



TECNOLOGÍA
en marcha

Revista trimestral

Marzo 2020

Volumen 33

ISSN 0379-3982 / ISSN-E 2215-3241

Número especial

CARLA 2019

LATIN AMERICA HIGH PERFORMANCE
COMPUTING CONFERENCE

**TURRIALBA
COSTA RICA**



TEC | Tecnológico
de Costa Rica

Publicación y directorio en catálogos

latindex

Dialnet
DOAJ



Comisión Editorial

Ana Ruth Vilchez Rodríguez. Directora.
Instituto Tecnológico de Costa Rica

Juan Antonio Aguilar Garib
Facultad de Ingeniería Mecánica y Eléctrica
Universidad Autónoma de Nuevo León.
México

Carlos Andrés Arredondo Orozco
Facultad de Ingenierías
Universidad de Medellín. Colombia

Lars Köhler
Experimenteller Botanischer Garten
Georg-August-Universität Göttingen.
Alemania

Jorge Solano Jiménez
Instituto Costarricense del Cemento
y del Concreto

Edición técnica

Alexa Ramírez Vega

Revisión filológica

Esperanza Buitrago Poveda

Diseño gráfico

Felipe Abarca Fedullo

Diagramación

Asesoría en Ediciones gráficas

Diseño de cubierta

Felipe Abarca Fedullo

Datos de catalogación en publicación

Tecnología en Marcha / Editorial Tecnológica
de Costa Rica. - Vol. 33, especial. Marzo
(2020) –Cartago: la Editorial, 2020 –
Trimestral
ISSN-E 2215-3241

1. Ciencia y Tecnología –
Publicaciones periódicas CDD:600

**TEC** | Tecnológico
de Costa Rica

Apdo 159-7050 Cartago, Costa Rica
Tel.:(506) 2550-2297, 2550-2618

Correo electrónico: editorial@itcr.ac.cr

Web: editorial.tec.ac.cr

http://revistas.tec.ac.cr/tec_marcha



Editorial Tecnológica
de Costa Rica

TEC | Tecnológico
de Costa Rica

La Editorial Tecnológica de Costa Rica es una dependencia especializada del Instituto Tecnológico de Costa Rica. Desde su creación, en 1978, se ha dedicado a la edición y publicación de obras en ciencia y tecnología. Las obras que se han editado abarcan distintos ámbitos respondiendo a la orientación general de la Institución.

Hasta el momento se han editado obras que abarcan distintos campos del conocimiento científico-tecnológico y han constituido aportes para los diferentes sectores de la comunidad nacional e internacional.

La principal motivación de la Editorial es recoger y difundir los conocimientos relevantes en ciencia y tecnología, llevándolos a los sectores de la comunidad que los requieren.

La revista *Tecnología en Marcha* es publicada por la Editorial Tecnológica de Costa Rica, con periodicidad trimestral. Su principal temática es la difusión de resultados de investigación en áreas de Ingeniería. El contenido de la revista está dirigido a investigadores, especialistas, docentes y estudiantes universitarios de todo el mundo.

Publicación y directorio en catálogos



Revista trimestral.
Especial 2020.
6th Latin America High
Performance Computing
Conference (CARLA)

ISSN 0379-3982 / ISSN-E 2215-3241

TECNOLOGÍA *en marcha*

Contenido

Contribuciones especiales a la Conferencia 6th Latin America High Performance Computing Conference (CARLA) 2019	
<i>Juan Luis Crespo-Mariño, Andrés Segura-Castillo, Esteban Meneses-Rojas</i>	3
Gaussian mixture analysis of basic meteorological parameters: Temperature and relative humidity Análisis de mixturas gaussianas de parámetros meteorológicos básicos: Temperatura y humedad relativa	
<i>Mariela Abdalah-Hernández, Javier Rodríguez-Yáñez, Daniel Alvarado-González</i>	5
Instance segmentation for automated weeds and crops detection in farmlands Segmentación de instancias para detección automática de malezas y cultivos en campos de cultivo	
<i>Adán Mora-Fallas, Hervé Goëau, Alexis Joly, Pierre Bonnet, Erick Mata-Montero</i>	13
MediaTIC: A Social Media Analytics Framework For the Costa Rican News Media MediaTIC: Una plataforma analítica de medios digitales costarricenses en redes sociales	
<i>Cristina Soto-Rojas, Carlos Gamboa-Venegas, Adriana Céspedes-Vindas</i>	18
Evaluating Resilience of Deep Learning Models Evaluando la Resiliencia de Modelos de Deep Learning	
<i>Elvis Rojas, Bogdan Nicolae, Esteban Meneses</i>	25
Proposal for metabolic flux pathways comparison Propuesta para la comparación de flujos metabólicos	
<i>Esteban Arias-Méndez, Alonso Montero-Marín, Francisco J. Torres-Rojas</i>	31
Estimating the redshift of galaxies from their photometric colors using machine learning methods Estimación del corrimiento al rojo para galaxias a partir de sus colores fotométricos usando métodos de aprendizaje automático	
<i>Felipe Meza-Obando</i>	38
Executing and Pausing Distributed Applications Running on Desktop Clouds by Global Snapshots Ejecutando y Pausando Aplicaciones Distribuidas Corriendo sobre Desktop Clouds Mediante Snapshots Globales	
<i>Carlos E. Gómez, Jaime Chavarriaga, David C. Bonilla, Harold E. Castro</i>	44

Understanding Variable Performance on Deep MIL Framework for the Acoustic Detection of Tropical Birds Entendiendo el Desempeño Variable en el Marco de Trabajo MIL Profundo para la Detección Acústica de Aves Tropicales <i>Jorge Castro, Roberto Vargas-Masis, Danny Alfaro-Rojas</i>	49
Motus: A Framework for Human Motion Classification in a Not-controlled Moving Environment Motus: Marco de trabajo para la clasificación de Captura de Movimiento humano en ambientes no controlados <i>Joselyn Rodríguez-González, María Hernández-López, Francisco Siles-Canales</i>	55
Initial Approach on Soccer Match's Scene Classification by Players' Field Spatial Distribution Abordaje inicial en la clasificación de escenas de partidos de fútbol a partir de la distribución espacial de los jugadores sobre la cancha <i>Lennon Núñez-Meño, Marco Villalta, Francisco Siles-Canales</i>	60
Reducing the two dimensional Green functions: Fourier mode decomposition Reduciendo las funciones de Green bidimensionales: Descomposición en modos de Fourier <i>Juan Pablo Mallarino-Robayo, Alejandro Ferrero-Botero</i>	66
Phylogenetic analysis of ITS data from Endophytic fungi using Massive Parallel Bayesian Tree Inference with Exabayes Análisis Filogenético de Secuencias ITS Provenientes de Hongos Endófitos Utilizando Inferencia Bayesiana Paralela de Árboles con Exabayes <i>Maripaz Montero-Vargas, Jean Carlo Umaña-Jiménez, Efraín Escudero-Leiva, Priscila Chaverri-Echandi</i>	74
A first approach to Acoustic Characterization of Costa Rican Children's Speech Un primer acercamiento a la caracterización acústica del habla de niños costarricenses <i>Marvin Coto-Jiménez, Maribel Morales-Rodríguez, Daniel Vargas-Díaz</i>	80
Serialization of a 3D Human Body based on MoCap Data in a BVH File for Sequence Comparison Serialización de un Cuerpo Humano Tridimensional basado en datos MoCap en un archivo BVH para la Comparación de Secuencias <i>Natalia Abarca-Jiménez, Francisco Siles-Canales</i>	85
Validation-data Generation for Brightfield Microscopy Cell Tracking using Fluorescence Samples Generación de Datos de Validación para Rastreo Celular en Microscopía de Campo Claro usando Muestras Fluorescentes <i>Patricia Quinde-Cobos, Steve Quirós, Francisco Siles-Canales</i>	91
A Biocomputational Platform for Template-based Protein-protein Docking Plataforma computacional de acoplamiento de proteínas basado en plantillas <i>Ricardo Román-Brenes, Francisco Siles-Canales, Daniel Zamora-Mata</i>	96
Design of a prototype of hand orthosis with pneumatic actuators Diseño de un prototipo de órtesis para mano con actuadores neumáticos <i>Pablo Enrique Tortós-Vinocour, Sofía Valverde-Gutiérrez, Marta Eugenia Vilchez-Monge</i>	101
Optimization of data in systems of nitrate water monitoring systems Data optimization in nitrate water monitoring systems <i>Laura Hernández-Alpizar, Arys Carrasquilla-Batista, Lilliana Sancho-Chavarría</i>	106
Advanced Computing National Collaboratory HPC Infrastructure, Kabré Infraestructura de HPC del Colaboratorio Nacional de Computación Avanzada, Kabré <i>Isaac Eduardo Gómez-Sánchez, Jean Carlo Umaña-Jiménez, Melissa Arce-Montero</i>	112

Contribuciones especiales a la Conferencia 6th Latin America High Performance Computing Conference (CARLA) 2019

Juan Luis Crespo-Mariño¹, Andrés Segura-Castillo²,
Esteban Meneses-Rojas³

Durante los días 23 a 27 de septiembre del 2019 tuvieron lugar en San José (Costa Rica) y Turrialba (Costa Rica) las actividades que constituyeron la sexta edición de la conferencia 6th Latin America High Performance Computing Conference (CARLA) 2019. Esta es la primera vez que dicha conferencia se celebró en Costa Rica (anteriormente se llevaron a cabo ediciones de la misma en Chile, Brasil, México, Argentina-Uruguay y Colombia). La organización de la conferencia (y de las actividades paralelas a la misma) corrió a cargo de un grupo de profesores e investigadores de diferentes universidades públicas costarricenses, así como del Colaboratorio Nacional de Computación Avanzada (CNCA), adscrito al Centro Nacional de Alta Tecnología (CENAT).

La idea que ha guiado esta conferencia ha sido la exposición de resultados de investigación en computación de alto rendimiento y áreas afines, generados por la actividad de grupos de investigación y laboratorios en todo el continente. En esta ocasión, CARLA 2019, no sólo se ha referido a los temas de investigación más estrechamente relacionados con la computación de alto rendimiento, ó HPC por sus siglas en inglés, sino que ha contado con una sesión especial dedicada a Procesamiento Bioinspirado (BIP). Este último entendido como la aproximación transdisciplinar entre las ingenierías tecnológicas y las ciencias básicas, la cual, consideramos, es un eje de trabajo adecuado e inherente a la realidad latinoamericana. Así mismo, se ha buscado que la conferencia fomentara la creación de vínculos sólidos y fortificados en la relación entre academia y sociedad, así como también la realización de actividades que ayuden a la formación de jóvenes investigadores en la región.

Por ello, durante los días 23 y 24 tuvieron lugar una serie de talleres y tutoriales en la sede del CENAT en San José, donde se abordaron temáticas como los fundamentos de la informática urbana y la movilidad, el modelado computacional de procesos biológicos por medio del software COPASI o el lenguaje de programación PyCOMPSs, entre otros. Se organizaron también talleres dedicados al rol de las mujeres latinoamericanas en la comunidad de HPC o a las buenas practicas profesionales en el área. De estas dos últimas actividades se generaron dos artículos que forman parte del presente número especial.

1 Área Académica de Ingeniería Mecatrónica. Tecnológico de Costa Rica Coeditor invitado del número especial.

2 Laboratorio de Investigación Tecnológica. Universidad Estatal a Distancia de Costa Rica. Coeditor invitado del número especial

3 Colaboratorio Nacional de Computación Avanzada. Centro Nacional de Alta Tecnología de Costa Rica. Coordinador General de la Conferencia CARLA 2019

Igualmente, durante los días de la conferencia propiamente dicha (25 a 27 de septiembre), en la sede de la misma en Turrialba, se llevó a cabo una sesión especial de posters científicos generados por investigadores en formación. Estos trabajos corresponden a proyectos (de grado en algunos casos, de posgrado -Maestría y Doctorado- en la mayoría de ellos) en realización. La idea de la sesión es someter a la persona joven investigadora en formación a la experiencia de difundir los resultados de su trabajo, para lo cual debían escribir un resumen ejecutivo (extended abstract), el cual sería analizado por medio de un mecanismo de revisión por pares anónimos, y sólo después de ser aceptado daría lugar al póster que se exhibió en la conferencia, recibiendo además una segunda realimentación por parte de los asistentes a la misma. Las versiones finales de dichos resúmenes ejecutivos, como verdaderos artículos científicos, son el resultado final de dicho proceso de formación y son los que se recojen igualmente en dicho número especial.

La creación de este número especial no es, en nuestra opinión, el final de un camino, sino de su primer paso. Consideramos que no sólo la difusión de resultados de investigaciones consolidadas, sino la creación de actividades de gran calidad y alto valor añadido en la formación de vocaciones y capacidades investigadoras en la juventud, son factores claves para la constitución de sociedades adaptadas para evolucionar en la nueva economía.

Cartago, noviembre de 2019

Gaussian mixture analysis of basic meteorological parameters: Temperature and relative humidity




Análisis de mixturas gaussianas de parámetros meteorológicos básicos: Temperatura y humedad relativa

Mariela Abdalah-Hernández¹, Javier Rodríguez-Yáñez²,
Daniel Alvarado-González³

Abdalah-Hernández, M; Rodríguez-Yáñez, J; Alvarado-González, D. Gaussian mixture analysis of basic meteorological parameters: Temperature and relative humidity. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 5-12.

 <https://doi.org/10.18845/tm.v33i5.5068>



- 1 Research assistant. National Advanced Computing Collaboratory, National Center for High Technology. Chemical Engineering student. Chemical Engineering School, University of Costa Rica. Costa Rica. Email: mariela.abdalah@ucr.ac.cr.
 <https://orcid.org/0000-0002-9790-2689>
- 2 Master of Environment, Chemical Engineer and researcher. Urban Ecology Laboratory, Universidad Estatal a Distancia. Costa Rica. Email: jrodriguez@uned.ac.cr.
 <https://orcid.org/0000-0001-5539-3153>
- 3 Master of Computer Science and researcher. National Advanced Computing Collaboratory, National Center for High Technology. Costa Rica. Email: dalvarado@cenat.ac.cr.
 <https://orcid.org/0000-0003-3290-690X>

Keywords

Temperature; relative humidity; Gaussian mixtures; meteorology.

Abstract

Gaussian mixture modelling was applied to describe the annual distribution of two important meteorological variables, temperature and relative humidity, inside the Costa Rican Central Valley from 2010 to 2017. A fixed number of components of Gaussian mixtures were used to fit data to a general mixture curve that represented data behavior throughout the year, this was performed through specific functions of Scikit-learn and SciPy libraries of Python language. Low values of approximation error were obtained when modelling temperature data and the relationship between its distribution and hourly variability was observed, finding high values around noon. For relative humidity, the Gaussian mixture model presented issues when fitting values greater than 90 %, as a result of this variable saturation limit at 100 %. The relationship with time was not clearly determined due to the many mixture components used to model, but a tendency of low values between the late morning and early afternoon was visualized. Iterative minimization of the error was considered as a future approach to achieve a better fit with Gaussian mixtures of these and other meteorological variables.

Palabras clave

Temperatura; humedad relativa; mixtura gaussiana; meteorología.

Resumen

Se aplicó el modelado por mixturas gaussianas para describir la distribución anual de dos variables meteorológicas importantes, temperatura y humedad relativa, dentro del Valle Central de Costa Rica desde el 2010 hasta el 2017. Se utilizó un número fijo de componentes gaussianas para ajustar los datos a una curva de mixtura general que representara el comportamiento durante todo el año, esto se realizó a través de funciones específicas de las bibliotecas Scikit-learn y SciPy del lenguaje Python. Al modelar los datos de temperatura se obtuvieron valores bajos del error de aproximación y se observó una relación entre su distribución y la variabilidad horaria, estableciendo altas temperaturas alrededor del mediodía. Para la humedad relativa, el modelo de mixturas gaussianas presentó problemas en el ajuste de valores mayores al 90 %, como resultado del límite de saturación de esta variable en el 100 %. La relación respecto al tiempo no fue claramente determinada debido a la cantidad de componentes de la mixtura usadas para modelar la humedad relativa, pero se apreció una tendencia de valores bajos entre el final de la mañana e inicios de la tarde. La minimización iterativa del error fue considerada como una aproximación futura para alcanzar un mejor ajuste con mixturas gaussianas para estas y otras variables meteorológicas.

Introduction

The region with the highest population and anthropogenic activity concentration in Costa Rica is the Central Valley. This is the reason why it is necessary to have a more accurate weather model for this area. Meteorological data analysis helps to understand climate behavior, the way it changes and how it affects human activity. This work is encompassed by a larger project that pretends to study the effects of contamination on the climate and on man-made structures, mainly

focusing on the corrosion of metallic structures. In this first stage, the main objective is to model the distribution of temperature and relative humidity with Gaussian mixtures. Visualizing these distributions will allow to improve the models for weather forecast, and even more importantly, pollutant transport and particle deposition [1], [2].

These parameters were chosen because they are included inside of the typical meteorological measures in all climatic stations, which means that there is a large amount of data available for these variables (over 95 % of the annual data from the year 2010 to 2017). Additionally, they have a direct relationship with the subject of corrosion [3], [4]. It is generally considered that relative humidity and temperature affect corrosion when having values above 80 % and 0 °C [1], [2], [3]. By isolating these conditions per area, it could be determined which regions could have higher corrosion levels in order to take it into account for the construction of metallic structures.

The main focus of this work was on the western part of the valley, limited to the northwest by the Central Volcanic Mountain Range, to the east by the hills of Ochomogo and to the southwest by the Talamanca Mountain Range [5].

Methodology

The selected parameters were temperature, which is a continuous function with no bounds, and relative humidity, bounded between 0 % and 100 %. These data were obtained from weather stations all around the Central Valley. For simplicity, in this work only three representative weather stations were taken, two from opposite sides within the valley (northwest, and southwest) and one located in the mountains.

The first step of the analysis consisted in generating visualizations of the frequency of values for each parameter. Histograms were made where each category size corresponded to a band of 1 °C for temperature and 1 % for humidity. Gaussian mixture modelling was performed to obtain a set of curves to approximate the real data in order to study the behavior throughout the year. Also, the absolute error was calculated as the difference between the value of the approximation function and the real frequency or density value. This error was used as a guidance to modify the parameters of the mixture modelling function and the number of components, in order to achieve a higher accuracy with the model. Time series was plotted, differentiating the points according to the ranges determined by the means of the mixture components.

Calculations were performed with Python 3, Pandas, SciPy and Scikit-learn, specifically the `sklearn.mixture` for the process of representing with Gaussian curves. The Pandas library was used to manipulate the large amount of data, distributed among different files and provided by the National Meteorological Institute of Costa Rica.

Results and discussion

In general, it was simpler to model temperature because there was not a maximum physical limit in this variable. A good level of approximation was obtained utilizing a couple of Gaussians, as it is shown in Figure 1. Errors found in the approximation curves were low, according to Figure 2, demonstrating the acceptable precision of the modelling.

A relationship between time series, represented as hours of the day, and the components was observed too. The value of the mean of each Gaussian sets the limit of each colored zone in Figure 3, where the purple zone corresponds to the overlapping area between both components.

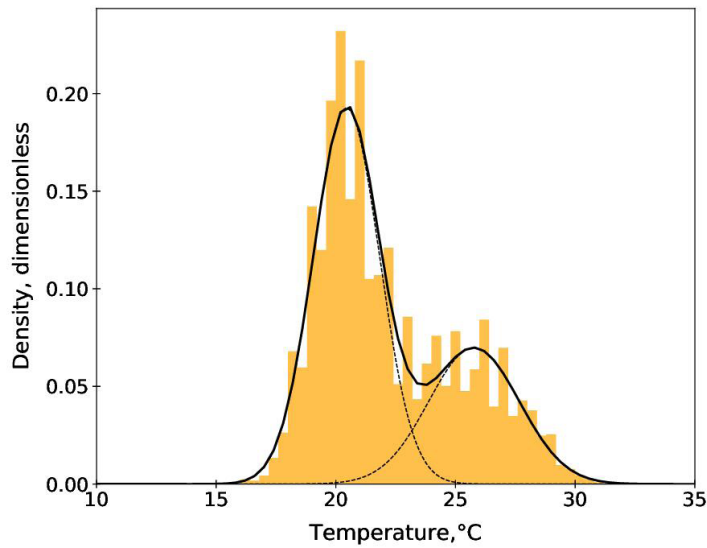


Figure 1. Temperature density with a Gaussian mixture approximation in a station located in the northwest of the valley.

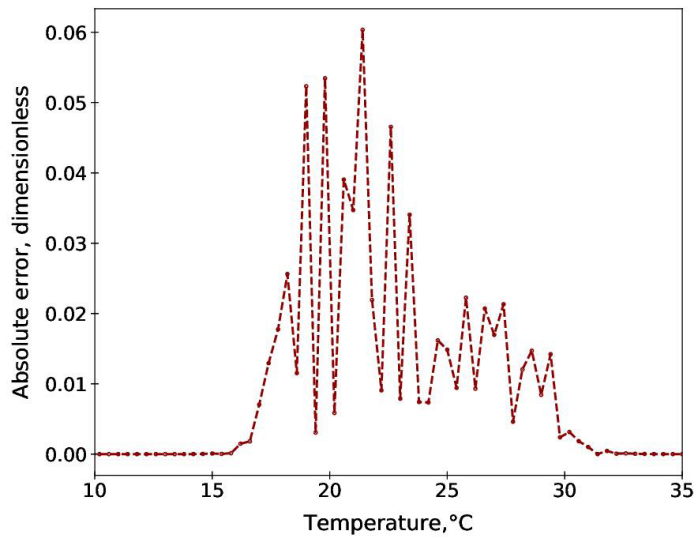


Figure 2. Error of the Gaussian mixture modelling of temperature in a station located in the northwest of the valley.

Data of the first half of the first curve in Figure 1 was below 20.5 °C (light blue region in Figure 3) and occurred around the morning and the afternoon. As it was expected, values above 25.5 °C or the second half of the second curve, were measured around noon (red region) between 09 h and 17 h.

Relative humidity distributions were more complicated to model, this was because it was necessary to consider multiple curves in the mixture and due to the saturation value. The above generated issues when fitting cases where the frequency of data above 90 % was higher. For this range of values the error increased. This can be shown in Figures 4 and 5.

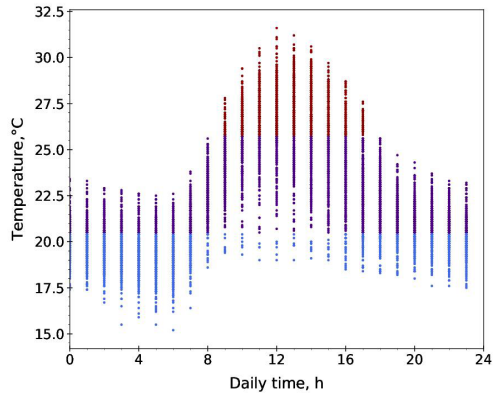


Figure 3. Temperature distribution throughout the day in a station located in the northwest of the valley.

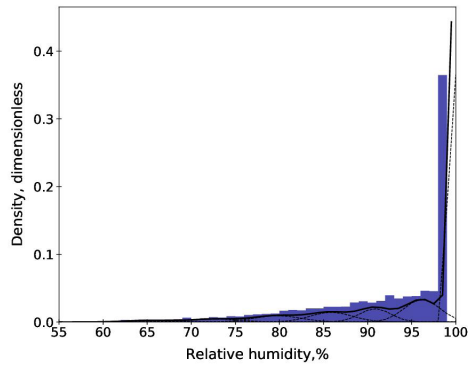


Figure 4. Relative humidity density with a Gaussian mixture approximation in a station located in the mountains of the valley.

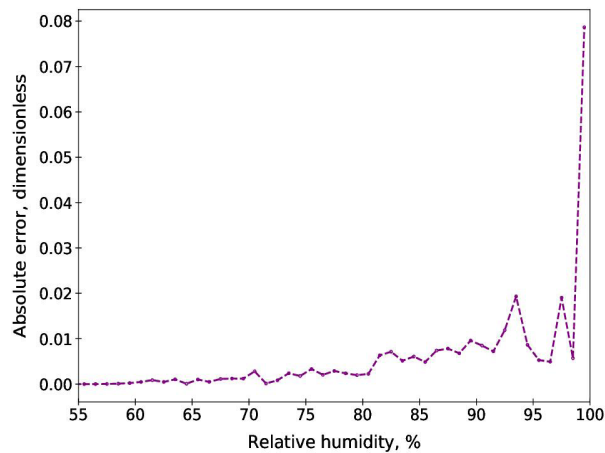


Figure 5. Error of the Gaussian mixture modelling of relative humidity in a station located in the mountains of the valley.

As can be seen in Figure 6, the achieved fit in other stations with lower humidity values was better in comparison with the situation in the mountains. Figure 7 shows that absolute error was lower in relative humidity than in temperature for this case, but this was obtained as a consequence of requiring more approximation curves for an acceptable model, making the interpretation difficult.

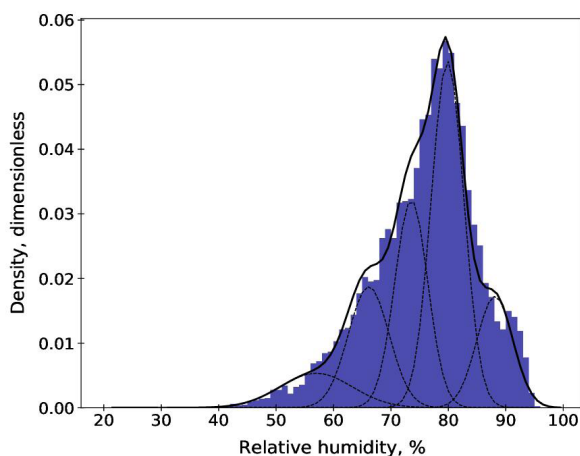


Figure 6. Relative humidity density with a Gaussian mixture approximation in a station located in the southwest of the valley.

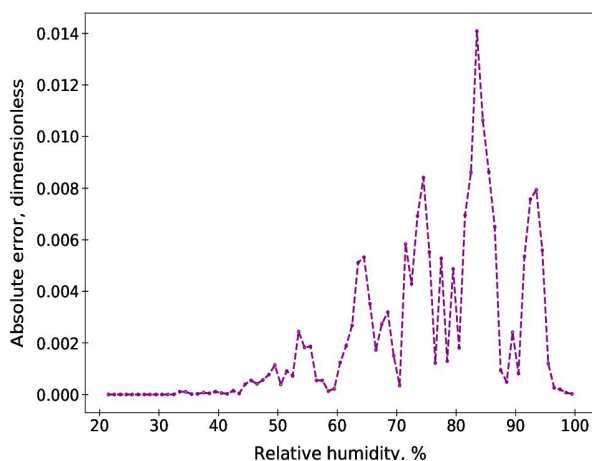


Figure 7. Error of the Gaussian mixture modelling of relative humidity in a station located in the southwest of the valley.

Analyzing time series, the quantity of components made the relationship unclear. However, taking only the first and the last components the Figure 8 was obtained. The first half of the first Gaussian was below 55 % approximately (light blue region), while high values in the second part of the last curve were above 89 % (red region). Intermediate data located inside the other components is grouped inside the purple area. In general terms, it was observed that low values occurred between 08 h and 15 h.

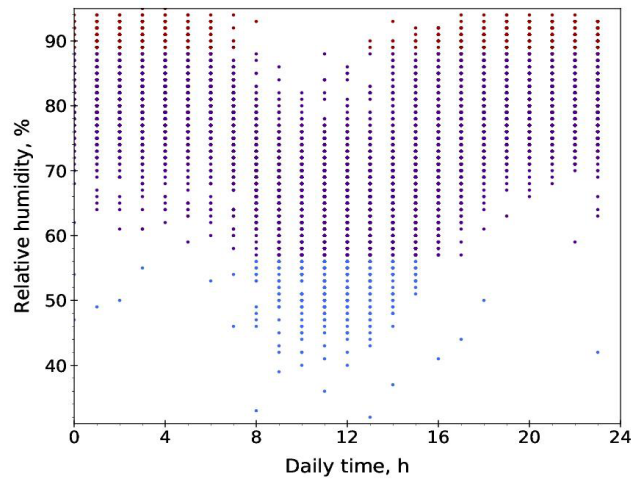


Figure 8. Relative humidity distribution throughout the day in a station located in the southwest of the valley.

Conclusions

The approximation with Gaussian mixtures gave acceptable results for temperature, but presented problems when modelling relative humidity near the 100 % value. The iterative minimization of the error was an adequate strategy to visualize the issues with fitting relative humidity, therefore it can be used with another techniques to improve modelling of this and other variables. The subject of lack of precision near the saturation limit in relative humidity still has to be solved. On the other hand, for temperature a good relationship between the distribution and time series was found, while for relative humidity this was not clearly established due to the number of components.

Modelling meteorological parameters with multiple Gaussians allows to notice associations with time. These simplified models are fundamental for the development of subsequent models or more complex dispersion algorithms. An example is the dispersion of pollutants in the air, which depends on the meteorological variables and, in some cases, the physicochemical interactions associated to the pollutants and the air components. These later models allow to optimize the environmental control nets in urban or equivalent areas, as the western Central Valley is.

Acknowledgments

To the Meteorological National Institute for providing the data for this study.

To the Special Funds of Superior Education (FEES) of the National Council of Rectors (CONARE) for financing the project.

References

- [1] L. Garita, J. Rodríguez, and J. Robles, "Modelado de la Velocidad de Corrosión de Acero de baja aleación en Costa Rica," *Revista Ingeniería*, vol. 24, no. 2, pp. 79-90, 2014.
- [2] M. Morcillo, E. Almeida, B. Rosales, J. Uruchurtu, and M. Marrocos, *Corrosión y Protección de Metales en las Atmósferas de Iberoamérica, Parte I: Mapas Iberoamericanos de Corrosión Atmosférica (MICAT)*. Madrid: Programa CYTED, 1998.

- [3] Corrosion of Metals and Alloys - Corrosivity of Atmospheres - Classification, ISO Standard 9223, 2012.
- [4] D. Singh, S. Yadav, and J. Saha, "Role of climatic conditions on corrosion characteristics of structural steels," *Corrosion Science*, vol. 50, pp. 93-110, 2008.
- [5] J. Solano, and R. Villalobos, *Regiones y Subregiones Climáticas de Costa Rica*. San José: Instituto Meteorológico Nacional, 2000.
- [6] A. Gómez, "Modelos de mixturas finitas para la caracterización y mejora de las redes de monitorización de la calidad del aire," Master's Thesis, Statistics and Operative Investigation Department, University of Granada, Granada, 2014.
- [7] Gaussian mixture models, Scikit-learn Project. [Online]. Available in: <https://scikit-learn.org/stable/modules/mixture.html>
- [8] Statistical functions (scipy.stats), SciPy Project. [Online]. Available in: [https:// docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html#scipy.stats.norm](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html#scipy.stats.norm)



Instance segmentation for automated weeds and crops detection in farmlands

Segmentación de instancias para detección automática de malezas y cultivos en campos de cultivo

Adán Mora-Fallas¹, Hervé Goëau², Alexis Joly³, Pierre Bonnet⁴,
Erick Mata-Montero⁵

Mora-Fallas, A; Goëau, H; Joly, A; Bonnet, P; Mata-Montero, E.
Instance segmentation for automated weeds and crops detection in farmlands. Tecnología en Marcha. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 13-17.

 <https://doi.org/10.18845/tm.v33i5.5069>

- 1 School of Computing, Costa Rica Institute of Technology, Cartago, Costa Rica. E-mail: adamora@ic-itcr.ac.cr.  <https://orcid.org/0000-0002-0893-1884>
- 2 AMAP, Univ Montpellier, CIRAD, CNRS, INRA, IRD, Montpellier, France. E-mail: herve.goeau@cirad.fr.
- 3 INRIA Sophia-Antipolis - ZENITH team, LIRMM - UMR 5506 - CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France. E-mail: alexis.joly@inria.fr
- 4 CIRAD, UMR AMAP, Montpellier, France. E-mail: pierre.bonnet@cirad.fr.  <https://orcid.org/0000-0002-2828-4389>
- 5 School of Computing, Costa Rica Institute of Technology, Cartago, Costa Rica. E-mail: emata@tec.ac.cr.  <https://orcid.org/0000-0001-5471-164X>



Keywords

Deep learning; instance segmentation; computer vision; precision agriculture; biodiversity informatics; weed detection; species identification.

Abstract

Based on recent successful applications of Deep Learning techniques in classification, detection and segmentation of plants, we propose an instance segmentation approach that uses a Mask R-CNN model for weeds and crops detection on farmlands. We evaluated our model performance with the MSCOCO average precision metric, contrasting the use of data augmentation techniques. Results obtained show how the model fits very well in this context, opening new opportunities to automated weed control solutions, at larger scales.

Palabras clave

Aprendizaje profundo; segmentación de instancias; visión por computadora; agricultura de precisión; bioinformática; detección de malezas; identificación de especies.

Resumen

Con base en las recientes aplicaciones exitosas de técnicas de Aprendizaje Profundo en la clasificación, detección y segmentación de plantas, proponemos un enfoque de segmentación de instancias utilizando un modelo Mask R-CNN para la detección de malezas y cultivos en tierras de cultivo. Evaluamos el rendimiento de nuestro modelo con la métrica de precisión promedio de MSCOCO, contrastando el uso de técnicas de aumento de datos. Los resultados obtenidos muestran cómo el modelo se adapta muy bien en este contexto, abriendo nuevas oportunidades para soluciones automatizadas de control de malezas a gran escala.

Acknowledgements

The authors would like to thank the french ANR (Agence Nationale de la Recherche), which has supported this research activity (under grant N°ANR-17-ROSE-0003).

Introduction

Biologists, agronomists, and other experts in the area of agro-biodiversity science are experiencing a digital era that provides access to massive amounts of visual data. For instance, natural history collections worldwide have recently largely digitized and made freely available large amounts of digital images from their vast collections. In addition to the big data currently available, new technological developments such as autonomous engines applied to agricultural field management lead the way for the development of new research aimed at identifying computational approaches to solve agro-biological problems. Thus, a critical current challenge for computer scientists is to take advantage of this large amount of primary data and new technologies to increase the efficiency of agricultural field management processes.

Fortunately, impressive recent advances in Artificial Intelligence open up new hopes for extracting knowledge automatically from large amounts of visual data. In particular, the use of Deep Learning has had a notable success in the automated classification of images, achieving levels of accuracy of around 90% [13] for the identification of species from plant leaf images.

Furthermore, challenges such as PlantCLEF (conducted in the context of the LifeCLEF2 Lab, [6] organized since 2011, have demonstrated the power of Deep Learning for plant species identification under more demanding conditions, for example, with large species datasets (both in terms of number of species and number of images) and using noisy images of many plants components (e.g., leaves, flowers, and fruits).

Nowadays, a large number of farmers and agronomists use herbicides to deal with weeds that affect growing stages on crops due to the competition for resources (water, light, nutriment). Nevertheless, the use of herbicides, in spite of its efficiency, is dangerous not only for the environment, but also for human health.

In this work, we propose a system capable of automatically detecting and segmenting weed and crop species from photographic material taken *in situ* in farmlands that can be used by an autonomous agent, such as a robot, for weed control tasks with eco-friendly methods.

Related work

Machine learning technologies have recently and largely transformed several aspects of agricultural activities [8], [7]. This includes, for example, the development of new approaches for plant diseases identification on isolated plants species such as maize [17], apple [11], wheat [5], or potato [14]; in addition to yield production [1] and crops quality evaluation [18]. As weed control has a major impact on agricultural production, several studies have been conducted to improve their detection, such as [15]. Nevertheless, the majority of these studies focus on a few crops or weed species (such as in [2]) and do not try to identify various weed species, in various agricultural systems. Our approach tries to solve that problem, in order to increase the benefit of the use of new deep learning technologies in agriculture.

Methodology

Dataset description

This research was conducted on two crop species, namely, *Zea mays* (corn) and *Phaseolus vulgaris* (green bean), and the following four common weed species: *Brassica nigra*, *Matricaria chamomilla*, *Lolium perenne*, and *Chenopodium album*. The dataset comprises more than 4.000 photos that were collected either manually with smartphones and other digital cameras, or automatically with digital cameras mounted on a robot. This dataset was produced in such way that it reflects different combinations of crop and weed species, illustrating the whole plant at different angles, distances, plant growth stages, and at different times of the year.

Deep Learning Approach

We propose deep model based on the Mask R-CNN architecture [3] due to its robustness and demonstrated efficiency in instance segmentation tasks and challenges such as MSCOCO, which stands for Microsoft Common Objects in Context [10].

We selected the Facebook's Mask R-CNN benchmark [12] with the official implementation on Pytorch [16]. This benchmark offers a set of pretrained ready-to-use models on the MSCOCO dataset, with different configurations for CNN backbone architectures and approaches for bounding box and segmentation estimation. In this work we chose a ResNet 50 [4] and the Feature Pyramid Networks [9]. Our experiments ran on a NVIDIA Geforce RTX 2080 Ti that uses CUDA 10 in Linux.

Experiments and Results

For our first experiments, we annotated manually 129 images following the MSCOCO annotation format, for the 6 classes of plants, where we took 80% for training and the remaining 20% for testing using the average precision metric as described on MSCOCO challenge, common in instance segmentation tasks.

We applied data augmentation techniques, random horizontal flip, random rotations and random variations on color contrast, saturation, brightness and hue color values; to ensure a more dynamic dataset that avoids an overfitted model.

Figure 1 shows results for this approach. We trained the model with 4.000 iterations, where each iteration represents one batch of images and its annotations. We tested how data augmentation affects the model's performance by applying or not the techniques described before. We reached high performance with an average precision between 0.4 and 0.5 in both cases, segmentation and bounding boxing tasks. Some predictions are shown in Figure 2, in which we can appreciate how precise the prediction is for each detected plant.

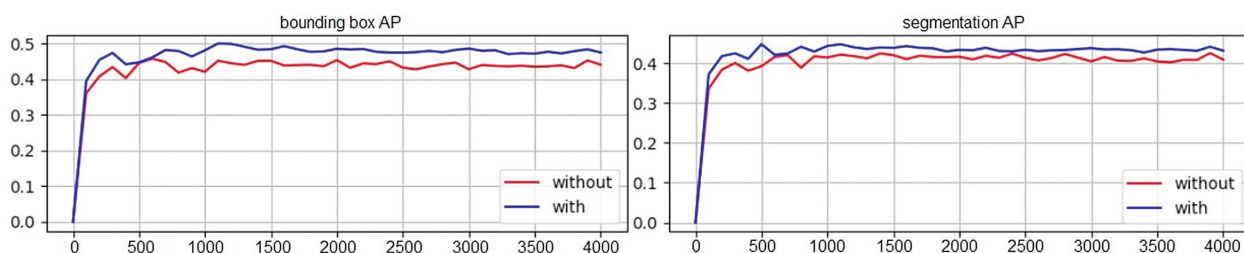


Figure 1. Testing results from the model measured by average precision with 4.000 iterations, red line represents training without data augmentation, blue line, by contrast, with data augmentation.

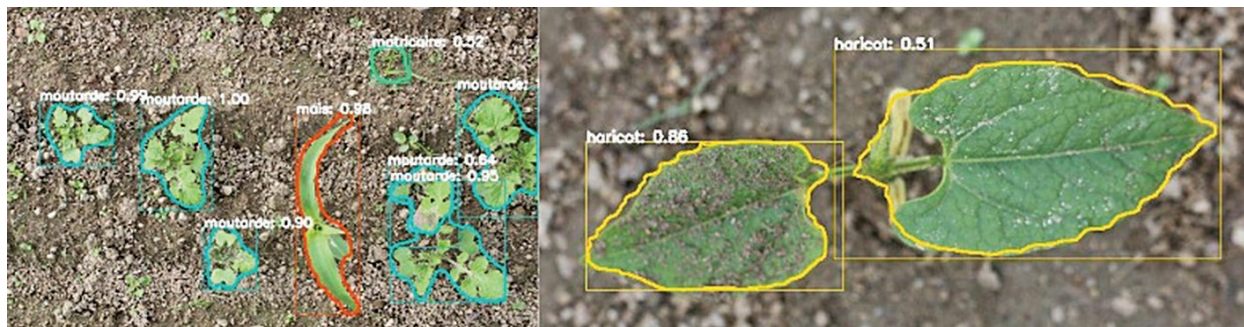


Figure 2. Predictions taken from testing images. Each prediction shows the mask contour (segmentation), bounding box, class predicted, and its individual score that evaluates how good the prediction is.

Conclusions and Future work

We find the results shown on figure 1 and 2 very promising, instance segmentation is a singularly demanding task where the state-of-art performance in MSCOCO reaches 0.5 of average precision. Particularly, the use of Mask R-CNN as baseline fits very well for weeds and crops detection tasks reaching similar results to the state-of-art. As we can see in figure 2, in some cases, the predicted mask seems better than a manual human-made mask. The experiments have shown the importance of using data augmentation and how it enhances the model's

performance, therefore, we want to test more data augmentation techniques and more model configurations in order to get a model with the best performance to be used in a real situation by an autonomous robot.

During the annotation process, we faced a common problem, namely, the highly time-consuming task of annotating massive data, especially in this context, where the annotation process involves drawing manually masks for each object of interest in each image. We are working on an automatic active learning system to deal with this problem and annotate more training data in much less time with less human interaction, as we have more than 4.000 images available.

References

- [1] S. Amatya, M. Karkee, A. Gongal, Q. Zhang, & M. D. Whiting. Detection of cherry tree branches with full foliage in planar architecture for automated sweet-cherry harvesting. *Biosystems engineering*, 2016, 146, 3-15.
- [2] A. Binch, & C. W. Fox. Controlled comparison of machine vision algorithms for Rumex and Urtica detection in grassland. *Computers and Electronics in Agriculture*, 2017, 140, 123-138.
- [3] K. He, G. Gkioxari, P. Dollár, & R. Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961-2969.
- [4] K. He, X. Zhang, S. Ren & J. Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
- [5] A. Johannes, A. Picon, A. Alvarez-Gila, J. Echazarra, S. Rodriguez-Vaamonde, A. D. Navajas & A. Ortiz-Barredo. Automatic plant disease diagnosis using mobile capture devices, applied on a wheat use case. *Computers and electronics in agriculture*, 2017, 138, 200-209.
- [6] A. Joly, H. Goëau, H., C. Botella, H. Glotin, P. Bonnet, W. P. Vellinga, ... & H. Müller. Overview of LifeCLEF 2018. In *Experimental IR meets multilinguality, multimodality, and interaction*, In *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018*. Springer.
- [7] A. Kamilaris, & F. X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 2018, 147, 70-90.
- [8] K. Liakos, P. Busato, D. Moshou, S. Pearson, & D. Bochtis, D. Machine learning in agriculture: A review. *Sensors*, 2018, 18(8), 2674.
- [9] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, & S. Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117-2125.
- [10] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, ... & C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, Springer, Cham, 2014, pp. 740-755.
- [11] B. Liu, Y. Zhang, D. He, & Y. Li. Identification of apple leaf diseases based on deep convolutional neural networks. *Symmetry*, 2017, 10(1), 11.
- [12] F. Massa, & R. Girshick. Maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. <https://github.com/facebookresearch/maskrcnn-benchmark>. 2018.
- [13] E. Mata-Montero, & J. Carranza-Rojas. Automated plant species identification: Challenges and opportunities. In *IFIP World Information Technology Forum*. Springer, Cham, 2016, pp. 26-36.
- [14] D. Oppenheim, & G. Shani. Potato disease classification using convolution neural networks. *Advances in Animal Biosciences*, 2017, 8(2), 244-249.
- [15] X. E. Pantazi, A. A. Tamouridou, T. K. Alexandridis, A. L. Lagopodi, J. Kashefi, & D. Moshou, Evaluation of hierarchical self-organising maps for weed mapping using UAS multispectral imagery. *Computers and Electronics in Agriculture*, 2017, 139, 224-230.
- [16] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, ... & A. Lerer. Automatic differentiation in pytorch. In *31st Conference on Neural Information Processing Systems, NIPS, 2017, Long Beach, CA, USA*.
- [17] T. Wiesner-Hanks, E. L. Stewart, N. Kaczmar, C. DeChant, H. Wu, R. J. Nelson, ... & M. A. Gore. Image set for deep learning: field images of maize annotated with disease symptoms. *BMC research notes*, 2018, 11(1), 440.
- [18] M. Zhang, C. Li, & F. Yang, F. Classification of foreign matter embedded inside cotton lint using short wave infrared (SWIR) hyperspectral transmittance imaging. *Computers and Electronics in Agriculture*, 2017, 139, 75-90.

MediaTIC: A Social Media Analytics Framework For the Costa Rican News Media

MediaTIC: Una plataforma analítica de medios digitales costarricenses en redes sociales

Cristina Soto-Rojas¹, Carlos Gamboa-Venegas²,
Adriana Céspedes-Vindas³

Soto-Rojas, C; Gamboa-Venegas, C; Céspedes-Vindas, A.
MediaTIC: A Social Media Analytics Framework For the Costa Rican News Media. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 18-24.

 <https://doi.org/10.18845/tm.v33i5.5070>

1 University of Costa Rica (UCR), National Center of High Technology (CeNAT).
2 National Center of High Technology (CeNAT). Email: cgamboa@cenat.ac.cr.
3 Universidad Estatal a Distancia (UNED). Email: acespedesv@uned.ac.cr.



Keywords

Social Media; Big Data; MongoDB; Spark; R; Facebook.

Abstract

Social media sites such as Facebook are tools that democratize contents. These tools facilitate the creation of news and ideas. In Costa Rica, we designed MediaTIC framework to track social media outlets communities by accessing their posts and user interactions. In order to do this, we created a big data store and designed a software architecture that uses high computing hardware (cluster) to manage software escalation, data volume, and future requirements. We used MongoDB, Spark and R technologies to build the platform and manage the information. This platform, allowed us to run faster queries, improving the performance time in 110%-180% approximately. The main objective is to provide visualizations and information for Costa Rican social media analysts that can give value to Social Communication Science in Costa Rica by using high technology underneath.

Palabras clave

Redes sociales; Big Data; MongoDB; Spark; R; Facebook.

Resumen

Las redes sociales como Facebook son herramientas democratizadoras de contenido. Estas herramientas facilitan la creación y visualización de ideas y noticias, favoreciendo la creación de comunidades virtuales. En Costa Rica, diseñamos MediaTIC como una plataforma para examinar estas comunidades y su comportamiento utilizando la información de las publicaciones y las interacciones entre usuarios. Para lograr esto, implementamos un repositorio *big data*, el cual trabaja sobre una infraestructura de alto rendimiento (cluster computacional) lo que permite que esta plataforma sea un sistema escalable que puede trabajar con grandes volúmenes de datos. En la implementación de esta plataforma y el manejo de los datos se utilizaron MongoDB, Spark y R. Esta plataforma nos permite ejecutar consultas a mayor velocidad, mejorando el tiempo de respuesta en un 110%-180% aproximadamente. El objetivo principal es proveer visualizaciones y estadísticas que brinden información de valor a la investigación en Comunicación Social en Costa Rica.

Introduction

The usage of social media has increased the generation and collection of data. This data contains more information that can be analyzed and used to respond to different questions. This project focuses on social media outlets with facebook pages. We record posts, comments and interactions between users in these pages. This information is stored in a platform that will allow users to determine media behavior in Costa Rica during certain months, seasons or major events.

MediaTIC is a computational platform for the analysis and visualization of big data produced by the principal digital media outlets in Costa Rica on facebook. The main objective of the project is to develop a computer system that not only collect information but also applies information retrieval algorithms and social network analysis to visualize the most relevant information through a web interface. The working group is composed of professionals in the area of computing and communication from three departments: Laboratory of Research and Technological Innovation (LIIT) at UNED, Communication Research Center (CICOM) at UCR, and Advanced Computing Laboratory (CNCA) at CeNAT (National Center for High Technology).

Social Media Analytics Framework

There are many studies related to social media analytics, with data collection and analysis being some of the most important steps. Social media analytics framework is seen as a guide to find and resolve conflicts [1] and it helps to identify challenges and design solutions [1], [2], [3]. In Figure 1 we present an analog framework for our project.

First, consider the tracking step. The tracking is a fundamental step that explains the process to recollect the data [1]. In our case, the data was collected manually through Netvizz tool, which allows data to be downloaded two weeks after the date they were published. The selected media are: La Nación, CR Hoy, Telenoticias, Repretel, Semanario Universidad, El Financiero, Noticias Monumental, Prensa Libre, Diario La Extra and Amelia Rueda.

The tool provides two plain-text files, one that contains all the posts with the publication date, text, and reactions (like, love, sadness, wow, angry). The other file contains the comments for each post with the publication date, text, reactions and the order chain between the comments.

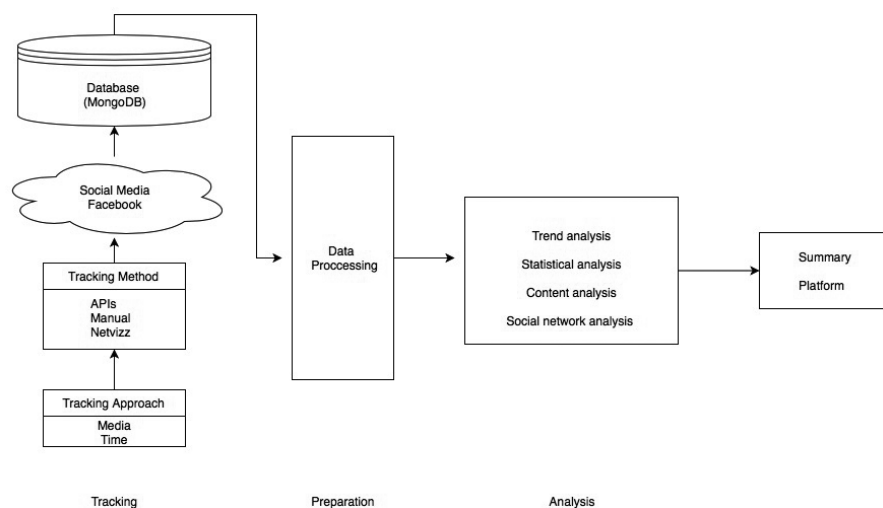


Figure 1. Social Media Analytic Framework Scheme

Preparation

Once we have the data collection step, the next step is the data storage [1]. We have selected MongoDB [4], a document-based database chosen in virtue of its facility to store data with different structures and its efficiency in queries. This database model makes possible to have different collections in the same storage space. We created a collection for posts and another collection for posts comments.

This Mongo database is located in the cluster Kabré. Hosted at CeNAT, Kabré is a multidisciplinary supercomputer, with 32 Intel KNL nodes dedicated to bioinformatics, simulation, and others. 4 k40 NVIDIA GPU nodes assigned to machine learning projects, and 3 nodes for big data applications. Each one of the big data nodes has 64 GB of memory and two Intel Xeon E5-2650 processors at 2.20 GHz, all together add 70 TB of storage and are connected through a 10 Gigabit Ethernet network which makes possible fast communication and low latency in data transmission [5][6].

To access the database a web platform was implemented, which we called “Mediatic-CMS” for Control Management System. The main goal of this platform is to provide a tool for queries and

the manual assignment of post categories (politics, sports, nationals, international). There is an automatic alternative to do this category assignment as an option. Users can utilize the platform to upload netviz data and enrich the collection. There are two main modules, one for posts, and one for comments. In the post section, users can look up for information using different criteria such as: media name, dates, category (main topic), tone, format, reporter, if it is a public or nonpublic affair, and keywords. In the comment section, users can filter information by media, dates, public or non-public news and keywords. In both sections, users have the option to download the information as Comma-Separated-Value files for complementary analysis.

Processing and analysis

The next step is analysis [1]. To compute the queries for the analysis we used Mongo Spark Connector [7], which allows us to optimize the queries as we can see in the comparative Table 1. Once the data had the structure as need it after the queries the next step is the visualization of these results.

As we can see in Table 1, the performance using mongodb and spark shows an increase in performance for Spark in 110%-180% depending on the query type. Spark offers a considerably shorter response time in this first phase of the project.

Table 1. Comparison time between MongoDB and Spark in the following operations: Aggregation, Sort and Group

Time in ms		
Query	MongoDB	Spark
1: Aggregation	339	3
2: Sort	332	3
3: Group	183	1

Data visualization

Data Visualization has been worked using R programming language. We are constructing an interactive site with Shiny that allows users to change the graphics in real-time. The type of graphics are word clouds, heat maps and time series. The user can filter by month, year and media.

Word Clouds of Post Contents

Word clouds are considered a text-mining technique. They work by highlighting the frequency of words in an input. This process requires clean data, that means that punctuation, numbers and special characters are removed. Words that match the stop-words language list (such as: articles, prepositions, and others) are also removed. Finally, the text, once clean, goes through a stemming-process that allow counting the words with the same root. The most frequent words are graphed. The bigger the word, the most frequent it is in the text.

In MediaTIC words clouds are used to visualize the most frequent words that a media outlet used during a specific period. For example, we can identify the topic of the month by the media. In figure 2, there is a word cloud sample with December's posts most frequent content.

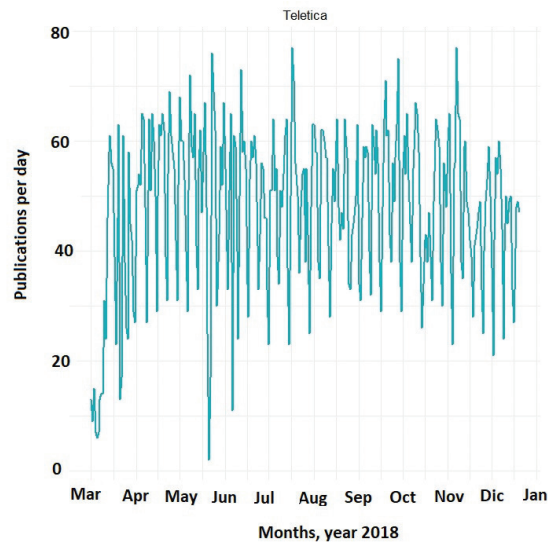


Figure 4. Time Series of Telenoticias, 2018 and the number of publications per day.

Conclusions and future work

Data provided by Netviz is enough to obtain a good variable set for analysis. Many permutations of data can be done to provide value to final users using advanced computing techniques.

Different visualizations are possible due to library packages that an environment like R provides. These visualizations enable diffusion and understanding of the data collected. High Computing infrastructure establishes a good set of resources that helps with the extraction and manipulation of big data with a very good performance.

Data retrieval is optimized by the use of the big data cluster infrastructure and MongoDB-Spark Connector. The time used to run the query is considerably better with spark. The performance is expected to be more notorious as the volume of data increases.

Since we have temporal and text variables, the analysis of temporal series and sentimental analysis are interesting future approaches applying studies from Rodrigues [8] and Nodarakis [9]. In temporal series, the next step is to analyze the stationarity and tendencies to make predictions. In sentimental analysis, since the language is Spanish there need to build the dictionary, and then proceed to analyze the posts and the comments of the gathered news.

At the current state of the project, we are working on the connection of our front-end (R-Shiny) and back-end (MongoDB) using Spark as a sort of middleware to allow us with the volume and processing time. It is important to consider that Netviz tool for data extraction from Facebook has the risk to stop working, and we need to accomplish the collecting step with other set of tools and sources different from facebook.

References

- [1] Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. "Social media analytics – Challenges in topic discovery, data collection, and data preparation". *International Journal of Information Management*, Vol. 39, no. 2018, pp 156–168, April, 2018. <https://doi.org/10.1016/j.ijinfomgt.2017.12.002>.
- [2] Labrinidis, A., Papakonstantinou, Y., Patel, J. M., & Ramakrishnan, R. "Big Data and Its Technical Challenges". *Communications of the ACM*. Vol. 57, no 7, pp 86-94, July, 2014. <https://doi.org/10.1145/2611567>.

- [3] Verma, J. P., Agrawal, S., Patel, B., & Patel, A. "Big Data Analytics: challenges and applications for text, audio, video, and social media data". *International Journal on Soft Computing, Artificial Intelligence and Applications*, Vol 5, No 1, pp 41–51, February, 2016. <https://doi.org/10.5121/ijdps.2017.8101>.
- [4] Chodorow, K and Dirolf, M. "MongoDB: The Definitive Guide", O'Reilly Media, Inc. 2010
- [5] Intel. Hadoop* Clusters Built on 10 Gigabit Ethernet, 2012, [Online]. Available: https://www.arista.com/assets/data/pdf/Whitepapers/Hadoop_WP_final.pdf
- [6] Arista. 10 Gigabit Ethernet: Enabling Storage Networking for Big Data, 2016, [Online]. Available: https://www.arista.com/assets/data/pdf/Whitepapers/AristaStorageNetworkingWhitepaper_v.6.0_GF.pdf
- [7] Shoro, A. G., & Rahim, S. T. "Big Data Analysis: Ap Spark Perspective". *Global Journal of Computer Science and Technology: C Software & Data Engineering*, Vol 15, No 1, 2015. Vol 15. No 1. pp 7-14. Retrieved from <http://www.computerresearch.org/index.php/computer/article/viewFile/1137/1124>.
- [8] Rodrigues, A. P., Chiplunkar, N. N., & Rao, A. "Sentiment analysis of social media data using Big Data Processing Techniques". *International Journal of Computer Applications*, Vol 22, No. 6, pp 56, 2016.
- [9] Nodarakis, N., Tsakalidis, A., Sioutas, S., & Tzimas, G. (2016). "Large scale sentiment analysis on Twitter with Spark". *CEUR Workshop Proceedings*, Bordeaux, France, 2016, Vol 1558.

Evaluating Resilience of Deep Learning Models

Evaluando la Resiliencia de Modelos de Deep Learning

Elvis Rojas¹, Bogdan Nicolae², Esteban Meneses³

Rojas, E; Nicolae, B; Meneses, E. Evaluating Resilience of Deep Learning Models. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 25-30.

 <https://doi.org/10.18845/tm.v33i5.5071>

1 Associate Professor, School of Informatics, National University of Costa Rica. Dr-Ing Student, Doctor of Engineering program, Costa Rica Institute of Technology. Costa Rica. E-mail: erojas@una.ac.cr.

 <https://orcid.org/0000-0002-4238-0908>

2 Computer Scientist, Mathematics and Computer Science, Argonne National Laboratory, United States. E-mail: bogdan.nicolae@acm.org.

 <https://orcid.org/0000-0002-0661-7509>

3 Director, Advanced Computing Collaboratory, Costa Rica National High Technology Center. Associate Professor, School of Computing, Costa Rica Institute of Technology. Costa Rica. E-mail: emeneses@cenat.ac.cr.

 <https://orcid.org/0000-0002-4307-6000>



Keywords

Resilience; fault tolerance; deep learning; fault injection.

Abstract

Deep learning applications have become a valuable tool to solve complex problems in many critical areas. It is important to provide reliability on the outputs of those applications, even if failures occur during execution. In this paper, we present a reliability evaluation of three deep learning models. We use an ImageNet dataset and a homebrew fault injector to make all the tests. The results show there is a difference in failure sensitivity among the models. Also, there are models that despite an increase in the failure rate can keep the resulting error values low.

Palabras clave

Resiliencia; tolerancia a fallas; deep learning; inyección de fallos.

Resumen

Los modelos de Aprendizaje Profundo se han convertido en una valiosa herramienta para resolver problemas complejos en muchas áreas críticas. Es importante proveer confiabilidad en las salidas de la ejecución de estos modelos, aún si se producen fallos durante la ejecución. En este artículo presentamos la evaluación de la confiabilidad de tres modelos de aprendizaje profundo. Usamos un conjunto de datos de ImageNet y desarrollamos un inyector de fallos para realizar las pruebas. Los resultados muestran que entre los modelos hay una diferencia en la sensibilidad a los fallos. Además, hay modelos que a pesar del incremento en la tasa de fallos pueden mantener bajos los valores de error.

Introduction

Machine learning (ML) has become a growing research area and it has applications from financial services to image recognition. Much of this growth has been fueled by the rise of deep neural networks (DNN). DNNs automatically learn based on a training dataset to recognize patterns to classify objects like images or audio signals. With the increase of interest in DNNs many deep learning (DL) frameworks have been developed (TensorFlow, Pytorch, MXNet, DeepLearning4j and Keras) to facilitate the development and implementation of DNNs. It is crucial to provide reliability to the outputs of DNNs to ensure the accuracy and correctness of the results, even if there are failures in the execution of the model. Therefore, there is a growing need in understanding the resilience of DNN models to analyze how they are affected by failures. The goals of this evaluation are: i) to analyze the behavior of real DNN models in the presence of failures and ii) provide results that can be used by designers or developers to improve the fault tolerance mechanisms of their models.

Methodology

In this paper, we present the evaluation of the resilience of three pretrained DNN models. Two of them are convolutional neural networks (CNN) and one is a residual neural network (RNN). These models were implemented in Python using the Pytorch framework, since Pytorch tensors can be used and manipulated like NumPy arrays and therefore more amenable to our research purposes. Also, Pytorch provides the possibility of implementing pretrained models with the ImageNet dataset.

Failure Injection

To test the resilience of DNN models, we developed a program capable to inject failures into a DNN model. There are many applications or APIs to perform fault injection (FI) [1][2][3] in programs, but these are not specifically designed to perform the FI in DNN models. Nevertheless, there are other studies that have developed and implemented FI mechanisms to understand the relationship between fault rate and model accuracy, and to test the resilience of DNN models [4][5][6][7]. Our work differs from these papers due to the fact that we focus on the development of experiments with different failure rates and types of perturbations using our failure injection program. Also, we perform the evaluation using a modern and popular deep learning framework, a different image dataset and different deep learning models to those presented in the literature.

The fault injector was developed to inject failures into the weights of the layers of a model. We get the layers of the DNN to manipulate later the tensors that they contain. To determine where to perform the injection, the program randomly selects the layers and the weights. The fault injector uses three parameters to perform the injection: i) model in which the failures are injected, ii) failure rate, is the factor to get the total number of failures to be injected, iii) perturbation type, which is the type of failure to be injected.

The fault injector determines the number of layers of each model. For each layer there was an extraction of a tensor that contains all the pretrained weights. Also, it is necessary to extract the shape of the layer to determine the size of its components (number of output channels and the number of filters with its weight and height). With the size, we can determine in which item to perform the injection and we can calculate the total amount of weights of a model to apply the failure rate factor.

Parameters of the experiment

We implemented the failure injection in three DNN models: VGG19, RESNET50, and InceptionV3. These models were pretrained with the ImageNet dataset. We used two factors in the experiment to change the behavior of the fault injector: i) The failure rate factor (*frate*). We used three different *frates* 0.00001, 0.00003 and 0.00005 to determine the number of failures to be injected based on the total weights of a model. These *frates* were selected based on a previous experiment, in which we tested the sensibility of a model with different failure injections. ii) The perturbation factor, which may be either all values set to zero or all values set to random values.

Experiment execution

ImageNet dataset was used with 1,000 different images to perform each experiment and replicas. We executed 21 experiments, 3 without failures (1 experiment per model) and 18 experiments with failure injection in a different factors combination (3 models times 3 *frates* times 2 perturbations). For each of the 18 experiments, we performed 10 replicas.

Results Analysis

To analyze the results of the experiments we used descriptive statistics and a bit of analysis of variance (ANOVA). Descriptive statistics help us to determine the dispersion and the symmetry of the results (error data) and to find outliers that could perturb the analysis. The analysis of variance provides us with a statistical test of whether two or more population means are equal to determine differences among the models regarding its failure resilience.

Experimental results

After the experiment execution, we got 183 mean error values (3 from model execution without failure and 180 from replicas). Figure 1 shows the behavior of the three models without failure and with three different frates values. Note that in all cases InceptionV3 has the lowest error value, showing that is the model with the highest classification error on average. Also, we can see that the models were susceptible to the increment of the frate and only the InceptionV3 model kept the error stable with frates 0.00003 and 0.00005. Note, that the InceptionV3 model is not affected in the same way by the frate as with the other two models. Also, we can see that the VGG19 and ResNet50 models are very similar. These two models have a similar behavior regarding the increase in failures.

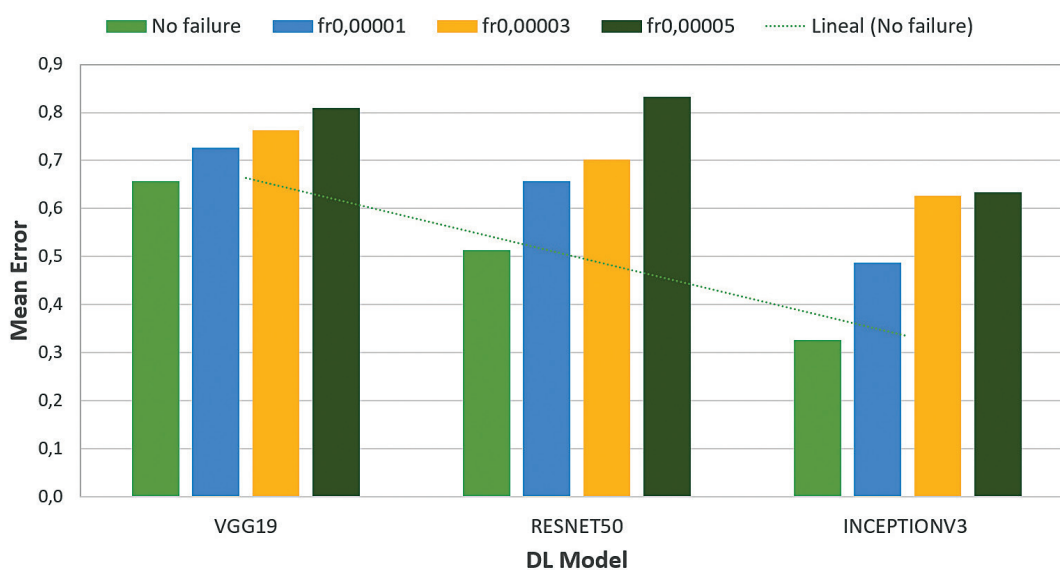


Figure 1. Model sensitivity to failures

We also analyzed the mean error of the replicas per each model and frate. Figure 2 shows that the VGG19 model has the lowest dispersion. The mean error values of VGG19 in all replicas were similar, unlike the ReNet50 and InceptionV3 that show high dispersion, especially with the frate 0.0003. The low dispersion in VGG19 shows that the mean error results of each replica did not differ significantly. Also, note that in with frate=0000.1 there is a negative skew in the interquartile range (IQR) and the minimum and maximum are similar, showing that the failure injection with a frate=0.00001 did not significantly impact in the classification process of all models. The InceptionV3 model shows a positive skew with the frates 0.00003 and 0.00004. Despite the InceptionV3 has the lowest error, it shows that the results with the FI can significantly vary between them.

Figure 3 shows standard deviation ranges that are part of the test for equality of variances. We can see that the frate intervals overlap on each model. That is an indicator that the standard deviation of the results in the three models was not significantly different. Also, we determine that there is equality of variances among the three frates of the ResNet50 and InceptionV3 models. The test for VGG19 determines that there was a statistically significant difference to accept the null hypothesis (all the variances are equal) due that the p-value < 0.05. Based on the former premise, we can conclude that the ResNet50 and Inception Models have a similar behavior through different FI rates.

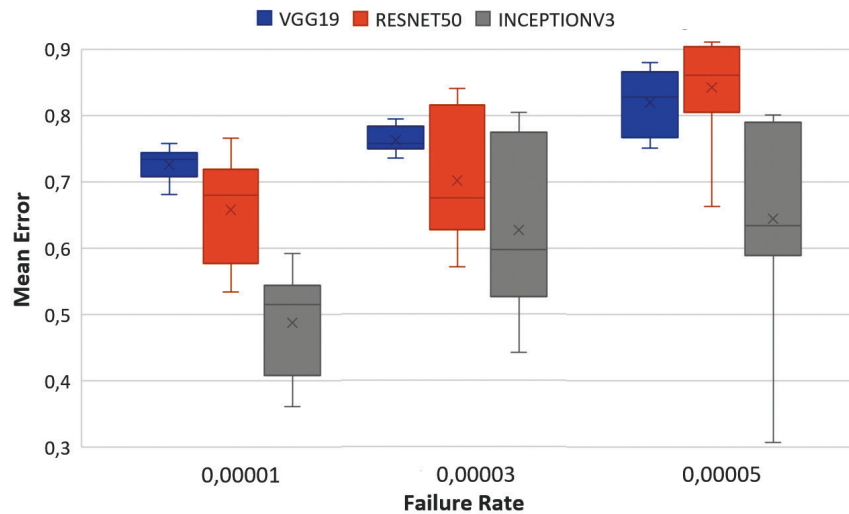


Figure 2. Model reliability according to frates

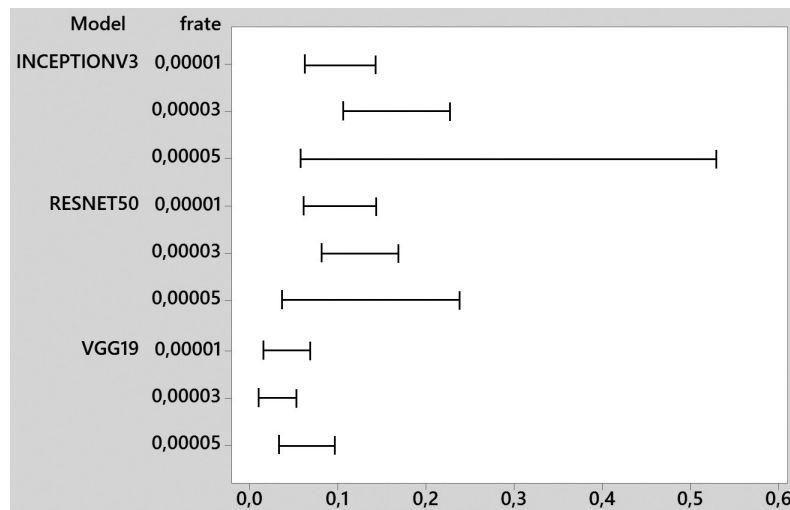


Figure 3. Standard deviation ranges

Final remarks

The experiments provide a useful review of the behavior of three real DNN models under three different failure injection rates. The use of descriptive statistics and the analysis of variance helps us to analyze the results, which can be used to deeply analyze why the deep learning models are affected differently in the presence of failures. Also, the statistical analysis could help other researchers develop methods and techniques to improve the resilience of deep learning models.

We conclude that the models were susceptible to the increase of failures in the classification process and there are models that despite the failures increase can keep the resulting error values low. We plan on extending the current analysis, expanding the statistical analysis with implementing more complex experiments. Also, it is important to analyze other scenarios, increasing the number of models, types of perturbations and the number of replicas.

References

- [1] Qining Lu, Mostafa Farahani, Jiesheng Wei, Anna Thomas, and Karthik Pattabiraman. Lfi: an intermediate code-level fault injection tool for hardware faults. In Software Quality, Reliability and Security (QRS), 2015 IEEE International Conference on Software Quality, Reliability and Security.
- [2]. K. S. Hari, S. Adve, H. Naeimi, and P. Ramachandran, "Relyzer: exploiting application-level fault equivalence to analyze application resiliency to transient faults,". ACM SIGARCH Computer Architecture News , vol. 40, p. 123, 04 2012.
- [3] U. Schiffel, A. Schmitt, M. Süßkraut and C. Fetzer, "Slice Your Bug: Debugging Error Detection Mechanisms Using Error Injection Slicing," 2010 European Dependable Computing Conference, Valencia, 2010, pp. 13-22.
- [4] TensorFI: A configurable fault injector for TensorFlow Applications", Guanpeng Li, Karthik Pattabiraman and Nathan DeBardeleben, 8th IEEE International Workshop on Software Certification (WoSoCER), 2018.
- [5] B. Reagen, U. Gupta, L. Pentecost, P. Whatmough, S. K. Lee, N. Mulholland, D. Brooks, and G.-Y. Wei, "Ares: A framework for quantifying the resilience of deep neural networks," in Proceedings of the 55th Annual Design Automation Conference, ser. DAC '18. New York, NY, USA: ACM, 2018, pp. 17:1–17:6.
- [6] Y. Liu, L. Wei, B. Luo and Q. Xu, "Fault injection attack on deep neural network," 2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), Irvine, CA, 2017, pp. 131-138.
- [7] G. Li, S. K. S. Hari, M. Sullivan, T. Tsai, K. Pattabiraman, J. Emer, and S. W. Keckler, "Understanding error propagation in deep learning neural network (dnn) accelerators and applications," in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, ser. SC '17. New York, NY, USA: ACM, 2017, pp. 8:1–8:12.

Proposal for metabolic flux pathways comparison

Propuesta para la comparación de flujos metabólicos

Esteban Arias-Méndez¹, Alonso Montero-Marín²,
Francisco J. Torres-Rojas³

Arias-Méndez, E; Montero-Marín, A; Torres-Rojas, F. Proposal for metabolic flux pathways comparison. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 31-37.

 <https://doi.org/10.18845/tm.v33i5.5072>



- 1 Escuela de Computación, Instituto Tecnológico de Costa Rica, Campus Cartago, Costa Rica. Profesor. Correo electrónico: esteban.arias@tec.ac.cr.
 <https://orcid.org/0000-0002-5600-8381>
- 2 Instituto Tecnológico de Costa Rica, Campus Cartago, Costa Rica. Estudiante. Escuela de Computación. Correo electrónico: luimontero@ic-itcr.ac.cr.
 <https://orcid.org/0000-0002-8206-5312>
- 3 Escuela de Computación, Instituto Tecnológico de Costa Rica, Campus Local San José, Costa Rica. Profesor. Correo electrónico: ftorres@tec.ac.cr.
 <https://orcid.org/0000-0001-7075-3305>

Keywords

Graph traversal; weighted graphs; Needleman-Wunsch algorithm; global alignment; Smith-Waterman algorithm; local alignment.

Abstract

Metabolic flux pathway analysis can provide important information for a better understanding of life and all its processes directly benefiting areas like medicine, agronomy, pharmacy and others alike. Some of the main tools used to study and analyze metabolic pathways have been based on the idea of pathway comparison, using graph data structures. Some of those tasks are considered hard computational problems. On the other hand, those comparisons have not yet taken into consideration the metabolic flux as part of the pathway or metabolic process. It means, to consider how much of a metabolite passes through a reaction system over time. We propose here a simple way to compare metabolic pathways using its related flux information by a simple metabolic pathway comparison method introduced in 2017 and adjusting it to weighted graphs. The algorithms analyze the associated weighted graphs of metabolic flux pathways and provide a fast scoring of its flux similarities in the first place and a list of similarities and differences between the given flux pathways, listed as pathways. We provide some insights into the analysis follow to get a good score system when comparing metabolic pathways related weighted graphs in a low-cost computation.

Palabras clave

Recorrido de grafos; grafos ponderados; algoritmo Needleman-Wunsch; alineamiento global; algoritmo Smith-Waterman; alineamiento local.

Resumen

El análisis de flujos metabólicos puede proporcionar información importante para una mejor comprensión de la vida y todos sus procesos, beneficiando directamente a áreas como la medicina, la agronomía, la farmacia y otras. Algunas de las principales herramientas utilizadas para estudiar y analizar las rutas metabólicas se han basado en la idea de la comparación de rutas metabólicas, utilizando estructuras de datos como grafos. Algunas de esas tareas se consideran problemas computacionales difíciles. Por otro lado, esas comparaciones aún no han tenido en cuenta el flujo metabólico como parte de la vía o proceso metabólico. Es decir, considerar cuánta cantidad de un metabolito pasa a través de un sistema de reacción con el tiempo. Proponemos aquí una forma simple de comparar rutas metabólicas utilizando su información de flujo relacionada mediante un método simple de comparación de rutas metabólicas introducido en 2017 y ajustándolo a grafos ponderados. Los algoritmos analizan los grafos ponderados asociados de las rutas de flujo metabólico y proporcionan una puntuación rápida de sus similitudes de flujo en primer lugar y una lista de similitudes y diferencias entre las rutas de flujo dadas, enumeradas como rutas. Proporcionamos algunas ideas sobre el análisis a continuación para obtener un buen sistema de puntuación al comparar grafos ponderados relacionados con las rutas metabólicas en un cálculo de bajo costo.

Introduction

Metabolic flux is the passage of a metabolite through a reaction system over time, and flux analysis is the combination of time-course methodologies in metabolomics and computational modeling of pathways [1]. A metabolic pathway is an ordered sequence of biochemical reactions between various actors named metabolites, these are substrates that are transformed into a product through a series of reactions catalyzed by enzymes [2], [3].

In graph theory, a graph is a structure consisting of a set of objects in which some pairs of the objects are in some sense “related”. The objects correspond to mathematical abstractions called vertices (also called nodes or points) and each of the related pairs of vertices is called an edge (also called link or line) [4]. Additional info can be added to a graph by assigning a weight to each edge of the graph. Weighted graphs are used to represent structures in which pairwise connections have some numerical values, usually representing a value of the strength or cost of the relationship between a pair of nodes.

Graph comparison is usually referred by the formal notion of *graph isomorphism* which captures the informal notion that some graphs have a *similar structure*. In graph theory, an isomorphism of graphs G and H is a bijection between the vertex sets of G and H , such that any two vertices u and v of G are adjacent in G if and only if $f(u)$ and $f(v)$ are adjacent in H . A problem with isomorphism alone is that it is usually specified for graphs with similar count of nodes and edges.

That being said, it means that graphs of different sizes cannot be well compared using this measure; however, in this present work we would like to present an application to compare weighted graphs representing metabolic fluxes, which had not been previously considered in many analysis about the comparison of metabolic pathways itself. Weighted graphs are very closely related to the topic of metabolic flux pathways, there is a great interest in its study and analysis, and weighted graphs have been used to describe the result of the flux discovery or prediction [5]. Flux analysis can be use also in other fields measuring different processes and its related behavior inside them. We present weighted graph comparison from the point of view of fluxes, determining how similar (structurally: nodes and its related weighted relations) a given pair of flux pathways are.

Many different techniques have been developed for the alignment and comparison of general graphs and the interesting routes inside them. Graph comparison is a computationally difficult task [6], [7], most of these comparisons can be represented as problems in the NP-Complete complexity class, which means there are currently no efficient algorithms for solving them. In some works, like [8] [9], [10], some heuristic techniques have been applied in order to reduce the time taken by graph-alignment algorithms. However, none have considered weighted graphs.

Algorithms

On a previous work [11], two different low-cost approaches were developed as simple mechanisms for the comparison of two metabolic pathways that can be used as a previous step to a deeper and more time-consuming analysis to be applied for the graph comparison associated to the pathways. Later, as an extension of the original algorithms, they were applied to a more general kind of graphs [12]. We propose here a simple way to compare metabolic flux pathways, using weighted graphs, based on the ideas of the previous works. These later works took into consideration the way to compare graphs with different base structures and allow us to look for subgraphs or parts of graphs similar between them.

Methodology

Graphs representing any metabolic pathway or metabolic flux pathway are special graphs, for two key reasons: 1) since the information on the graph is not represented just as a regular directed or weighted graph but representing the order of a reaction system or process. 2) the nodes on each graph are not any but distinguished nodes and we use this to simplify the extraction of information from the corresponding graphs and the comparison process.

Algorithm A - Transformation from 2D to 1D and alignment: we must convert the 2D metabolic flux pathway related graph into a linear version of it, now taking into consideration the weights related. For this graph traversal we don't look for the first metabolite in the reaction, but we start with the nodes associated to the edge with the higher value and then continuing with the lower values while writing down the nodes visited following this path in a decreasing order. The process is applied to both metabolic flux pathways associated graphs to be compared and then regular alignment algorithms are applied (global and local). Exactly the opposite is applied when the interest is to start from the edge with less weight or lower value to the heaviest or highest one. In this case we don't look necessarily to visit all the nodes in the graph, but the most important (heaviest) pathway.

Algorithm B - List the Differences: we look for the differences between the metabolic flux pathways. What we do is to take out the edges with equal value between the same couple nodes in both graphs, so we can say it is the same relation. A correctness adjustment value can be introduced in this comparison in order to allow a gain or loss of an amount or value indicated by the user, like a +/- 5 value for instance. So, if the edge is similar inside that weight range (and its related nodes) then the relation is considered similar and annotated. Also, the user can indicate if the weights of the graphs should be normalized or no; this means, all the weights in the graph must sum 100.00 or 1.00 so, all the weights represent its correspondent percentage of the total value of the graph. This is to make possible to compare metabolic flux pathways with different weight scales, but that might have similar flux structures. A list of the found differences (weighted reactions) is given to the user.

Discussion

To provide insights on how our algorithms works we provide a sample case to show the outputs of the algorithms and its implementations. Let's consider a pair of metabolic flux pathways shown on figure 1 and figure 2, in an intuitive way, they look very similar and we want to reinforce that with the data extracted.

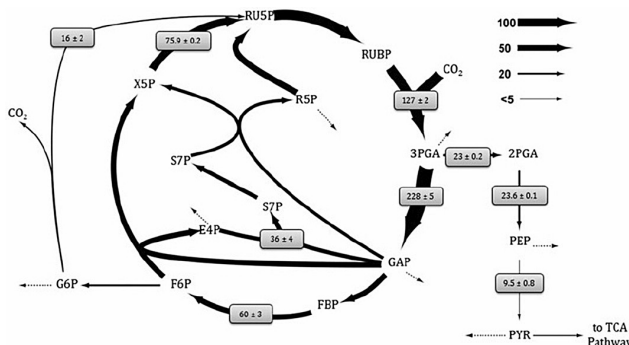


Figure 1. Metabolic Flux Pathway: MFP 1. This flux map shows the estimated fluxes associated with glycolysis and the Calvin cycle for a *Synechocystis* INST-MFA study. From [5].

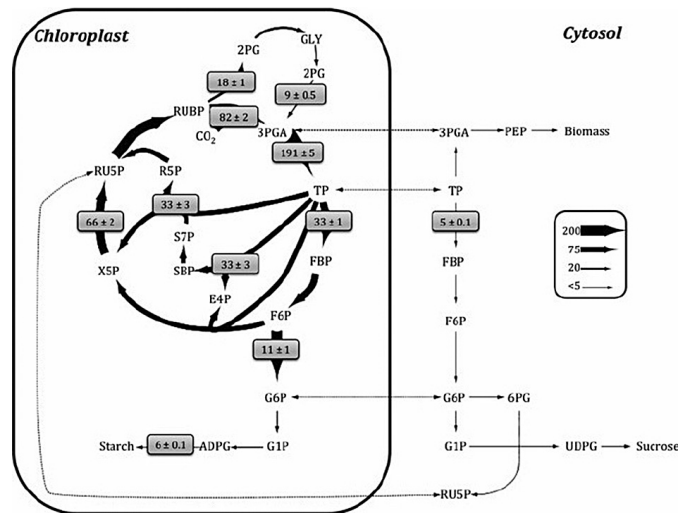


Figure 2. Metabolic Flux Pathway: MFP 2. This flux map shows a hypothetical flux map associated with a plant INST-MFA study involving multiple subcellular compartments. From [5].

For the first algorithm we show the traversal outputs of the associated graphs to the metabolic flux pathways MFP 1 and MFP 2 based on the weights only, from the highest value to the minimum. Not all the values of the original pathway may be considered, since some of the metabolite flux passages can produce cycles or some other fluxes might go to some other reactions. We want to focus on the major flux.

MFP 1 traversal: RUSP - RUBP - 3PGA - 6AP - FBP - F6P - X5P

MFP 2 traversal: RUSP - RUBP - 3PGA - TP - S7P - X5P

Then, for these outputs the corresponding Global and Local alignments of pathways MFP 1 and MFP 2, using standard values match=+1, mismatch=-1, gap=-2, we obtained a Global score = 0, and Local score = 3, with alignment: RUSP RUBP 3PGA. We can observe here that the local score value shows a similarity between pathways MFP 1 and MFP 2 between the nodes RUSP to 3PGA.

For the second algorithm we want to find the differences between the weighted paths in the graphs, to discover the broader similar subgraph from the point of view of the associated weights to each pathway, so we can compare better the flux between the given graphs. Let's consider that here we focus in the similar relationship between the nodes and its associated weight, so that an edge connecting two nodes is considered relevant for the comparison if it shows equal or similar weight for the same relation in the other graph. We consider in our algorithm a delta value as a range difference permitted when comparing a couple edge's weights. So, if the same relation is found between the same two nodes in both graphs, but the weights associated to each edge in each graph are not equal or they differ on a maximum delta value given, we don't consider this edge as similar.

Comparing then the pathways MFP 1 and MFP 2 in this way we get the outputs below:

MFP 1 - MFP 2 - Similar weighted paths: RUSP-RUBP-3PGA, FBP-F6P-X5P-RUSP, S7P- X5P-RUSP, R5P-RUSP

MFP 1 - MFP 2 - Differences: 3PGA-2PG-GLY, 3PGA-2PGA-PEP-PYR, 3PGA-GAP-RSP, 3PGA-GAP-S7P, 3PGA-GAP-E4P, 3PGA-GAP-FBP-F6P-G6P-RUSP

Conclusions

As we can see in the outputs obtained, we don't look to propose which metabolic flux is more efficient or try to get a perfect match between a pair of given pathways, but to provide some insights on the most relevant information on a couple hopefully related processes. This means provide the part of the reaction (sub-graph) more similar between the pathways that matches the related metabolites nodes and its associated weights, for the first case, and the list of main different parts of the pathways, for the second algorithm.

As explained before, the fact that the graphs represents metabolic information is considered: the specific order in the reactions (edges) and corresponding metabolites (nodes), and that each node represents a distinguished metabolite; all this is used to minimize the amount of comparisons made and simplify the extraction of information from the graphs, and to analyze that information. For metabolic flux pathway comparison, we established that the mechanism proposed by the first approach transforming a 2D structure to 1D structure following its weights for later alignment and evaluation can be used as a testing evaluation to predict good comparison results in case a deeper analysis is desired. For the second approach we want to offer to the expert an additional point of view for evaluations about the metabolic flux pathways being compared. In this case, no score is provided but the listed similarities and differences.

The proposed algorithms are fast and of relatively low computational cost, it is possible to provide relevant information for the comparison study about metabolic flux pathways of interest and some other derived analyzes. On the other hand, there are currently not many information available about metabolic flux pathways on databases, most of this information comes on papers analyzing and interpreting the meaning of the fluxes on a metabolic process in study. So, it's our interest to provide this proposal for a future necessity of comparing more discovered metabolic flux pathways.

Acknowledgment

We would like to thank the support provided by the Computer Science School of Instituto Tecnológico de Costa Rica and the Vicerectory of Research and Extension.

References

- [1] S. Padmanabhan, *Handbook of pharmacogenomics and stratified medicine*. Academic Press. 2014.
- [2] J. M. Lee, E. P. Gianchandani, J. A. Eddy, & J. A. Papin, "Dynamic analysis of integrated signaling, metabolic, and regulatory networks". *PLoS computational biology*, e1000086, 2008.
- [3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, & P. Walter, "Molecular biology of the cell." *Garland Science*. 2007.
- [4] R. J. Trudeau, R. J. (2013). *Introduction to graph theory*. Courier Corporation.
- [5] J. D. Young, A. A. Shastri, G. Stephanopoulos, & J. A. Morgan, "Mapping photoautotrophic metabolism with isotopically nonstationary ^{13}C flux analysis". *Metabolic Engineering*, pp. 656-665, 2011.
- [6] G. Abaka, T. Biyikoglu, & C. Erten, "CAMPways: constrained alignment framework for the comparative analysis of a pair of metabolic pathways." *Bioinformatics*, pp. 145-153. 2013.
- [7] F. Ay, M. Kellis, & T. Kahveci, "SubMAP: aligning metabolic pathways with subnetwork mappings". *Journal of Computational Biology*, pp. 219-235, 2011.
- [8] O. Kuchaiev, & N. Przulj, "Integrative network alignment reveals large regions of global network similarity in yeast and human". *Bioinformatics*, pp. 1390-1396, 2011.
- [9] R. Patro, & C. Kingsford, "Global network alignment using multiscale spectral signatures". *Bioinformatics*, pp. 3105-3114, 2012.
- [10] R. Y. Pinter, O. Rokhlenko, E. Yeger-Lotem, & M. Ziv-Ukelson, "Alignment of metabolic pathways". *Bioinformatics*, pp. 3401-3408, 2005.

- [11] E. Arias-Mendez, & F. Torres-Rojas, "Alternative low cost algorithms for metabolic pathway comparison." *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, Funchal, Portugal: IEEE. pp. 1-9, 2017.
- [12] E. Arias-Mendez, A. Montero-Marin, D. Chaves-Chaves, & F. J. Torres-Rojas, "Simple Graph Comparison Inspired on Metabolic Pathway Correlation". *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)* San Carlos, Costa Rica: IEEE. pp. 1-8, 2018.


Estimating the redshift of galaxies from their photometric colors using machine learning methods

Estimación del corrimiento al rojo para galaxias a partir de sus colores fotométricos usando métodos de aprendizaje automático

Felipe Meza-Obando¹

Meza-Obando, F. Estimating the redshift of galaxies from their photometric colors using machine learning methods. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 38-43.

 <https://doi.org/10.18845/tm.v33i5.5073>

¹ Doctorate in Engineering Candidate, Artificial Intelligence for Natural Science Lab (LIANA) and PAttern Recognition and Machine Learning (PARMA)- Instituto Tecnológico de Costa Rica (Thanks to Vicerrectoría de Investigación y Extensión (VIE) at Instituto Tecnológico de Costa Rica for supporting this research at LIANA.). E-mail: fmeza@tec.ac.cr.
 <https://orcid.org/0000-0003-4239-3116>



Keywords

Universe; expansion; redshift; galaxies; svm; decision trees; ada boost; random forest.

Abstract

The determination of the redshift, a factor also known as z , is obtained from variations in the wavelength's spectrum of galaxies or distant objects, such variation is basically the difference between the wavelength measure on Earth of the element present in the galaxy and the direct measure of the same element on the object by the use of spectroscopy. From the value z , it's possible to obtain the values of the object's distance and the speed at which it moves away from us. Obtaining spectroscopic data directly from astronomical objects, is not always an easy task to run and the use of color index become a more accessible alternative for many researchers. In this work we present the preliminary results of several machine learning methods, using regression based algorithms. The goal will be to obtain the value of z , from the photometric colors.

Palabras clave

Universo; expansión; desplazamiento al rojo; galaxias svm; árboles de decisión; bosque al azar.

Resumen

La determinación del corrimiento al rojo, factor conocido como z , se obtiene a partir de las variaciones en la longitud de onda del espectro de la galaxia u objeto lejano, dicha variación se da entre la medición en la Tierra del elemento presente en la galaxia y la medición directamente en el objeto mediante espectroscopia. A partir del valor z , es posible obtener los valores de la distancia del objeto y la velocidad a la que se aleja de nosotros. La obtención de datos espectroscópicos en el objeto, no siempre resultan fáciles de obtener y los índices de color se convierten en una alternativa más accesible para muchos investigadores, en este trabajo se muestran los resultados preliminares de diversos métodos de aprendizaje automático, donde como un problema de regresión y a partir de los índices fotométricos podemos estimar el valor de z .

Introduction

Since the beginning of our universe 14 billions of years ago, the space around it has been expanding, for this reason, the galaxies and other objects seem to be moving away from us at specific rates. Due to the uniformity of such expansion, a direct relationship is present between the speed of expansion of those galaxies and the distance from us. The resulting movement of the expansion, as shown if figure 1, causes a shift in the frequency of photons that can be visualized in the spectrum of the distant galaxies, since the universe is expanding away from us, the shift is towards lower frequencies i.e red side of the spectrum, the effect is called redshift.

Spectral analysis consists in the measurement of the emission of photons at certain wavelengths that can be represented as spectral lines in specific positions in the resulting spectrum, from those lines one can determine the elements present in the object [2]. When astronomers measure spectral lines in distant objects such as galaxies, lines appear to be shifted toward the red side of the spectrum due to the difference between the wavelength measured on earth and the one observed, the redshift is then defined by a value called z :

$$z = \frac{\lambda_{obs} - \lambda_{emit}}{\lambda_{emit}}$$

From the redshift value z , it's possible to obtain the velocity v at which the galaxy is moving away from us using the relation $v = c \times z$, where c is the speed of light. Finally, with the value of v it's also possible to estimate the distance of the galaxy using the Hubble Constant ($H_0=72 \text{ Km/s/Mpc}$) and the following relation $d=v/H_0$

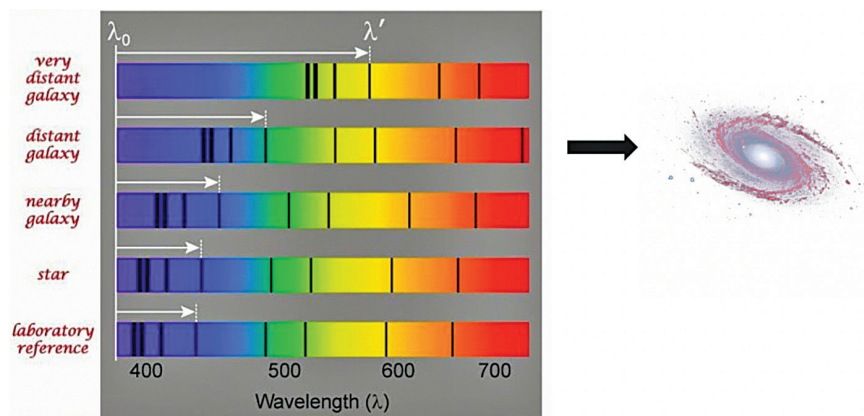


Figure 1. Comparison of different redshifts (Source: <https://www.universetoday.com/>)

The conventional way to do this analysis is by the observation of the samples one by one of the resulting lines in the spectrums and calculate the observed shift, however with large amounts of data this process can be very time consuming and subjected to errors. On the other hand, not all the galaxies have all the spectral values related to the shift, for those reasons the use of machine learning results in a powerful and convenient option to estimate the values of z , from photometric colors.

Machine Learning (ML) Algorithms

We've chosen four basic supervised methods to estimate the value of z from the photometrics filters [3], which results in a very convenient way to estimate z since the observational data i.e color index is generally more available and easier to be obtained than the spectral data for such galaxies. The methods selected are briefly described [1], [5]:

- *Support Vector Machines (SVM)*: In this case is called Support Vector Regression (SVR), essentially a support vector (SV) is generated to follow the data trend and create the regression model.
- *Decision Trees (DT)*: This method maps a set of inputs features to their corresponding output targets, thanks to a series of individual decisions where each decision is represented by a node of the tree. Each decision node involves specific values, such values are determined from the training data in the learning algorithm based on the principle of "the most effective way to fit the data".
- *Random Forest (RF)*: The goal is to have several random DT, each tree is created from random sets of samples from the data and each node is also created by selecting randomly the features to generate the best data fit.
- *AdaBoost (AB)*: Adaptive Boosting is based on the idea of having several weak classifiers or regressors to form together a much stronger and effective classifier or as in this case, a regressor.

Methodology and data

We use the spectrums from Sloan Digital Sky Survey catalogue (SDSS) [4], it's of particular interest the magnitude of the flux received in five bands (u, g, r, i and z). In astrophysics the color index is the result of the difference between the magnitude of two filters i.e u-g, g-r, r-i and i-z, those values will be our features. The u filter is near the blue side of the spectrum and the z filter near the red side of the spectrum, if the spectrum is shifted to the red side then the value of z will be larger, meaning that the galaxy is moving away from us. In figure 2 the blue spectrum corresponds to the measurement at $z=0$ in other words our reference, however as the galaxy is moving away the z value will be increased towards the red side, resulting in a speed v and a distance d of the galaxy from our perspective. Our database was made up of samples with values of flux in the u, g, r, i and z bands and is composed of 50000 galaxies; table 1 shows the first and last three values of our dataset.

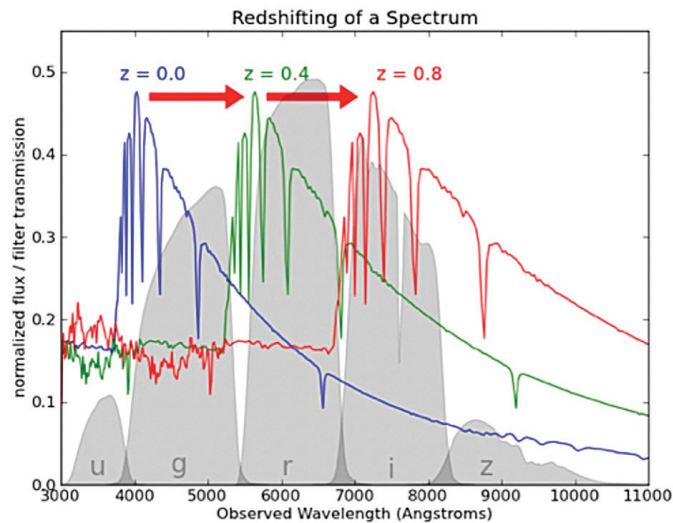


Figure 2. Redshifting of a spectrum (Source: <https://scikit-learn.org/>)

With the goal of making use of conventional ML algorithms to solve this regression problem, we based our analysis in the measurement of the mean error as our metric:

$$med_{diff} = median(|Y_{i,predicted} - Y_{i,target}|)$$

Table 1. Samples from the database (Source: original dataset from SDSS)

ID	u	g	r	i	z	redshift
0	19.84132	19.52656	19.46946	19.17955	19.10763	0.539301
1	19.86318	18.66298	17.84272	17.38978	17.14313	0.164570
2	19.97362	18.31421	17.47922	17.07440	16.76174	0.041900
49997	19.82667	18.10038	17.16133	16.57960	16.19755	0.078459
49998	19.98672	19.75385	19.57130	19.27739	19.25895	1.567295
49999	18.00024	17.80957	17.77302	17.72663	17.72640	0.474945

Results

In the experiments, we used scikit-learn framework to run a comparison between the four proposed methods. In the case of SVM, the kernel used was rbf (SVM-rbf), with decision trees (DT) we used the depth as our hyperparameter, even with the use of several depth values the results remain almost the same. With Random Forrest we used two references values for the `n_estimators` hyperparameter, 20 and 100 (RF-20 and RF-100), beyond the value of 200 the results seems to have no major variations, finally with Ada Boost we did the same setting as with RF to create two models (ADA-20 and ADA-100). Considering that the value of `z` is mostly in a range between 0 and 1 in our database, the mean error obtained for the all the methods varied from 0.001 to 0.003, in the order of 10^{-3} a variation between 10 and 30, that represent a good result considering this as our first approach, however RF algorithm was the one that delivered the best result of 0.001, for this reason is considered the best result. In the case of ada boost method, one particular experiment using the hyperparameter `n_estimators` with a value of 10, resulted in the poorest mean error in the margin of 0.008, not bad at all but the lowest obtained from the experiments. As expected SVM reached the highest computational cost, due to the nature of the algorithm. The results obtained in regards to the mean error are shown in figure 3. In terms of the time consumed per algorithm, all the measurements were taken using the same computer: Apple MacBook Pro 2015 with processor 2.5 GHz Intel Core i7 and 16 GB 1600 MHz DDR3 memory, as expected SVM-rbf was the highest in terms of computational cost and RF was not just the best in terms of the mean error but also one of the most effective (RF-20) in terms of computational cost.

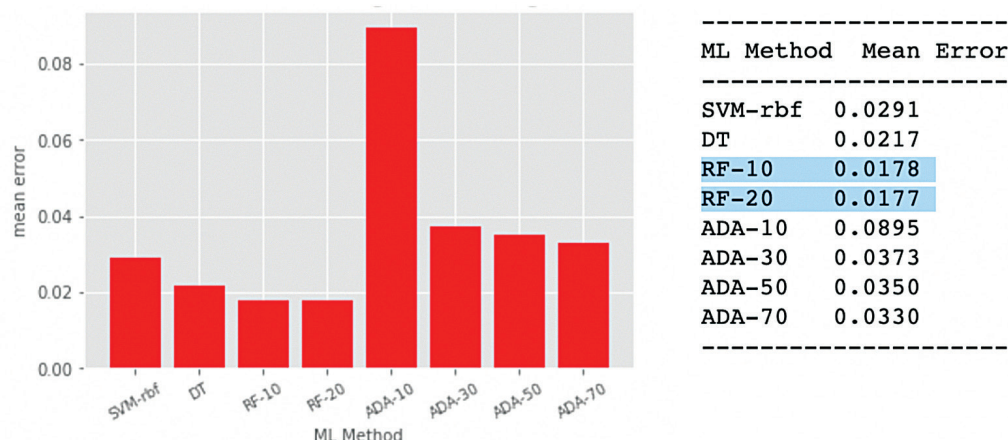


Figure 3. Mean error of the redshift in the galaxies, using the ML proposed methods.

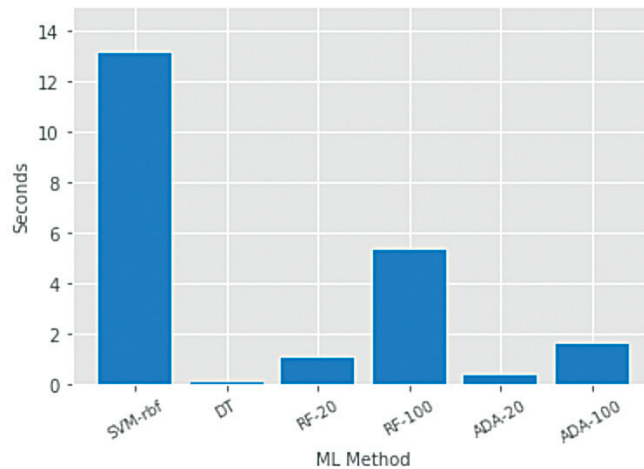


Figure 4. Time consumed per each ML method.

Conclusions and future work

The regression methods selected for the series of experiments resulted in a good approach to estimate the redshift value z , the use of mean error represented also a good metric due to the nature of the regression model. Even with the good results obtained with the proposed methods, actual analysis is concentrated in the minimization of the mean error through the use of artificial neural networks, another potential improvement to the model consist in the inclusion of more data to the current dataset as it become available. Other tasks are concentrated in the analysis of variations in the splitting of the dataset in order to validate and measure any overfitting effects on the results.

References

- [1] S. Marsland, "Machine Learning: An Algorithmic Perspective", 2nd Edition, C&H/CRC, 2014.
- [2] J. Rich, "Fundamentals of Cosmology", Springer-Verlag Berlin Heidelberg, 2010.
- [3] C. Davis, et al., "Accurate redshift estimation from photometric colors". Physics Stanford. 2013.
- [4] Sloan Digital Sky Survey, Data Access, Retrieved from <https://www.sdss.org/> , January 2019.
- [5] E. Alpaydin, "Introduction to Machine Learning", 2nd Edition, The MIT Press, 2016.

Executing and Pausing Distributed Applications Running on Desktop Clouds by Global Snapshots

Ejecutando y Pausando Aplicaciones Distribuidas Corriendo sobre Desktop Clouds Mediante Snapshots Globales

Carlos E. Gómez¹, Jaime Chavarriaga², David C. Bonilla³, Harold E. Castro⁴

Gómez, C. E.; Chavarriaga, J; Bonilla, D. C; Castro, H. E.
Executing and Pausing Distributed Applications Running on Desktop Clouds by Global Snapshots. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 44-48.

 <https://doi.org/10.18845/tm.v33i5.5074>

- 1 Systems and Computing Engineering Department Universidad de los Andes, Bogotá, Colombia and Universidad del Quindío, Armenia, Colombia. E-mail: ce.gomez10@uniandes.edu.co.
 <https://orcid.org/0000-0002-5202-1167>
- 2 Systems and Computing Engineering Department Universidad de los Andes, Bogotá, Colombia and Universidad del Quindío, Armenia, Colombia. E-mail: ja.chavarriaga908@uniandes.edu.co.
 <https://orcid.org/0000-0002-8372-667X>
- 3 Systems and Computing Engineering Department Universidad de los Andes, Bogotá, Colombia and Universidad del Quindío, Armenia, Colombia. E-mail: dc.bonilla10@uniandes.edu.co.
 <https://orcid.org/0000-0002-3834-4736>
- 4 Systems and Computing Engineering Department Universidad de los Andes, Bogotá, Colombia and Universidad del Quindío, Armenia, Colombia. E-mail: hcastro@uniandes.edu.co.
 <https://orcid.org/0000-0002-7586-9419>



Keywords

Reliability; Fault tolerance; Checkpointing; Global snapshot; Desktop clouds.

Abstract

Desktop Clouds rely on volatile computing resources. For instance, platforms such as cuCloud and UnaCloud run scientific applications in virtual machines exploiting idle resources harvested in computer labs. Regretfully, these resources can be claimed by users, turned off and faulted at any time. The application running on these platforms suffer interference and interruptions that do not occur in dedicated platforms. We have been researching how to deal with these interruptions to increase the platform reliability and support applications running for large periods of time. This paper describes an application of our Global Snapshot Protocol, which can be employed for executing and pausing distributed applications running on desktop clouds. We found that, in these environments, the number of failures caused by desktop users is greater than the caused by hardware and communications. There, when a distributed system running in the virtual machines of a desktop cloud is paused, it can be restored in the same desktops, and successfully finish the application execution.

Palabras clave

Confiabilidad; Tolerancia a fallas; Checkpointing; Snapshot Global; Desktop clouds.

Resumen

Los desktop clouds dependen de recursos computacionales volátiles. Por ejemplo, plataformas como cuCloud y UnaCloud ejecutan aplicaciones científicas en máquinas virtuales que aprovechan recursos ociosos en salas de cómputo y laboratorios. Lamentablemente, estos recursos pueden ser reclamados por los usuarios, apagados o presentar fallas en cualquier momento. La aplicación que se ejecuta en estas plataformas sufre interferencias e interrupciones que no ocurren en plataformas dedicadas. Nosotros hemos estado investigando cómo enfrentar estas interrupciones para aumentar la confiabilidad de la plataforma y soportar aplicaciones que se ejecutan durante largos períodos de tiempo. Este artículo describe una aplicación de nuestro Protocolo de Snapshot Global, el cual puede emplearse para ejecutar y pausar aplicaciones distribuidas que se ejecutan en desktop clouds. Nosotros encontramos que, en estos entornos, la cantidad de fallas causadas por los usuarios de los computadores de escritorio es mayor que la causada por el hardware y las comunicaciones. Allí, cuando se detiene un sistema distribuido que se ejecuta en las máquinas virtuales de un desktop cloud, nosotros podemos reanudar la ejecución usando los mismos computadores y finalizar exitosamente la ejecución de las aplicaciones.

Introduction

Desktop clouds (DC) offer cloud computing services using idle resources on common desktop computers [1]. Typically, DCs offer infrastructure services (IaaS) such as running virtual machines (VMs) and virtual clusters. For instance, researchers can run Bag of Tasks (BoT) applications that divide the work in chunks and distribute them on virtual machines running on multiple desktops. In addition, they can use platforms such as cuCloud [2] and UnaCloud [3] to run clusters of customized virtual machines to run more complex distributed applications. All these platforms help researchers to run scientific applications at lower costs than using dedicated datacenters.

In a DC, computing resources are very volatile because the platform harvest idle resources from desktop computers. A desktop user may claim these resources, kill some processes or turn off the physical machines (PM). DCs are more susceptible to failures at runtime than other cloud platforms. For instance, when the users run BoT applications, these failures can halt the execution of some chunk of work. There, the platform can solve these interruptions by assigning the failed chunks to other virtual machine and starting again its processing. However, when DC users run distributed applications, these failures are hard to manage. Distributed applications, such as GROMACS and other MPI-based, fail if one of the nodes is not accessible. Any failure of a virtual machine may result in the lost of many hours of work. Existing platforms do not provide strategies nor tools to offer warranties for executing distributed applications during large periods of time.

We have been researching how to improve reliability on desktop clouds. In a previous paper, we proposed a global snapshot protocol and develop a software tool to obtain a consistent global snapshot for a general distributed system running on virtual machines [4]. The software tool maintains the semantics of the distributed system without modifying applications running on virtual machines or hypervisors. In this paper, we present a use case the global snapshot system that allows the desktop cloud user to pause the execution of the distributed system that runs on virtual machines and resume its execution. This use case is important, especially when we have distributed applications whose execution is carried out for long periods of time, for example, weeks or months, it may be necessary to pause its execution, without losing the work done so far.

Global Snapshot System

A Global Snapshot creates a checkpoint for a general distributed system, preserving the state of the participant nodes and the communications among them. Using these snapshots, it is possible to resume the execution of both the processing and the communications in progress.

In [4] we presented a Global Snapshot Protocol and a software tool that can be applied to save the state of collections of VMs running on multiple computers. Here the virtual machines communicate with each other using normal network devices instead of virtual networks and it is not possible to take “almost at the same time” snapshots of VMs and the communication channels. Creating a checkpoint requires to coordinate local snapshots in each node and provide means to resume the in-progress transmissions among them.

Our implementation is able to create snapshots of distributed systems on DCs without requiring modifications on the applications, the VMs or the hypervisors. The concepts used in the implementation are the following: TCP reliability mechanisms (if applicable); colors in the packages in the nodes, in the network layer; packet filtering, again, in the network layer and we adapt a simple coordination protocol to obtain a global snapshot at any time.

The protocol is an adaptation of the two-phase commit protocol to coordinate the participating desktops. The first phase consists of a previous verification in which the coordinator (one of the participants) consults each process if the VMs corresponding to the system are in execution. If all VMs are running, we go to the second phase, which consists of taking the global snapshot. Applying the concept of colors to identify datagrams, we use colorless datagrams when the global snapshot protocol is not in operation, that is, before starting and after finishing, and we assume that a global snapshot will not start before finish the previous one. In addition, the protocol uses the operating system running in the VMs to process the network communications by modifying control fields, marking the outgoing datagrams and filtering out some of the incoming ones. The process has been designed based on the expected behavior of the TCP and UDP protocols. In this way, a datagram sent that does not reach its destination, will be retransmitted after resuming

execution, if the transport layer protocol is TCP. On the contrary, if the protocol is UDP, the packet will not be retransmitted, as it happens in this type of applications.

On the other hand, the system has a resume protocol to coordinate the process when it is desired to continue the execution of a distributed system from a global snapshot. In this case, it cannot be guaranteed that when resuming the execution of the VMs from a global snapshot, they do so at the same time. Therefore, we create three-phase protocol, to coordinate the restoration of the system, looking for communications to be restored and that the system can successfully complete its execution.

The first phase is to verify that the physical machines, where VMs are hosted, are running. If any of the PMs is not ready, a timer is activated and after a timeout, the verification starts again. When all the PMs have answered to the coordinator, we will go to phase 2. In the second phase the restore point is established in a specific snapshot (typically the last one), although it could be any. The coordinator sends a message to the other participating processes. When all the PMs have answered the coordinator, we move on to phase 3. The third phase is similar to the second one. The coordinator sends the other participants a message and they give an answer. The coordinator determines that the resumption has been successful when he receives the response from all participants.

Executing and pausing distributed applications: Experience report

In our UnaCloud, we are running GROMACS processes that spent several days, using clusters of tens of physical machines. Although we can deploy clusters with hundreds of virtual machines, the applications may run slower on them because the use of non-dedicated network connections.

We perform multiple functional and performance tests to determine the effectiveness of the tool developed. We used a virtual image with the Ubuntu Server 16.04 operating system, which occupied 3.51 GB to implement virtual machines with 1 GB of RAM, 5 GB of virtual hard disk and 1 processing core. The virtual machines were run on physical machines with an Intel Core i7-4770 processor, 20 GB of RAM and 500 GB of hard disk. We use a computer lab with 20 desktop computers, connected via a 1GB Ethernet network. We use a GROMACS-MPI benchmark in the lab, running 1 VM on each PM and 2 processes for each node. The test took approximately 11 hours to complete, and we took snapshots every 1 hour.

It resumed from snapshot # 7 and from snapshot # 11 (the last one). In both cases, the test ended successfully.

It is very important consider that probability of failure in computer labs helps to decide to run and pause the distributed system because we found that, in our case, the number of failures caused by desktop users is greater than the caused by hardware and communications. The computer labs are upgraded every three or four years and the hardware typically do not fault. According to official statistics, there were only 11 incidents for faulty hardware during the last 5 years. Our studies about failures that occur in a DC [5] shows that desktop users are the main cause of problems. If necessary, we can use our global snapshot tool to pause system execution and resume it at the desired time, avoiding losing the work done so far.

References

- [1] A. Alwabel, R. J.Walters, and G. B.Wills, "A View at Desktop Clouds," in International Workshop on Emerging Software as a Service and Analytics (ESaaS 2014), pp. 55–61, 2014.

- [2] T. M. Mengistu, A. M. Alahmadi, Y. Alsenani, A. Albuai, and D. Che, “cucloud: Volunteer computing as a service (vcaas) system,” in *International Conference on Cloud Computing*, pp. 251–264, Springer, 2018.
- [3] E. Rosales, H. Castro, and M. Villamizar, “UnaCloud: Opportunistic Cloud Computing Infrastructure as a Service,” *Cloud Computing*, pp. 187–194, 2011.
- [4] C. E. Gómez, H. E. Castro, and C. A. Varela, “Global Snapshot of a Distributed System Running on Virtual Machines,” in *2017 29th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, pp. 169–176, IEEE, 2017.
- [5] C. E. Gómez, J. Chavarriga, and H. E. Castro, “Fault characterization and mitigation strategies in desktop cloud systems,” in *Latin American High Performance Computing Conference*, (Cham), pp. 322–335, Springer International Publishing, 2019.

Understanding Variable Performance on Deep MIL Framework for the Acoustic Detection of Tropical Birds

Entendiendo el Desempeño Variable en el Marco de Trabajo MIL Profundo para la Detección Acústica de Aves Tropicales

Jorge Castro¹, Roberto Vargas-Masís², Danny Alfaro-Rojas³

Castro, J; Vargas-Masís, R; Alfaro-Rojas, D. Understanding Variable Performance on Deep MIL Framework for the Acoustic Detection of Tropical Birds. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 49-54.

 <https://doi.org/10.18845/tm.v33i5.5075>



- 1 Advanced Computing Laboratory. Costa Rica National High Technology Center. Email: jcastro@cenat.ac.cr.
 <https://orcid.org/0000-0003-1553-0461>
- 2 Laboratorio de Investigación e Innovación Tecnológica. Vicerrectoría de Investigación, Universidad Estatal a Distancia, Costa Rica. E-mail: rovargas@uned.ac.cr.
 <https://orcid.org/0000-0003-1244-4381>
- 3 Escuela de Ciencias Exactas y Naturales. Universidad Estatal a Distancia, Costa Rica. E-mail: soloard89@gmail.com.
 <https://orcid.org/0000-0001-7694-7194>

Keywords

Deep Learning; Multiple Instance Learning; Bioacoustic; Bird Detection.

Abstract

Many audio detection algorithms have been proposed to monitor birds using their vocalizations. Among these algorithms deep learning based techniques have taken the lead in terms of performance at large scale. However, usually a lot of manual work has to be done to correctly label bird vocalizations in large datasets. One way to tackle this limitation is using the Multiple Instance Learning (MIL) framework, which models each recording as a bag of instances, i.e., a collection of audio segments that is associated with a positive label if a bird is present in the recording. In this work, we modified a previously proposed Deep MIL network to predict the presence or absence of birds in audio field recordings of one minute. We explore the behavior and performance of the network when using different number of Mel-Frequency Cepstral Coefficients (MFCC) to represent the recordings. The best configuration found achieved a 0.77 F-score over the validation dataset.

Palabras clave

Aprendizaje profundo; Aprendizaje de instancias múltiples; Bioacústica; Detección de Aves.

Resumen

Se han propuesto muchos algoritmos de detección de audio para monitorear aves usando sus vocalizaciones. Entre estos algoritmos, las técnicas basadas en el aprendizaje profundo han tomado la delantera en términos de rendimiento a gran escala. Sin embargo, usualmente se requiere de mucho trabajo manual para etiquetar correctamente las vocalizaciones de aves en grandes conjuntos de datos. Una forma de abordar esta limitación es usar el marco de trabajo de aprendizaje de instancias múltiples (MIL), que modela cada grabación como una bolsa de instancias, es decir, una colección de segmentos de audio que se asocia con una etiqueta positiva si un pájaro está presente en la grabación. En este trabajo, modificamos una red profunda MIL propuesta previamente, para predecir la presencia o ausencia de aves en grabaciones de campo de un minuto. Exploramos el comportamiento y el rendimiento de la red cuando utilizamos un número diferente de coeficientes cepstrales de frecuencia de mel (MFCC) para representar las grabaciones. La mejor configuración encontrada logró un valor F de 0.77 sobre el conjunto de datos de validación.

Introduction

Birds are key to assess environmental health as indicators of anthropogenic changes [1]. Noninvasive bioacoustic monitoring methodologies present the challenge of developing algorithms to detect birds in a large number of acoustic recordings [2].

Many algorithms have been proposed to classify bird species, bird songs, and individuals [3]. Usually, bird vocalizations are segmented to improve the performance of the classifier. However, these segmentation algorithms are commonly too simple for real conditions in the field or follow a supervised learning scheme where a lot of manual work has to be done to label the vocalizations used for training [4].

The Multiple Instance Learning (MIL) scheme reduce the manual work by using higher abstraction-level labels associated to audio recordings (“bags”) instead of each individual

vocalization. The audio segments that compose a recording are known as “instances”. This approach was previously used to classify 13 bird species in ten-seconds recordings [5].

Since Deep Neural Networks (DNNs) have outperformed most of previous classification algorithms and require large training datasets with vocalizations labels [6], [7], it is naturally desirable to combine DNNs with MIL framework (Deep MIL) to reduce the amount of manual work needed. Deep MIL architecture has successfully been applied for image [8] and video [9] classification tasks. In this paper we adapted the Deep MIL architecture to predict the presence or absence of tropical birds (binary classification) in acoustic field recordings.

Deep MIL Architecture

Let $\{(X_1, Y_1), \dots, (X_N, Y_N)\}$ be the training set, where each recording X_i is composed of l audio segments $\{x_{(p,1)}, \dots, x_{(p,l)}\}$ and each output Y_i is a binary label $\{1,0\}$ that indicates the presence or absence of bird vocalizations on recording X_i . Thus, the goal is to predict Y_i for any unseen recording X_i . The Deep MIL architecture used is based on the works presented by Jiajun et al. [10] and Gao et al. [11] and it is shown in figure 1. The input of the network is a bag of l instances of size F . First, we extract features of each instance using a 1×1 convolutional layer (CL) with F input channels and P output channels, followed by a batch normalization (BN) and a rectified linear unit (ReLU). Then, we applied a second 1×1 CL with P input channels and P output channels, also followed by a BN and a ReLU. After that, another 1×1 CL with P input channels and one output channel is applied, followed by a BN to obtain one unique value per instance. Finally, a max pooling operation is applied over the instances to obtain a score for the bag. Only if the score is higher than zero birds are detected.

To train the network we use the binary cross entropy with logits loss function and the Adam optimizer. We set empirically the batch size to 50, the learning rate α to 0.001, and to 1024. We reduce α each 25 epochs by multiplying it by a factor of 0.96.

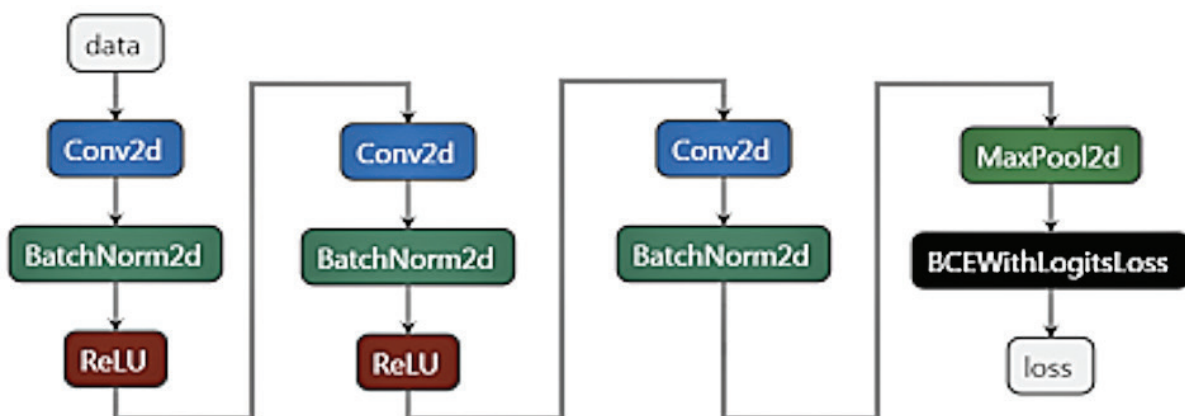


Figure 1. Deep MIL Network architecture.

Data Collection

The acoustic landscape of the upper part of the micro watershed of the Bermúdez River in Heredia Costa Rica (10°3'58.57"N; 84°4'36.39"O) was monitored using Audiomoth recorders in “wav” format, sampling rate of 48 kHz and 16 bits resolution [10]. One-minute recordings were

made over seven continuous days at ten-minute intervals. We tagged 2000 recordings where only 1206 presented any vocalization produced by 46 bird species.

Experimental Setup

Each recording was divided into 119 one-second audio segments with 50% of overlap to represent each instance. Each audio segment was also divided into 99 frames of 20 ms with 50% overlap. Then, we obtain four different instance representations based on 15, 30, 60, and 120 Mel-frequency cepstral coefficients (MFCCs) per frame. We chose MFCCs since they have been popularly used for the detection of bird sounds, especially in field conditions where variability in sound quality, diversity of species and intrinsic variability of the vocalizations increase the detection complexity and analysis [11].

We used librosa Python package [12] version 0.6.3 to implement all audio functions. To train the network we randomly divide the 2000 recordings dataset into 80% training, 10% validation, and 10% test. We used the F1-score (harmonic mean of precision and recall) to measure the performance on the training and validation datasets.

Results

The loss function decreased further as the number of MFCCs used increased, as shown in figure 2. However, the F1-score values for both training and validation datasets presented high oscillations for all the number of MFCC used, as shown in figure 3.

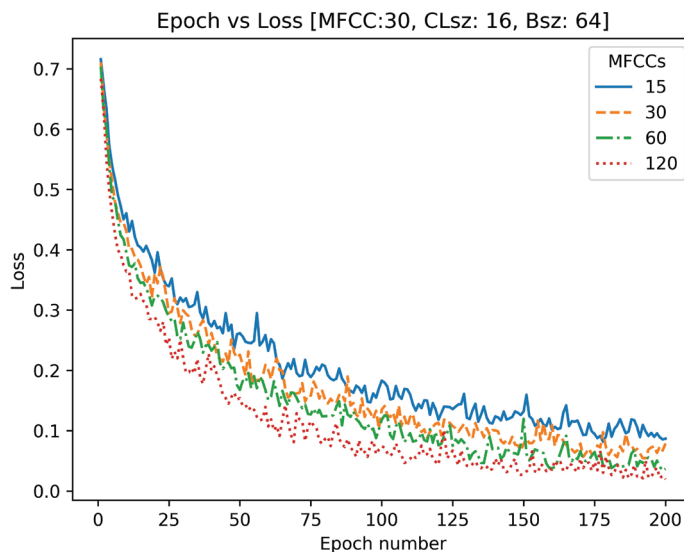


Figure 2. Loss function reduction when using 15, 30, 60 and 120 MFCCs.

As expected, we observed higher F1-score values for the training set, reaching a maximum of 0.98 at epoch 15 when using 120 MFCCs. On the other hand, the highest F1-score achieved for the validation set was 0.77 at epoch 20 when using 60 MFCCs.

Discussion and Future Work

Although the loss function is reduced further when we increase the number of MFCCs used, there are still two main issues to be addressed: the high variability in the F1-score and the high variance error.

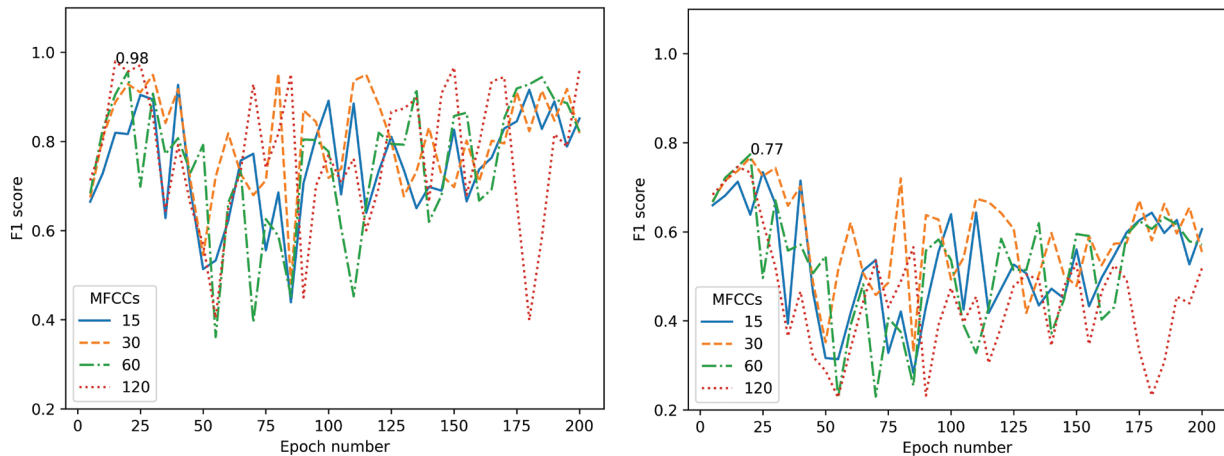


Figure 3. F1-score values in training (left) and validation (right) datasets for each MFCCs over the 200 training epochs. The F1-score was computed every 5 epochs.

The high variability in F1-score performance could be an effect of using longer recordings (60 seconds) with more instances (119) than those used in previous MIL bird classifiers (10 s recordings with an average of 19 instances) [13]. Longer training experiments could also be performed to explore if this behavior continues.

The high variance error between the training and validation sets (about 0.21 comparing their highest F1-score values) could not be reduced using dropout techniques or L2 normalization. Thus, it could also be an effect of using longer recordings with more instances as the majority of instances do not correspond to bird vocalizations and could be triggering the detection of positive bags.

Our future work is focused on exploring other data representations and network architectures to improve the F1-score performance (currently a maximum F-score of 0.77 on the validation set) and reduce its variance error. Furthermore, our long term objective is to generalize this solution using the Multiple Instance Multiple Label (MIML) framework to detect different bird species present in our dataset.

Considering the large amount of acoustic information that can be collected nowadays using autonomous recorders, it is important to develop automatic tools for the conservation and monitoring of biodiversity worldwide.

Acknowledgements

This research was partially supported by a machine allocation on Kabré supercomputer at the Costa Rica National High Technology Center.



References

- [1] R. D. Gregory and A. van Strien, "Wild Bird Indicators: Using Composite Population Trends of Birds as Measures of Environmental Health," *Ornithol. Sci.*, vol. 9, no. 1, pp. 3–22, 2010.
- [3] E. C. Knight, K. C. Hannah, G. J. Foley, C. D. Scott, R. M. Brigham, and E. Bayne, "Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs," *Avian Conserv. Ecol.*, vol. 12, no. 2, p. art14, 2017.
- [4] V. Morfi and D. Stowell, "Deep Learning for Audio Event Detection and Tagging on Low-Resource Datasets," *Appl. Sci.*, vol. 8, no. 8, p. 1397, 2018.
- [5] J. F. Ruiz-Muñoz, M. Orozco-Alzate, and G. Castellanos-Dominguez, "Multiple instance learning-based bird-song classification using unsupervised recording segmentation," in *IJCAI International Joint Conference on Artificial Intelligence*, 2015, vol. 2015-Janua, pp. 2632–2638.
- [6] E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, "Audio based bird species identification using deep learning techniques," in *CEUR Workshop Proceedings*, 2016, vol. 1609, pp. 547–559.
- [7] J. Salamon, J. P. Bello, A. Farnsworth, and S. Kelling, "Fusing shallow and deep learning for bioacoustic bird species classification," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017.
- [8] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07-12-June, pp. 3460–3469.
- [9] R. Gao, R. Feris, and K. Grauman, "Learning to Separate Object Sounds by Watching Unlabeled Video," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11207 LNCS, pp. 36–54.
- [10] J. L. Deichmann, A. Hernández-Serna, J. A. Delgado C., M. Campos-Cerqueira, and T. M. Aide, "Soundscape analysis and acoustic monitoring document impacts of natural gas exploration on biodiversity in a tropical forest," *Ecol. Indic.*, vol. 74, pp. 39–48, 2017.
- [11] S. Fagerlund, "Bird species recognition using support vector machines," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, pp. 1–8, 2007.
- [12] B. McFee et al., "librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–24.
- [13] F. Briggs et al., "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *J. Acoust. Soc. Am.*, vol. 131, no. 6, pp. 4640–4650, 2012.

Motus: A Framework for Human Motion Classification in a Not-controlled Moving Environment

Motus: Marco de trabajo para la clasificación de Captura de Movimiento humano en ambientes no controlados

Joselyn Rodríguez-González¹, María Hernández-López²,
Francisco Siles-Canales³


Rodríguez-González, J; Hernández-López, M; Siles-Canales, F. Motus: A Framework for Human Motion Classification in a Not-controlled Moving Environment. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 55-59.

 <https://doi.org/10.18845/tm.v33i5.5076>

1 Centro de Investigación de Ingeniería en Sistemas (CIIS), Escuela de Ingeniería en Sistemas, Universidad Latina de Costa Rica, Costa Rica. Email: joselyn.rodriguez@ulatina.cr.

 <https://orcid.org/0000-0003-4326-5036>

2 PRIS-Lab: Pattern Recognition and Intelligent System Laboratory, Department of Electrical Engineering, School of Engineering, Universidad de Costa Rica, Costa Rica. Email: maria.hernandezlopez@ucr.ac.cr.

 <https://orcid.org/0000-0003-2548-0277>

3 PRIS-Lab: Pattern Recognition and Intelligent System Laboratory, Department of Electrical Engineering, School of Engineering, Universidad de Costa Rica, Costa Rica. Email: francisco.siles@ucr.ac.cr.

 <https://orcid.org/0000-0002-6704-0600>



Keywords

MoCap; classification; segmentation; feature selection; data cleansing.

Abstract

This work introduces a framework proposal based on various algorithms, processes, and methods to classify Motion Capture (MoCap) data. To provide a generalized model for MoCap data classification, the approach is defined step by step: data collecting, data cleansing, segmentation, data pre-processing, feature selection, model selection, and validation. For each step, we selected and evaluated algorithms, process and methods have shown good performance in previous studies, all of them were proved and validated in BVH databases, but in not freely moving environment.

Palabras claves

MoCap; clasificación; segmentación; selección de características; limpieza.

Resumen

Este trabajo presenta la propuesta para un marco de trabajo (Motus) basado en varios algoritmos, procesos y métodos para la clasificación de archivos de captura de movimiento. El objetivo es proveer un modelo generalizado para la clasificación de movimiento, el enfoque es defino bajo los siguientes pasos: recolección y limpieza de datos, pre-procesamiento de la información, selección de características, segmentación y la selección del modelo y la validación. Para casa uno de los pasos, se selección y evaluó algoritmos que han demostrado un buen rendimiento, en ambientes controlados, en estudios previos, todos ellos han sido probados en archivos BVH, pero no en ambientes no controlados.

Introduction

The classification of human body motion from MoCap data is a difficult problem [1]. Also, the automatic segmentation of image sequences containing more than one class of motion is particularly challenging [1].

Although comparing two motion sequences is an easy task for a person, an automatic comparison is hard due to enormous numerical differences between two similar motion sequences. Spatial variations are mostly due to almost rigid transformations among similar postures. Temporal variations are due to non-linear differences in the dynamic of one action when performed by different subjects or even by two different performances of the same subject [2]. Because of these issues, it is highly important to define an appropriate procedure for a validating framework. It is then essential to define an automated technique that can segment MoCap data into homogeneous intervals, classify each interval into a motion type and produce minimal errors [3].

Related Work

After the MoCap data has been recorded, there are several essential steps that one must take with it in order to make it into an analyzable format [4].

To generate the continuous MoCap data usable for analyses, it needs to be imported with the video material into a data management program in which it can be segmented into more processable chunks [4].

More than a few methods have been previously investigated in the pursuit of this task; such as the weighted principal component analysis [5] which allows finding a particular motion in a large movement sequence; in addition, image-based reconstruction has also played an important role in motion classification, this may be done by creating a comprehensive set of training data for motion recognition methods with views of motion from all angles, or by converting non-orthogonal views taken from a single camera into orthogonal views for recognition [6].

It is important to mention that the literature reviewed for this investigation is based on very controlled motion capture; as for the present work, it is necessary to implement and compare both algorithms in a more freely moving environment.

Methods

Validation Design

The framework was designed considering every single step to motion capture automatic classification, for collecting data we take bvh files, some of them were controlled environment and human movement samples from the motion capture database of Carnegie Mellon University (CMU), and the others came from our records in a not-controlled sceneries. Our data set need to be processing by methods to cleaning data, reduce noising, and normalize, in this research we propose some procedures to correct it. In order to segment motion, we chose two algorithms for each one of the following process: classify data, feature selection and segmentation, all of them with a high performance.

Subjects and data set

For recording the Motion Capture sequences students were selected between the ages of 20 and 24 years old. A mixture of males and females was generated in order to obtain a more wide-range set of results. A total of three sessions were recorded, each one with its own BVH file set.

Results

Framework proposal

Collecting data

Data were obtained by recording interacting sessions between volunteer subjects. There was a total of 8 volunteers and the sessions were divided as follow: The first session involving an interaction between a male and a female of 21 and 23 years old respectively, with a duration of approximately 30 minutes. The environment was set so that joint attention would be initiated by any one of the subjects. In order to do that, no instructions were given to neither subject, but conversation cards were placed in a table to promote interaction between each other. Session two was developed in a very similar manner except that it involved three males of ages between 20 and 24 years old; while session three was a reiteration of session one except that subjects were replaced with new volunteers.

Cleaning data & preprocess data

This variability suggests that factors such as lighting conditions, camera angle, participant body shape, clothing and stance have a noticeable impact upon the quality of the recording.

These types of 'Glitches' are common in motion capture causing legs or arms to twitch rapidly into violently bent angles [7]. A solution is the algorithm (figure 1) for the detection, location and correction of glitches in a BVH file, Castresana [8] implemented in Python an algorithm for glitches reduction and will be used in the validation of the framework.

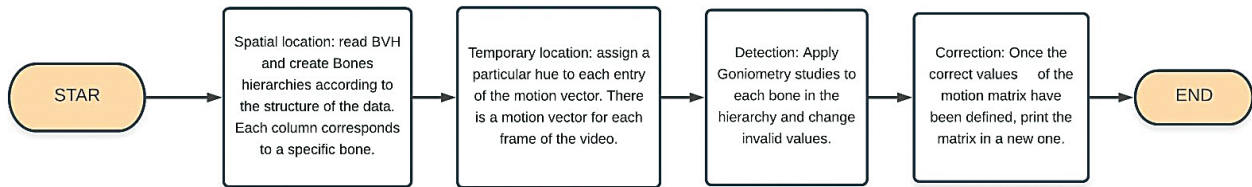


Figure 1. Algorithm Architecture [8]

Segmentation

A genetic algorithm approach to human MoCap data segmentation is applied to obtain the optimal solution, accuracy and efficiency of segmentation in controlled data. They conduct sparse learning on the raw motion data to derive a dictionary of representative postures and convert the raw motion sequence into a symbolic sequence [9]. As well, Quaternion Watershed Transform in Segmentation of Motion Capture Data controlled has a good performance, but in this case, they use the following segmentation hierarchical has still acceptable accuracy, which exceeds 91% [10]. Compare both methods will be the next step.

Feature selection (see figure 2):

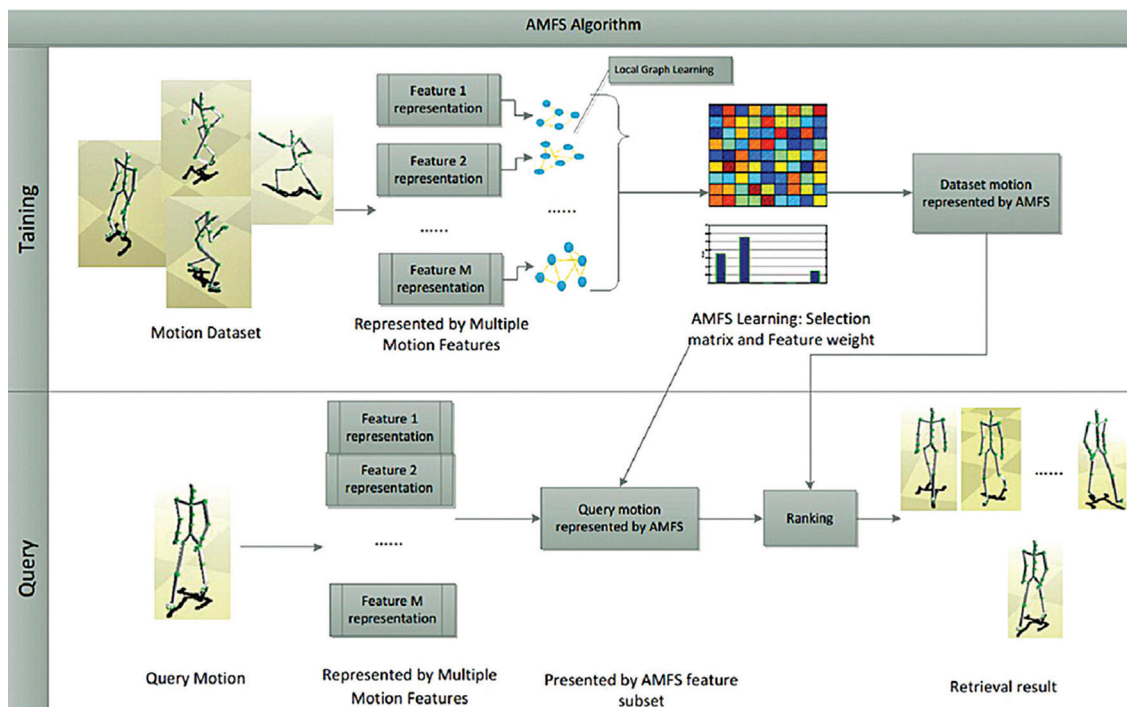


Figure 2. Adaptive multi-view feature selection (AMFS) [11]

Model Selection:

Two-Step SVM Fusion: Combine tree-structured vector quantization and Multiple binary Support Vector Machine classifiers [3]. The proposed algorithms using the 5-fold cross validation procedure and obtained a correct classification rate of 99.6% [3].

Future Work

The framework presented in this work is part of ongoing research, which intends to development a combination of algorithms, processes and methods mentioned in the previous sections and apply it, in order to validate a methodology that helps to classify not controlled motion capture and complex poses.

References

- [1] J. Rittscher and A. Blake, "Classification of human body motion," Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 1999, pp. 634-639 vol.1. doi: 10.1109/ICCV.1999.791284
- [2] A. W. Vieira, T. Lewiner, W. R. Schwartz and M. Campos, "Distance matrices as invariant features for classifying MoCap data," Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, 2012, pp. 2934-2937.
- [3] H. Kadu and C. - J. Kuo, "Automatic Human Mocap Data Classification," in IEEE Transactions on Multimedia, vol. 16, no. 8, pp. 2191-2202, Dec. 2014. doi: 10.1109/TMM.2014.2360793
- [4] Jantunen, T., Buerger, B., De Weerd, D., Seilola, I., & Wainio, T. Experiences Collecting Motion Capture Data on Continuous Signing. Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon, 75-82. 2012
- [5] Forbes, K. & Fiume, E. An efficient search algorithm for motion data using weighted PCA. Eurographics/ACM SIGGRAPH Symposium on Computer Animation. USA, 2005 p.p. 67-76, Doi: 10.1145/1073368.1073377
- [6] Bodor, R., Drenner, A., Fehr, D., Masoud, O., & Papanikolopoulos, N. View-independent human motion classification using image-based reconstruction. Image and Vision Computing, 27(8), 1194–1206. July, 2009. Doi: 10.1016/j.imavis.2008.11.008
- [7] Gillies, M., Brenton, H., Yee-King, M., Grimalt-Reynes, A., & d' Inverno, M. Sketches vs skeletons. Proceedings of the 2nd International Workshop on Movement and Computing - MOCO '15. Canada, August, 2015, p.p. 104-111 doi:10.1145/2790994.2790995
- [8] M. Castresana and F. Siles, "Goniometry-based Glitch-Correction Algorithm for Optical Motion Capture Data," 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOB), San Carlos, 2018, pp. 1-8. doi: 10.1109/IWOB.2018.8464198
- [9] Lv F., Nevatia R. Recognition and Segmentation of 3-D Human Action Using HMM and Multi-class AdaBoost. In: Leonardis A., Bischof H., Pinz A. (eds) Computer Vision – ECCV 2006. ECCV 2006. Lecture Notes in Computer Science, vol 3954. Berlin, 2006- Doi: 10.1007/11744085_28
- [10] Świtoński A., Michalczuk A., Josiński H., Wojciechowski K. (2019) Quaternion Watershed Transform in Segmentation of Motion Capture Data. In: Nguyen N., Gaol F., Hong TP., Trawiński B. (eds) Intelligent Information and Database Systems. ACIIDS 2019. Lecture Notes in Computer Science, vol 11432. March, 2019 . Doi: 10.1007/978-3-030-14802-7_49
- [11] Zhao Wang, Yinfu Feng, Tian Qi, Xiaosong Yang, Jian J. Zhang, Adaptive multi-view feature selection for human motion retrieval, Signal Processing, Volume 120, 2016, p.p. 691-701, ISSN 0165-1684, doi: 10.1016/j.sigpro.2014.11.015.


Initial Approach on Soccer Match's Scene Classification by Players' Field Spatial Distribution

Abordaje inicial en la clasificación de escenas de partidos de fútbol a partir de la distribución espacial de los jugadores sobre la cancha

Lennon Núñez-Meño¹, Marco Villalta², Francisco Siles-Canales³

Núñez-Meño, L; Villalta, M; Siles-Canales, F. Initial Approach on Soccer Match's Scene Classification by Players' Field Spatial Distribution. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 60-65.

 <https://doi.org/10.18845/tm.v33i5.5077>

- 1 PRIS-Lab: Pattern Recognition and Intelligent Systems Laboratory, Department of Electrical Engineering, School of Engineering and Postgraduate Studies in Electrical Engineering, Universidad de Costa Rica (UCR). CNCA: National Advanced Computing Collaboratory, CeNAT: National Center for High Technology. Costa Rica. E-mail: lennon.nunez@ucr.ac.cr
- 2 PRIS-Lab: Pattern Recognition and Intelligent Systems Laboratory, Department of Electrical Engineering, School of Engineering and Postgraduate Studies in Electrical Engineering, Universidad de Costa Rica (UCR). E-mail: marco.villalta@ucr.ac.cr
- 3 PRIS-Lab: Pattern Recognition and Intelligent System Laboratory, Department of Electrical Engineering, School of Engineering, Universidad de Costa Rica, Costa Rica. Email: francisco.siles@ucr.ac.cr
 <https://orcid.org/0000-0002-6704-0600>



Keywords

Soccer player; object tracking; scene classification; occlusion; stretch index; player dispersion; surface area.

Abstract

Soccer player tracking is moving towards the use of high-cost algorithms; however, these are implemented to the whole video recording of interest with no flexibility on the type of scene being evaluated. The current work proposes to classify the scenes according to the players' distribution over the field, designing a metric for players' Dispersion Level that would allow implementing a more flexible tracking method that selects a light algorithm for well-behaved scenes and a heavy and robust algorithm for bad-behaved scenes. This work pushes towards considering different existent metrics, as Voronoi diagrams for surface area analysis, in future works to come.

Palabras clave

Jugador de fútbol; rastreo de objetos; clasificación de escenas; oclusión; índice de estrechés; dispersión de los jugadores; área superficial.

Resumen

El rastreo de jugadores de fútbol se inclina al uso de algoritmos de alto costo computacional, sin embargo, estos suelen ser implementados de manera pareja al video de interés sin consideración de los tipos de escena evaluada. El trabajo presente propone clasificar las escenas con respecto a la distribución espacial de los jugadores sobre el campo por medio de una métrica del nivel de dispersión de los jugadores, con la cual implementar métodos de rastreo flexibles en cuanto a la selección de algoritmos livianos para escenas de 'buen carácter', y algoritmos pesados y robustos para escenas de 'mal carácter'. En investigaciones a venir, se pretende evaluar distintas métricas, como el uso de diagramas de Voronoi para el análisis del área superficial del espacio entre jugadores.

Introduction and Related Work

In many sport disciplines, video recordings are used to extract statistics on performance, the coach or technical director is in charge of studying that information, or even to extract it himself, in order to design the best strategy and help the athletes to improve; modern techniques include algorithms on object tracking to automate this process. In the case of soccer teams, players tracking provides information of the position of the player at any time during the whole match, with it, it is possible to tell different performance statistics: total time played, distance traveled, fatigue and placement of the player during different key plays [8]. Thus, players' performance and tactical evaluation can be assessed based on information from object tracking, making it a valuable tool for sports analysis.

When implementing multiple object tracking, one of the biggest problems is the occlusion event, especially when using a single point of view, as it only requires two overlapping players for it to happen. Different algorithms have been developed to solve this issue [3,1], commonly with high computational cost, the common approach is to apply these algorithms to the whole video recording, usually distributing the frames into different blocks that are merged to give the final result, there is no decision over which algorithm is best for a specific scene as there is no current

information of what's going on until the tracking results are obtained. Many platforms rely on the human interpretation of the different event scenes to classify them by game plays, as automatically semantic interpretation is still in early development [5]. If the scenes were automatically classified before implementing the tracking algorithm, different decisions could be made according to the scene type, as when to use a heavy algorithm to treat occlusion cases or when to use a light algorithm for 'Well Behaved' scenes.

Automatic scene classification started as a method to aid the tracking on sports broadcasting, where there are many changes in perspective (camera changes, zoom-in, and zoom-out) as well as sequence loss (commercials, replays) without prior knowledge of when any of these will occur. The broadcasting methods rely on color-based filtering to recognize the scenes [10,7], where there is a noticeable difference in color properties. In an Ultra High Definition recording of the whole field, all these special cases are non-existent as the view is maintained throughout the match; there is no need to classify scenes into zoom-in or zoom-out, commercials or match events.

As mentioned for the examples above, trends focus on broadcast videos, therefore, the scene classification types are accordingly, separating the events as shots. [9] classifies the shots in classes as transitional effects with different span times, overlays, long-distance and close up shorts. However, the proposed scene classification focuses on acquiring the ability to detect possible occlusion events. On this topic, [7] implements the Occlusion Alarm Probability (OAP) to detect players' proximities. This solution is an approach into detecting occlusion events, but there is not a classification or indexing of the players proximity events.

Currently, scene classification by game plays is performed fully manually or human aided, moreover, in occlusion matters, the scenes are identified as occlusion risky scenes just after the tracking information is obtained, as it is easier to detect when there's a blob merge or split. Therefore, the aim of the current effort is to design a metric for scenes evaluation prior to obtaining the tracking results, with which to classify the game match scenes by players congestion and distribution over the soccer field, classification that will aid decision making on different tracking approaches based on the type of scene in process, specially to decide between a light or a heavy algorithm if there is a high chance of occlusion.

Methodology

The occlusion events occur as the players get closer to each other, therefore, finding a metric that can numerically represent the distribution of the players over the game field can help to implement flexible tracking methods depending on the agglomeration complexity of the scene, or, in other words, the players dispersion level over the field. As the event to measure is agglomeration, i.e. how close the players are from each other, distance player to player is evaluated. Different alternatives are considered, [2,4] mention three main metrics, Stretch Index, Surface Area, and Team Length. The Stretch Index compares the average distance of a team's player and geometrical center, being a measure of how compact or stretched is the team. Taking $P_{axis,n}(k)$ as the n'th player's position on $axis = X, Y$ on the frame k , $GC_{axis}(k)$ the position of the geometrical center of the team on frame k , and N the total number of team players on the frame k , the Stretch Index (SI) is calculated per equation 1.

$$SI = \frac{\sum_n^N \sqrt{(P_{X,n}(k) - GC_X(k))^2 + (P_{Y,n}(k) - GC_Y(k))^2}}{N} \quad \text{Equation 1}$$

The equation for SI is, therefore, the average of magnitudes of the vectors *cluster center to player* for each player of the current team. This metric is designed taking into account that the players are separated into two teams, we intend to see all the players as a whole single group, as occlusion may occur with players from opposing teams, thus, calculating a distribution of all the individuals over the field is required.

A useful metric for measuring dispersion is the standard deviation, its value can help to assess how close the players are in comparison to a common reference, an approach similar to SI with the difference that it considers the average of the measurements as the reference for dispersion. Given the frame independent reference R with position (X, Y) , the vector player-to-reference $PR_n(k) = P_n(k) - R$, with $P_n(k)$ the player's position from the origin of the image's frame, and the mean of the vector's magnitudes μ , the standard deviation σ of the frame k for a single reference is calculated by the equation 2.

$$\sigma = \sqrt{\frac{1}{N} \sum_n (PR_n(k) \vee -\mu)^2} \quad \text{Equation 2}$$

Considering that SI is the average of magnitudes with a specific reference ($GC(k)$), the standard deviation can be described by $SI(R)$ with R as the desired reference, taking the equation 2 and replacing μ with $SI(R)$. We take into account four different references (each corner of the game field) and calculate the standard deviation for each, the result is four standard deviations.

Experiments and Preliminary Results

One soccer match was picked from the PRIS-Lab's video database, selected by the quality of the panoramic take (middle line alignment, view perspective), from which the scenes with a more noticeable change on dispersion are manually picked for the experiment. The four edges of the game area are used as references. Making the supposition that players tend to be somehow dispersed over the field, a trend is calculated to define the ranges of a Dispersion Level (DL), this to classify the scenes in three classes that would tell if the frames have a high probability to present occlusion: High, Medium and Low Level of dispersion, or, Well, Medium Well and Bad-behaved scenes, respectively. It is expected that for the frames where the players are visually close to each other, the standard deviation is lower than in those frames where the players are spaced or spread along the field. The above assumption can be observed in figure 1, which describes three main scenarios: the match starts with a big players agglomeration (low standard deviation, bad-behaved scene), the players start to disperse but all move following the ball (standard deviation increases, medium well-behaved scene), and the players space along the field (standard deviation reaches the highest values, well-behaved scene). The standard deviation for the first 41 seconds (in frames) of the match using the four corners of the field as references is shown in figure 2. The first 300 frames of video (lowest standard deviation) correspond to Bad-Behaved scene (figure 1.c), the middle of the match (frames about 300 to 900) are Medium Well Behaved scene (figure 1.b), and last frames are Well Behaved scene (Figure 1.a) when ball gets out of play and players disperse back to the center.

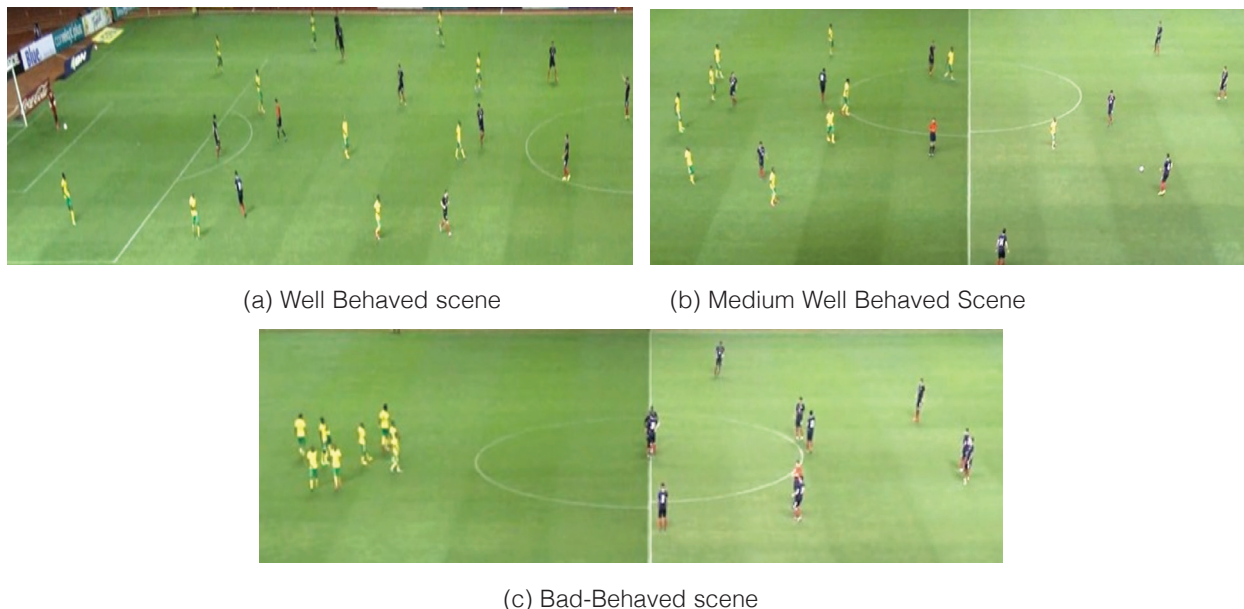


Figure 1. (a) Players are spread along the field. (b) Players are spread along the field but closer to each other. (c) Players are agglomerated in different clusters.

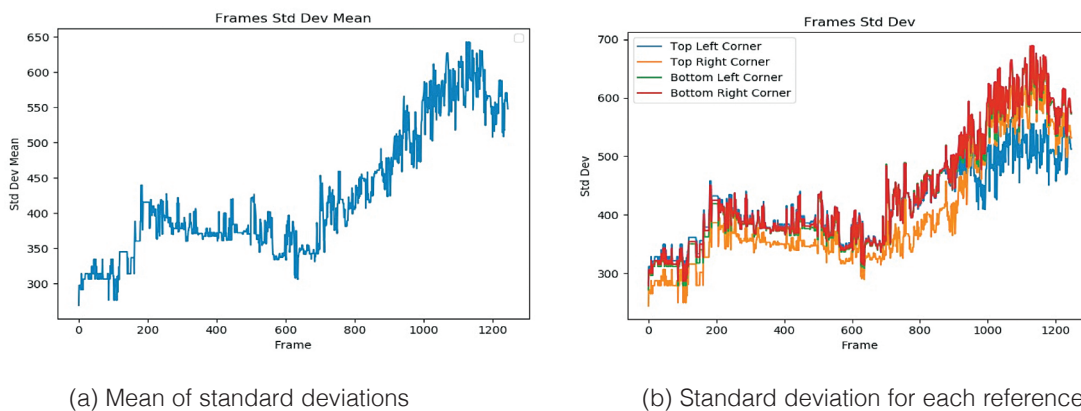


Figure 2. Standard Deviation of all the players on a soccer match using the four corners of the soccer field as reference

Conclusions and future work

The presented metric is a simple approach independent from the object tracking results, as it requires only the blob's 2D positions, regardless of which player is which or to which team it corresponds. However, more experiments are required to provide a quantitative result on its capacity to properly distinguish the required scenes, including longer video clips and more study cases. Also, as the tracking information is not used, spurious blobs will affect the calculations, a situation that would also affect even having the tracking information as a no-error fully automatic tracking platform doesn't exist yet. This metric cannot tell if there is or not occlusion on the image, it gives a hint on which frames have a high probability of presenting tracking issues (mostly occlusion) due to a low players' Dispersion Level; even in a Well Behaved scene, there could be

a two players occlusion event. Thus, the need for testing different metrics to obtain a more robust method to detect these special cases. One consideration is to use surface area metrics based on Voronoi diagrams, with which to evaluate the size of the area spaces between players inside the soccer field, the smaller the area in between, the more agglomerated the players are.

References

- [1] Amditis, A., Thomaidis, G., Maroudis, P., Lytrivis, P., Karaseitanidis, G. (2012). Multiple Hypothesis Tracking Implementation, chap. 10. Laser Scanner Technology, IntechOpen.
- [2] Clemente, M., Sequeiros, J., Correira, A., Sylva, F., Laurenço, F. (2012). Computational Metrics for Soccer Analysis: Connecting the dots, chap. 5. SpringerBriefs in Applied Sciences and Technologies, Springer.
- [3] Dearden, A., Demiris, Y., Grau, O. (2006). Tracking football player movement from a single moving camera using particle filters. In: CVMP-2006. pp. 29–37.
- [4] Folgado, H., Lemmink, K., Frencken, W., Sampaio, J. (2014). Length, width and centroid distance as measures of teams' tactical performance in youth football. *European Journal of Sport Science* (14), pp. 487–492.
- [5] Kapela, R., McGuinness, K., O'Connor, Noel E. (2017). Real-time field sports scene classification using colour and frequency space decompositions. *Journal of Real-Time Image Processing* 13(4), 725–737.
- [6] Mei, T., Ma, Y.-F., Zhou, H.-Q., Ma, W.-Y., Zhang, H.-J. (2005). Sports video mining with mosaic. In: 2005 International Conference on Multimedia Modelling. pp. 107–114.
- [7] Ok, H.-W., Seo, Y., Hong, K.-S. (2002). Multiple soccer players tracking by condensation with occlusion alarm probability. Tech. rep., IIP Lab., Pohang University of Science and Technology (POSTECH), Republic of Korea.
- [8] Radakovic, R. Dopsaj, M. Vulovic, R. (2015). The reliability of motion analysis of elite soccer players during match measured by the tracking motion software system. In: 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE).
- [9] Siles, F. (2014). Automated Semantic Annotation of Football Games from TV Broadcast. Ph.D. thesis, Technische Universität München.
- [10] Zhong, D., Chang, S. (2004). Real-time field sports scene classification using color and frequency space decompositions. *Journal of Visual Communication and Image Representation* 15(3), pp. 330–347.

Reducing the two dimensional Green functions: Fourier mode decomposition


Reduciendo las funciones de Green bidimensionales: Descomposición en modos de Fourier

Juan Pablo Mallarino-Robayo¹, Alejandro Ferrero-Botero²


Mallarino-Robayo, J. P.; Ferrero-Botero, A. Reducing the two dimensional Green functions: Fourier mode decomposition. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 66-73.

 <https://doi.org/10.18845/tm.v33i5.5078>

1 Coordinador Laboratorio Computacional HPC, Facultad de Ciencias, Universidad de los Andes, Bogotá, Colombia. Email: jp.mallarino50@uniandes.edu.co.

 <https://orcid.org/0000-0002-7424-3000>

2 Docente de tiempo completo, Departamento de Ciencias Básicas, Universidad Católica de Colombia, Bogotá, Colombia. Email: aferrero@ucatolica.edu.co.

 <https://orcid.org/0000-0001-8318-8886>



Abstract

Often we encounter high dimensional differential equations. A clever representation of a generalized solution could be procured in certain cases using Green functions. We show how this representation could be achieved and via a clever Fourier mode decomposition for the particular disc case resulting in a highly correlated set of functions that transforming into a discrete representation – via a classical second order finite difference approximation – can be ultimately represented as a linear equation for matrices embedding all boundary conditions in the structure of such objects. The resulting problem could be solved using stochastic gradient descent with an additional on-the-fly optimization reducing required computation resources substantially.

Resumen

Comúnmente encontramos ecuaciones diferenciales en espacios de alta dimensionalidad. Una representación útil de la solución generalizada puede ser expresada en ciertos casos usando funciones de Green. Se muestra como esta representación se puede lograr por medio de una descomposición en modos de Fourier para el caso particular de un disco dando origen a un conjunto altamente correlacionado de funciones que transformando a una representación discreta – a través de una típica aproximación de segundo grado por métodos finitos – puede ser representado como una ecuación lineal para matrices que contienen las condiciones iniciales de tales objetos. El problema resultante se puede resolver por medio del método de *Gradient Descent* estocástico cuyos componentes se calculan sobre la marcha para optimizar el uso de recursos computacionales.

Introduction

We often encounter ourselves solving complex multidimensional second order differential equations. A clever artifact for this purpose is the known *Green function* construction that has been employed in a cornucopia of areas in Physics throughout history. This scenario is clearly seen in Quantum and Statistical Physics [1–8].

From a mathematical point of view, the Green function problem can be conveniently defined via a differential operator also known as the Liouville operator as follows

$$\hat{\mathcal{L}}_{\{\mathbf{r}\}} \square = \left(\vec{\nabla}_{\{\mathbf{r}\}} + \vec{f}(\mathbf{r}) \right) \cdot \left(\vec{\nabla}_{\{\mathbf{r}\}} \square \right) + g(\mathbf{r}) \square, \quad (1)$$

acting on a scalar field \square in \mathfrak{R}^d , with d the dimension of the system – *i.e.* $\mathbf{r} \in \mathfrak{R}^d$.

In particular, our aim is to tackle the two dimensional case to allow a reduction in dimensionality using a Fourier expansion. Two dimensional systems are of great interest in material sciences, quantum computing, experimental and high energy physics, ionic fluids, theoretical mathematics, and many others. Numerical methods have been described to compute Green's functions, however, due to its increased complexity some remarkable efforts have been made for establishing a precise formulation of boundary conditions for their numerical calculation [9] and, more interestingly, for elaborating solutions in periodic systems in similar work by [10].

The key for introducing Green's function – or more accurately, distribution – formalism is writing the overall problem as the in-homogeneous equation

$$\hat{\mathcal{L}}\psi(\mathbf{r}) = \phi(\mathbf{r}), \quad (2)$$

with $\psi(\mathbf{r})$ and $\phi(\mathbf{r})$ two scalar functions. It is proven that existence of Green's distribution relies on the likewise existence of a weight function introduced to guarantee symmetrical convolution in Hilbert's space \mathcal{H} – also known as the space of functions.

The Green function formalism starts from the identity,

$$\hat{\mathcal{L}}_{\{\mathbf{r}\}}G(\mathbf{r}, \mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}'), \quad (3)$$

with $\delta(\mathbf{r} - \mathbf{r}')$ the \mathfrak{R}^d Dirac delta distribution, where the solution to the in-homogeneous equation (2) is postulated by the convolution identity from the distribution,

$$\psi(\mathbf{r}) = \int_{\mathbf{r}'} G(\mathbf{r}, \mathbf{r}')\phi(\mathbf{r}') d\mathbf{r}' + \text{b.c.}, \quad (4)$$

with either Dirichlet or Neumann boundary conditions (b.c).

Fourier modes analysis and reduction

In polar coordinates eq. (3) is given by

$$\left[\frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \theta^2} + f_1(r, \theta) \frac{\partial}{\partial r} + f_2(r, \theta) \frac{1}{r} \frac{\partial}{\partial \theta} + g(r, \theta) \right] G(\mathbf{r}, \mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}'), \quad (5)$$

where $f_1(r, \theta)$ and $f_2(r, \theta)$ are the components of \vec{f} in the radial and angular directions respectively.

The presence of the delta function $\delta(\mathbf{r} - \mathbf{r}') = \frac{1}{r} \delta(r - r')\delta(\theta - \theta')$ suggests an expansion for the Green function of the form $\frac{1}{2\pi} \sum_l e^{il\theta} G_l(r - r')$ for the angular term where a separable solution can be procured by symmetry. However, in the most general case, an approximate method could be formulated to evaluate the l -modes G_l .

Completeness of the Fourier expansion and a divergence free system (f_1 and f_2 are well behaved within the domain), the effective one-dimensional system equation for r yields – multiplying eq. (5) by r^2 to avoid a pathological behavior at $r = 0$ –,

$$\sum_l e^{il(\theta-\theta')} \left[r^2 \frac{d^2}{dr^2} + A(r, \theta) \frac{d}{dr} + B_l(r, \theta) \right] G_l = \sum_l e^{il(\theta-\theta')} r \delta(r - r'). \quad (6)$$

In terms of the functions introduced in eq. (5) we have the relations $A(r, \theta) = r(1 + rf_1(r, \theta))$ and $B(r, \theta) = -l^2 + ilrf_2(r, \theta) + r^2g(r, \theta)$. Following the same recipe for expansion in Fourier representation of both $A(r, \theta)$ and $B(r, \theta)$ yields in simplified form,

$$r^2 G_n'' + \sum_m A_m(r) e^{im\theta'} G'_{n-m} + \sum_m B_{n-m,m}(r) e^{im\theta'} G_{n-m} = r \delta(r - r'), \quad (7)$$

after averaging over θ as $\int_0^{2\pi} d\theta e^{-in\theta} \dots$

In terms of the Fourier modes of the original functions $f_1(r, \theta)$, $f_2(r, \theta)$, and $g(r, \theta)$ (the modes of a function $f(r, \theta)$ are defined, as usually, as $f_m(r) = \frac{1}{2\pi} \int_0^{2\pi} f(r, \theta) e^{-im\theta} d\theta$ last equation takes the form

$$r^2 G_l'' + r(1 + r f_{1,0}) G_l' - (l^2 - ir l f_{2,0} - g_0) G_l + \sum_{m \neq 0} [ir(l - m) f_{2,m} + r^2 g_m] G_{l-m} + \sum_{m \neq 0} r^2 f_{1,m} G_{l-m}' = r \delta(r - r') e^{il\theta}. \quad (8)$$

This last equation represents an infinitely coupled system of differential equations. An exact solution for the effective one dimensional system can be found by adding the infinite modes of m . However, a numerical calculation of such problem is impractical. We are then forced to approximate the result by expanding over a finite number of modes. The rationale behind this cutoff is also justified by the limitations of numerical precision in floating point operations. Thus, the number of modes to be taken will depend entirely on how fast the functions $f_1(r, \theta)$, , and $g(r, \theta)$ decay on m . Textbook Fourier analysis shows that whenever a function $f(\theta)$ is at least twice differentiable, its Fourier series converges uniformly to the function and its coefficients decay as m^2 [11]. Additional convergence requirements in a numerical platform sets an upper bound to the utmost value of $|m|$ virtue of the first and second order derivatives approximation in Finite Elements.

Approximate solution using finite elements

A discretization of the radial variable, for both r and r' allows us to find an approximate solution to eq. (8) using the three-point-stencil. We now define a step size given by $h = R/N$, where N is the number of points along the grid. As the Green function for each mode has two degrees of freedom (r and r'), the Green function for each mode will be represented by an $N + 1 \times N + 1$ matrix, where we locate r and r' along the rows and columns of such matrix respectively.

Using a Taylor expansion [12–14] up to second order in h^2 to express the first and second derivatives of $G_l(r, r')$, eq. (8) can be written as

$$A_l^j G_l^{jj'} + B_l^j G_l^{j+1j'} + C_l^j G_l^{j-1j'} + \sum_{m \neq 0}^M D_{lm}^j G_{l-m}^{jj'} + \sum_{m \neq 0}^M E_{lm}^j (G_{l-m}^{j+1j'} - G_{l-m}^{j-1j'}) = F_l^j \delta^{jj'}, \quad (9)$$

where A_l^j , B_l^j , C_l^j , D_{lm}^j , E_{lm}^j and F_l^j are matrix elements, which can be written in terms of the discretized variable $r^j = hj$ and the Fourier modes of the functions: f_{1m}^j , f_{2m}^j , and g_m^j .

As the truncate the maximum number of Fourier l -modes by a value of L , we have a linear system (by blocks) of the form

$$\mathbf{P}_{(2L+1)(N+1) \times (2L+1)(N+1)} \cdot \mathbf{G}_{(2L+1)(N+1) \times 1} = \mathbf{V}_{(2L+1)(N+1) \times 1} \cdot$$

More explicitly, the block linear system can be written as

$$\begin{pmatrix} \mathbf{P}_{-L-L} & \mathbf{P}_{-L-L+1} & \cdots & \mathbf{P}_{-LL-1} & \mathbf{P}_{-LL} \\ \mathbf{P}_{-L+1-L} & \mathbf{P}_{-L+1-L+1} & \cdots & \mathbf{P}_{-L+1L-1} & \mathbf{P}_{-L+1L} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{P}_{L-1-L} & \mathbf{P}_{L-1-L+1} & \cdots & \mathbf{P}_{L-1L-1} & \mathbf{P}_{L-1L} \\ \mathbf{P}_{L-L} & \mathbf{P}_{L-L+1} & \cdots & \mathbf{P}_{LL-1} & \mathbf{P}_{LL} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{G}_{-L} \\ \mathbf{G}_{-L+1} \\ \vdots \\ \mathbf{G}_{L-1} \\ \mathbf{G}_L \end{pmatrix} = \begin{pmatrix} \mathbf{V}_{-L} \\ \mathbf{V}_{-L+1} \\ \vdots \\ \mathbf{V}_{L-1} \\ \mathbf{V}_L \end{pmatrix}, \quad (10)$$

where the diagonal block matrices \mathbf{P}_{ll} and the off-diagonal block matrices \mathbf{P}_{lm} are $N + 1 \times N + 1$ matrices of the form

$$\mathbf{P}_{ll} = \begin{pmatrix} \text{BC} & \text{BC} & \cdots & 0 & 0 \\ G_l^1 & A_l^1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & A_l^{N-1} & B_l^{N-1} \\ 0 & 0 & \cdots & \text{BC} & \text{BC} \end{pmatrix}, \quad \mathbf{P}_{lm} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ -E_{lm}^1 & D_{lm}^1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & D_{lm}^{N-1} & E_{lm}^{N-1} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}. \quad (11)$$

In last expression, BC are elements associated with the boundary conditions chosen (either Dirichlet or Neumann). On the other hand, the same notation as in eq. (9) has been used. Similarly

$$\mathbf{G}_l = \begin{pmatrix} G_l^{00} & G_l^{01} & \cdots & G_l^{0N-1} & G_l^{0N} \\ G_l^{10} & G_l^{11} & \cdots & G_l^{1N-1} & G_l^{1N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ G_l^{N-10} & G_l^{N-11} & \cdots & G_l^{N-1N-1} & G_l^{N-1N} \\ G_l^{N0} & G_l^{N1} & \cdots & G_l^{NN-1} & G_l^{NN} \end{pmatrix}, \quad \mathbf{V}_l = \begin{pmatrix} \text{BC} & 0 & \cdots & 0 & 0 \\ 0 & F^1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & F^{N-1} & 0 \\ 0 & 0 & \cdots & 0 & \text{BC} \end{pmatrix}. \quad (12)$$

Matrix inversion

A solution to the individual modes and hence, to the whole system (which can be found by adding the contribution of the different modes) can be found by obtaining the inverse of the full matrix P, or, likely, finding G *s.th* eq. (10). Although well-established numerical algorithms exist, such algorithms become unviable for *large* systems due to limitations on memory resources and time constraints. Particularly, these matrices are potentially highly sparse. In addition to, approximating via the 3 or 5 point stencil reflects an underlying structure of these matrices that suggests that an alternate solution method could be orchestrated.

An alternate method to find the required solution is an optimization scheme by means of the *gradient descent* method – and its faster sibling the *stochastic gradient descent* [15–17]. While this method does not directly find the inverse of the matrix P, it attempts to find the solution for the matrix G directly by the minimization of a cost function, usually the mean square error,

$$J = \frac{1}{N + 1} \sum_{j,k=0}^D [V_{jk} - V_{jk}^{(i)}]^2, \quad (13)$$

where $V^{(i)}$ is an ansatz satisfying the relation $\mathbf{P} \cdot \mathbf{G}^{(i)} = \mathbf{V}^{(i)}$. Notice that by starting from an educated guess of $\mathbf{G}^{(i)}$ we obtain a very crude estimate of $\mathbf{V}^{(i)}$ distant from the expected V measured by the cost function J. Since the profile is quadratic, we can potentially improve $\mathbf{G}^{(i)}$ iteratively by making multiple fractional updates in the opposite direction of the gradient to the hyper-surface profile for the cost function. The update is given by $\vec{G}_{(i)}^a = \vec{G}_{(i-1)}^a - \eta_{(i)} \vec{\nabla} J_{(i-1)}$. While this method

is guaranteed to converge – provided that the resulting matrix is invertible – it may require a significant number of updates, thus undermining its efficiency. For that matter, scientists often resource to Stochastic Gradient Descent to accelerate convergence. However, it is well known that high accuracy is not an outstanding trait of it. Lately, research on variations of the Stochastic Gradient Descent methods for higher accuracy have been developed which can be utilized in place such as variations of Kaczmarz’s algorithm for solving systems of linear equations among others [17–19].

Results

In order to test the validity of the method, we will show a particular example in which the analytical solution to a second order differential equation is known. Let us suppose that we want to find the Green function associated with the Helmholtz equation

$$(\nabla^2 - m^2)G(\mathbf{r}, \mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}'). \quad (14)$$

We will use units such that $m = 1$ and confine the system in a large disk of radius $R = 10$. Since the system can be solved by separation of variables and so the modes do not couple to each other $C_{lm} = D_{lm} = 0$. Additionally, $A_l^j = 2r_j^2 + h^2(l^2 + r_j^2)$, $B_l^j = r_j(r_j + \frac{h}{2})$, $C_l^j = r_j(r_j - \frac{h}{2})$, and $F^j = -hr_j$. The analytical solution to eq. (14) with Dirichlet boundary conditions is

$$G(\mathbf{r}, \mathbf{r}') = -\frac{1}{2\pi} \sum_l e^{il(\theta-\theta')} \left[I_l(r_{<})K_l(r_{>}) - \frac{K_l(R)}{I_l(R)} I_l(r)I_l(r') \right], \quad (15)$$

where $r_{<}$ and $r_{>}$ represent the smaller and the larger radii between r and r' respectively and $I_l(r)$ and $K_l(r)$ are the modified Bessel function of first and second kind. In the case in which $\theta = \theta'$ the numerical and analytical solutions are shown in figure 1. The average mean square error $\langle MSE \rangle$, representing the accuracy of the numerical solution is shown in table 1.

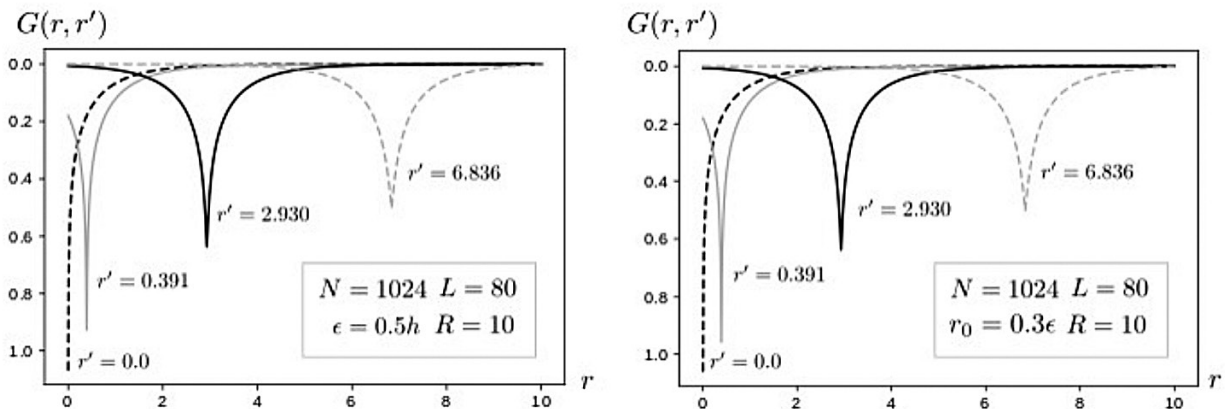


Figure 1: Left: numerical solution to eq. (14). Right: analytical solution provided by eq. (15). In each plot from left to right: $r' = 0$, $r' = 0.391$, $r' = 2.930$, and $r' = 6.836$. In both cases we took a maximum number of $L = 80$ modes; some other important parameters are shown in the plots. The parameters ϵ and r_0 are necessary values introduced to avoid divergences, as the Green function diverges as a consequence of the Dirac Delta distribution.

Table 1. Mean square error associated with the numerical solution shown in figure 1. MSE_{max} represents the maximum error and r_{max} the value of r' at which it occurs. As expected, the maximum error takes place at the location of the peaks. While the height of the peaks is formally infinite, the cutoffs ϵ and r_0 are introduced to avoid such divergences.

r'	$\langle MSE \rangle$	MSE_{max}	r_{max}
0.000	1.44×10^{-7}	8.69×10^{-5}	0.000
0.391	9.26×10^{-7}	8.58×10^{-4}	0.391
2.930	2.21×10^{-10}	7.74×10^{-8}	2.930
6.836	6.11×10^{-9}	7.33×10^{-7}	6.836

Conclusions

We synthesized a prescription for finding solutions to the general Green functions formalism problem in two dimensions by using the well-known Fourier expansion arriving to a set of infinite countable coupled differential equations. In the process of describing solutions to G, we conceded in finding an alternate solution to an optimization problem where existence is guaranteed when we scale up to the structural properties of the Liouville operator. Ergo, a solution could always be found. Ultimately, this problem is paramount in an arena where the evolution of technology is highly dependent on the behavior and dynamics of two dimensional systems in material science.

References

- [1] J. Schwinger, "Brownian motion of a quantum oscillator," *Journal of Mathematical Physics*, vol. 2, no. 3, pp. 407–432, 1961.
- [2] J.-S. Wang, B. K. Agarwalla, H. Li, and J. Thingna, "Nonequilibrium green's function method for quantum thermal transport," *Frontiers of Physics*, vol. 9, no. 6, pp. 673–697, Dec. 2014.
- [3] S. Foster and N. Neophytou, "Effectiveness of nanoinclusions for reducing bipolar effects in thermoelectric materials," *Computational Materials Science*, vol. 164, pp. 91–98, Jun. 2019.
- [4] L. P. Kadanoff and G. Baym, "Quantum statistical mechanics. Green's function methods in equilibrium and nonequilibrium problems." New York: W. A. Benjamin, Inc. XI, 203 p. (1962)., 1962.
- [5] V. Lucarini, "Revising and extending the linear response theory for statistical mechanical systems: Evaluating observables as predictors and predictands," *Journal of Statistical Physics*, vol. 173, no. 6, pp. 1698–1721, Dec. 2018.
- [6] Z. Chen, J. de Gier, I. Hiki, and T. Sasamoto, "Exact confirmation of 1D nonlinear fluctuating hydrodynamics for a two-species exclusion process," *Phys. Rev. Lett.*, vol. 120, no. 24, p. 240601, Jun. 2018.
- [7] I. Brevik, P. Parashar, and K. V. Shajesh, "Casimir force for magnetodielectric media," *Phys. Rev. A*, vol. 98, no. 3, p. 032509, Sep. 2018.
- [8] F. Xu and J. Wang, "Statistical properties of electrochemical capacitance in disordered mesoscopic capacitors," *Phys. Rev. B*, vol. 89, no. 24, p. 245430, Jun. 2014.
- [9] F. Tonon, E. Pan, and B. Amadei, "Green's functions and boundary element method formulation for 3D anisotropic media," *Computers & Structures*, vol. 79, no. 5, pp. 469–482, 2001.
- [10] D. Duhamel, "Finite element computation of green's functions," *Engineering Analysis with Boundary Elements*, vol. 31, no. 11, pp. 919–930, 2007.
- [11] M. Adams and V. Guillemin, "Fourier analysis," in *Measure theory and probability*, Boston, MA: Birkhäuser Boston, 1996, pp. 118–177.
- [12] A. Tveito, H. P. Langtangen, B. F. Nielsen, and X. Cai, "Differential equations: The first steps," in *Elements of scientific computing*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 31–73.

- [13] J. Whiteley, *Finite element methods, a practical guide*. Springer, 2017.
- [14] T. Grätsch and F. Hartmann, "Duality and finite elements," *Finite Elements in Analysis and Design*, vol. 40, no. 9, pp. 1005–1020, 2004.
- [15] G. Montavon, G. Orr, and K.-R. Müller, *Neural networks: Tricks of the trade*, 2nd ed. Springer Publishing Company, Incorporated, 2012.
- [16] S. Shalev-Shwartz and S. Ben-David, "Stochastic gradient descent," in *Understanding machine learning: From theory to algorithms*, Cambridge University Press, 2014, pp. 150–166.
- [17] R. M. Gower, F. Hanzely, P. Richtárik, and S. U. Stich, "Accelerated stochastic matrix inversion: General theory and speeding up bfgs rules for faster second-order optimization," in *Proceedings of the 32Nd international conference on neural information processing systems*, 2018, pp. 1626–1636.
- [18] D. Needell, R. Ward, and N. Srebro, "Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm," in *Advances in neural information processing systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 1017–1025.
- [19] N. Loizou, M. Rabbat, and P. Richtárik, "Provably accelerated randomized gossip algorithms," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7505–7509.

Phylogenetic analysis of ITS data from Endophytic fungi using Massive Parallel Bayesian Tree Inference with Exabayes

Análisis Filogenético de Secuencias ITS Provenientes de Hongos Endófitos Utilizando Inferencia Bayesiana Paralela de Árboles con Exabayes

Maripaz Montero-Vargas¹, Jean Carlo Umaña-Jiménez²,
Efraín Escudero-Leiva³, Priscila Chaverri-Echandi⁴


Montero-Vargas, M; Umaña-Jiménez, J; Escudero-Leiva, E; Chaverri-Echandi, P. Phylogenetic analysis of ITS data from Endophytic fungi using Massive Parallel Bayesian Tree Inference with Exabayes. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 74-79.

 <https://doi.org/10.18845/tm.v33i5.5079>


1 Biologist. Centro Nacional de Computación Avanzada (CNCA), CeNAT-CONARE, Costa Rica. Email: mmontero@cenat.ac.cr .

 <https://orcid.org/0000-0002-6562-4231>


2 Computing Engineer. Centro Nacional de Computación Avanzada (CNCA), CeNAT-CONARE, Costa Rica. Email: jumana@cenat.ac.cr .

 <https://orcid.org/0000-0003-0857-6007>

3 Biologist. Centro Nacional de Innovaciones Biotecnológicas (CENIBiot), CeNAT-CONARE, Costa Rica & Centro de Investigaciones en Productos Naturales (CIPRONA), Universidad de Costa Rica, Costa Rica.

 <https://orcid.org/0000-0003-4440-4296>

4 Associate Professor. Universidad de Costa Rica, Escuela de Biología & Centro de Investigaciones en Productos Naturales (CIPRONA), Costa Rica.

 <https://orcid.org/0000-0002-8486-6033>



Keywords

Fungi; ITS; Exabayes; Phylogenetics; Parallelization; Biodiversity.

Abstract

Ecological studies of fungal communities have been favored thanks to the emergence and improvement of independent culture techniques that use the ITS region as a molecular marker. This has allowed a more accurate identification compared to traditional culture-dependent methods.

Next-generation sequencing techniques have increased the amount of data available for the understanding of endophytic fungal communities. An important part of this process is the phylogenetic inference to decipher how the different taxa are related and interact, however, this may become one of the bioinformatic analysis that demands more time.

In response to this, the bioinformatics along with high-performance computing offer solutions to accelerate and make more efficient the tools available for data processing through the implementation of supercomputers and the parallelization of tools

In this study we carried out the processing of ITS sequences to then use the parallelization of Exabayes, software specialized in the analysis and creation of phylogenetic trees.

Thanks to the use of this technique, it was possible to reduce the running time of Exabayes from more than 400 hours to 6 hours, which demonstrates the benefits of the use of high-performance computing platforms.

Palabras clave

Hongos; ITS; Exabayes; Filogenética; Paralelización; Biodiversidad.

Resumen

Los estudios ecológicos de las comunidades fúngicas se han visto favorecidos gracias a la aparición y mejora de técnicas independientes de cultivo que utilizan la región ITS como marcador molecular. Esto ha permitido una identificación más precisa en comparación con los métodos tradicionales dependientes de la cultura.

Las técnicas de secuenciación de próxima generación han aumentado la cantidad de datos disponibles para la comprensión de las comunidades de hongos endofíticos. Una parte importante de este proceso es la inferencia filogenética para descifrar cómo se relacionan e interactúan los diferentes taxones, sin embargo, este puede convertirse en uno de los análisis bioinformáticos que exige más tiempo.

En respuesta a esto, la bioinformática junto con la informática de alto rendimiento ofrecen soluciones para acelerar y hacer más eficientes las herramientas disponibles para el procesamiento de datos a través de la implementación de supercomputadoras y la paralelización de herramientas.

En este estudio llevamos a cabo el procesamiento de secuencias ITS para luego utilizar la paralelización de Exabayes, software especializado en el análisis y creación de árboles filogenéticos.

Gracias al uso de esta técnica, fue posible reducir el tiempo de ejecución de Exabayes de más de 400 horas a 6 horas, lo que demuestra los beneficios del uso de plataformas informáticas de alto rendimiento.

Introduction

Endophyte fungi have been found in almost all vascular plant species examined to date and they are considered important components of fungal biodiversity [1]. They can have profound effects on their host physiology, influence multitrophic networks and entire ecosystems [2]. Fungi and particularly endophytes are a very promising source of novel biologically active compounds [3] so studying them may help to implement new techniques for new biotechnological applications.

An essential part of ecological studies of endophytic fungi is the taxonomic classification and phylogenetic inference [4], [5]. Fungi taxonomy is certainly complex, even now, what was believed to be one species can be in fact be an assemblage of productively isolated lineages. In mycology, species were defined based on the morphology of asexual and sexual reproductive structures, unfortunately, the number of characters is limited generating insecurity in the identification of species [6].

The easiest way to identify endophytic fungi is through molecular methods. ITS region is accepted as a barcode for fungi because of the higher amplification success rate for many fungal groups [7] [8]. Thanks to the next generation sequences the amount of data obtained in the ecological studies has increased exponentially through the years [9], [5], which in turn implies a challenge for the processing and analysis of the information.

Recently is common to join computational efforts, the development of bioinformatics tools and the and specialized databases to make the analysis of the large sets of biological data more fast and efficient [10], [11]. High performance and high throughput computing technologies permit that the processing of such datasets could be automated to accelerate bioinformatics processes [12].

One of the most time-consuming analyses is the Bayesian Tree inference [13]. This is why a critical step in the ecological analysis is to use parallel tools that fulfill this task.

Exabayes is an is a software package that computes Bayesian tree posterior probability using the Markov Chain Monte Carlo sampling approach. This tool applies commonly used evolutionary models and can handle big data sets efficiently. An important aspect of this tool is that it allows the use of Message-passing Interface (MPI) to parallelize the analysis using a computer cluster so that the only limit for the analysis of large data sets is the memory held by the cluster [14].

Exabayes works by performing a calculation of the posterior probability of sampling trees. This tool can divide its analysis on multiple independent runs which in turn analyze multiple chains responsible for sampling the parameters for stochastic simulation to obtain a sample from the posterior distribution of trees [15], [14].

In this paper, we implement the analysis of ITS sequences of endophytic fungi of coffee plants to achieve the reconstruction of their phylogeny through the parallelization of exabayes with the objective of improving the performance of this tool using a computational cluster. These techniques may help ecologist or micologist to process their data in a efficient way, improving the use if computational resources and generating more accurate results.

Methodology and results

For the following experiments, a set of sequence data in Fastq format was used, which were obtained from endophyte fungi extracted from Costa Rica coffee plants. Then a quality control of the sequences was carried out using the FastQC [16] tool for the identification of poor quality bases. All bioinformatic processing was carried out in the Kabre cluster of the National Center of High Technology in San Jose, Costa Rica

The preprocessing of the sequences was made with the Seqtk tool [17], using two of its functions. First, the Trimfq function was used to remove the poor quality bases at both ends of the sequences, identified with a Phred score of less than 20. Then the Seq function was used to change the format of the Fastq sequences to FASTA.

Then duplicate sequences were removed with the USEARCH [18] tool with which a single sequence multifasta file was obtained with 331 sequences of 1331 each. The identification of the Operational Taxonomic Units (OTUs) was then carried out using the UNITE database [10] through the Blast tool of the National Center for Biotechnology Information [19].

For the phylogenetic analysis, an alignment of the sequences was performed using the MUSCLE tool [20]. The resulting file in phylip format is used as input for the Tree inference with Exabayes.

Initially, the analyzes in Exabayes were carried out using nodes, composed of Intel Xeon Phi KNL nodes processors, each one with 64 cores @ 1.3 GHz and 96 GB (architecture A). Historically. A sequential test was performed using 256 ranks and dividing the analysis into four runs and these in turn into four chains using four swaps between chains. This test lasted 447 hours.

To increase efficiency, parallelization of runs and chains was implemented using MPI the parameters described in the documentation: number of runs executed in parallel (-R), number of chains per run executed in parallel (-C) and number of swap attempts between chains per generation.

From these experiments, the best performance was obtained using 64 ranks, and dividing the analysis into 2 runs with 2 strings each. This same distribution was used for parallel runs and parallel chains, with 4 swaps per generation. This test lasted 20:29:01 (hh: mm: ss) reaching statistical significance with $p = 0.0131916$ (table 1).

These tests were also performed with a different architecture (architecture B) using nodes c composed of nodes with Intel (R) Xeon (R) CPUs E3-1225 v5 @ 3.30GHz and 16GB RAM. In the same way, as with the previous tests, the use of the parameters was compared in parallel and better and statistically significant performance ($p = 0.0199349$) was obtained using 4 ranks and dividing the analysis into 2 runs and 2 chains. It was also specified that the two runs and the two strings will be executed in parallel. With this the best result was obtained with a run time with 6:25:31 (hh: mm: ss) (table 1).

Analysis

As shown in the previous results, Exabayes efficiency improved substantially by balancing the number of runs and chains in which the analysis is divided with the number of runs and chains executed in parallel.

In this way, the better performance of the traditional test tool is observed to test 1. Thus, the number of ranks can be better distributed among the groups and achieve better performance.

With these tests, it was also found that by decreasing swapping among coupled chains, the distribution of resources becomes more efficient by decreasing communication between groups (of runs and chains that are being executed in parallel).

On the other hand, when comparing the performance of Exabayes between the different architectures, there is a significant improvement in the tests carried out on nodes B with respect to nodes A. This indicates that it is more advantageous for Exabayes, to use fewer CPUs but with more clock speed (3.3 GHz), than using a node that has a lot of processors with a very low clock speed (1.3 GHz).

Table 1. Results obtained from the evaluation of efficiency in the use of Exabayes in parallel.

Architecture	Number of Ranks	Runs in parallel	Chains in parallel	Runs	Chains	swaps	Wall time (hh:mm:ss)	asdfs
A	256	NA	NA	4	4	4	447:36:22	0.014756
A	64	2	2	2	2	2	9:44:05	0.0156383
B	4	2	2	2	2	2	6:25:31	0.0199349

Conclusions

The use of the parallel modality of the Exabayes tool is shown as an excellent alternative to improve the efficiency of tree inference with the Bayesian method, allowing better run times with statistically significant precision. In this matter, is important to balance the number of runs and the number of chains runs in parallel so that a homogeneous distribution of the workload of the groups in the different ranks is achieved

We recommend making an evaluation not only of the input data to be used during the analyzes but also on the available computational resources that can give Exabayes more efficiency for phylogenetic inference.

For future studies, it is recommended to scale the data set to a more complex one in terms of number of taxa and length of the sequence. A next step would be to apply data level parallelism in alignments with multiple partitions, so that the scalability of the balance of the parameters used in this experiment can be tested.

References

- [1] J. Rodríguez, J. Elissetche and S. Valenzuela, "Tree Endophytes and Wood Biodegradation". *Endophytes of Forest Trees*, pp.81-93, 2011.
- [2] M. Unterseher, "Diversity of fungal endophytes in temperate forest trees". In *Endophytes of forest trees*. Springer, Dordrecht. pp. 31-46, 2011.
- [3] T. Larsen, J. Smedsgaard, K. Nielsen, M. Hansen and J. Frisvad, "Phenotypic taxonomy and metabolite profiling in microbial drug discovery". *Natural Product Reports*, vol. 22, no. 6, pp. 672, 2005.
- [4] J. Fouquier, et al. "Ghost-tree: creating hybrid-gene phylogenetic trees for diversity analyses". *Microbiome*, vol. 4, no. 1, 2016.
- [5] S. Tibpromma, "Identification of endophytic fungi from leaves of Pandanaceae based on their morphotypes and DNA sequence data from southern Thailand". *MycKeys*, vol. 33, pp.25-67, 2018.
- [6] C. Grünig, V. Queloz and T. Sieber, "Structure of Diversity in Dark Septate Endophytes: From Species to Genes". *Endophytes of Forest Trees*, pp.3-30, 2011.
- [7] C. Schoch, C., et al., "Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi". *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, pp. 6241-6246, 2012.
- [8] U. Kõljalg, et al. (2013). "Towards a unified paradigm for sequence-based identification of fungi". *Molecular Ecology*, vol. 22, no. 21, pp.5271-5277, 2013.
- [9] J. Zoll, E. Snelders, P. Verweij and W. Melchers, "Next-Generation Sequencing in the Mycology Lab". *Current Fungal Infection Reports*, vol.10, no. 2, pp. 37-42, 2016.
- [10] K. Abarenkov, et al., "The UNITE database for molecular identification of fungi - recent updates and future perspectives". *New Phytologist*, vol. 186, no. 2, pp.281-285, 2010.
- [11] C. Wurzbacher, "Introducing ribosomal tandem repeat barcoding for fungi". *Molecular Ecology Resources*, vol. 19, no. 1, pp.118-127, 2018.

- [12] A. Welivita, I. Perera, D. Meedeniya, A. Wickramarachchi and V. Mallawaarachchi, "Managing Complex Workflows in Bioinformatics: An Interactive Toolkit With GPU Acceleration". *IEEE Transactions on NanoBioscience*, vol. 17, no. 3, pp.199-208, 2018.
- [13] G. Altekar, S. Dwarkadas, J. Huelsenbeck and F. Ronquist, "Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference". *Bioinformatics*, vol. 20, no. 3, pp.407-415, 2004.
- [14] A. Aberer, K. Kobert and A. Stamatakis, (2019). ExaBayes: Massively Parallel Bayesian Tree Inference for the Whole-Genome Era.
- [15] B. Rannala and Z. Yang, "Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference". *Journal of Molecular Evolution*, vol. 43, no. 3, pp. 304-311, 1996.
- [16] S. Andrews, (2010). FastQC: a quality control tool for high throughput sequence data.
- [17] H. Li, (2012). seqtk Toolkit for processing sequences in FASTA/Q formats.
- [18] R. Edgar, "Search and clustering orders of magnitude faster than BLAST". *Bioinformatics*, vol. 26, no. 19, pp. 2460-2461, 2010.
- [19] G. Boratyn, et al., "BLAST: a more efficient report with usability improvements". *Nucleic Acids Research*, vol. 41, no. W1, pp. W29-W33, 2013.
- [20] R. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput". *Nucleic Acids Research*, vol. 32, no. 5, pp.1792-1797, 2004.

A first approach to Acoustic Characterization of Costa Rican Children's Speech


Un primer acercamiento a la caracterización acústica del habla de niños costarricenses

Marvin Coto-Jiménez¹, Maribel Morales-Rodríguez²,
Daniel Vargas-Díaz³


Coto-Jiménez, M; Morales-Rodríguez, M; Vargas-Díaz, D.
A first approach to Acoustic Characterization of Costa Rican Children's Speech. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 80-84.

 <https://doi.org/10.18845/tm.v33i5.5080>

1 PRIS-Lab: Pattern Recognition and Intelligent Systems Laboratory, Department of Electrical Engineering, School of Engineering, Universidad de Costa Rica (UCR). E-mail: marvin.coto@ucr.ac.cr.

 <https://orcid.org/0000-0002-6833-9938>

2 Department of Orientation and Special Education, Universidad de Costa Rica (UCR). E-mail: maribel.moralesrodriguez@ucr.ac.cr.

 <https://orcid.org/0000-0002-3426-5192>

3 PRIS-Lab: Pattern Recognition and Intelligent Systems Laboratory, Department of Electrical Engineering, School of Engineering, Universidad de Costa Rica (UCR). E-mail: daniel.vargasdiaz@ucr.ac.cr.

 <https://orcid.org/0000-0002-9015-2328>



Keywords

Formants; Signal processing; speech technologies.

Abstract

As human interaction with computers becomes more pervasive, the value of developing automatic speech recognition, text-to-speech synthesis, and related speech technologies become more important for people of all ages, accents, and conditions.

One of the groups that represent bigger challenges is children, due to the difficulties in recording enough speech, and the lack of characterization of their speech, which is particular of every language and accent. This paper presents the first approach to acoustic analyses of Costa Rican children aged from six to twelve years. These analyses aimed to achieve a better understanding of the characteristics of speech produced by this group, in terms of providing future development and enhancement of automatic speech recognizers and speaker identification systems.

For this purpose, we record the speech consisting of isolated words of three children, and compare the results with three adults, in terms of the vowel's formants. The formants give information about the vocal track of the speaker, and it is an important method to provide the first analysis of these signals. Results show noticeable differences between the children and adults and may provide useful information about future trends to adapt and develop the current speech technologies for this population.

Palabras clave

Formantes; procesamiento de señales; tecnologías del habla.

Resumen

A medida que la interacción de las personas con las computadoras se hace más extendida, se vuelve más importante el desarrollo de tecnologías para el reconocimiento automático de la voz, la síntesis de voz, así como otras tecnologías relacionadas, considerando a personas de todas las edades, acentos y condiciones. Uno de los grupos humanos que representan desafíos más grandes es el de los niños, debido a las dificultades para grabar suficientes recursos de habla, y la falta de caracterización de su forma de hablar, la cual es particular de cada idioma y acento. Este artículo presenta una primera aproximación para el análisis acústico de niños costarricenses de entre seis y doce años. Estos análisis tienen como objetivo lograr una mejor comprensión de las características del habla producida por este grupo en particular, en términos de propiciar el desarrollo y mejora de los reconocedores automáticos del habla y los sistemas de identificación del hablante.

Para este propósito realizamos grabaciones de habla de tres niños, las cuales consistieron en palabras aisladas, y comparamos los resultados con tres adultos en términos de los formantes de las vocales. Los formantes proporcionan información sobre el tracto vocal del hablante, y es un parámetro importante para establecer un primer análisis de estas señales. Los resultados muestran diferencias notables entre los niños y los adultos y pueden brindar información útil para futuros estudios en términos de adaptar y desarrollar las tecnologías del habla para esta población.

Introduction

Human interaction with computers and technological devices of all kind has become more extensive in recent years. Being the speech the main form of human communication, the value of speech technologies, such as Automatic Speech Recognition (ASR), Text-to-speech synthesis and related technologies has increased.

This remarkable importance has still many challenges in building robust, and more natural systems for all people, including the elderly, children and people with disabilities. In the case of children, using speech technology is a field underdeveloped in certain contexts, as in the case of Costa Rica.

Several previous analyses have been focused on acoustic analysis of certain sounds, being the most frequent the vowels. Moreover, since this has been performed not only to age-dependent characteristics but also for specific accents [1] or conditions, such as Parkinson's disease [2].

For example, several references have reported the differences between acoustic and linguistic characteristics of children's speech [3-4]. In the English language, children's speech is characterized by higher pitch, and formants occur at higher frequencies [5]. This has an impact if exists on automatic speech recognition or analysis in the presence of perturbations or degradation of the signal, such as bandwidth reduction. Also, in the English language, children below the age of 10 exhibits a wider range of vowel durations relative to older children and adults, and wider variability in formant locations.

If automatic speech recognition systems are trained using acoustic models from adult speech and tested against speech from children, show performance degradation with decreasing age. On average, the word error rates are two to five times worse for children speech than for adult speech [6].

In this work, we perform a first approach to acoustic characterization of Costa Rican children speech, to achieve a better understanding of this particular age group.

Materials and methods

Like other Latin American variants of the Spanish language, the Costa Rican Spanish has five main categories of vowels [7]: <a> (open), <e> (open), <i> (closed), <o> (open), <u> (closed). The open and close categories refer to the general classification of vowels based on whether the sound is produced with the tongue far from the roof of the mouth (open) or with the tongue touching the roof of the mouth (closed).

To provide the inputs for the children voice database, several recording sessions were made during this study. The first session was held on January 2019 and had the participation of 3 children between 6 and 12 years old. The gender and ages of the participants are described in tables 1 and 2.

Table 1. Age of participants (children)

Gender	Age (years)
Male	6
Female	8
Female	12

Table 2. Age of participants (adults)

Gender	Age (years)
Male	18
Female	20
Female	23

Results

Previous references have related the changes in the formant of vowels with the characteristics and form of the vocal tract [8]. The vowel space is used to compare the range and variation in the position of the two first formants.

Figure 1 shows the mean position of the first two formants for two female children and two female adults of the dataset, using the same scale on both axes. The children have variable patterns in the polygon of formants. Whereas the adults demonstrate that they have a wider range regarding the formants and form a more homogeneous shape in this polygon.

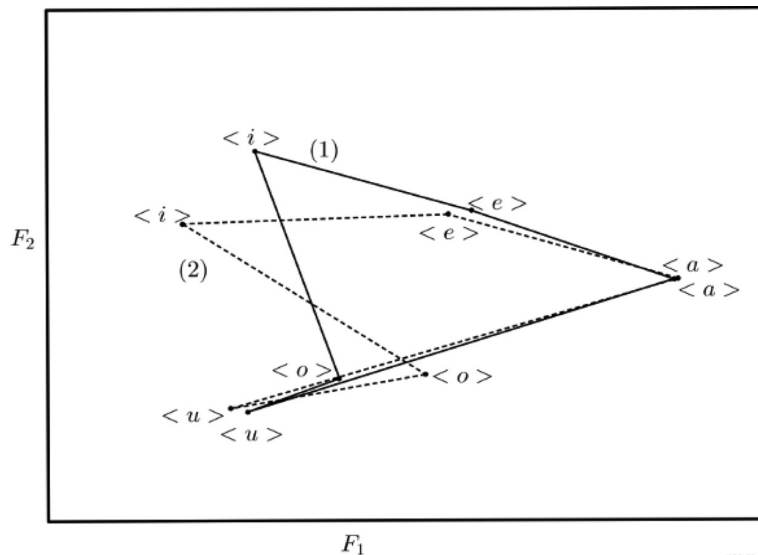


Figure 1: Changes in F1-F2 vowel space as a function of age (Female children (2) and adult speakers (1))

Table 3 shows the mean value of the formants for the <u> vowel. It is noticeable how different the values are between children and adults, even more for the second formant.

Conclusions

In this paper, the first approach to acoustic analyses of Costa Rican children's speech was presented. These analyses were focused on formants. The first group of children 6-12 years was recorded using isolated words, and the words were replied in a group of adults, for comparison purposes.

The formant polygons show different patterns between adults and children. These results are valuable for the knowledge of the speaker characteristics for this population and allow future research in areas such as gender/age voice recognition based on acoustic parameters such as formants. Additionally, the characterization of speech for this age and geographical region may allow the implementation of algorithms to improve the performance of automatic speech recognizers or other speech technologies for this variant of Spanish.

Table 3. Mean value of first and second formants for the vowel <u>. All the values in Hertz.

Speaker	First formant	Second formant
Child (Male, 6)	459,58	1268,76
Child (Female, 8)	441,83	1254,75
Child (Female, 12)	432,94	1192,64
Mean values (children)	444,78	1238,72
Adult (Male, 18)	362,94	1152,33
Adult (Female, 20)	409,20	1103,10
Adult (Female, 23)	444,24	1054,30
Mean values (adults)	405,46	1103,24

Acknowledgments

This work was supported by the University of Costa Rica (UCR), Project No. 322-B9-105, and Project No. ED-3416.

References

- [1] T. Leinonen, *An acoustic analysis of vowel pronunciation in Swedish dialects*. Groningen: [Rijksuniv.], 2010.
- [2] S. Skodda, W. Visser and U. Schlegel, "Vowel Articulation in Parkinson's Disease", *Journal of Voice*, vol. 25, no. 4, pp. 467-472, 2011. Available: 10.1016/j.jvoice.2010.01.009.
- [3] J. McKechnie, B. Ahmed, R. Gutierrez-Osuna, P. Monroe, P. McCabe and K. Ballard, "Automated speech analysis tools for children's speech production: A systematic literature review", *International Journal of Speech-Language Pathology*, vol. 20, no. 6, pp. 583-598, 2018. Available: 10.1080/17549507.2018.1477991.
- [4] L. Perry, M. Perlman, B. Winter, D. Massaro and G. Lupyan, "Iconicity in the speech of children and adults", *Developmental Science*, vol. 21, no. 3, p. e12572, 2017. Available: 10.1111/desc.12572.
- [5] S. Safavi, M. Najafian, A. Hanani, M. Russell, P. Jancovic, P., M. Carey, "Speaker recognition for children's speech". *arXiv preprint*. Available: arXiv:1609.07498.
- [6] A. Potamianos, S. Narayanan, "Robust recognition of children's speech", *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 603-616.
- [7] F. Martínez-Licona, J. Goddard-Close, A. Martínez-Licona and M. Coto-Jiménez, "Models and Analysis of Vocal Emissions for Biomedical Applications", 2013, pp. 235-238.
- [8] J. Eichhorn, R. Kent, D. Austin and H. Vorperian, "Effects of Aging on Vocal Fundamental Frequency and Vowel Formants in Men and Women", *Journal of Voice*, vol. 32, no. 5, pp. 644.e1-644.e9, 2018. Available: 10.1016/j.jvoice.2017.08.003.

Serialization of a 3D Human Body based on MoCap Data in a BVH File for Sequence Comparison



Serialización de un Cuerpo Humano Tridimensional basado en datos MoCap en un archivo BVH para la Comparación de Secuencias

Natalia Abarca-Jiménez¹, Francisco Siles-Canales²

Abarca-Jiménez, N; Siles-Canales, F. Serialization of a 3D Human Body based on MoCap Data in a BVH File for Sequence Comparison. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 85-90.

 <https://doi.org/10.18845/tm.v33i5.5082>



- 1 PRIS-Lab: Pattern Recognition and Intelligent Systems Laboratory. Department of Electrical Engineering, School of Engineering, Universidad de Costa Rica. E-mail: natalia.abarcajimenez@ucr.ac.cr
 <https://orcid.org/0000-0001-6445-750X>
- 2 PRIS-Lab: Pattern Recognition and Intelligent System Laboratory, Department of Electrical Engineering, School of Engineering, Universidad de Costa Rica, Costa Rica. Email: francisco.siles@ucr.ac.cr.
 <https://orcid.org/0000-0002-6704-0600>

Keywords

BVH; Motion Capture; Serialization; Sequence comparison.

Abstract

This work focuses on the use of sequence alignment algorithms as validation of the results obtained in the serialization of tridimensional body using BVH file, from obtained a sequence 3D to the symbology necessary for the validation that represents the data obtained in the Pattern Recognition and Intelligent Systems Laboratory (PRIS-Lab). The procedure use steradians as the method to obtain the symbology for the sequence for the Needleman Wunsch algorithm for alignment sequence, for validation the alignment algorithm score is used.

Palabras claves

BVH; Captura de Movimientos; Serialización; Comparación de Secuencias.

Resumen

Este trabajo se enfoca en el uso de algoritmos de alineamiento de secuencias como la validación de los resultados obtenidos de la serialización de cuerpos tridimensionales utilizando archivos BVH, desde obtener una secuencia 3D hasta la simbología necesaria para la validación de los datos obtenidos en el Laboratorio de Reconocimiento de Patrones y Sistemas Inteligentes (PRIS-Lab). El procedimiento utiliza esteroreadianes como el método para la obtención de la simbología necesaria para utilizar el algoritmo Needleman Wunsch para las secuencias, y como métrica de validación se utiliza el resultado numérico del algoritmo.

Introduction and Related Work

Biomechanics studies the living beings, looking for relations between magnitudes and looking for behavioral explanations and observations. (Aguilar, 2000) The structures and forces allow the study of the movements from an anatomical or structural point of view, so that the movements are deduced from the structure in movement (skeleton, joints, etc.) applying physiological and physical properties. This science is in development due to the difficulty of establishing a normalization of the human body.

The main complication during the biomechanical analysis of the human body is comparing two different actors performing the same action, since the movement depends of the upper or lower extremities, largely on their dimensions and the mass distribution, the results obtained in the same movement can differ greatly from one person to another (Ceccarelli, M.J. Gómez García and C. Castejón Sisamón and J.C. García Prada and G. Carbone and M., 2011). Therefore, a normalization is required to be able to compare the actions of two people without the results being affected by their dimensions.

Nowadays, in order to capture the performance of people doing sports and other activities in real-time, several MOtion CAPture systems (or MoCap for short) are commercially available like OptiTrack, Qualysis, Codamotion or Vicon that offer cameras for indoor or outdoor, virtual reality systems and the accessories necessary for the task. Motion capture is a process in which the real movements of humans are obtained and used to map these to a virtual representation. Motion Capture involves the detection, digitization, and capture of the actions made by the actors to obtain the three-dimensional representation of it. (Parent, 2012).

There are different MoCap technologies, where the most commonly used is optical capture. This kind of system is based on the measurement of reflected light, mainly using passive infrared markers placed on the body to track. The main advantage of this type of system is that by using the images created from the sensors, valuable information such as the position of every sensor can be extracted. Although this information usually has a high computational cost to be able to deduce it, and its main disadvantage is that the sensors require a direct line of light from the light emission (Welch, G.F. and Foxlin, Eric, 2002). Ramos created a system to extract movement patterns using kinematic variables at the points of an individual's body using the Kinect camera. This is a depth camera system that uses an infrared emitter and for processing uses the captured images and the OpenNI software that determines the depth of the body. (Gutiérrez, 2013)

Methodology

First, it is necessary to obtain a tridimensional representation of the human movement, which is done with the optical motion capture system of PRIS-Lab. This requires the use of a suit to put passive infrared markers to obtain the information with OptiTrack Prime 13 cameras, then this representation is saved in the BVH file.

This file passes through the normalization algorithm that consist in:

- Read the file and obtain the *OFFSET* information from the BVH file.
- Obtain the biggest magnitude in the information from part one.
- Make a new file BVH normalized.

This way a new BVH file is obtained as output of the algorithm. With this file, the information from 20 points is obtained and use for the creation of a sequence that describes the movement.

Sequence N
 Head, LeftShoulder, RightShoulder, LeftElbow, RightElbow,
 LeftHand, RightHand, LeftKnee, RightKnee, LeftToe, RightToe.

Each component of the sequence is obtained by a vector sum of the points of the BVH file, then each of this need to be converted to a symbol, because the sequence alignment algorithm use symbols for the sequence not a vector. For this conversion steradians are used; this is the unit of measurement of the solid angle and can be obtained at drawing a sphere of radius R with center P, this intersects the conic surface generating a surface sector S that is described by Equation 1.

$$S = R^2$$

Equation 1. Steradian equation.

For each component, the correspondent steradian is calculated, this way a sequence of eleven symbols is obtained that represented the human movement. The alignment algorithm uses letters as part of the sequence, because it is used mainly for ADN and ARN; so for better results is necessary to obtain a sequence that uses letters as the symbol. As the dimensions of the body are normalized, the maximum results of the steradian are 1, we can use the alphabet letters and each of those has a value of 1/27 (using ñ). The results can be validated using the metric of the sequence alignment algorithm. If both sequences are the same the result of the Needleman Wunsch algorithm is zero.

Experimental Results and Discussion

For this experiment we take data from two different subjects, the subject N has a height of 1.7m and the subject R of 1.5m (figure 1). Using the software Blender to obtain the graphical representation and the Ruler tool to measure the height.

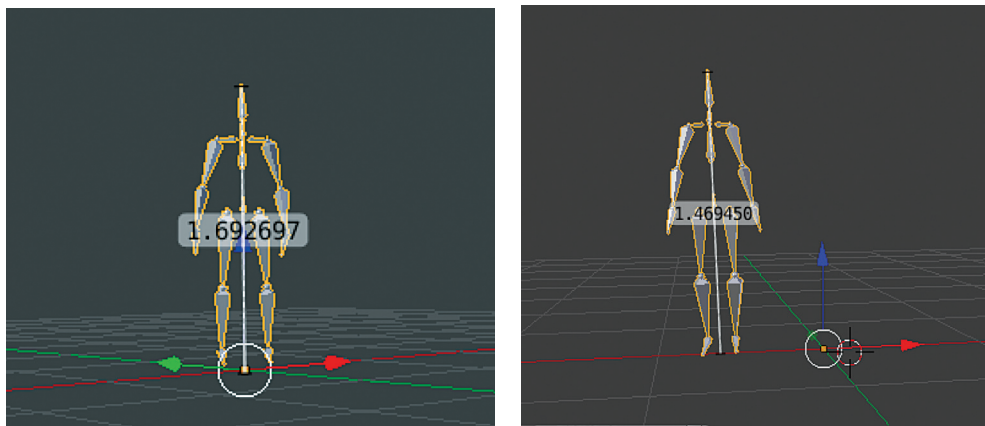


Figure 1. Normal position for the subjects.

Each of the representations was passed through the algorithm, as maximum magnitude for the normalization and creation of the new file 0,447347 was obtained for the Subject N and 0,365281 for the Subject R. Using each of these maximum the new data for each subject and its graphical representation shows in the figure 2.

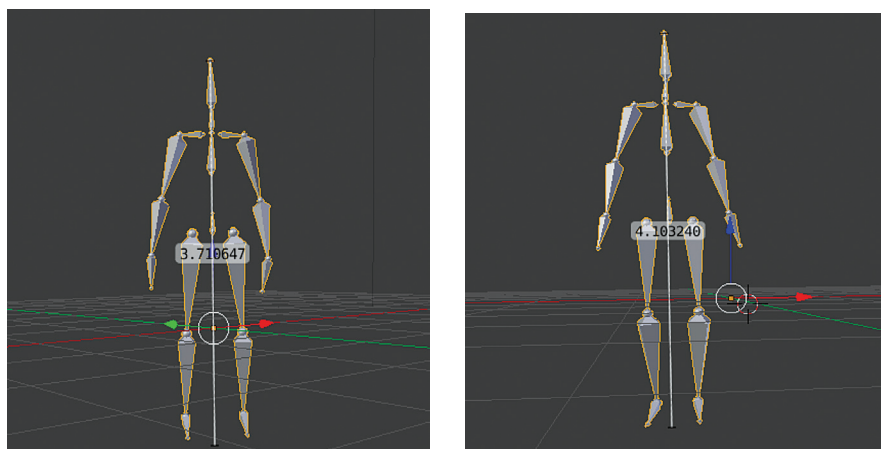


Figure 2. Normal position for the subjects after the algorithm

For a quantitative result in the comparison of the movement, after the creation of the new BVH file is necessary to define the N sequence. The file contains the position of 20 points of the body, with these points the sequence was defined which is obtained by a vector sum. This results in a sequence composed of vectors, but the alignment algorithm only allows sequences composed of letters to obtain desired results. Therefore, it was decided to use steradians as a symbol for each vector. The sequence using the corresponding symbology is shown in table 1.

Table 1. Sequence N for each subject.

Subject N	Subject R
0,0	0,0
0,459548	0,465821
0,459283	0,465811
0,620894	0,56875
0,620317	0,568725
0,77471	0,65134
0,773961	0,651305
0,457493	0,359617
0,457493	0,359617
0,919838	0,791834
0,919838	0,791834

Each of the components of the N sequence is matched with the corresponding letter, using this convention the N sequence is shown in table 2.

Table 2. Sequence N using letters for each subject.

Subject N
A M M P P T T M M X X
Subject R
A M M O O Q Q J J U U

When both sequences are obtained the sequence alignment algorithm is used to obtain a quantitative result, when passing it through the Needleman Wunsch a result of 8 is obtained.

Conclusions and Future Work

The serialization of the human body was used to create the normalization data. For each BVH file, the maximum magnitude was found to perform the normalization and create the output BVH file. Then the creation of the N sequence was made with the use of the steradian as symbols and each of them was matched with its respective letter. The metric of the Needleman Wunsch sequence alignment algorithm was used to obtain a quantitative result. This algorithm creates a symbology to pass it through to have a quantitative comparison of a movement made by two people with different physical characteristics. The quantitative result gives us an idea that the difference between the movements are bigger than what was expected, that is zero because the movements are the same. One way that this can change is using another symbology for the sequence instead of the steradians, and compare the results obtained with different symbology. Using different symbology allows the comparison of them to obtain the smallest result with the sequence alignment algorithm for the same movement.

References

- Aguilar, G. M. (2000). Biomecánica: física y fisiología. Madrid: Instituto de Ciencia de los Materiales, Consejo Superior de Ingestigaciones Científicas.
- Ceccarelli, M.J. Gómez García and C. Castejón Sisamón and J.C. García Prada and G. Carbone and M. (2011). Motion capture systems for human walk. Analysis and comparative. Obtenido de XIX CONGRESO NACIONAL DE INGENIERÍA MECÁNICA: <http://www.xixcnim.uji.es/CDActas/Documentos/ComunicacionesPosters/11-03.pdf>
- Gutiérrez, D. R. (2013). Estudio cinemático del cuerpo humano mediante kinect. Madrid: EUIT Telecomunicación.
- Parent, R. (2012). Computer Animation: Algorithms and Techniques. Waltham, MA: Elsevier Science.
- Welch, G.F. and Foxlin, Eric. (2002). Motion Tracking: No Silver Bullet, but a Respectable Arsenal. Computer Graphics and Applications, IEEE, 24-38.




Validation-data Generation for Brightfield Microscopy Cell Tracking using Fluorescence Samples

Generación de Datos de Validación para Rastreo Celular en Microscopía de Campo Claro usando Muestras Fluorescentes

Patricia Quinde-Cobos¹, Steve Quirós²,
Francisco Siles-Canales³

Quinde-Cobos, P; Quirós, S; Siles-Canales, F. Validation-data Generation for Brightfield Microscopy Cell Tracking using Fluorescence Samples. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 91-95.

 <https://doi.org/10.18845/tm.v33i5.5083>

- 1 PRIS-Lab: Pattern Recognition and Intelligent Systems Laboratory. Department of Electrical Engineering, School of Engineering, Universidad de Costa Rica (UCR). Postgraduate Studies in Electrical Engineering, UCR. CNCA: National Advanced Computing Collaboratory, CeNAT: National Center for High Technology. E-mail: lesly.quinde@ucr.ac.cr
 <https://orcid.org/0000-0002-1042-6545>
- 2 Tumoral Chemosensitivity Laboratory, Research Center on Tropical Diseases, School of Microbiology, UCR. E-mail: steve.quirros@ucr.ac.cr
 <https://orcid.org/0000-0002-9377-0199>
- 3 PRIS-Lab: Pattern Recognition and Intelligent Systems Laboratory. Department of Electrical Engineering, School of Engineering, Universidad de Costa Rica (UCR). E-mail: francisco.siles@ucr.ac.cr
 <https://orcid.org/0000-0002-6704-0600>



Keywords

Brightfield microscopy; cancer; fluorescence microscopy; pattern recognition.

Abstract

This work focuses on the use of fluorescent cancer cell images as data to validate the results obtained in segmenting brightfield cancer cell images, as the latter's current validation consists of manual annotation of cells in the original images. The procedure uses pattern recognition and starts with preprocessing the fluorescent samples to ensure cell detection, focused on area and intensity value. As the fluorescent images are segmented, each cell's nucleus is detected and counted, with a high success rate as each nucleus's contour was detected with its original shape. As each image's density is calculated, they can be clustered according to their density value and used for cell detection in brightfield samples.

Palabras clave

Cáncer; microscopía de campo claro; microscopía de fluorescencia; reconocimiento de patrones.

Resumen

Este trabajo usa imágenes de fluorescencia de células cancerígenas como para validación de resultados obtenidos por segmentación de imágenes de campo claro de células cancerígenas, ya que actualmente la validación consiste en la anotación manual de las imágenes originales. Se usó reconocimiento de patrones y se inició con preprocesamiento de muestras fluorescentes para asegurar la detección de células, considerando área y valores de intensidad. Al segmentar las imágenes de fluorescencia el núcleo de cada célula es detectado, con su forma original. Al calcular la densidad de cada imagen, estas pueden ser agrupadas de acuerdo a su valor de densidad y usadas para la detección de células en muestras de campo claro.

Introduction

Cancer occurs when damaged cells are able to bypass checkpoints and replicate out of control, transmitting the damage. When cancer is considered as a possible diagnosis, part of the affected tissue is analyzed and maintaining the genetic material's integrity is essential.

The analysis of cell samples is a complex process that depends on the technique used to see them. The genetic material must retain its integrity to obtain valid results, so an appropriate technique to visualize the samples is brightfield microscopy. This technique is used as it is of low cost and the sample isn't damaged. The brightfield images studied for cancer make analysis difficult, they present low contrast and a density that increases through time, as a study can last about 92 hours. These images also have noise and illumination irregularities, the cell's contours can't be properly defined and the nuclei are not distinguishable.

Fluorescence microscopy is a technique that in these cancer samples shows nuclei, with different levels of absorption for each cell. This technique, although it allows the visibility of the cells to be studied, compromises their DNA integrity.

Considering that throughout the study, images in brightfield and fluorescence microscopy are obtained, the purpose of this project is to use microscopy segmented images as validation data to corroborate the results in bright field segmented images, and guarantee a cell tracking algorithm's validity. It also speeds up the process of obtaining validation data, as the current way involves hours of manually annotating cells in brightfield microscopy images.

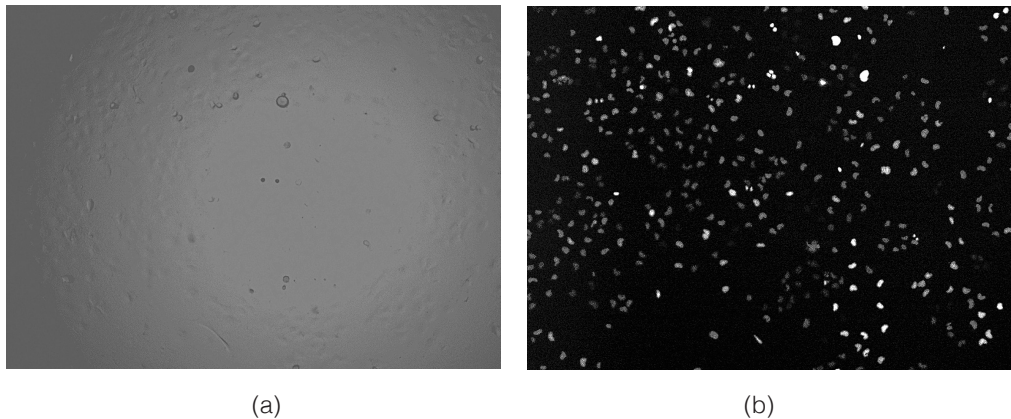


Figure 1. Samples obtained through (a) Brightfield microscopy and (b) Fluorescence microscopy

Related Work

Commercial solutions for analysis of cell samples do exist and most of them focus on using additives as staining or wave producing filters to obtain fluorescent samples [1]. The problem is that this method cannot guarantee or protect the integrity of the genetic material in the cells, so a complete analysis is difficult to perform.

According to Zuiderveld [2] and Kim [3], CLAHE Equalization divides the image in blocks and performs the histogram equalization, avoiding noise amplification. Some studies that used fluorescence microscopy show techniques as anisotropic smoothing, gradient mapping and the use of characteristics as size and shape [4], [5]. Other solutions [6] could be used for noise removal and edge preservation. The deceived bilateral filter allows noise reduction and contrast increase showing the benefit of a good preprocessing phase can overall help the tracking algorithm in its search of the image's cells [7].

For unstained cell detection as segmentation is needed, Espinoza worked on an alternative that used of a local threshold after the global threshold is estimated considering contours [8]. Operations in this and in [9] come with a computational price, as they are applied first to the whole image and then to different regions in the image separately and sequentially. Contour definition and center detection were considered ideal to segment cells separately from the cluster in high density images [10], [11].

Methods

For cell identification in fluorescent cell images, it is necessary to first implement a preprocessing phase, to guarantee the segmentation and detection of the cells in the images. The next steps are followed to ensure cell detection:

1. The bilateral filter for smoothing and erasing noise while preserving contours. The Canny filter with a Gaussian filter helps define the nuclei's contours in preparation for segmenting.
2. The segmentation using the Otsu technique can obtain a binary image, resulting in white objects and black background.

Detection of cells in brightfield microscopy and their nuclei in fluorescence microscopy considers their intensity and size. As the objects are detected, their positions in the images are preserved for future procedures. Through the analysis of different images throughout the study, we can classify the images according to their density, based on the number of nuclei detected.

$$density = \frac{cellpixels}{totalpixel}$$

Equation 1. Image density.

As this value is obtained for each image, they can be classified and obtain the segmentation's result according to how many cells each image has, and test the validity of the method depending on the image's density. The validation of the methodology followed must detect any error present in the procedure, as they translate into its utility. For a quantitative validation manually annotated cells fluorescent images are used to compare the number of cells found and their position in the image.

Experiments and Preliminary Results

To obtain the cells present in a Brightfield cell samples, the algorithm for pattern recognition is followed, with intensity, shape and size as special characteristics considered (Quinde-Cobos, 2020). These samples are characterized by their low contrast and increase in density as the study advances. Because of this, the complexity of the algorithm increases, which has to be covered in the detection and tracking algorithm.

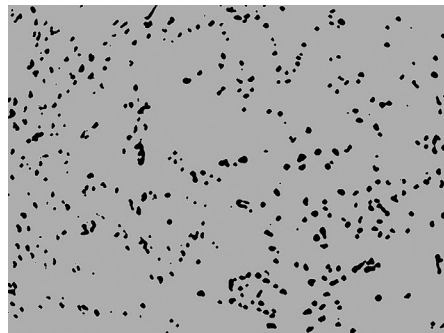


Figure 2. Objects detected for Brightfield image

Figure 2 shows a preliminary result of Brightfield sample's segmentation of high density. In comparison to the manually annotated image, a significant percentage of the cells were not detected, which showed the necessity of the fluorescent sample's segmentation as validation data to make easier the analysis of results as shown in the image.

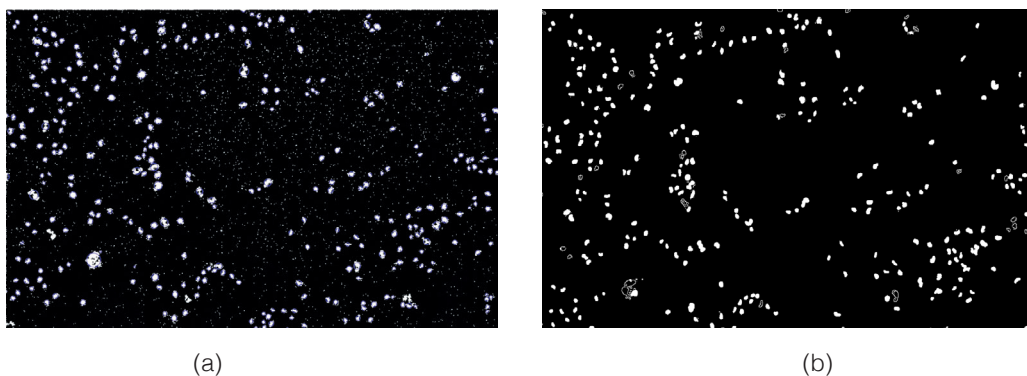


Figure 3. Objects detected for fluorescent images

The figures 3.a and 3.b show objects detected for an image obtained through fluorescence microscopy. In 3.a, each detected cell has a blue circle over it to show that it was properly detected. Morphology operations were applied which modified the contours' original shape. In 2.b, each cell was identified because of their contour, which allows to maintain their original shape but limits their identification to the detection of the contours. In comparison to their annotated images, the second method provided a better detection for the objects present, which represent each cell's nucleus.

Conclusions and Future Work

The basic structure of the pattern recognition process was used to create each segmentation's algorithm. These algorithms, created for these specific cancer cell samples, try to eliminate noise or other hindrances present in the images to facilitate the object detection process. For each image, the present cells were obtained and their position in the image saved for validation and future procedures.

With the detection of the objects present in the fluorescent images, an analysis of the density ratio according to the frame number must be performed, to detect clustering in density values and divide them into groups, resulting in a classification of the images based on their density, which can help create a personalized method according to this measurement. These classification can be applied to an algorithm for brightfield cell detection and tracking to add robustness and validate the results and objects found in each sample.

Measurements of density may vary depending on the type of cancer studied and how fast cells replicate according to the type analyzed. Because of that, to each study analyzed the density values must be obtained to ensure the images classification according to the number of objects present.

The data managed in this work differ from the data analyzed in previous studies, given the complexity of their features and variations along the samples examined. Even though the method followed may appear outdated, according to the literature consulted it gives the desired result.

References

- [1] Thermofisher. *Thermofisher*. Available: www.thermofisher.com/, 2019.
- [2] K. Zuiderveld, "Contrast limited adaptive histogram equalization," Academic Press Inc., 1994.
- [3] S. Kim *et al.*, "Determining parameters in contrast limited adaptive histogram equalization," *ASTL*, vol. 21, pp. 204-207, 2013.
- [4] U. Adiga *et al.*, "High-throughput analysis of multispectral images of breast cancer tissue," *IEEE Transactions on Image Processing*, vol. 18, no. 8, pp. 2259-2268, 2006.
- [5] A. Stajduhar, *et al.*, "3d localization of neurons in brightfield histological images," *International Symposium ELMAR*, vol. 60, pp. 75-78, 2018.
- [6] C. Tomasi, "Bilateral filtering for gray and color images," *ICCV*, vol. 6, pp. 257-286, 1998.
- [7] S. Calderon *et al.*, "Dewaff: A novel image approach to improve the performance of a cell tracking system," *IWOBI*, pp. 81-88, 2015.
- [8] E. Espinoza *et al.*, "Cell cluster segmentation based on global and local thresholding for in-situ microscopy," *ISBI: Macro to Nano*, vol. 3, pp. 542-545, 2006.
- [9] E. Andrade *et al.*, "Region-based analysis and retrieval for tracking semantic objects and provision of augmented information in interactive sport scenes," *IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1084-1096, 2005.
- [10] C. Zhang *et al.*, "Yeast cell detection and segmentation in brightfield microscopy," *IEEE*, pp.1267-1270, 2014.
- [11] A. Sheehy *et al.*, "Region and contour based cell cluster segmentation algorithm for in-situ microscopy," *CCE*, vol. 5, pp. 168-172, 2008.

A Biocomputational Platform for Template-based Protein-protein Docking


Plataforma computacional de acoplamiento de proteínas basado en plantillas

Ricardo Román-Brenes¹, Francisco Siles-Canales²,
Daniel Zamora-Mata³


Román-Brenes, R; Siles-Canales, F; Zamora-Mata, D. A Biocomputational Platform for Template-based Protein-protein Docking. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 96-100.

 <https://doi.org/10.18845/tm.v33i5.5084>


1 PRIS-Lab: Pattern Recognition and Intelligent Systems Laboratory School of Electrical Engineering, Faculty of Engineering, Universidad de Costa Rica (UCR). Costa Rica.
E-mail: ricardo.roman@ucr.ac.cr.

 <https://orcid.org/0000-0002-6104-7561>

2 PRIS-Lab: Pattern Recognition and Intelligent Systems Laboratory School of Electrical Engineering, Faculty of Engineering, Universidad de Costa Rica. (UCR). Costa Rica.
E-mail: francisco.siles@ucr.ac.cr.

 <https://orcid.org/0000-0002-6704-0600>

3 PRIS-Lab: Pattern Recognition and Intelligent Systems Laboratory School of Electrical Engineering, Faculty of Engineering, Universidad de Costa Rica (UCR). Costa Rica.
E-mail: daniel.zamoramata@ucr.ac.cr.

 <https://orcid.org/0000-0001-8213-4974>



Keywords

Clustering; protein-protein docking; template-based docking; HPC.

Abstract

We propose the creation of a Biocomputational Platform for template-based protein-protein docking that aims reduce computational time by clustering data before the rigid body alignment. Using data from the Dockground project, models will be created using multiple clustering methods that will annotated each protein into a class, such that when performing the match search, not all of the databank needs to be inspected but just the class that resembles the most to the studied protein. This will reduce the time that conformation matching requires without incurring in lower precision.

Palabras clave

Agrupamiento; acoplamiento de proteínas; Acoplamiento basado en plantillas; HPC.

Resumen

Se propone la creación de una Plataforma Biocomputacional para el acoplamiento de proteínas que reduce el tiempo computacional al agrupar los datos previo al alineamiento de cuerpos rígidos. Utilizando datos del proyecto Dockground, se crearán modelos usando múltiples métodos de agrupación que asignarán cada proteína a un grupo para que, al realizar la búsqueda de pares, no se realice una búsqueda a fuerza bruta, sino una acotada. Esto reducirá el tiempo que requiere una conformación en ser procesada sin perder precisión.

Introduction

Proteins are macromolecules that serve as catalyst for virtually all the cell's functions. In order to perform these functions, proteins need to interact with each other to form functional complexes [1]. Protein-protein docking (PPD) is an important area of study in molecular biology due to its importance in fields like drug discovery or precision medicine. This process can be done in silico or in vivo. The two general methods that exist for in silico PPD are de novo and template-based. The latter technique is called template-based protein docking (TBPD).

TBPD consists in finding out if two proteins form a complex by matching them by homology in a databank of complexes. If it is required to know if protein A and B can dock and form a functional complex, they would be iteratively aligned to all the proteins P in the databank. These alignments typically yield a score so that good matches can be selected. Finally, if protein A matches to a part of a complex and B matches to the other part of the complex, then it is said that A and B can dock [1].

In any case, the computational methods for docking proteins is very demanding on CPU resources [2] [3].

We propose here a computational platform with interchangeable modules to utilize a different set of descriptors, namely 3D geometrical descriptors [4] [5] to cluster, classify and validate TBPD in order to improve overall response time. The platform will use annotated data from the Dockground project [6] (full structure docking templates v1.1 databank) and run in the PRIS-Lab supercomputer TARÁ [7].

Related Work

The use of clustering techniques for PPD is not a something new. There are numerous studies regarding this topic.

The Critical Assessment of Predicted Interactions (CAPRI) [8] is a community-wide experiment to measure the capabilities of protein-docking methods. In this experiment several teams try their algorithms to assess which is the best one for the task at hand. In 2005, most of the methods, 13 out of 20, used some sort of clustering technique in one point or another of their workflow [9].

By far the most common use for clustering is to discard decoys [10] [2] [11] [12] [13] [14] [15] [16]. Despite this extensive use of clustering, as far as this study goes, the importance of computational time has been relegated to a second plane of importance. We did not find any study that uses clustering to generate models prior to the mass comparison of each template candidate rigid body, in order to speed up the overall process.

Methodology and Experiments

The platform will consist of several components. Items can be added to these components

- A set of paired template candidates $TC = \{(tc00, tc01), (tc10, tc11), \dots, (tcn0, tcn1)\}$ that form a functional complex.
- A set of 3D geometrical feature extraction methods $F = \{fe0, fe1, \dots, fen\}$ that will generate the features to be used in the clustering of the Dockground databank. Features like protein area, protein volume, protein circularity, 3D Zerkine descriptor for the protein (Daberdaku & Ferrari, 2018) and the 3 PCA components of the protein (en R3 solo hay 3 PCA components) will be used.
- A set of clustering methods $C = \{c0, c1, \dots, cn\}$ that will form groups based on the 3D features. All the candidates TC will be clustered with each method in order to generate groups. Clustering methods like k-means, x-means, OPTICS and BIRCH will be used.
- A set of classification methods $K = \{k0, k1, \dots, kn\}$ with which new proteins will be matched with subgroups in order to reduce run time. For starters the classifications methods to be used are SVM, k-nearest neighbors and LDA.

Each component of a set can be used interchangeably so that if for instance c0 does a better job than c3, the user or programmer can choose which one to use. A parallel and distributed implementation of the platform will be done so that multiple methods can be run simultaneously, since the process of each Fe, C, K, alignment and low-resolution docking by GRAMM don't have any kind of data dependency.

Each stage can be run by a pool of process or threads in TARÁ.

The Platform will have at least two modes of operation: databank clustering and docking prediction.

For the docking prediction, the normal workflow would be as follows

- Obtain the PDB files for the 2 proteins that want to be docked, p0 and p1.
- Choose one or more of the already clustered databanks generated with the features Fe extracted from TC using C methods.
- Using the groups generated in the previous step, choose one or more classification methods to find the best match for p0 and p1 in TC. The similarity of the match will be measured with TM-align (Zhang & Skolnick, 2005).

- The Platform shows, ordered by TM-align, the list of template candidates and the visualization of both the templates and p0 and p1.

For the databank clustering, the procedure would be

- Choose one of the clustering methods C.
- Select which databank to use, initially Dockground will be used. This databank will be split, first, in two sets for work and validation, and then the training set will be split in three, for training, testing and confirmation.
- Choose which 3D features from F will be used to perform the clustering.
- The Platform shows the results of the clustering and performs validation using a 10-fold cross-validation against the already annotated entries of the validation set from Dockground.
- If the user is satisfied with the results, the models will be saved as annotated data for
- classification, for future use in the docking prediction.

Aside from the automatization of the process of protein docking, run-time tests as well as a comparison of the docking solutions will be performed to ensure that the Platform behaves well against the current methods for TBPD, in particular using TM-align [17] and GRAMM-X [18]. The Platform's parallel implementation will ensure at least a higher throughput of data.

From an algorithmic point of view, this process can be seen as several time functions, all dependent on the size of the protein databank n . $T(n)$ where n is the size of the protein databank. These functions can be seen in equationa 1, 2 and 3

$$T_0(n) = 2 \times \left(\sum_{i=1}^n lr(i) + \sum_{i=1}^n a(i) \right)$$

Equation 1. Time function of normal TBPD process, where $lr(i)$ is the time taken to transform a PDB into a low-resolution representation and $a(i)$ is the time taken to align a protein to another one in the databank.

$$T_1(n) = 2 \times \frac{\left(\sum_{i=1}^n lr(i) + \sum_{i=1}^n a(i) \right)}{P}$$

Equation 2. Time function of normal TBPD method proposed here, after the databank has been annotated with clustering. Here P is the size of the largest cluster, $lr(i)$ is the time taken to transform a PDB into a low-resolution representation and $a(i)$ is the time taken to align a protein to another one in the databank.

$$T_2(n) = \left(\sum_{i=1}^n fe(i) + \sum_{i=1}^n c(i) + \sum_{i=1}^n k(i) \right)$$

Equation 3. Time function for the clustering of the protein databank. Here $fe(i)$ refers to the time taken by one of the feature extraction methods, $c(i)$ is the take by a clustering method and $k(i)$ the time taken by a classification method

Expected Results

We expect that this study shows that run-times can be lowered if annotated data from clustering is used prior to the rigid body docking. The clustering model creation time can be significant but since this is a task that will be performed few times, it is negligible in the overall picture. We

don't expect that the docking of one particular pair of proteins will be faster than using TM-align and GRAMM-X but using our system for batch process large amounts of proteins will give much better response times than others. This means that Equation 1 might not take much more time than Equation 2, but after performing an amortized time analysis, the response-time will be much lower.

References

- [1] A. Szilagyi and Y. Zhang, "Template-based structure modeling of protein-protein interactions.," *Current Opinion in Structural Biology*, pp. 10-23, 2014.
- [2] S. B. Abhishek K, "Protein-Protein Docking Using MultiDimensional Spherical Basis Functions on High Performance Computing Platform," *International Journal of Innovative Technology and Exploring Engineering*, 2019.
- [3] M. H. Avinash Mishra, "Computational Issues of Protein-Ligand Docking," *Journal of Biomolecular Research & Therapeutics*, 2018.
- [4] S. Daberdaku and C. Ferrari, "Exploring the potential of 3D Zernike descriptors and SVM for protein-protein interface prediction," *BMC Bioinformatics*, 2018.
- [5] I. Budowski-Tal, R. Kolodny and Y. Mandel-Gutfreund, "A Novel Geometry-Based Approach to InterProtein Interface Similarity," *Scientific Reports*, 2018.
- [6] P. J. Kundrotas, I. Anishchenko, T. Dauzhenka, I. Kotthoff, D. Mnevets, M. M. Copeland and I. A. Vakser, "Dockground: A comprehensive data resource for modeling of protein complexes," *Protein Science*, pp. 172-181, 2018.
- [7] PRIS-Lab, "TARÁ - Cluster HPC PRIS-Lab," PRIS-Lab/UCR, 1 8 2019. [Online]. Available: <https://wiki.prislab.org/doku.php?id=tara>. [Accessed 1 8 2019].
- [8] H. K. Janin J., "CAPRI: a Critical Assessment of PRedicted Interactions.," *Proteins*, 2003.
- [9] R. Méndez, R. Laplae, M. F. Lensink and S. J. Wodak, "Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures," *Proteins*, 2005.
- [10] G. G.-P. Varela-Salinas, "Visual Clustering Approach for Docking Results from Vina and AutoDock," *Hybrid Artificial Intelligent Systems*, 2017.
- [11] S. R. Comeau, D. W. Gatchell, S. Vajda and C. J. Camacho, "ClusPro: a fully automated algorithm for protein-protein docking," *Nucleic Acids Research*, pp. 96-99, 2007.
- [12] J. J. Gary, S. W. C. S.-F. O. Moughon, B. Kuhlman, C. A. Rohl and D. Baker, "Protein-Protein Docking with Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations," *Journal of Molecular Biology*, pp. 281-299, 2003.
- [13] D. Kozakov, K. H. Clodfelter, S. Vajda and C. J. Camacho, "Optimal Clustering for Detecting Near-Native Conformations in Protein Docking," *Biophysical Journal*, pp. 867-875, 2005.
- [14] S. Vajda and D. Kozakov, "Convergence and combination of methods in protein-protein docking.," *Current Opinion in Structural Biology*, pp. 164-170, 2009.
- [15] M. Torchala, I. H. Moal, R. A. Chaleil, J. Fernandez-Recio and P. A. Bates, "SawrmDock: a server for flexible protein protein docking," *Bioinformatics*, pp. 807-809, 2013.
- [16] S. Lorenzen and Y. Zhang, "Identification of near-native structures by clustering protein docking conformations," *Proteins: Structure, Function and Bioinformatics.*, pp. 187-194, 2007.
- [17] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Research*, pp. 2302-2309, 2005.
- [18] A. Tovchigrechko and I. A. Vakser, "GRAMM-X public web server for protein-protein docking," *Nucleic Acids Research*, 2006.

Design of a prototype of hand orthosis with pneumatic actuators


Diseño de un prototipo de órtesis para mano con actuadores neumáticos

Pablo Enrique Tortós-Vinocour¹, Sofia Valverde-Gutiérrez²,
Marta Eugenia Vílchez-Monge³


Tortós-Vinocour, P; Valverde-Gutiérrez, S; Vílchez-Monge, M, N; Siles, F. Design of a prototype of hand orthosis with pneumatic actuators. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 101-105.

 <https://doi.org/10.18845/tm.v33i5.5085>

1 Simulation Bioengineering Laboratory (SIBILA), Mechatronics Engineering Academic Área, Costa Rica Institute of Technology, Cartago, Costa Rica. E-mail: pablortos1@hotmail.com.

 <https://orcid.org/0000-0001-7236-9374>

2 Simulation Bioengineering Laboratory (SIBILA), Mechatronics Engineering Academic Área, Costa Rica Institute of Technology, Cartago, Costa Rica. E-mail: sofiavalverde8@gmail.com.

 <https://orcid.org/0000-0003-0280-2472>

3 Professor, Simulation Bioengineering Laboratory (SIBILA), Mechatronics Engineering Academic Área, Costa Rica Institute of Technology, Cartago, Costa Rica. E-mail: mvilchez@itcr.ac.cr.

 <https://orcid.org/0000-0002-3271-2569>





Keywords

Orthosis; pneumatic actuation; cerebral palsy; soft robotics.

Abstract

In Costa Rica, there has been an increasing need for prosthetic and orthotic devices in the last years, especially in populations with diabetes, cerebral palsy, and people over 65 years old, some of the devices are developed in the country, however, others such as hand orthoses are rarely produced in the region. The use of soft robotics in the replica of human movement has been growing. By considering the work done by Deimel and Brock[3] and the Simulation Bioengineering Laboratory(SIBILA) it is intended to develop a soft hand orthosis for a patient with cerebral palsy, who has problems grasping objects with his right hand. This project aims to improve grip strength and finger position for the patient. The project is divided in 3 parts: replica of the patient's hand, elaboration of pneumatic muscles and the elaboration of a voice control system. The result is a pneumatic actuation that allows to grip objects with different shapes and gives strength to the fingers, in addition to this, the orthosis fits the model of the patient's hand. In conclusion, it is recommended to optimize the design of the muscles with a silicone that is more resistant and a thicker reinforcement helix. Also, it is recommended to start carrying out the tests of the orthosis with the model of the hand inside it.

Palabras clave

Órtesis; actuación neumática; parálisis cerebral; robótica suave.

Resumen

En Costa Rica, en los últimos años se registró un aumento en la necesidad de dispositivos protésicos y ortésicos, en poblaciones con diabetes, parálisis cerebral y adulta mayor, algunos de los dispositivos son elaborados en el país, sin embargo, otros como las ortesis de mano son raramente producidos en la región. El uso de robótica suave en la réplica del movimiento humano ha venido en crecimiento, por lo que tomando en cuenta lo desarrollado por Deimel and Brock[3] y en el Laboratorio de Simulación y Bioingeniería(SIBILA) se pretende crear una órtesis de mano con robótica suave para un paciente con parálisis cerebral, el cual tiene problemas para asir objetos con su mano derecha. Se busca mejorar la fuerza de agarre y la posición de los dedos. El proyecto se divide en 3 partes: desarrollo de una réplica de la mano del paciente, elaboración de los músculos neumáticos y elaboración de un sistema de control por voz. El resultado es una órtesis de actuación neumática que permite asir objetos de distintas formas y brinda fuerza a los dedos, además la órtesis se ajusta al modelo de la mano del paciente. Se recomienda utilizar un silicon e hilo más resistente, además de realizar pruebas con el guante colocado sobre la réplica de la mano.

Introduction

In Costa Rica, there has been an increasing need for prosthetic and orthotic devices in the last years. In 2017 the CCSS (Caja Costarricense de Seguro Social) received 2254 request for prostheses and made 890 of these, this need has increased 2.7 times in the last 6 years [2]. The need of orthotic devices is also expected to grow in the coming years. For example, current research establishes that up to 74% of stroke patients need of long-term assistance [8]. 75% of patients with stroke are people over 65 years old, a population that has been growing at an accelerated rate the last few years. In Costa Rica in 2015, 7.36% of the population was

composed of people over 65 years old [6], and it is expected to triple by 2040 [5]. This population is expected to keep growing in the following years which means an increase in prospective users of orthotic devices. This, devices are however rarely made in Costa Rica (in the case of hand orthosis). Other populations that are prospective users of these devices are the people with cerebral palsy and diabetic people. In 2017, around 9.6% of the population older than 20 years old in Costa Rica had diabetes, this percentage is expected to grow to around 12.6% in 2045 [4]. 40% of these patients tend to have problems with their hand movements [7].

The use of soft robotics has been growing as a medium for replicating natural movements. For example, Deimel and Brock[3] have developed a soft actuated hand that is able to grasp successfully objects with different shapes. At the Simulation Bioengineering Laboratory (SIBILA) 2 previous attempts have been made to develop a soft orthosis. Of this 2, an interesting first approach was made by Arrieta[1] who made an initial attempt to develop a soft orthosis actuated with pneumatic muscles. In the current project, we attempt to use the pneumatic muscle design made by Deimel and Brock[3] to develop a soft hand orthosis, taking into account the results of the 2 previous iterations done at the laboratory.

This project in particular was developed for a patient with cerebral palsy. This patient has problems grasping objects with his right hand, this is because he has difficulty holding his fingers in the adequate position. As a means to help him, it is born the idea of designing an orthosis that is able to provide his fingers with support by increasing the strength with which the patient is able to grab objects as well as helping him maintain the position of his fingers.

Materials and Methods

For the elaboration of this project is needed the following materials:

- Epoxy resin
- Torsion springs
- Silicone: Dragon Skin FX-Pro®
- Polyester thread
- Polyester fabric
- Raspberry PI 3B+
- Sensors: Force Sensitive Resistors.
- USB Microphone

The elaboration of this project can be separated in 3 different parts:

First of all, a replica of the patient's hand was made using epoxy resin in order to be able to verify that the orthosis complies with the geometry of the hand of the patient. This model is articulated with torsion springs and it is able to flex to the same angles that the patient uses for grabbing objects. Said replica can be seen in figure 1.

The second part was the elaboration of the actuators used to mobilize the fingers of the patient. This actuators need to be made of a soft material that can be in contact with skin. Said material also has to be able to provide the necessary strength for the orthosis and have a good lifespan. This is why it was decided to use pneumatic muscles that are made of silicone. This material plus the design applied to the muscles allow them be soft to the contact with skin as well as having the necessary rigidity to hold down the fingers of the patient. This muscles are made by using molds and a cape of silicone spread over polyester screen. Once the muscles have hardened, a reinforcement helix made of sewing thread is wrapped around them, this allows the

muscles to have the appropriate movement. The muscles are inflated using a 12V air pump. The silicone used for the design of the muscles was Dragon Skin FX-Pro®. Other types of silicone were also tested. The model used for the design of this muscles was proposed by Deimel and Brock (2016) and the final result for each silicone used can be seen in figure 2. Number 3 is the one made with Dragon Skin FX-Pro®.



Figure 1. Articulated Hand Replica



Figure 2. Pneumatic muscles made with different silicones.

The third part was the elaboration of a force sensitive voice control system. A Raspberry PI 3B+ with Snowboy® was used for the control system. Snowboy® is a software with libraries that can be used with Python® and it is designed to allow the use of voice control via the use of Hotwords. For this project the commands that were used were “Ortesis Abre” y “ Ortesis Cierra” (Spanish for “Orthosis, open” and “Orthosis, close” respectively). This commands were used to indicate the controller which function to carry out. Also, to control the force of grasping of the orthosis, the sensors used were Force Sensitive Resistors. This sensors were used to feedback the Raspberry PI with the grasping force values that each finger is making so that it is able to know when the object is being successfully grabbed. Tests with volunteers were done to know the necessary force values for the successful grabbing of different objects.

Results

The result of this project is a soft orthosis that is able to grab objects with different forms and hold them up in the air. Said orthosis was capable of grabbing bottles, cans and rectangular

juice boxes. Also, it was observed that the orthosis inputs strength mainly in the thumb but also in all the other fingers. It was also verified that the model of the hand of the patient can fit in the orthosis. The results can be seen in figure. 3.



Figure 3. Hand Orthosis Grabbing a Bottle

Recommendations

For the next steps of this project it is recommended to optimize the design of the muscles with a silicone that is more resistant and a thicker reinforcement helix. Also, it is recommended to start carrying out the tests of the orthosis with the model of the hand inside it. Another important next step is the implementation of a flexible thumb, this is because the one used right now only has one direction of movement. The implementation of a flexible thumb will be an important step in achieving that the orthosis is able to grab a bigger amount of objects with different forms.

References

- [1] Arrieta, S. Propuesta de diseño de un primer prototipo de ortesis para paciente con parálisis cerebral: diseño neumático.2018
- [2] Coto, D. Producción de prótesis de la CCSS creció 2.7 veces en los últimos cinco años. 2019
- [3] R. Deimel and O. Brock, "A novel type of compliant and underactuated robotic hand for dexterous grasping," *The International Journal of Robotics Research*, vol. 35, (1-3), pp. 161-185, 2016.
- [4] International Diabetes Federation . (2017). IDF diabetes atlas eighth edition 2017.
- [5] Instituto Nacional de Estadística y Censos. La población adulta mayor se triplicaría en los próximos 40 años. 2015
- [6] Junta de Pensiones y Jubilaciones del Magisterio Nacional. Informe azul del régimen de capitalización colectiva. 2016
- [7] Servicio de Cirugía Ortopédica y Traumatología, Hospital de la Santa Cruz y San Pablo, Universidad Autónoma de Barcelona. La mano diabética. 2015
- [8] Vinstrup, J., Calatayud, J., Jakobsen, M. D., Sundstrup, E., Jørgensen, J. R., Casaña, J., & Andersen, L. L. Hand strengthening exercises in chronic stroke patients: Dose-response evaluation using electromyography. *Journal of Hand Therapy*, 31(1), 111-121. 2018.



Optimización de datos en sistemas de monitoreo hídrico de nitratos

Data optimization in nitrate water monitoring systems

Laura Hernández-Alpizar¹, Arys Carrasquilla-Batista²,
Lilliana Sancho-Chavarría³

Hernández-Alpizar, L; Carrasquilla-Batista, A; Sancho-Chavarría, L. Optimización de datos en sistemas de monitoreo hídrico de nitratos. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 106-111.

 <https://doi.org/10.18845/tm.v33i5.5086>

1 Instituto Tecnológico de Costa Rica, Escuela de Química. Costa Rica.
Correo electrónico: lahernandez@tec.ac.cr.

2 Instituto Tecnológico de Costa Rica, Ingeniería Mecatrónica. Costa Rica.
Correo electrónico: acarrasquilla@tec.ac.cr.

3 Ingeniera en Computación, Máster en Computación, Doctora en Ingeniería. Instituto Tecnológico de Costa Rica, Ingeniería en Computación. Costa Rica.
Correo electrónico: lsancho@tec.ac.cr.



Palabras clave

Disparador de frecuencia; Internet de las Cosas; monitoreo de nitratos.

Resumen

Actividades antropogénicas, tal como la fertilización intensiva, generan un aumento en la concentración de nitratos de los sistemas hídricos que puede provocar contaminación en aguas de consumo humano y eutrofización en aguas superficiales. El análisis de muestras discretas revela diferencias espaciales en la concentración, pero un análisis continuo proporciona más información acerca del origen, la dinámica hidrológica, el transporte y procesamiento de los nitratos. Sin embargo, la frecuencia, el periodo y la calidad de los datos deben ser optimizados de acuerdo con el objetivo de investigación, ya que, un monitoreo continuo implica un alto consumo instrumental y una gran generación de datos que puede no aportar información relevante para el objetivo. La espectroscopía UV con análisis en flujo continuo es una técnica que de forma directa cuantifica la concentración de nitratos y se adapta bien a un monitoreo de alta resolución. En este trabajo se plantea el diseño de un sistema que utiliza este tipo de análisis acoplado a un sensor de conductividad, el cual se utiliza como disparador de frecuencia de muestreo. Además, se implementa el uso de Internet de las Cosas (IoT) tanto para realizar procesos de configuración en la toma de datos como para el accionamiento electromecánico remoto lo que permite un ajuste manual o automático en la obtención de datos y consecuentemente, de la información temporal y espacial requerida para el estudio de nitratos en el recurso hídrico.

Keywords

Frequency triggers; Internet of Things; nitrates monitoring.

Abstract

Anthropogenic activities, such as intensive fertilization, generate an increase in the concentration of nitrates in water systems that can cause contamination in waters for human consumption and eutrophication in surface waters. Discrete sample analysis reveals spatial differences in concentration, although continuous analysis provides more information about the origin, hydrological dynamics, transport, and nitrates bioprocessing. Nevertheless, the frequency, the period and the data quality must be optimized according to the research objective, since continuous monitoring implies high instrumental consumption and a large generation of data that may not provide relevant information for the objective. UV spectroscopy with continuous flow analysis is a technique that directly quantifies nitrate concentration and is well suited to high resolution monitoring. In this work, the design of a system that uses this type of analysis is proposed coupled to a conductivity sensor, as a trigger for the sampling frequency. Furthermore, the use of the Internet of Things (IoT) is implemented both to carry out configuration processes in data collection and for remote electromechanical actuation, which allows manual or automatic adjustment in obtaining data and, consequently, the information temporal and spatial required for the study of nitrates in the water resource.

Introducción

El exceso de nitratos que entra a los sistemas hídricos es un problema creciente de contaminación de aguas de consumo humano y también de eutrofización de aguas superficiales a nivel global y su principal origen se encuentra en actividades antropogénicas tales como la



fertilización intensiva de cultivos o la ganadería [1]–[3]. Estas actividades son de manejo, intensidad e impacto variable y una toma poco frecuente de muestras (por ej. mensual) puede que brinde información de la distribución espacial de la contaminación, particularmente si es durante un largo periodo [4], pero no es útil para comprender los procesos dinámicos de origen, transporte y procesamiento de los nitratos, lo cual es un conocimiento necesario para la toma de decisiones y acciones de mitigación. Para estos fines, es más apropiado un sistema de monitoreo de alta frecuencia, el cual puede dar información de tendencia o líneas base, distinguir impactos, efectos o cambios ya sea en corrientes de agua superficiales, océanos, embalses hidroeléctricos o acuíferos [5], [6]. Aunque la información requerida para alcanzar un objetivo de investigación también puede provenir de un equilibrio entre el periodo de muestreo y la frecuencia de toma de muestras u obtención de datos [7].

En la actualidad hay sistemas de análisis continuo de nitratos de alta precisión, especificidad que permiten la obtención de datos con una frecuencia ideal para la construcción de sistemas de monitoreo continuo, estos son los sistemas de flujo continuo de la muestra o de sondas sumergibles con detección por espectroscopia UV [8]–[12]. Pero la recolección de una gran densidad de datos tiene un costo computacional e instrumental elevado y no necesariamente aporta información [13], [14]. Además, los sistemas continuos pueden introducir nuevas fuentes de error analítico que afectan la calidad de los datos [6], [15]. De manera que plantear una estrategia instrumental y computacional para el ajuste de la obtención de datos es un requerimiento en el diseño de sistemas continuos para análisis cuantitativo, en este caso, del agua [11].

Existen diversas formas de optimizar el volumen y calidad de datos que se obtienen de un monitoreo continuo de nitratos. Una estrategia para ajustar la frecuencia de muestreo de un analito, de forma adaptativa, es utilizando disparadores de frecuencia [13]. Esta es posible de implementar si existe una propiedad u otro analito medible que se relaciona de alguna manera y con una varianza similar al analito de interés. Por ejemplo, la precipitación, la turbulencia, el nivel de agua o la conductividad eléctrica (CE) pueden relacionarse con variaciones de concentración de nitratos [13]. También se puede efectuar un tratamiento y reducción estadística de datos [7], utilizar información científica previa acerca del comportamiento del analito de interés [5], [6] u observar gráficamente cambios repentinos o tendencias que indiquen hacer ajustes manuales o automáticos, e idealmente de forma remota puesto que el monitoreo continuo debe tener un máximo de autonomía [16].

En este trabajo se consideran las diversas opciones antes mencionadas para la evaluación de nitratos en el recurso hídrico con un sistema de monitoreo continuo y se describe el uso de dos innovaciones desarrolladas en el Instituto Tecnológico de Costa Rica (TEC), integradas en una propuesta de diseño instrumental que permite el monitoreo de nitratos con una frecuencia ajustable de muestreo y de la obtención de datos.

Método

Diseño del sistema de muestreo y análisis

En investigaciones previas se desarrolló una innovación en la metodología para la calibración y medición en línea de nitratos en presencia de interferencias [17, 18]. Para la ejecución del este método se requiere efectuar operaciones de muestreo, filtración, calibración, medición y limpieza. Estas operaciones se esquematizan para su inclusión en accionamientos remotos de válvulas y actuadores.

El sistema incluye el desarrollo de un software para el procesamiento de datos, calibración y medición con posibilidades de almacenamiento remoto de los datos, denominado MOLABS. EL

espectrómetro puede generar una ráfaga de datos por milisegundo, en unidades de absorbancia (U.A.) entre las longitudes de onda 200.00-400.00 nm (± 0.10). Sin embargo, el recorrido de una completamente nueva toma de muestra hasta el detector puede tardar entre 30 s y 1 min, por lo que se toma un minuto como el mínimo de frecuencia entre la obtención de datos.

En otra investigación desarrollada en el Instituto Tecnológico de Costa Rica (TEC), se diseñó un sensor de conductividad que se encuentra en proceso de patente [19] y tiene la particularidad de que puede permanecer sumergido por largos periodos de tiempo sin necesidad de mantenimiento o calibración, una característica deseable para un sistema de monitoreo continuo. Para el ajuste de la frecuencia de muestro se conecta este sensor de conductividad al sistema y, para optimizar aún más la generación de datos se utiliza la capacidad del software de promediar las señales de absorbancia en un tiempo determinado por el usuario.

Almacenamiento de datos y IoT

El servicio de almacenamiento datos debe cumplir requerimientos mínimos de disponibilidad (de al menos 95%), seguridad, facilidad de integración a distintas plataformas, capacidad de almacenamiento flexible y autoajutable, facilidad de interconexión a dispositivos COTS (Circuits-off-the-shelf, por sus siglas en inglés), acceso a direcciones IP elásticas para el manejo de dinámico de las aplicaciones en la nube y por último, que las instancias del servidor sea compatible con los sistemas operativos comerciales y abiertos. Estas capacidades se encuentran en diversos servicios comerciales.

En el diseño final del sistema se utiliza IoT para el sistema de control, almacenamiento y acceso remoto de datos. Este sistema ya ha sido implementado en aplicaciones de medición de variables ambientales [20].

Resultados

En la figura 1 se muestra un esquema conceptual del manejo del sistema para monitoreo de nitratos MOLABS por medio de IoT.

En el bloque de control, los datos son adquiridos y almacenados localmente en una plataforma IoT, tipo Raspberry pi3, con conectividad a un servidor WEB mediante telefonía celular. La configuración permite el control local y remoto del sistema de muestreo y análisis. Los datos almacenados y disponibles en la nube pueden ser accedidos tanto en una aplicación WEB como con software de funcionamiento local, con el cual se puede efectuar procesamiento y visualización gráfica de los datos, evaluación de la calibración analítica, tendencia de los resultados o cambios. Con la información generada se puede cambiar frecuencia de muestreo y realizar operaciones instrumentales. Sin embargo, es necesario incluir criterios estadísticos de valoración inmediata de desviaciones, calidad, varianza para el ajuste automático de frecuencia y obtención de datos.

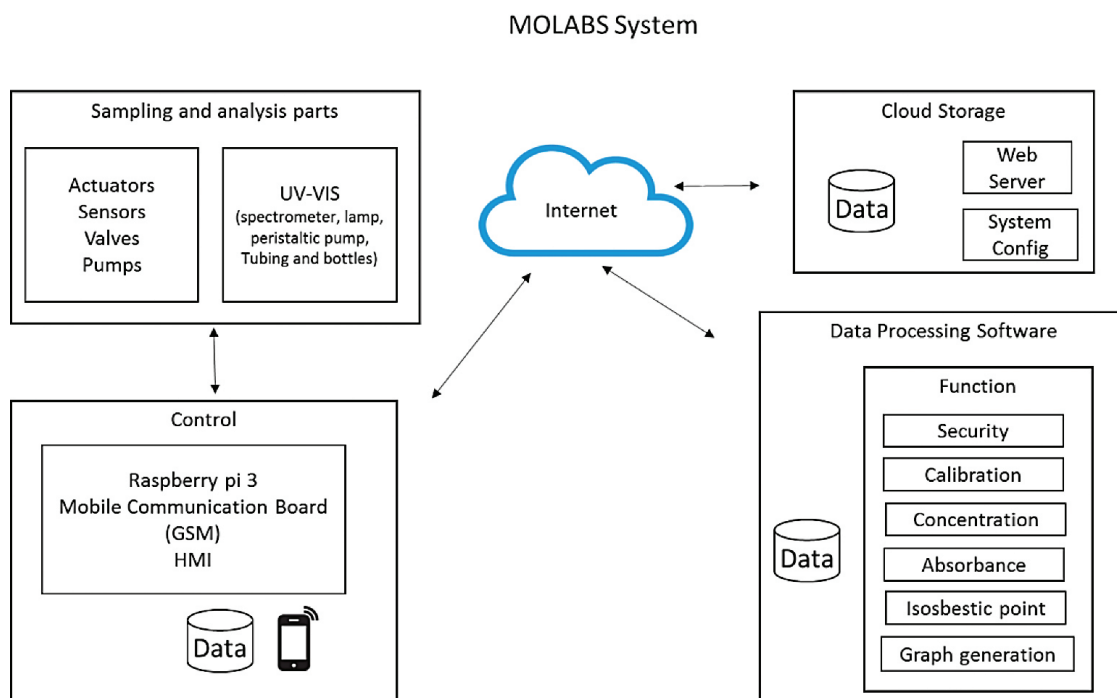


Figura 1. Sistema de monitoreo remoto de nitratos

Conclusión

La optimización del volumen de datos es una parte integral e importante del diseño para el sistema de monitoreo hídrico de nitratos propuesto, ya que permite cambiar su resolución en función del objetivo de medición, el aseguramiento de la calidad y representatividad de los datos en función de la variación. Las características instrumentales del módulo de análisis, el disparador de muestreo, así como la gestión y control del sistema por medio de IoT se eligieron considerando el potencial de autonomía y automatización del sistema. Se recomienda incluir en las próximas fases del diseño instrumental criterios estadísticos de evaluación de calidad y pertinencia científicas de los datos. Adicionalmente, se planifica la inclusión en el software de la estimación de las capacidades instrumentales tales como energía, duración de la intensidad de la lámpara, duración de las acciones de mantenimiento y limpieza en función de la duración del periodo de muestreo y su frecuencia así como el control de calidad de la medición y el funcionamiento del sistema mediante la introducción periódica de una muestra de validación.

Referencias

- [1] D. E. Canfield, A. N. Glazer, and P. G. Falkowski, "REVIEW The Evolution and Future of Earth's Nitrogen Cycle," *Science* (80-.), vol. 330, pp. 192–196, 2010.
- [2] J. W. Erisman, A. Bleeker, J. Galloway, and M. S. Sutton, "Reduced nitrogen in ecology and the environment," *Environ. Pollut.*, vol. 150, no. 1, pp. 140–149, 2007.

- [3] D. Fowler et al., "The global nitrogen cycle in the twenty-first century The global nitrogen cycle in the twenty-first century," 2013.
- [4] T. P. Burt, N. J. K. Howden, F. Worrall, and M. J. Whelan, "Long-term monitoring of river water nitrate : how much data do we need ?", *J. Environ. Monit.*, pp. 71–79, 2010.
- [5] M. P. Miller et al., "Quantifying watershed-scale groundwater loading and in-stream fate of nitrate using high-frequency water quality data," *Water Resour. Res.*, 52, pp. 330-347, 2016.
- [6] B. A. Pellerin et al., "Mississippi River Nitrate Loads from High Frequency Sensor Measurements and Regression-Based Load Estimation," *Environmental Science & Technology*, 48(21), pp. 12612–12619, 2014.
- [7] Y. Guo, M. Markus, and M. Demissie, "Uncertainty of nitrate-N load computations for agricultural watersheds," *Water Resources Research*, vol. 38, no. 10, 2002.
- [8] V. Cerdá, J. Avivar, and A. Cerdá, "Laboratory automation based on flow techniques," *Pure Appl. Chem.*, vol. 84, no. 10, pp. 1983–1998, 2012.
- [9] M. Trojanowicz and K. Kołacińska, "Recent advances in flow injection analysis," *Analyst*, vol. 141, no. 7, pp. 2085–2139, 2016.
- [10] Y. C. Moo, M. Z. Matjafri, H. S. Lim, and C. H. Tan, "New development of optical fibre sensor for determination of nitrate and nitrite in water," *Optik (Stuttg)*., vol. 127, no. 3, pp. 1312–1319, 2016.
- [11] T.M. Mathany, J. F. Saraceno, and J. T. Kulongoski "Guidelines and Standard Procedures for High-Frequency Groundwater-Quality Monitoring Stations — Design, Operation, and Record Computation Techniques and Methods 1 – D7." USGS. 2019.
- [12] B. A. Pellerin, B. A. Bergamaschi, B. D. Downing, J. F. Saraceno, J. D. Garrett, and L. D. Olsen. Chapter "Optical Techniques for the Determination of Nitrate in Environmental Waters : Guidelines for Instrument Selection , Operation , Deployment , Maintenance , Quality Assurance , and Data Reporting." USGS, 2013.
- [13] P. J. Blaen, K. Khamis, C. E. M. Lloyd, C. Bradley, D. Hannah, and S. Krause, "Real-time monitoring of nutrients and dissolved organic matter in rivers: Capturing event dynamics, technological opportunities and future directions," *Sci. Total Environ.*, vol. 569–570, pp. 647–660, 2016.
- [14] A. J. Horowitz, "A Review of Selected Inorganic Surface Water Quality-Monitoring Practices : Are We Really Measuring What We Think , and If So , Are We Doing It Right ?", 2013.
- [15] R. D. Harmel, R. J. Cooper, R. M. Slade, R. L. Haney, and J. G. Arnold, "Cumulative uncertainty in measured streamflow and water quality data for small watersheds". *Transactions of the ASABE*, vol. 49, no. 3, pp. 689–701, 2006.
- [16] ASTM. D3864 – 12, "Standard Guide for On-Line Monitoring Systems for Water Analysis 1," pp. 1–14, 2014.
- [17] L. Hernández-Alpizar and R. Coy-Herrera, "Cuantificación de nitratos en agua potable para análisis en línea" *Tecnología en Marcha*. vol. 28, pp. 86–93, 2015.
- [18] L. Hernández-alpizar and R. Coy-herrera. Dispositivo y método de calibración interpolativo en análisis cuantitativo de flujo continuo. 2019. CR20180476 (A). https://worldwide.espacenet.com/searchResults?submitted=true&locale=en_EP&DB=EPODOC&ST=singleline&query=CR20180476&Submit=Search
- [19] A. Carrasquilla-Batista. "Sensor de conductividad resistente a medios acuosos altamente salinos," noviembre 2017, CR 20170516 (A), Patente pendiente. [Online]. https://worldwide.espacenet.com/searchResults?AB=&AP=&CPC=&DB=EPODOC&IC=&IN=&PA=&PD=&PN=CR20170516A&PR=&ST=advanced&TI=&bclId=1&locale=en_EP&page=0&return=true
- [20] A. Carrasquilla-Batista, A. Chacón-Rodríguez, M. Solorzano-Quintana. Internet of Things for the Global Community (IoTGC-2017), "IoT applications : on the path of Costa Rica ' s commitment to becoming carbon-neutral," Portugal, 2017.

Advanced Computing National Collaboratory HPC Infrastructure, Kabré

Infraestructura de HPC del Colaboratorio Nacional de Computación Avanzada, Kabré

Isaac Eduardo Gómez-Sánchez¹, Jean Carlo Umaña-Jiménez²,
Melissa Arce-Montero³

Gómez-Sánchez, I; Umaña-Jiménez, J; Arce-Montero, M.
Advanced Computing National Collaboratory HPC Infraestructure, Kabré. *Tecnología en Marcha*. Edición especial 2020.
6th Latin America High Performance Computing Conference (CARLA). Pág 112-117.

 <https://doi.org/10.18845/tm.v33i5.5087>

1 Electrical engineering student, Costa Rica National Center for High Technology. Costa Rica. E-mail: isaac.gomez@ucr.ac.cr.

 <https://orcid.org/0000-0001-6991-2460>

2 Computer engineer, Costa Rica National Center for High Technology. Costa Rica. E-mail: jumana@cenat.ac.cr.

 <https://orcid.org/0000-0003-0857-6007>

3 Electrical engineering student, Costa Rica National Center for High Technology. Costa Rica. E-mail: melissa.arcemontero@ucr.ac.cr

 <https://orcid.org/0000-0001-7553-6369>



Keywords

Cluster; high performance computing; parallel computing.

Abstract

Kabré supercomputer is a High Performance Computing (HPC) cluster that provides tools that are needed to execute specific projects. In this paper, we present the infrastructure of such supercomputer settled at Advanced Computing National Collaboratory (CNCA), how its structure is defined by user needs and some recommendations in order to enhance the use and development of an HPC cluster infrastructure for scientific research purposes.

Palabras clave

Cluster; computación de alto rendimiento; computación paralela.

Resumen

La supercomputadora Kabré es un clúster de computación de alto rendimiento (HPC) que proporciona las herramientas necesarias para ejecutar proyectos específicos. En este documento, presentamos la infraestructura de dicha supercomputadora establecida en Colaboratorio Nacional de Computación Avanzada (CNCA), cómo se define su estructura según las necesidades del usuario y algunas recomendaciones para mejorar el uso y el desarrollo de una infraestructura de clúster HPC para fines de investigación científica.

Introduction

In science, throughout history, existed many problems which meant a significant and repetitive work in order to prove a theorem, validate a method, or simply to achieve a specific result. With the birth of HPC, many research centers saw potential in parallelization as a tool that speedup repetitive, mechanical and hard to compute tasks. Examples of these are presented in [9] with the use of TensorFlow for image recognition, [8] with a comparison of three different genome assemblers (Velvet, ABySS and SOAPdenovo) implementing de Bruijn graphs, [7] with an analysis and design of an LP and tremor location application, based on amplitude decay for seismic studies, and many others. These efforts wouldn't be more than that without a computational infrastructure that allows such power. The setup of a cluster infrastructure is required in order to solve such issues. For instance [1], [5] and [11] present an infrastructure focused on solving this requirement. But, as said in [11], such a cluster is necessarily defined by the users requirements, as defining it without demand, could be an expensive way of planning.

In Costa Rica, those efforts are not left behind. In Costa Rica National Center for High Technology (CeNAT) a computational cluster is in constant development. In this paper, we show how the users requirements shaped such infrastructure, its hardware constitution to fulfill those requirements and some tools that ease the administration, development and use of this one.

Infrastructure

Kabré supercomputer, cluster of CeNAT, is an HPC platform created to provide users with tools for data analysis, modeling and visualization. It has a lot of compute servers/nodes networked together that work in parallel raising the processing speed to give high performance computing.

In table 1 can be seen the current devices that conforms Kabré. Such devices has four main areas of specialization: simulation, machine learning, big data and bioinformatics. Use examples in these areas are presented in [2], [3], [4], [6], [7] and [8]. Below is shown a brief history of the origin of each area, as well as the hardware and software infrastructure that conforms Kabré, and some of its respective trends.

Hardware infrastructure

From the beginning, a focus on simulation to be the main area was sought. Xeon Phi were used, which has a large number of cores that are not very strong, but are fast enough to facilitate the execution of parallel projects. Currently, it is the strongest area of the cluster with 32 nodes. Over time, after listening to users and knowing which projects were they working on, different aspects were polished to create an infrastructure that adapts to their needs. That's why the second area arose, machine learning. There was a need of buying equipment with graphics cards capable of executing signal analysis projects. The graphic cards used are the NVIDIA Tesla K40 and the V100. These facilitated the execution of machine learning and deep learning programs. Also new projects arose with the need to run Big Data projects, therefore Data Science appeared as a new area. It works with more data and information, so another infrastructure was built with the necessary characteristics to facilitate the execution of these projects.

Because of its current faculties, the cluster is used by many users from different science areas. Currently it has at least 266 users. The main areas in which the cluster is used are presented in table 1. Most of the users that are presented in the table come from the public universities of Costa Rica, associated to Consejo Nacional de Rectores (CONARE). With such universities as well as with other institutions like Universidad de la República in Uruguay, or University of Chicago, there are about 270 different institutions around the world that uses this cluster. Also, Kabré is connected to RedCLARA, which interconnect many institutions along the globe with a 2.5Gbps link.

Software infrastructure

All these hardware would be useless without a software platform that ease the user experience and allow the cluster infrastructure at high level. For this reason, Kabré use a software infrastructure that optimize, and enables the use of each of the aforementioned areas of specialization. The main software platform used is Slurm. As said in [10]: "Slurm is an open source, fault-tolerant, and highly scalable cluster management and job scheduling system for large and small Linux clusters". Something that experience showed is the necessity of managing the resources available in a fair and smart way. With this tool, users jobs can be scheduled using weight paradigms, setup a maximum time for each job and, of course, distribute the resources among the users.

Other tools used in Kabré are Spark and Hadoop. These two platforms are frameworks which ease the big data work done with the cluster. The Spark platform can be used with Scala, Python, Java, R programming languages. The main difference among these is the algorithms for processing data: while Hadoop use MapReduce, Spark implements its own way of processing.

Trends

New necessities appeared with time. One of them is to increase the cluster infrastructure. It's expected to buy servers for the most used and growing areas: 1 node for Bioinformatics, 1 for Big Data and 2 for Machine Learning. Technical specifications about Bioinformatics node are brought in table 1, while the extra nodes of the other 2 areas are intended to keep with similar technical specifications than originals. Also, with the desire of ease the management of all the infrastructure, it's expected to implement new software tools such as CHEF to manage server

configuration details and Spack for package management. In order to improve performance and ease the cluster management, new methods, tools and equipment are constantly researched and implemented.

Recommendations

When talking about the creation of a data center, there are many variables to take into account, and many times these are ignored. For example, the conditions of the room where the equipment will be housed must have adequate space to be able to place the racks with their respective servers and leaving space to work properly. Another very important factor is the electrical connection, making sure it's enough to support the load of all the equipment to be used. In addition to this, is required to have adequate cooling equipment to prevent the equipment from overheating and reduce its life.

Also, when a data center is starting and does not have much money, is good to avoid making very risky bets such as buying some very specific technologies or betting on not so common trends. It would be best to buy something more standard that meets the goal for which the infrastructure is created. As an example, Xeon Phi technology was purchased at CNCA. However, soon after, this was discontinued, which implies a problem if in the future it's desired to invest in the infrastructure that uses this technology.

Table 1. Technical specifications.

Purpose	RAM (GB)	Storage (TB)	Processor	Nodes	Science areas of use
Simulation	92	4.256	Xeon Phi 7210 @ 1.30GHz	32	Mathematical models, Molecular dynamics, Computational physics, Quantum physics, Condensed matter, Lattice Field Theories, HPC, Climate modeling, Air pollutants modeling, Meteorology and atmospheric physics, Computational Neuroscience and Simulations, Computational Seismology, Geomatics.
	94	6	Xeon Phi 7230F @ 1.30GHz		
Big Data	64	8	Xeon E5-2650 v4 @ 2.20GHz	4	Data Mining, Big Data, Statistical Analysis.
	64	1.2	Xeon E5-2650 v4 @ 2.20GHz		
	32	1.8	Xeon E5-2620 v4 @ 2.10GHz		
	62	11.018	Xeon E5-2650 v4 @ 2.20GHz		
Machine Learning	16	1	Xeon E3-1225 v5 @ 3.30GHz	4	Machine Learning, Deep Learning, Pattern Recognition, Computational Neuroscience and Simulations.
			Graphic Card: NVIDIA Tesla K40c		
Bioinformatics (coming soon)	1024	240	Xeon Gold 6154 @ 3.00GHz	1	Bioinformatics, Biocomputing, Genetics and Genomics, Microbiology, Biotechnology, Population's genetics.

Another good practice that must be carried out is that of a good network infrastructure design. The purchase of network switches, network interfaces and cables, as well as the way in which internal connections are defined (routers, gateways, etc.) must be optimized so that the speed of the connection, both inside and outside the system, is as quick as possible. A good practice would be to have a 10Gb Base T infrastructure as a minimum, to reduce latency.

In addition, within the good practices that it is advisable to have, is the use of failure response mechanisms. The use of RAIDs, from 1 onwards; make backups and allocate equipment only for that purpose; virtualization of the administration system, since it allows you to easily create snapshots and rollback; and keep track of the installed packages, as well as their installation process, in case it's required to reinstall them or install similar packages. These are some of the practices that can be implemented as an error containment mechanism. However, it is good to constantly improve this area and seek and implement new methodologies.

Finally, two elements that have been found important to implement are the modularization of user environments and infrastructure documentation. The first one is to allow loading the packages as modules so that each user has their personalized session with the packages and their respective versions according to their needs. And the second one is nothing more than the proper handling of updated and complete documentation. It has been found that the latter is of great importance, as this can speed up the process of debugging errors, as well as the introduction of new personnel to the support area.

Conclusion

The scientific task requires the power of HPC to solve problems that require great computational power. To solve this need, a computing infrastructure with sufficient power to carry out the tasks entrusted is essential. In the CNCA there is a computer platform (Kabré) in constant development capable of solving this need through the power of HPC. It has a hardware and software infrastructure that allows its users to develop scientific works using simulation tools, machine learning, big data, bioinformatics and storage.

Defining an HPC computing platform to be used in scientific research without proper knowledge of what the user's needs will be is complicated, since it's possible to make rash decisions that result in the acquisition of equipment that quickly becomes obsolete. For this reason, it is always advisable to study the technology market and choose equipment with good support, as well as one that adapts to the user's requirements. Is also required to implement proper methods of failure containment, documentation and design of the workspace for the proper function of an HPC infrastructure purposed for scientific research.

References

- [1] A. Adakin et. al., "Building a high performance computing infrastructure for novosibirsk scientific center," *Journal of Physics: Conference Series* (Vol. 331, No. 5, p. 052020), IOP Publishing, 2011.
- [2] D. Alvarado, A. de la Osa & M. Frutos, "Parallelization of a Magnetohydrodynamics Model for Plasma Simulation," *IEEE International Work Conference on Bioinspired Intelligence (IWOB)*, 2018.
- [3] S. Calderón, J. Castro & M. Zumbado, "DNLM-MA-P: A Parallelization of the Deceived Non Local Means Filter with Moving Average and Symmetric Weighting," *IEEE International Work Conference on Bioinspired Intelligence (IWOB)*, 2018.
- [4] M. Calvo, D. Jiménez & E. Meneses, "Analyzing Communication Features and Community Structure of HPC Applications," *IEEE International Work Conference on Bioinspired Intelligence (IWOB)*, 2018.
- [5] C. Carley, L. Sells, B. McKinney, C. Zhao & H. Neeman, "Using a shared, remote cluster for teaching HPC," *IEEE International Conference on Cluster Computing (CLUSTER)*, 1-6, 2013.

- [6] J. Castro & E. Meneses, "Parallelization of a Denoising Algorithm for TonalBioacoustic Signals using OpenACC Directives," *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 2018.
- [7] G. Cornejo, L. Van der Laat, E. Meneses, J. Pacheco & M. Mora, "Using parallel computing for seismo-volcanic event location based on seismic amplitudes," *IEEE 38th Central America and Panama Convention (CONCAPAN XXXVIII)*, 2018.
- [8] C. Gamboa & E. Meneses, "Comparative Analysis of de Bruijn Graph Parallel Genome Assemblers," *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 2018.
- [9] G. Ramirez-Gargallo, M. Garcia-Gasulla, & F. Mantovani, "TensorFlow on State-of-the-Art HPC Clusters: A Machine Learning use Case," *2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, 526-533, 2019.
- [10] Slurm.com, "Quick Start User Guide", 2017. [Online] Available: <https://slurm.schedmd.com/quickstart.html>. [Accessed: 12-Aug- 2019].
- [11] S. Varrette, P. Bouvry, H. Cartiaux & F. Georgatos, "Management of an academic HPC cluster: The UL experience," *International Conference on High Performance Computing & Simulation (HPCS)*, 959-967, 2014.