho mais abrangente, deixando como sugestão para trabalhos futuros o uso de técnicas de visão computacional, como *Haar Cascade* e *Deep Learning*.

# REFERENCES

AGOSTINHO, A. A.; PELICICE, F. M.; JR., H. F. J. . Biodiversidade e introdução de espécies de peixeis: Unidades de conservação. In: ____. **Unidades de Conservação Ações para valorização da Biodiversidade**. Instituto Ambiental do Paraná, 2006. p. 90–177. Disponível em: <http://www.terrabrasilis.org.br/ecotecadigital/pdf/unidades-de-conservacao>. Acesso em: 26 out 2016.

BIOINFORMáTICA EXAMES. **A computação por trás dos exames de sequenciamento genético**: A era do big data na medicina genômica. 4 2015. Disponível em: <https://www.genomika.com.br/blog/a-computa%C3%A7%C3%A3o-por-tr%C3%A1s-dos-exames-de-sequenciamento-gen%C3%A9tico-a-era-do-big-data-na-medicina-gen%C3%B4mica/>. Acesso em: 26 out 2016.

GUERRA, M.; SOUZA, M. J. de. **Como observar cromossomos**: Um guia de técnicas em citogenética vegetal, animal e humana. FUNPEC - Editora, 2002. ISBN 85-87528-38-6. Disponível em: <http://www.ensp.fiocruz.br/portal-ensp/_uploads/documentos-pessoais/documento-pessoal_52172.pdf>. Acesso em: 26 out. 2016.

POVH, J.; MOREIRA, H.; RIBEIRO, R.; PRIOLI, A.; VARGAS, L.; BLANCK, D.; GASPARINO, E.; JR, D. S. . Estimativa da variabilidade genética em linhagens de tilápia do nilo (oreochromis niloticus) com a técnica de rapd. **Acta Scientiarum. Animal Sciences**, v. 27, n. 1, p. 1–10, 2005. ISSN 1807-8672. Disponível em: <http://revistas.bvs-vet.org.br/actascianimsci/article/view/10755>. Acesso em: 25 out. 2016.