

ACTAS

DE LAS

XXXVIII Jornadas de Automática

Gijón · Palacio de Congresos · 6, 7 y 8 de Septiembre de 2017



Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo



CEA
Comité Español
de Automática

Colabora

Gijón

Convention Bureau

Actas de

XXXVIII

Jornadas de Automática

© 2017 Universidad de Oviedo
© Los autores

Servicio de Publicaciones de la Universidad de Oviedo
Campus de Humanidades. Edificio de Servicios. 33011 Oviedo (Asturias)
Tel. 985 10 95 03 Fax 985 10 95 07
[http: www.uniovi.es/publicaciones](http://www.uniovi.es/publicaciones)
servipub@uniovi.es

DL AS 2749-2017

ISBN: 978-84-16664-74-0

Todos los derechos reservados. De conformidad con lo dispuesto en la legislación vigente, podrán ser castigados con penas de multa y privación de libertad quienes reproduzcan o plagien, en todo o en parte, una obra literaria, artística o científica, fijada en cualquier tipo y soporte, sin la preceptiva autorización.

Prefacio

Las *Jornadas de Automática* se celebran desde hace **40 años** en una universidad nacional facilitando el encuentro entre expertos en esta área en un foro que permite la puesta en común de las nuevas ideas y proyectos en desarrollo. Al mismo tiempo, propician la siempre necesaria colaboración entre investigadores del ámbito de la Ingeniería de Control y Automática, así como de campos afines, a la hora de abordar complejos proyectos de investigación multidisciplinares.

En esta ocasión, las Jornadas estarán organizadas por la Universidad de Oviedo y se han celebrado del 6 al 8 de septiembre de 2017 en el Palacio de Congresos de Gijón, colaborando tanto la Escuela Politécnica de Ingeniería de Gijón (EPI) como el Departamento de Ingeniería Eléctrica, Electrónica de Computadores y de Sistemas del que depende el Área de Ingeniería de Sistemas y Automática.

Además de las habituales actividades científicas y culturales, esta edición es muy especial al celebrarse el **50 aniversario de la creación de CEA**, Comité Español de Automática. Igualmente este año se conmemora el 60 aniversario de la Federación Internacional del Control Automático de la que depende CEA. Así se ha llevado a cabo la presentación del libro que se ha realizado bajo la coordinación de D. Sebastián Dormido, sobre la historia de la Automática en España en una sesión en la que han participado todos los ex-presidentes de CEA conjuntamente con el actual, D. Joseba Quevedo.

Igualmente hemos contado con la presencia de conferenciantes de prestigio para las sesiones plenarias, comunicaciones y ponencias orales en las reuniones de los 9 grupos temáticos, contribuciones en formato póster. Se ha celebrado también el concurso de CEABOT, así como una nueva Competición de Drones, con el ánimo de involucrar a más estudiantes de últimos cursos de Grado/Máster.

En el marco de las actividades culturales programadas se ha podido efectuar un recorrido en el casco antiguo situado en torno al Cerro de Santa Catalina y visitar la Laboral.

Gijón, septiembre de 2017

Hilario López
Presidente del Comité Organizador

Program Committee

Antonio Agudo	Institut de Robòtica i Informàtica Industrial
Rosa M Aguilar	University of La Laguna.
Luciano Alonso	University of Cantabria
Ignacio Álvarez García	Universidad de Oviedo
Antonio Javier Artuñedo García	Centre for Automation and Robotics (CSIC-UPM)
José M. Azorín	Miguel Hernandez University of Elche
Pedro Balaguer	Universitat Jaume I
Antonio Javier Barragán Piña	Universidad de Huelva
Alfonso Baños	Universidad de Murcia
Guillermo Bejarano	University of Seville
Gerardo Beruvides	Centro de Automática y Robótica
Carlos Bordons	University of Seville
Jose Manuel Bravo	University of Huelva
Jose Luis Calvo-Rolle	University of A Coruña
Fernando Castaño Romero	Centro de Automática y Robótica (UPM -CSIC)
José Luis Casteleiro-Roca	University of Coruña
Alvaro Castro-Gonzalez	Universidad Carlos III de Madrid
Ramon Costa-Castelló	Universitat Politècnica de Catalunya
Abel A. Cuadrado	University of Oviedo
Arturo De La Escalera	Universidad Carlos III de Madrid
Emma Delgado	Universidad de Vigo
Jose-Luis Diez	Universitat Politecnica de Valencia
Manuel Domínguez	Universidad de León
Juan Manuel Escaño	Universidad de Sevilla
Mario Francisco	University of Salamanca
Maria Jesus Fuente	Universidad de Valladolid
Juan Garrido	Universtiy of Cordoba
Antonio Giménez	Universidad de Almeria
Evelio Gonzalez	Universidad de La Laguna
José-Luis Guzmán	Universidad de Almería
Rodolfo Haber	Center for Automation and Robotics (UPM-CSIC)
César Ernesto Hernández	Universidad de Almería
Eloy Irigoyen	UPV/EHU
Agustin Jimenez	Universidad PolitÁcnica de Madrid
Emilio Jiménez	University of La Rioja
Jesus Lozano	Universidad de Extremadura
Jorge Luis Madrid	Centro de Automática y Robótica
Luis Magdalena	Universidad Politécnic de Madrid
David Martin Gomez	Universidad Carlos III de Madrid
Fernando Matia	Universidad Politecnica de Madrid
Joaquim Melendez	Universitat de Girona
Juan Mendez	Universidad de La Laguna
Luis Moreno	Universidad Carlos III de Madrid
María Dolores Moreno Rabel	Universidad de Extremadura
David Muñoz	Universidad de Sevilla
Antonio José Muñoz-Ramirez	Universidad de Málaga
Jose Luis Navarro	Universidad Politecnica de Valencia
Manuel G. Ortega	University of Seville
Andrzej Pawlowski	UNED
Mercedes Perez de La Parte	University of La Rioja
Ignacio Peñarrocha	Universitat Jaume I de Castelló, Spain
José Luis Pitarch	Universidad de Valladolid

Daniel Pérez	University of Oviedo
Emilio Pérez	Universitat Jaume I
Juan Pérez Oria	Universidad de Cantabria
Miguel Ángel Ridao	Universidad de Sevilla
Gregorio Sainz-Palmero	Universidad de Valladolid
Antonio Sala	Universitat Politecnica de Valencia
Ester Sales-Setién	Universitat Jaume I
Jose Sanchez	UNED
Javier Sanchis Saez	Universitat Politecnica de Valencia (UPV)
José Pedro Santos	ITEFI-CSIC
Matilde Santos	Universidad Complutense de Madrid
Alvaro Serna	University of Valladolid
José Enrique Simó	Universidad Politécnica de Valencia
José A. Somolinos	ETS I Navales. Universidad Politecnica de Madrid
Fernando Tadeo	Univ. of Valladolid
Alejandro Tapia	Universidad de Loyola Andalucía
David Tena	Universitat Jaume I
Jesús Torres	Universidad de La Laguna
Pedro M. Vallejo	Universidad de Salamanca
Guilherme Vianna	Universidad de Sevilla
Alejandro Vignoni	AI2 - UPV
Ramón Vilanova	UAB
Francisco Vázquez	Universidad de Cordoba
Jesús M. Zamarreño	University of Valladolid

Revisores Adicionales

Al-Kaff, Abdulla

Balbastre, Patricia
Beltrán de La Cita, Jorge
Bermudez-Cameo, Jesus
Blanco-Claraco, Jose-Luis
Blanes, Francisco
Bonin-Font, Francisco

Cancela, Brais

Ferraz, Luis

Garita, Cesar
Gimenez, Antonio
Gruber, Patrick
Guindel, Carlos

Hernandez Ruiz, Alejandro
Hernandez, Daniel

Jardón Huete, Alberto

López, Amable

Marin, Raul
Marín Plaza, Pablo
Mañanas, Miguel Angel
Morales, Rafael
Moreno, Francisco-Angel

Núñez, Luis Ramón

Ponz Vila, Aurelio
Posadas-Yague, Juan-Luis
Poza-Luján, Jose-Luis
Pumarola, Albert

Raya, Rafael
Revestido Herrero, Elías
Rocon, Eduardo
Ruiz Sarmiento, José Raúl
Ruiz, Adria

Torres, Jose Luis

Vaquero, Victor

Table of Contents

Ingeniería de Control	
<hr/>	
TÚNEL DE AGUA PARA PRUEBAS Y CARACTERIZACIÓN DE DISEÑOS EXPERIMENTALES DE TURBINAS HIDROCINÉTICAS	1
<i>Eduardo Alvarez, Manuel Rico-Secades, Antonio Javier Calleja Rodríguez, Joaquín Fernández Francos, Aitor Fernández Jiménez, Mario Alvarez Fernández and Samuel Camba Fernández</i>	
Reduction of population variability in protein expression: A control engineering approach.	8
<i>Yadira Boada, Alejandro Vignoni and Jesús Picó</i>	
CONTROL ROBUSTO DEL PH EN FOTOBIORREACTORES MEDIANTE RECHAZO ACTIVO DE PERTURBACIONES	16
<i>José Carreño, Jose Luis Guzman, José Carlos Moreno and Rodolfo Villamizar</i>	
Control reset para maniobra de cambio de carril y validación con CarSim	23
<i>Miguel Cerdeira, Pablo Falcón, Antonio Barreiro, Emma Delgado and Miguel Díaz-Cacho</i>	
Maniobra de aterrizaje automática de una Cessna 172P modelada en FlightGear y controlada desde un programa en C	31
<i>Mario de La Rosa, Antonio Javier Gallego and Eduardo Fernández</i>	
Alternativas para el control de la red eléctrica aislada en parques eólicos marinos	38
<i>Carlos Díaz-Sanahuja, Ignacio Peñarrocha, Ricardo Vidal-Albalade and Ester Sales-Setién</i>	
CONTROL PREDICTIVO DISTRIBUIDO UTILIZANDO MODELOS DIFUSOS PARA LA NEGOCIACIÓN ENTRE AGENTES	46
<i>Lucía Fargallo, Silvana Roxani Revollar Chavez, Mario Francisco, Pastora Vega and Antonio Cembellín</i>	
Control Predictivo en el espacio de estados de un captador solar tipo Fresnel	54
<i>Antonio Javier Gallego, Mario de La Rosa and Eduardo Fernández</i>	
Control predictivo para la operación eficiente de una planta formada por un sistema de desalación solar y un invernadero	62
<i>Juan Diego Gil Vergel, Lidia Roca, Manuel Berenguel, Alba Ruiz Aguirre, Guillermo Zaragoza and Antonio Giménez</i>	
Depuración de Aguas Residuales en la Industria 4.0	70
<i>Jesus Manuel Gomez-De-Gabriel, Ana María Jiménez Arévalo, Laura Eiroa Mateo and Fco. Javier Fernández-De-Cañete-Rodríguez</i>	
Control robusto con QFT del pH en un fotobioreactor raceway	77
<i>Ángeles Hoyo Sánchez, Jose Luis Guzman, Jose Carlos Moreno and Manuel Berenguel</i>	
Revisión sistemática de la literatura en ingeniería de sistemas. Caso práctico: técnicas de estimación distribuida de sistemas ciberfísicos	84
<i>Carmelina Ierardi, Luis Orihuela Espina, Isabel Jurado Flores, Álvaro Rodríguez Del Nozal and Alejandro Tapia Córdoba</i>	
Desarrollo de un Controlador Predictivo para Autómatas programables basado en la normativa IEC 61131-3	92
<i>Pablo Krupa, Daniel Limon and Teodoro Alamo</i>	
Diseño de un emulador de aerogenerador de velocidad variable DFIG y control de pitch ...	100
<i>Manuel Lara Ortiz, Juan Garrido Jurado and Francisco Vázquez Serrano</i>	

Observación de la fracción de agua líquida en pilas de combustible tipo PEM de cátodo abierto.....	108
<i>Julio Luna and Ramon Costa-Castelló</i>	
Control Predictivo Basado en Datos.....	115
<i>José María Manzano, Daniel Limón, Teodoro Álamo and Jan Peter Calliess</i>	
Control MPC basado en un modelo LTV para seguimiento de trayectoria con estabilidad garantizada.....	122
<i>Sara Mata, Asier Zubizarreta, Ione Nieva, Itziar Cabanes and Charles Pinto</i>	
Implementación y evaluación de controladores basados en eventos en la norma IEC-61499.	130
<i>Oscar Miguel-Escrig, Julio-Ariel Romero-Pérez and Esteban Querol-Dolz</i>	
AUTOMATIZACIÓN Y MONITORIZACIÓN DE UNA INSTALACIÓN DE ENSAYO DE MOTORES.....	138
<i>Alfonso Poncela Méndez, Miguel Ochoa Vega, Eduardo J. Moya de La Torre and F. Javier García Ruíz</i>	
OPTIMIZACIÓN Y CONTROL EN CASCADA DE TEMPERATURA DE RECINTO MEDIANTE SISTEMAS DE REFRIGERACIÓN.....	146
<i>David Rodríguez, José Enrique Alonso Alfaya, Guillermo Bejarano Pellicer and Manuel G. Ortega</i>	
Diseño LQ e implementación distribuida para la estimación de estado.....	154
<i>Álvaro Rodríguez Del Nozal, Luis Orihuela, Pablo Millán Gata, Carmelina Ierardi and Alejandro Tapia Córdoba</i>	
Estimación de fugas en un sistema industrial real mediante modelado por señales aditivas.	160
<i>Ester Sales-Setién, Ignacio Peñarrocha and David Tena</i>	
Advanced control based on MPC ideas for offshore hydrogen production.....	167
<i>Alvaro Serna, Fernando Tadeo and Julio. E Normey-Rico</i>	
Transfer function parameters estimation by symmetric send-on-delta sampling.....	174
<i>José Sánchez, María Guinaldo, Sebastián Dormido and Antonio Visioli</i>	
An Estimation Approach for Process Control based on Asymmetric Oscillations.....	181
<i>José Sánchez, María Guinaldo Losada, Sebastian Dormido, José Luis Fernández Marrón and Antonio Visioli</i>	
Robust PI controller for disturbance attenuation and its application for voltage regulation in islanded microgrid.....	189
<i>Ramon Vilanova, Carles Pedret and Orlando Arrieta</i>	
Infraestructura para explotación de datos de un simulador azucarero.....	197
<i>Jesús M. Zamarréño, Cristian Pablos, Alejandro Merino, L. Felipe Acebes and De Prada César</i>	
<hr/> Automar <hr/>	
INFRAESTRUCTURA PARA ESTUDIAR ADAPTABILIDAD Y TRANSPARENCIA EN EL CENTRO DE CONTROL VERSÁTIL.....	203
<i>Juan Antonio Bonache Seco, José Antonio Lopez Orozco, Eva Besada Portas and Jesús Manuel de La Cruz</i>	
ARQUITECTURA DE CONTROL HÍBRIDA PARA LA NAVEGACIÓN DE VEHÍCULOS SUBMARINOS NO TRIPULADOS.....	211
<i>Francisco J. Lastra, Jesús A. Trujillo, Francisco J. Velasco and Elías Revestido</i>	

Exploración y Reconstrucción 3D de Fondos Marinos Mediante AUVs y Sensores Acústicos	218
<i>Oscar L. Manrique Garcia, Mario Andrei Garzon Oviedo and Antonio Barrientos</i>	
AUTOMATIZACIÓN DE MANIOBRAS PARA UN TEC DE 2GdL	226
<i>Marina Pérez de La Portilla, José Andrés Somolinos Sánchez, Amable López Piñeiro, Rafael Morales Herrera and Eva Segura</i>	
MERBOTS PROJECT: OVERALL DESCRIPTION, MULTISENSORY AUTONOMOUS PERCEPTION AND GRASPING FOR UNDERWATER ROBOTICS INTERVENTIONS	232
<i>Pedro J. Sanz, Raul Marin, Antonio Peñalver, David Fornas and Diego Centelles</i>	
<hr/> Bioingeniería <hr/>	
MARCADORES CUADRADOS Y DEFORMACIÓN DE OBJETOS EN NAVEGACIÓN QUIRÚRGICA CON REALIDAD AUMENTADA	238
<i>Eliana Aguilar, Oscar Andres Vivas and Jose Maria Sabater-Navarro</i>	
Entrenamiento robótico de la marcha en pacientes con Parálisis Cerebral: definición de objetivos, propuesta de tratamiento e implementación clínica preliminar	244
<i>Cristina Bayón, Teresa Martín-Lorenzo, Beatriz Moral-Saiz, Óscar Ramírez, Álvaro Pérez-Somarriba, Sergio Lerma-Lara, Ignacio Martínez and Eduardo Rocon</i>	
PREDICCIÓN DE ACTIVIDADES DE LA VIDA DIARIA EN ENTORNOS INTELIGENTES PARA PERSONAS CON MOVILIDAD REDUCIDA	251
<i>Arturo Bertomeu-Motos, Santiago Ezquerro, Juan Antonio Barios, Luis Daniel Lledó, Francisco Javier Badesa and Nicolas Garcia-Aracil</i>	
Sistema de Visión Estereoscópico para el guiado de un Robot Quirúrgico en Operaciones de Cirugía Laparoscópica HALS.....	256
<i>Carlos Castedo Hernández, Rafael Estop Remacha, Eusebio de La Fuente López and Lidia Santos Del Blanco</i>	
Head movement assessment of cerebral palsy users with severe motor disorders when they control a computer thought eye movements.....	264
<i>Alejandro Clemotte, Miguel A. Velasco and Eduardo Rocon</i>	
Diseño de un sensor óptico de fuerza para exoesqueletos de mano.....	270
<i>Jorge Diez Pomares, Andrea Blanco Ivorra, José María Catalan Orts, Francisco Javier Badesa Clemente, José María Sabater and Nicolas Garcia Aracil</i>	
POSIBILIDADES DEL USO DE TRAMAS ARTIFICIALES DE IMAGEN MOTORA PARA UN BCI BASADO EN EEG	276
<i>Josep Dinarès-Ferran, Christoph Guger and Jordi Solé-Casals</i>	
EFFECTOS SOBRE LA ERD EN TAREAS DE CONTROL DE EXOESQUELETO DE MANO EMPLEANDO BCI.....	282
<i>Santiago Ezquerro, Juan Antonio Barios, Arturo Bertomeu-Motos, Luisa Lorente, Nuria Requena, Irene Delegido, Francisco Javier Badesa and Nicolas Garcia-Aracil</i>	
Formulación Topológica Adaptada para la Simulación y Control de Exoesqueletos Accionados con Transmisiones Harmonic Drive.....	288
<i>Andres Hidalgo Romero and Eduardo Rocon</i>	

Identificación de contracciones isométricas de la extremidad superior en pacientes con lesión medular incompleta mediante características espectrales de la electromiografía de alta densidad (HD-EMG)	296
<i>Mislav Jordanic, Mónica Rojas-Martínez, Joan Francesc Alonso, Carolina Migliorelli and Miguel Ángel Mañanas</i>	
Diseño de una plataforma para analizar el efecto de la estimulación mecánica aferente en el temblor de pacientes con temblor esencial	302
<i>Julio S. Lora, Roberto López, Jesús González de La Aleja and Eduardo Rocon</i>	
DEFINICIÓN DE UN PROTOCOLO PARA LA MEDIDA PRECISA DEL RANGO CERVICAL EMPLEANDO TECNOLOGÍA INERCIAL	308
<i>Álvaro Martín, Rafael Raya, Cristina Sánchez, Rodrigo Garcia-Carmona, Oscar Ramirez and Abraham Otero</i>	
SISTEMA BRAIN-COMPUTER INTEFACE DE NAVEGACIÓN WEB ORIENTADO A PERSONAS CON GRAVE DISCAPACIDAD.....	313
<i>Víctor Martínez-Cagigal, Javier Gómez-Pilar, Daniel Álvarez, Eduardo Santamaría-Vázquez and Roberto Hornero</i>	
ESTRATEGIAS DE NEUROESTIMULACIÓN TRANSCRANEAL POR CORRIENTE DIRECTA PARA MEJORA COGNITIVA	320
<i>Silvia Moreno Serrano, Mario Ortiz and José María Azorín Poveda</i>	
COMPARATIVA DE ALGORITMOS PARA LA DETECCIÓN ONLINE DE IMAGINACIÓN MOTORA DE LA MARCHA BASADO EN SEÑALES DE EEG	328
<i>Marisol Rodriguez-Ugarte, Irma Nayeli Angulo Sherman, Eduardo Iáñez and Jose M. Azorin</i>	
DETECCIÓN, MEDIANTE UN GUANTE SENSORIZADO, DE MOVIMIENTOS SELECCIONADOS EN UN SISTEMA ROBOTIZADO COLABORATIVO PARA HALS	334
<i>Lidia Santos, José Luis González, Eusebio de La Fuente, Juan Carlos Fraile and Javier Pérez Turiel</i>	
BIOSENSORES PARA CONTROL Y SEGUIMIENTO PATOLOGÍAS REUMATOIDES	340
<i>Amparo Tirado, Raúl Marín, José V Martí, Miguel Belmonte and Pedro Sanz</i>	
Assessment of tremor severity in patients with essential tremor using smartwatches	347
<i>Miguel A. Velasco, Roberto López-Blanco, Juan P. Romero, M. Dolores Del Castillo, J. Ignacio Serrano, Julián Benito-León and Eduardo Rocon</i>	
INTERFAZ CEREBRO-ORDENADOR PARA EL CONTROL DE UNA SILLA DE RUEDAS A TRAVÉS DE DOS PARADIGMAS DE NAVEGACIÓN	353
<i>Fernández-Rodríguez Álvaro, Velasco-Álvarez Francisco and Ricardo Ron-Angevin</i>	
<hr/> Control Inteligente <hr/>	
Aprendizaje por Refuerzo para sistemas lineales discretos con dinámica desconocida: Simulación y Aplicación a un Sistema Electromecánico	360
<i>Henry Diaz, Antonio Sala and Leopoldo Armesto</i>	
Diseño de sistemas de control en cascada clásico y borroso para el seguimiento de trayectorias	368
<i>Javier G. Gonzalez, Rodolfo Haber, Fernando Matia and Marcelino Novo</i>	

ANÁLISIS FORMAL DE LA DINÁMICA DE SISTEMAS NO LINEALES MEDIANTE REDES NEURONALES.....	376
<i>Eloy Irigoyen, Mikel Larrea, A. Javier Barragán, Miguel Ángel Martínez and José Manuel Andújar</i>	
Predicción de la energía renovable proveniente del oleaje en las islas de Fuerteventura y Lanzarote.	384
<i>G.Nicolás Marichal, Deivis Avila, Ángela Hernández, Isidro Padrón and José Ángel Rodríguez</i>	
Aplicación de Redes Neuronales para la Estimación de la Resistencia al Avance en Buques	393
<i>Daniel Marón Blanco and Matilde Santos</i>	
Novel Fuzzy Torque Vectoring Controller for Electric Vehicles with per-wheel Motors	401
<i>Alberto Parra, Martín Dendaluze, Asier Zubizarreta and Joshué Pérez</i>	
REPOSTAJE EN TIERRA DE UN AVIÓN MEDIANTE ALGORITMOS GENÉTICOS .	408
<i>Elías Plaza and Matilde Santos</i>	
VISUALIZACIÓN WEB INTERACTIVA PARA EL ANÁLISIS DEL CHATTER EN LAMINACIÓN EN FRÍO.....	416
<i>Daniel Pérez López, Abel Alberto Cuadrado Vega and Ignacio Díaz Blanco</i>	
BANCADA PARA ANÁLISIS INTELIGENTE DE DATOS EN MONITORIZACIÓN DE SALUD ESTRUCTURAL.....	424
<i>Daniel Pérez López, Diego García Pérez, Ignacio Díaz Blanco and Abel Alberto Cuadrado Vega</i>	
CONTROL DE UN VEHÍCULO CUATRIRROTOR BASADO EN REDES NEURONALES.....	431
<i>Jesus Enrique Sierra and Matilde Santos</i>	
CONTROL PREDICTIVO FUZZY CON APLICACIÓN A LA DEPURACIÓN BIOLÓGICA DE FANGOS ACTIVADOS.....	437
<i>Pedro M. Vallejo Llamas and Pastora Vega Cruz</i>	
<hr/> Educación en Automática <hr/>	
REFLEXIONES SOBRE EL VALOR DOCENTE DE UNA COMPETICION DE DRONES EN LA EDUCACIÓN PARA EL CONTROL.....	445
<i>Ignacio Díaz Blanco, Alvaro Escanciano Urigüen, Antonio Robles Alvarez and Hilario López García</i>	
Uso del Haptic Paddle con aprendizaje basado en proyectos	451
<i>Juan M. Gandarias, Antonio José Muñoz-Ramírez and Jesus Manuel Gomez-De-Gabriel</i>	
REPRESENTACION INTEGRADA DE ACCIONAMIENTOS MECANICOS Y CONTROL DE EJES ORIENTADA A LA COMUNICACIÓN Y DOCENCIA EN MECATRONICA	457
<i>Julio Garrido Campos, David Santos Esterán, Juan Sáez López and José Ignacio Armesto Quiroga</i>	
Construcción y modelado de un prototipo fan & plate para prácticas de control automático	465
<i>Cristina Lampon, Javier Martin, Ramon Costa-Castelló and Muppaneni Lokesh Chowdary</i>	

EDUCACION EN AUTOMATICA E INDUSTRIA 4.0 MEDIANTE LA APLICACIÓN DE TECNOLOGÍAS 3D	471
<i>Jose Ramon Llata, Esther Gonzalez-Sarabia, Carlos Torre-Ferrero and Ramon Sancibrian</i>	
Desarrollo e implementación de un sistema de control en una planta piloto hibrida.....	479
<i>Maria P. Marcos, Cesar de Prada and Jose Luis Pitarch</i>	
LA INFORMÁTICA INDUSTRIAL EN LAS INGENIERÍAS INDUSTRIALES	486
<i>Rogelio Mazaeda, Eusebio de La Fuente López, José Luis González, Eduardo J. Moya de La Torre, Miguel Angel García Blanco, Javier García Ruiz, María Jesús de La Fuente Aparicio, Gregorio Sainz Palmero and Smaranda Cristea</i>	
Ventajas docentes de un flotador magnético para la experimentación de técnicas control ..	495
<i>Eduardo Montijano, Carlos Bernal, Carlos Sagües, Antonio Bono and Jesús Sergio Artal</i>	
PROGRAMACIÓN ATRACTIVA DE PLC	502
<i>Eduardo J. Moya de La Torre, F. Javier García Ruíz, Alfonso Poncela Méndez and Victor Barrio Lángara</i>	
MODERNIZACIÓN DE EQUIPO FEEDBACK MS-150 PARA EL APRENDIZAJE ACTIVO EN INGENIERÍA DE CONTROL	510
<i>Perfecto Reguera Acevedo, Miguel Ángel Prada Medrano, Antonio Morán Álvarez, Juan José Fuertes Martínez, Manuel Domínguez González and Serafín Alonso Castro</i>	
INNOVACIÓN PEDAGÓGICA EN LA FORMACIÓN DEL PERFIL PROFESIONAL PARA EL DESARROLLO DE PROYECTOS DE AUTOMATIZACIÓN INDUSTRIAL A TRAVÉS DE UNA APROXIMACIÓN HOLÍSTICA.	517
<i>Juan Carlos Ríos, Zaneta Babel, Daniel Martínez, José María Paredes, Luis Alonso, Pablo Hernández, Alejandro García, David Álvarez, Jorge Miranda, Constantino Manuel Valdés and Jesús Alonso</i>	
Aprendiendo Simulación de Eventos Discretos con JaamSim	522
<i>Enrique Teruel and Rosario Aragüés</i>	
RED NEURONAL AUTORREGRESIVA NO LINEAL CON ENTRADAS EXÓGENAS PARA LA PREDICCIÓN DEL ELECTROENCEFALOGRAMA FETAL...	528
<i>Rosa M Aguilar, Jesús Torres and Carlos Martín</i>	
ANÁLISIS DEL COEFICIENTE DE TRANSFERENCIA DE MATERIA EN REACTORES RACEWAYS.....	534
<i>Marta Barceló, Jose Luis Guzman, Francisco Gabriel Acién, Ismael Martín and Jorge Antonio Sánchez</i>	
MODELADO DINÁMICO DE UN SISTEMA DE ALMACENAMIENTO DE FRÍO VINCULADO A UN CICLO DE REFRIGERACIÓN	539
<i>Guillermo Bejarano Pellicer, José Joaquín Suffo, Manuel Vargas and Manuel G. Ortega</i>	
Predictor Intervalar basado en hiperplano soporte	547
<i>José Manuel Bravo Caro, Manuel Vasallo Vázquez, Emilian Cojocarú and Teodoro Alamo Cantarero</i>	
Dynamic simulation applied to refinery hydrogen networks	555
<i>Anibal Galan Prado, Cesar De Prada, Gloria Gutierrez, Rafael Gonzalez and Daniel Sarabia</i>	

APROXIMACIÓN DE MODELOS ALGEBRAICOS MEDIANTE ALAMO Y ECOSIMPRO	563
<i>Carlos Gómez Palacín, José Luis Pitarch, Gloria Gutiérrez and Cesar De Prada</i>	
A Causal Model to Analyze Aircraft Collision Avoidance Deadlock Scenarios	569
<i>Miquel Àngel Piera Eroles, Julia de Homdedeu, Maria Del Mar Tous, Thimjo Koca and Marko Radanovic</i>	
ONLINE DECISION SUPPORT FOR AN EVAPORATION NETWORK	575
<i>José Luis Pitarch, Marc Kalliski, Carlos Gómez Palacín, Christian Jasch and Cesar De Prada</i>	
Predicción de la irradiancia a partir de datos de satélite mediante deep learning	582
<i>Javier Pérez, Jorge Segarra-Tamarit, Hector Beltran, Carlos Ariño, José Carlos Alfonso Gil, Aleks Attanasio and Emilio Pérez</i>	
MODELO DINÁMICO ORIENTADO AL TRATAMIENTO Y SEGUIMIENTO DE LA LEUCEMIA MIELOIDE CRÓNICA	589
<i>Gabriel Pérez Rodríguez and Fernando Morilla</i>	
Modelado y optimización de la operación de un sistema de bombeo de múltiples depósitos	596
<i>Roberto Sanchis Llopis and Ignacio Peñarrocha</i>	
DEVELOPMENT OF A GREY MODEL FOR A MEDIUM DENSITY FIBREBOARD DRYER IN ECOSIMPRO	604
<i>Pedro Santos, Jose Luis Pitarch and César de Prada</i>	
DETECCIÓN AUTOMÁTICA DE FALLOS MEDIANTE MONITORIZACIÓN Y OPTIMIZACIÓN DE LAS FECHAS DE LIMPIEZA PARA INSTALACIONES FOTOVOLTAICAS	611
<i>Jorge Segarra-Tamarit, Emilio Pérez, Hector Beltran, Enrique Belenguer and José Luis Gandía</i>	
Modelado de micro-central hidráulica para el diseño de controladores con aplicación en regiones aisladas de Honduras	618
<i>Alejandro Tapia Córdoba, Pablo Millán Gata, Fabio Gómez-Estern Aguilar, Carmelina Ierardi and Álvaro Rodríguez Del Nozal</i>	
FRAMEWORK PARA EL MODELADO DE UN LAGO DE DATOS	626
<i>J.M Torres, R.M. Aguilar, C.A. Martin and S. Diaz</i>	
SIMULADOR CARDIOVASCULAR PARA ENSAYO DE ROBOTS DE NAVEGACION AUTONOMA	633
<i>José Emilio Traver, Juan Francisco Ortega Morán, Ines Tejado, J. Blas Pagador, Fei Sun, Raquel Pérez-Aloe, Blas M. Vinagre and F. Miguel Sánchez Margallo</i>	
PLANIFICACION DE LA PRODUCCION BASADA EN CONTROL PREDICTIVO PARA PLANTAS TERMOSOLARES	641
<i>Manuel Jesús Vasallo Vázquez, José Manuel Bravo Caro, Emilian Cojocarú and Manuel Emilio Gegundez Arias</i>	
Evaluación multicriterio para la optimización de redes de energía	649
<i>Ascensión Zafra Cabeza, Rafael Espinosa, Miguel Àngel Ridao Carlini and Carlos Bordóns Alba</i>	
Percibiendo el entorno en los robots sociales del RoboticsLab	657
<i>Fernando Alonso Martín, Jose Carlos Castillo Montoya, Àlvaro Castro-Gonzalez, Juan José Gamboa, Marcos Maroto Gómez, Sara Marqués Villaroya, Antonio J. Pérez Vidal and Miguel Àngel Salichs</i>	

DISEÑO DE UNA PRÓTESIS DE MANO ADAPTABLE AL CRECIMIENTO	664
<i>Marta Ayats and Raul Suarez</i>	
COOPERATIVISMO BIOINSPIRADO BASADO EN EL COMPORTAMIENTO DE LAS HORMIGAS	672
<i>Brayan Bermudez, Kristel Novoa and Miguel Valbuena</i>	
PROCEDIMIENTO DE DISEÑO DE UN EXOESQUELETO DE MIEMBRO SUPERIOR PARA SOPORTE DE CARGAS	680
<i>Andrea Blanco Ivorra, Jorge Diez Pomares, David Lopez Perez, Francisco Javier Badesa Clemente, Miguel Ignacio Sanchez and Nicolas Garcia Aracil</i>	
Estructura de control en ROS y modos de marcha basados en máquinas de estados de un robot hexápodo	686
<i>Raúl Cebolla Arroyo, Jorge De Leon Rivas and Antonio Barrientos</i>	
USING AN UAV TO GUIDE THE TELEOPERATION OF A MOBILE MANIPULATOR	694
<i>Josep Arnau Claret and Luis Basañez</i>	
Estudio de los patrones de marcha para un robot hexápodo en tareas de búsqueda y rescate	701
<i>Jorge De León Rivas and Antonio Barrientos</i>	
SISTEMA DE INTERACCIÓN VISUAL PARA UN ROBOT SOCIAL	709
<i>Mario Domínguez López, Eduardo Zalama Casanova, Jaime Gómez García-Bermejo and Samuel Marcos Pablos</i>	
Mejora del Comportamiento Proxémico de un Robot Autónomo mediante Motores de Inteligencia Artificial Desarrollados para Plataformas de Videojuegos	717
<i>David Fernández Chaves, Javier Monroy and Javier Gonzalez-Jimenez</i>	
Micrófonos de contacto: una alternativa para sensado táctil en robots sociales	724
<i>Juan José Gamboa, Fernando Alonso Martín, Jose Carlos Castillo, Marcos Maroto Gómez and Miguel A. Salichs</i>	
Clasificación de información táctil para la detección de personas	732
<i>Juan M. Gandarias, Jesús M. Gómez-De-Gabriel and Alfonso García-Cerezo</i>	
Planificación para interceptación de objetivos: Integración del Método Fast Marching y Risk-RRT	738
<i>David Alfredo Garzon Ramos, Mario Andrei Garzon Oviedo and Antonio Barrientos</i>	
ESTABILIZACIÓN DE UNA BOLA SOBRE UN PLANO UTILIZANDO UN ROBOT PARALELO 6-RSS	746
<i>Daniel González, Lluís Ros and Federico Thomas</i>	
TELEOPERACIÓN DE INSTRUMENTOS QUIRÚRGICOS ARTICULADOS	754
<i>Ana Gómez Delgado, Carlos Perez-Del-Pulgar, Antonio Reina Terol and Victor Muñoz Martinez</i>	
CONTROL OF A ROBOTIC ARM FOR TRANSPORTING OBJECTS BASED ON NEURO-FUZZY LEARNING VISUAL INFORMATION	760
<i>Juan Hernández Vicén, Santiago Martínez de La Casa Díaz and Carlos Balaguer</i>	
PLATAFORMA BASADA EN LA INTEGRACIÓN DE MATLAB Y ROS PARA LA DOCENCIA DE ROBÓTICA DE SERVICIO	766
<i>Carlos G. Juan, Jose Maria Vicente, Alvaro Garcia and Jose Maria Sabater-Navarro</i>	

Estimadores de fuerza y movimiento para el control de un robot de rehabilitación de extremidad superior	772
<i>Aitziber Mancisidor, Asier Zubizarreta, Itziar Cabanes, Pablo Bengoa and Asier Brull</i>	
Definiendo los elementos que constituyen un robot social portable de bajo coste	780
<i>Marcos Maroto Gómez, José Carlos Castillo, Fernando Alonso-Martín, Juan José Gamboa, Sara Marqués Villarroya and Miguel Ángel Salichs</i>	
Interfaces táctiles para Interacción Humano-Robot	787
<i>Sara Marqués Villarroya, Jose Carlos Castillo Montoya, Fernando Alonso Martín, Marcos Maroto Gómez, Juan José Gamboa and Miguel A. Salichs</i>	
HERRAMIENTAS DE ENTRENAMIENTO Y MONITORIZACIÓN PARA EL DESMINADO HUMANITARIO	793
<i>Hector Montes, Roemi Fernandez, Pablo Gonzalez de Santos and Manuel Armada</i>	
Control a Baja Velocidad de una Rueda con Motor de Accionamiento Directo mediante Ingeniería Basada en Modelos	799
<i>Antonio José Muñoz-Ramírez, Jesús Manuel Luque-Bedmar, Jesus Manuel Gomez-De-Gabriel, Anthony Mandow, Javier Serón and Alfonso Garcia-Cerezo</i>	
SIMULACIÓN DE VEHÍCULOS AUTÓNOMOS USANDO V-REP BAJO ROS	806
<i>Cándido Otero Moreira, Enrique Paz Domonte, Rafael Sanz Dominguez, Joaquín López Fernández, Rafael Barea, Eduardo Romera, Eduardo Molinos, Roberto Arroyo, Luís Miguel Bergasa and Elena López</i>	
Cinemática y prototipado de un manipulador paralelo con centro de rotación remoto para robótica quirúrgica.....	814
<i>Francisco Pastor, Juan M. Gandarias and Jesús M. Gómez-De-Gabriel</i>	
ANÁLISIS DE ESTABILIDAD DE SINGULARIDADES AISLADAS EN ROBOTS PARALELOS MEDIANTE DESARROLLOS DE TAYLOR DE SEGUNDO ORDEN.....	821
<i>Adrián Peidro Vidal, Óscar Reinoso, Arturo Gil, José María Marín and Luis Payá</i>	
INTERFAZ DE CONTROL PARA UN ROBOT MANIPULADOR MEDIANTE REALIDAD VIRTUAL	829
<i>Elena Peña-Tapia, Juan Jesús Roldán, Mario Garzón, Andrés Martín-Barrio and Antonio Barrientos</i>	
Evolución de la robótica social y nuevas tendencias	836
<i>Antonio J. Pérez Vidal, Alvaro Castro-Gonzalez, Fernando Alonso Martín, Jose Carlos Castillo Montoya and Miguel A. Salichs</i>	
DISEÑO MECÁNICO DE UN ASISTENTE ROBÓTICO CAMARÓGRAFO CON APRENDIZAJE COGNITIVO	844
<i>Irene Rivas-Blanco, M Carmen López-Casado, Carlos Pérez-Del-Pulgar, Francisco García-Vacas, Víctor Fernando Muñoz, Enrique Bauzano and Juan Carlos Fraile</i>	
CÁLCULO DE FUERZAS DE CONTACTO PARA PRENSIONES BIMANUALES.....	852
<i>Francisco Abiud Rojas-De-Silva and Raul Suarez</i>	
Modelado del Contexto Geométrico para el Reconocimiento de Objetos.....	860
<i>José Raúl Ruiz Sarmiento, Cipriano Galindo and Javier Gonzalez-Jimenez</i>	
Estimación Probabilística de Áreas de Emisión de Gases con un Robot Móvil Mediante la Integración Temporal de Observaciones de Gas y Viento	868
<i>Carlos Sanchez-Garrido, Javier Monroy and Javier Gonzalez-Jimenez</i>	

MANIPULADOR AÉREO CON BRAZOS ANTROPOMÓRFICOS DE ARTICULACIONES FLEXIBLES	876
<i>Alejandro Suarez, Guillermo Heredia and Anibal Ollero</i>	
EVALUACIÓN DE UN ENTORNO DE TELEOPERACIÓN CON ROS	864
<i>David Vargas Frutos, Juan Carlos Ramos Martínez, José Luis Samper Escudero, Miguel Ángel Sánchez-Urán González and Manuel Ferre Pérez</i>	

Sistemas de Tiempo Real

GENERACIÓN DE CÓDIGO IEC 61131-3 A PARTIR DE DISEÑOS EN GRAFCET....	892
<i>María Luz Alvarez Gutierrez, Isabel Sarachaga Gonzalez, Arantzazu Burgos Fernandez, Nagore Iriondo Urbistazu and Marga Marcos Muñoz</i>	
CONTROL EN TIEMPO REAL Y SUPERVISIÓN DE PROCESOS MEDIANTE SERVIDORES OPC-UA	900
<i>Francisco Blanes Noguera and Andrés Benlloch Faus</i>	
Control de la Ejecución en Sistemas de Criticidad Mixta	906
<i>Alfons Crespo, Patricia Balbastre, Jose Simo and Javier Coronel</i>	
GENERACIÓN AUTOMÁTICA DEL PROYECTO DE AUTOMATIZACIÓN TIA PORTAL PARA MÁQUINAS MODULARES	913
<i>Darío Orive, Aintzane Armentia, Eneko Fernandez and Marga Marcos</i>	
DDS en el desarrollo de sistemas distribuidos heterogéneos con soporte para criticidad mixta	921
<i>Hector Perez and J. Javier Gutiérrez</i>	
ARQUITECTURA DISTRIBUIDA PARA EL CONTROL AUTÓNOMO DE DRONES EN INTERIOR	929
<i>Jose-Luis Poza-Luján, Juan-Luis Posadas-Yaguë, Giovanni-Javier Tipantuña-Topanta, Francisco Abad and Ramón Mollá</i>	
Ingeniería Conducida por Modelos en Sistemas de Automatización Flexibles	935
<i>Rafael Priego, Elisabet Estévez, Darío Orive, Isabel Sarachaga and Marga Marcos</i>	
Estudio e implementación de Middleware para aplicaciones de control distribuido	942
<i>Jose Simo, Jose-Luis Poza-Lujan, Juan-Luis Posadas-Yaguë and Francisco Blanes</i>	

Visión por Computador

Real-Time Image Mosaicking for Mapping and Exploration Purposes	948
<i>Abdulla Al-Kaff, Juan Camilo Soto Triviño, Raúl Sosa San Frutos, Arturo de La Escalera and José María Armingol Moreno</i>	
ALGORITMO DE SLAM UTILIZANDO APARIENCIA GLOBAL DE IMÁGENES OMNIDIRECCIONALES	956
<i>Yerai Berenguer, Luis Payá, Mónica Ballesta, Luis Miguel Jiménez, Sergio Cebollada and Oscar Reinoso</i>	
Medición de Oximetría de Pulso mediante Imagen fotopletismográfica.....	964
<i>Juan-Carlos Cobos-Torres, Jordan Ortega Rodríguez, Pablo J. Alhama Blanco and Mohamed Abderrahim</i>	
Algoritmo de captura de movimiento basado en visión por computador para la teleoperación de robots humanoides	970
<i>Juan Miguel Garcia Haro and Santiago Martinez de La Casa</i>	

COMPARACIÓN DE MÉTODOS DE DETECCIÓN DE ROSTROS EN IMÁGENES DIGITALES	976
<i>Natalia García Del Prado, Victor Gonzalez Castro, Enrique Alegre and Eduardo Fidalgo Fernández</i>	
LOCALIZACIÓN DEL PUNTO DE FUGA PARA SISTEMA DE DETECCIÓN DE LÍNEAS DE CARRIL	983
<i>Manuel Ibarra-Arenado, Tardi Tjahjadi, Sandra Robla-Gómez and Juan Pérez-Oria</i>	
Oculus-Crawl, a Software Tool for Building Datasets for Computer Vision Tasks	991
<i>Iván De Paz Centeno, Eduardo Fidalgo Fernández, Enrique Alegre Gutiérrez and Wesam Al Nabki</i>	
Clasificación automática de obstáculos empleando escáner láser y visión por computador ..	999
<i>Aurelio Ponz, Fernando Garcia, David Martin, Arturo de La Escalera and Jose Maria Armingol</i>	
T-SCAN: OBTENCIÓN DE NUBES DE PUNTOS CON COLOR Y TEMPERATURA EN INTERIOR DE EDIFICIOS	1007
<i>Tomás Prado, Blanca Quintana, Samuel A. Prieto and Antonio Adan</i>	
EVALUACIÓN DE MÉTODOS PARA REALIZAR RESÚMENES AUTOMÁTICOS DE VÍDEOS	1015
<i>Pablo Rubio, Eduardo Fidalgo, Enrique Alegre and Víctor González</i>	
SIMULADOR PARA LA CREACIÓN DE MUNDOS VIRTUALES PARA LA ASISTENCIA A PERSONAS CON MOVILIDAD REDUCIDA EN SILLA DE RUEDAS ..	1023
<i>Carlos Sánchez Sánchez, María Cidoncha Jiménez, Emiliano Pérez, Ines Tejado and Blas M. Vinagre</i>	
Calibración Extrínseca de un Conjunto de Cámaras RGB-D sobre un Robot Móvil	1031
<i>David Zúñiga-Nöel, Rubén Gómez Ojeda, Francisco-Ángel Moreno and Javier González Jiménez</i>	

Oculus-Crawl, a Software Tool for Building Datasets for Computer Vision Tasks

Iván de Paz Centeno, Eduardo Fidalgo Fernández, Enrique Alegre Gutiérrez, Wesam Al Nabki
 Dpto. Ingeniería Eléctrica y de Sistemas y Automática,
 Universidad de León, Campus de Vegazana s/n, 24071 León, Spain,
 ipazc@unileon.es, efidf@unileon.es, ealeg@unileon.es, mnab@unileon.es

Abstract

Building datasets for Computer Vision tasks require a source of a large number of images, like the ones provided by the Internet search engines, joined with automated scraping tools, to construct them in a reasonable time. In this paper it is presented Oculus-Crawl, a tool designed to crawl and scrape images from the search engines Google and Yahoo Images to build datasets of pictures, that is modular, scalable and portable. It is also discussed a benchmark for this crawler and an internal feature for storing and sharing big datasets, that makes it suitable for Computer Vision and Machine Learning tasks. In our tests we were able to crawl and fetch 11.555 images in less than 14 minutes, including also their meta-data description, showing that it might be well-suited for retrieving large datasets.

Key words: crawler, search engine, dataset, images, computer vision.

1 INTRODUCTION

Nowadays there exist a huge number of search engines that allow us to search content on the web including almost any type of resource, ranging from documents and pictures to sounds and videos. The nature of the web is to link multiple resources as hyper-links among them and, following the analogy, the process of reaching an end resource is done by crawling the interconnected nodes. Historically, the search engines have been fed by multiple web crawlers [4, 6, 9] that automatically track and follow the hyper-links from the content of the web, creating a database of entries that are usually formatted into a human-readable view in order to be presented to humans and to be read by humans. This adds an overhead in the automatic retrieval of content from search engines, as most of the times their results require to be analyzed and parsed from a markup language; in addition, the way to navigate through their content is usually handled dynamically by JavaScript code in form of AJAX calls [5], which requires of a sort of human intervention like scrolling down the content

or clicking on certain regions of the view, adding extra layers of complexity to the task of crawling those web sites. Even though most of them are not program-friendly in terms of extracting information, there have existed many successful attempts in retrieving useful information by automatically parsing the results from those search engines, like the framework Scrapy [15], the project icrawler [3] for python, or Apache Nutch [2], which takes advantage of big data tools such as Apache Hadoop [1]. Despite Computer Vision is one of the computer fields that most demand of large numbers of images, commonly required to solve specific classification or detection problems, crawling and scraping tools might be well suited for it. Even though most of the times Computer Vision problems leverage into public datasets, sometimes it exists the need to improve them or create new ones. For those situations where a distributed crawling is required and there is a lack of a distributed infrastructure, we provide an alternative named Oculus-Crawl, a crawler in the form of command line tool for images from the search engines Google and Yahoo, that can follow conveniently a distributed nature [10], which is isolated from underlying Big Data frameworks, and that can be shipped in the form of Docker containers [11]. It does not require to write code and is projected to be used as a source for Computer Vision and Machine Learning datasets.

The paper is structured in 4 sections, being the section 1 dedicated to the introduction; the section 2 is destined to an overview of the architecture of the solution which gives small insights about its main features; the section 3 summarizes the experiments and results we had with the tool applied to different topology configurations and tool options; and the section 4 dedicated to the conclusions and possible future works for this tool.

2 ARCHITECTURE

The tool has three roles out-of-the-box: a factory, a crawler and a client. Each role is performed by its correspondent entry point in the application, and they communicate each other through the factory, which exports an API-REST interface on a

specific port. The generation of a dataset comprises 5 stages:

1. The request to the factory for the generation of a dataset, done by the client.
2. The crawling of images, done by the crawlers.
3. The fetching of crawled images, done by the factory.
4. The packaging of the images into a single zip file with their crawled meta-data, done by the factory.
5. The publication of the dataset into a public directory, done by the factory.

An overview of the stages can be seen in Figure 1.

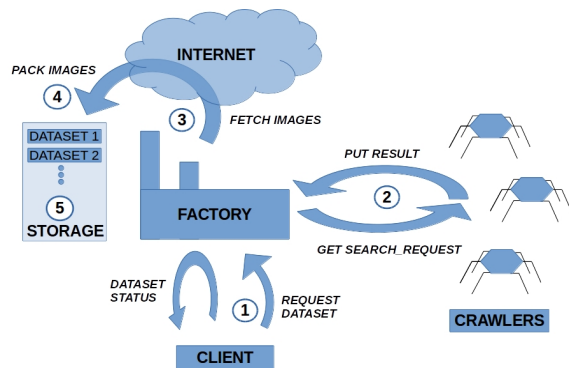


Figure 1: The 5 stages in the generation of a dataset of images.

The factory allows to create requests for generation of datasets within a single HTTP call. Each request for generation of a dataset is formed by a set of search requests for a specific search engine and a search words to be used, among other parameters; a visualization of this scheme can be found in Figure 2.

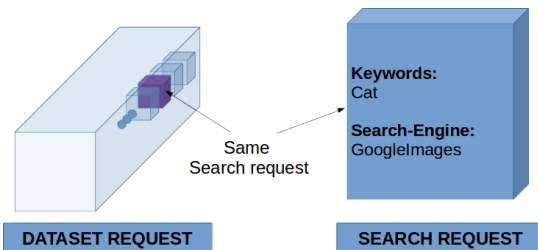


Figure 2: A dataset generation request is an aggregation of search requests. It is handled by the factory and the crawlers.

The crawlers, that might be running on any host, steadily poll the factory for requests of new datasets until one is retrieved, which initiates the

crawling process. This process consists of retrieving search requests from the dataset request, crawling their results and returning them back to the factory. The final scenario is a system on where the available crawlers compete to retrieve search requests and process them until all are processed. Then, they jump to the next available dataset request or stay idle waiting for new ones. This crawler’s behavior leads to a scalable distributed system, where increasing the number of crawlers reduces proportionally the overall time for crawling. Note that since the factory is an HTTP API-REST server, it can also be scaled up by load balancing it the same way a web server is usually scaled up.

When a dataset is completely crawled, the factory starts fetching all the crawled items in order to generate the final elements of the dataset, each consisting of the content of the image and its associated meta-data in JSON format. Note that the crawlers only gather the meta-data referenced by the search engine, including the URL to the images; and the factory, once the crawler process is finished, fetches their content.

2.1 THE SEARCH SESSION

The search session is the key feature in Oculus-Crawl as it preserves the whole dataset state in form of a serializable JSON structure that can be saved directly to a file. Therefore, each dataset request will have a search session attached that can be managed remotely, through the factory’s API-REST. It can be used to backup a process of dataset creation at any time and to restore this process remotely, from the client side.

The proposed key feature in Oculus-Crawl suits perfectly in the creation of Computer Vision datasets, as once this session is filled up, it can be used as a pre-fetching step on the creation of a dataset, avoiding the need to crawl again. Moreover, the serialization of the session state eases the sharing of the dataset over the network, hence, reduces the bandwidth, since its size is several times smaller than the complete fetched dataset, as shown in Figure 3.

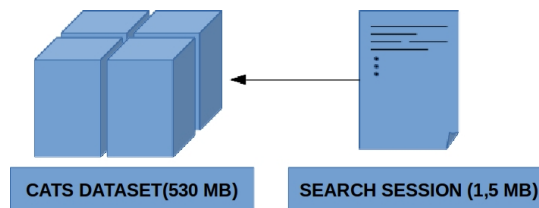


Figure 3: The search session representing an entire dataset. It contains information enough to rebuild the dataset without the need to crawl again.

2.2 THE SEARCH ENGINES

It is a common practice for search engines to associate meta-data to each of the elements they present to the final user. This meta-data might be useful for tagging the resources displayed by a search engine, as it is common to find resources with a descriptive text attached. Oculus-Crawl takes advantage of this behavior, storing the descriptive text with the width, height and file extension in the meta-data file of the final raw dataset. Currently, Google Images and Yahoo Images are officially supported by Oculus-Crawl. During the development phase of the tool, Bing Images was also supported, but a change in their presentation scheme left the search engine currently out of support. It is stated that search engines set limits on the number of elements they display for a single search request, as it is demonstrated that the behavior of most of the users is to look and use only the first entries of the results [16]. This leads the search engines to leave the less accurate elements they display on the least results they show, or to limit the number of results they provide e.g. Google Images is limited to 400 elements. When a search engine is not limited or its limit is too high, like Yahoo Images, the tool establishes a soft limit on approximately 500 results. This soft limit avoids to process an excessive number of pictures from a single Document Object Model (DOM) and also discards the least accurate resources. In Table 1 are shown the limits for each search engine.

Table 1: Search engines limits. Hard limit is imposed by the search engine. Soft limit is imposed by Oculus-Crawl.

Search Engine	Results Limit	Type
Google Images	400	Hard
Yahoo Images	500	Soft

In order to circumvent these limits and at the same time retrieve the most accurate results for a given topic, Oculus-Crawl follows a divide-and-conquer strategy, splitting each search request in multiple search requests, having each slight changes that consist of appending an adjective to the main search words. This task is accomplished by the client role of Oculus-Crawl, which accepts a set of adjectives in addition to the main search words, and automatically combines them, therefore, generating multiple and different search requests that forces the search engines to change the nature and order of the elements displayed for each. Hence, a list of adjectives compatible with the main search words should be manually provided during the invocation of the client.

2.3 USAGE OF ADJECTIVES

The Oculus-Crawl client accepts as input a set of adjectives in order to combine them with the original search words with the goal of increasing the number of pictures retrieved. Each combination leverages into a different set of results but sharing all of them the same inner semantic. A restriction, however, is that the adjectives chosen should be applicable to the search words context, e.g. a chair can be blue, beautiful or small; but can not be angry or thirsty. Even though the search engines always retrieve results regardless of the search keywords used, the results of using incompatible adjectives lead to an unknown or incorrect semantic, where most of the results are probably going to be out of the context of the original search words. This is explained by the fact that search engines associate key words to images, being the origin of these key words in the description that usually users attach next to the images in the HTML documents.

2.3.1 Number of adjectives to use.

The number of adjectives used for crawling affects the number of images retrieved. In order to know how many adjectives are needed to build a certain dataset of N images for a single topic, being L_i the limit for the search engine i , the Eq. (1) approximate it.

$$Adjv(N) = \frac{N}{\sum_{i=1}^{|L|} L_i} \quad (1)$$

Even though the number of images should be proportional to the number of adjectives used, the factory implements a deduplication mechanism of images, during the fetching stage, that may decrease the number of total pictures compared to the results retrieved with Eq. (1), the higher the number of adjectives used.

2.4 THE FETCHING STAGE

When the crawling process is finished, the factory fetches all the resources crawled. The fetching stage consists of a pool of 10 workers that downloads distributed the content of each of the crawled resources, which implies that up to 10 images can be downloaded simultaneously. Having increased this value might have forced the DNS servers to resolve too frequently the addresses of the hosts that contains the resources, which could potentially be blocked due to the Request Response Limit (RRL) of certain Domain Name Servers (DNS) [17], which can lead to a temporal

ban from the DNS resolver, hence, stopping the factory from successfully generating the dataset. Nonetheless, this parameter will eventually become configurable. Lastly, when a resource is requested to a host, the Oculus-Crawl factory sets a timeout of 15 seconds for the host to answer this request before it is marked as invalid and discarded from the dataset.

2.4.1 Deduplication of resources.

It is common for multiple search engines to refer to the same resources in certain number of results. This can be split into two different situations: 1) the same resource is hosted by two different hosts; 2) multiple search engines provide a reference to exactly the same host. In both cases, the end resource is the same, but the description used as meta-data might be different. A way to tackle this problem is to hash the resources in order to discard duplicates. In Oculus-Crawl, the hashing is done by using the MD5Hash [14] algorithm for each resource, which allows to retrieve the links and search engines that point to the same resources and storing them along with the meta-data element for each resource. For this reason, Oculus-Crawl might be also useful to catch hosts with duplication of resources. The reasons for choosing MD5Hash instead of a more secure hashing method is: 1) even though a collision of hashes is possible [18], in the case of a hash collision for different resources in high sized datasets, it is probably not going to cause a big trouble for the end dataset; 2) low sized datasets are not prone to present collisions and 3) because in scaled environments where a high sized dataset is required, the speed in hashing takes importance and MD5Hash is one of the fastest and reliable-enough hashing methods. However, it is common to have the same picture duplicated with different dimensions or formats each, a situation that MD5Hash or most of the hashing methods are not able to tackle. In this case, a more complex hash algorithm can be used like the Perceptual Hashing [12], which Oculus-Crawl will include in the future.

2.4.2 Inferring the extension of the pictures

When the crawling process is finished, the factory fetches all the resources crawled. In order to know the extension of the fetched picture, the name of the URL that points to it can not be trusted, as it does not necessarily point to a file-system file, e.g. a URL that apparently refers to an image because ends with ".jpg" might refer to an HTML document or a binary executable file instead. Hence, finding the correct extension requires of checking at the response headers of an HTTP HEAD call

to the remote server that is hosting the picture, and to process the MIME-type header that specifies what kind of resource it is returning. Even though this MIME-type header can not be completely trusted, as not all the web servers return a correct MIME-type header for the resources they send, it is the fastest method for inferring the resource's format in a reliable-enough way. Note that MIME-Type is the most reliable method just after the checking at the resource's content itself, and it is also used by the web browsers to correctly parse the retrieved content for the web pages they render. For this reason, Oculus-crawl follows an extension inferring procedure that, by priority, consists of: 1) retrieve the extension from the MIME-type; 2) use the URL name to inaccurately infer the extension when the first method fails. If none of both methods are able to report a valid picture extension, the file is stored in the dataset without extension.

2.5 TECHNOLOGIES USED

Oculus-Crawl has been built entirely in Python3. The project can be directly executed in any x86_64 architecture by using Docker with the *latest* Docker image for Oculus-Crawl¹, which contains all the dependencies satisfied.

2.5.1 Factory process.

The factory process uses the Python's library *Flask* [7] to expose an API-REST which allows standardized interactions for crawlers and clients with the datasets' sessions. Moreover, this functionality can be easily tested and consumed externally (e.g. using HTTP calls with the UNIX tool *cURL*) or wrapped and interfaced in a web view. This means that the creation of a dataset can be invoked, tracked, backup-ed or dumped at any time without the need to have explicitly a client; however, Oculus-Crawl bundles a specific client for managing these tasks. The factory is a multi-tasked process which uses the Python's library *urllib2* to fetch the crawled resources whenever the crawling stage has finished, distributed among processes by using the Python's library *multiprocess*. It also uses the Python's library *hashlib* to perform MD5Hash on each fetched resource in order to avoid exact duplications of resources. The final zipped dataset is generated by using the library *shutil* from Python.

¹<https://hub.docker.com/r/dkmivan/oculus-crawl/tags/>

2.5.2 Crawler process.

The crawler process takes advantage of the framework *Selenium* and its web-drivers [13] for the web-browser *Firefox*. This framework allows a direct interaction with the elements from the DOM and, at the same time, to perform common user's actions like clicking on buttons, performing scrolls or filling forms on the HTML view. Moreover, this scheme takes advantage of the JavaScript engine from the web-browser since the page gets rendered. This way of crawling through Selenium adds overhead on the processing of the HTML by increasing the load times due to rendering the web-page rather than only parsing the HTML, however this behavior reduces the probability for the crawler of getting detected as a bot. Even though it uses a graphical web-browser instead of direct HTTP calls, it can run in non-graphical environments by using the library *PyVirtualDisplay* to wrap the view in a virtual display. Furthermore, a crawler process can be split in several workers taking advantage of the Python's library *multiprocess*, behaving each as a single crawler instance and increasing the overall speed for crawling the resources within a single host.

2.5.3 Client process.

The client process wraps all the API-REST calls from the factory for the generation and tracking of datasets by using the Python's *requests* library. It is a simple client that generates a dataset request on the factory and steadily polls for its status until it is finished, showing a progress bar for each of the stages in the dataset generation.

3 EXPERIMENTS AND RESULTS

We tested the tool in two dedicated servers Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz, one dedicated server Intel(R) Xeon(R) CPU D-1531 @ 2.20GHz and one virtual private server Intel(R) Xeon(R) CPU E5-2650 0 @ 2.00GHz, each of them connected to different networks of 1 Gbps of connectivity. We followed different topologies, running each Oculus-Crawl role in different machines and also combining all the roles together in a single machine to measure the performance impact. In order to help in the measurements of our tests, we defined a measurement variable that we called *adjective_rate*, which represents the ratio of adjectives per crawler. We realized that, for a small *adjective_rate*, a computer with poor performance running a crawler does not improve significantly the overall performance of the crawling process when added to the crawlers pool, as shown in

the Table 2 for the cases A_1 and B_1 . Nonetheless, the performance increases only on situations where the *adjective_rate* is larger, as shown in the Table 2 for the cases A_2 and B_2 . This fact is explained because the search-requests are retrieved by the crawlers whenever they get freed rather than equally distributed among them; leading faster crawlers to process most of the requests, a situation that is best used in the case of a high number of adjectives.

Table 2: Benchmark of crawling same search words with 3 adjectives. A_i for the case of a single and fast crawler and B_i for the case of sharing the crawling process from the same fast crawler with an extra slow crawler.

Crwls	Adjv	Size	Imgs	Time
A_1	3	465 MB	2118	4m 20s
B_1	3	612 MB	2525	4m 17s
A_2	15	2,4 GB	11342	14m 47s
B_2	15	2,4 GB	11555	13m 40s

A_1 : 1 crawler x 3 workers

B_1 : 1 crawler x 3 workers + 1 slow crawler x 1 worker

A_2 : 3 crawlers x 3 workers

B_2 : 3 crawlers x 3 workers + 1 slow crawler x 1 worker

During the crawling process, Oculus-Crawl mixes the search words with each adjective in order to generate new search requests, which usually results in different images being displayed by the search engine. The number of images retrieved is proportional to the number of adjectives used for generating the dataset; however, as it can be stated in the Figure 4, the number of images is less than expected the more adjectives are used.

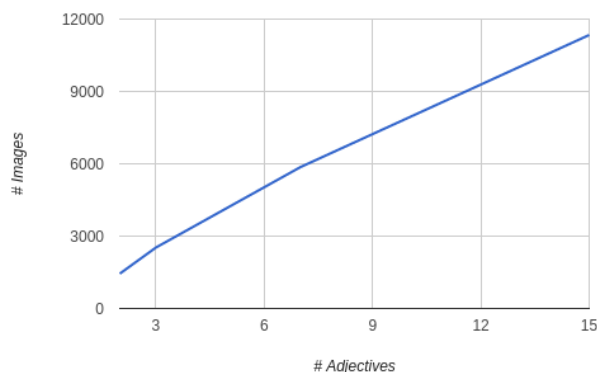


Figure 4: Relation between number of images retrieved and number of adjectives used to crawl.

The reason behind this distribution is that Oculus-Crawl hashes the images by their MD5Hash in order to discard duplicates, and it is more likely to

find more duplications or pictures out of service the more adjectives used for crawling. However, we noticed that when the crawlers were spread among servers located in different countries rather than a single country, the number of images retrieved was higher, as shown in the Table 2 in the case of B_1 and B_2 . This is explained by the fact that some search engines, like Google, displays different results for the same search words based on the geographic localization of the IP that made the request [8], which reduces the probability of duplication of images.

We also measured the time spent by Oculus-Crawl to process 4 adjectives using from 1 single-threaded crawler to 4 single-threaded crawlers spread on 4 different machines and networks, as shown in Figure 5.

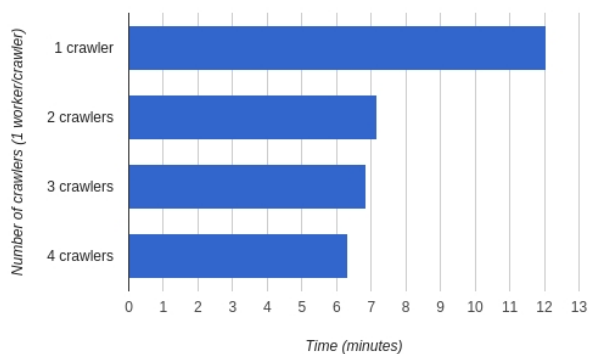


Figure 5: Time benchmark for 4 adjectives in function of the number of crawlers with a single worker each.

The measurement was done by using the UNIX *time* command preceding the invocation of the dataset creation, which gives an exact execution time. The highest increase rate in time performance was achieved by using two crawlers, which passed from 12 minutes to 7. However, the best performance was achieved by using as many crawlers as adjectives. When the crawling process is finished, the factory starts to fetch each resource from the references and finally to compress it in order to be published in a directory, which adds a static time independently of the number of crawlers working in the same pool. This static time depends on: 1) the connection speed of the factory’s host and the throughput of the factory’s hard disk; 2) the size, availability and connection speed of each crawled resource’s host. Under our tests, 4 adjectives leveraged into 3200 images, a total of 800 MB of size and it took 2 minutes and 20 seconds from the start of the fetching stage until the publishing stage.

Finally, we counted the extensions inferred by

Oculus-Crawl by checking the MIME-Type from the response headers and the URL names whenever a MIME-Type was missing, showing that the most used picture extensions belong to the JPEG extension group. In our tests, we realized that there were some files fetched by the factory that were executables, like shown in the Table 3. This implies that search engines for images sometimes might refer to resources that are not images, even though originally they were images, showing that a preprocessing of the files to ensure that they are images is desirable.

Table 3: Extension for images found by crawling 3 adjectives.

Extension	Count	Representation
.jpg	3190	93,91%
.png	117	3,44%
.gif	81	2,38%
.jpeg	4	0,12%
.html	2	0,06%
.jpg c200	1	0,03%
.bin	1	0,03%
.exe	1	0,03%

4 CONCLUSIONS AND FUTURE WORKS

We developed and presented Oculus-Crawl, a stand-alone alternative for existing crawling tools that serves for building Computer Vision datasets by crawling images from Google and Yahoo images. It was discussed its suitability for building large datasets due to its modular and scalable architecture and its capacity to circumvent the search engines limits by combining adjectives with search words. Within our tests, we were able to crawl and fetch 11.555 images in less than 14 minutes. We concluded that the best results are achieved by distributing crawlers’ workers among different countries, which leads to a different set of pictures being displayed for the same search words reducing the probability of duplications and increasing the quality and richness of the final dataset; also, the usage of as many crawlers as adjectives gives the best performance. We provided a relation between number of adjectives and number of images retrieved for a single search-word topic and a function to know an approximation of how many adjectives should be used to retrieve a specific number of images for a given topic. We also discussed about one benchmark for performance in function of the number of crawlers used and another benchmark for the impact of different speeds in multiple crawlers, In addition, we found a practice on certain hostings of swapping an original

indexed image with an executable, thus a check of image correctness before fully using the scraped dataset is advisable. During our tests we used high-end machines that vastly satisfied the needs of the tool; a much lesser specifications might be capable of achieving the same results.

The creation of a dataset of images is the first step in building a working model for Computer Vision and Machine Learning, but in some cases it is required to label each of the elements that compose the dataset; for this reason it might be useful to combine the results of this tool with some logic able to take advantage of the extracted meta-data for each element in order to infer a correct label for each resource.

Also, future directions point towards retrieving other kind of resources like sounds, music, documents and videos; and to increase the number of supported search engines. Even though this software is non graphical, it might be able to be interfaced as a web page. In addition, another way of improving this tool is to automate the generation of adjectives that are semantically valid with the main search words, e.g. by using Natural Language Processing (NLP) techniques based on the language that the main search words belong to.

Lastly, we propose a session scheme, which is a way to share datasets based on crawling, that contains references instead of the whole crafted dataset's content. This tool is able to use this session scheme to rebuild the same dataset in any other computer, easing the sharing process of a crawled dataset.

4.1 HOW TO CONTRIBUTE

This project is released as open-source under the GNU GPL v3 License. It can be located in a git repository within GitHub². Any contribution can be done by making pull requests to the repository or filling the issues tracker.

Acknowledgements

This research was funded by the framework agreement between the University of León and INCIBE (Spanish National Cybersecurity Institute) under addendum 22.

References

- [1] Apache Software Foundation. Hadoop. Version 2.8.0. Mar. 22, 2017. URL: <https://hadoop.apache.org>.
- [2] Apache Software Foundation. Nutch. Version 1.13.0. Apr. 22, 2017. URL: <http://nutch.apache.org>.
- [3] Chen, K. Python icrawler. Version 0.3.6. May. 8, 2017. URL: <https://github.com/hellok/icrawler>.
- [4] Desai Student, K., Devulapalli Student, V., Agrawal Asst, S., Kathiria Asst, P., and Patel Professor, A., (2017). Web Crawler : Review of Different Types of Web Crawler, Its Issues, Applications and Research Opportunities. *International Journal of Advanced Research in Computer Science*, 8(3).
- [5] Duda, C., Frey, G., Kossmann, D., Matter, R., and Zhou, C. (2009). AJAX crawl: Making AJAX applications searchable. In *Proceedings - International Conference on Data Engineering*, pages 78–89.
- [6] El-Ramly, N., Harb, H., Amin, M., and Tolba, A., (2004). More effective, efficient, and scalable web crawler system architecture. *International Conference on Electrical, Electronic and Computer Engineering, 2004. ICEEC '04.*, pages 120–123.
- [7] Grinberg, M., (2014). *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media, Inc., 1st edition.
- [8] Gupta, V., Gomes, B., Lamping, J., McGrath, M., Singhal, A., and Tong, S., (2008). System and method for providing preferred country biasing of search results. US Patent 7,451,130.
- [9] Hafri, Y. and Djeraba, C., (2004). High performance crawling system. *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 299–306.
- [10] Kausar, M. A., Dhaka, V. S., and Singh, S. K., (2013). Web Crawler: A Review. *International Journal of Computer Applications*, 63(2):975–8887.
- [11] Merkel, D., (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239):2.
- [12] Niu, X.-m. and Jiao, Y.-h., (2008). An overview of perceptual hashing. *Acta Electronica Sinica*, 36(7):1405–1411.
- [13] Razak, R. A. and Fahrurazi, F. R. (2011). Agile testing with selenium. In *Software Engineering (MySEC), 2011 5th Malaysian Conference in*, pages 217–219. IEEE.
- [14] Rivest, R., (1992). The md5 message-digest algorithm. *IETF Network Working Group, RFC 1321*.

²<https://github.com/ipazc/oculus-crawl>

- [15] Scrapy, A., (2016). Fast and powerful scraping and web crawling framework. *Scrapy.org*. *Np*.
- [16] Silverstein, C., Marais, H., Henzinger, M., and Moricz, M., (1999). Analysis of a very large web search engine query log. *ACM SIGIR Forum*, 33(1):6–12.
- [17] Vixie, P., (2014). Rate-limiting state. *Communications of the ACM: ACM Queue*, 12(2):10.
- [18] Wang, X., Feng, D., Lai, X., and Yu, H., (2004). Collisions for hash functions md4, md5, haval-128 and ripemd. *IACR Cryptology ePrint Archive*, 2004:199.