ACMCV'2023

10th Annual Catalan Meeting
on Computer Vision

MASTER THESES PRESENTATIONS | SCHEDULE

Monday, September 18

16:00 – 16:45

BERKAY ARPACI

3D Apple Detection from Large Point Clouds Using Deep Learning

Supervisors: Jordi Gené Mola (UdL), Javier Ruiz Hidalgo (UPC)

Abstract:
In recent years, with the increasing number of studies about autonomous vehicles, LiDAR sensor technology and 3D deep learning object detection models have kept improving steadily. As a result of these improvements, these models have started to be implemented in other fields, such as agriculture. One important potential implementation of 3D object detection is to determine various quantities related to fruits such as their number, position and size which can be utilized to optimize the harvesting. In this work, performance of a selected 3D deep learning object detection model for detection of apples in an orchard is examined to achieve in-field automatic fruit counting and localization. Additionally, the results of different types of data are compared to determine the most suitable data type for these tasks. Among many 3D object detection models, PointRCNN is selected as it supports point clouds having homogeneous characteristics in all 3 axes unlike other models using bird-eye-view or similar methods treating the height differently. Point clouds obtained from two different methods and a combination of the data obtained from them are used to see the effect of positional accuracy, cloud density and the effect of additional parameters such as color and reflectance. A LiDAR dataset is obtained by a mobile LiDAR sensor combined with a positioning sensor and merged to form a single point cloud for 11 trees. Photogrammetry data set is obtained from a set of color images using Structure from Motion (SfM) algorithm. The experimental results in this thesis show that photogrammetry data with color information yielded scores much higher than the scores obtained from the LiDAR data. The maximum mAP score for photogrammetry and LiDAR data is % 84.39 and % 32.50 respectively.

Tuesday, September 19

14:00 – 14:45

HICHAM EL MUHANDIZ AARAB

Empirical Study of Uncertainty Estimation Techniques for Zero-Shot Image Classification Pre-trained CLIP Models

Supervisors: Lluís Gómez Bigorda (UAB)

Abstract:
In this work, we present an empirical study that investigates various techniques for uncertainty estimation in zero-shot image classification models based on contrastive language-image pre-training (CLIP). The study aims to evaluate and compare these techniques without requiring re-training of the models, thus providing practical insights for real-world applications. We assess the performance of several well-established approaches, including some statistics over the softmax response (e.g. max probability, least confidence, predictive entropy, margin of confidence, ratio of confidence), Monte Carlo Dropout, and Deep Ensembles, across a diverse range of image classification tasks. We also evaluate the effect on uncertainty of two other techniques that are inherent in modern CLIP models: Patch-Dropout, and Prompt Embedding Space Ensembles. We utilize a collection of 15 datasets to comprehensively evaluate the effectiveness of these techniques using standard evaluation metrics.

ADAM SZUMMER confidential

Multi-person 3D Human Pose Estimation in Real-Time for Fitness and Rehabilitation

IÑIGO AUZMENDI IRIARTE confidential

Real-Time Optical Marker Detection on Edge Devices

JOHNNY NÚÑEZ CANO confidential

Quantifying Suspicion: Advanced Trajectory Analysis for Loitering Anomaly Detection

14:45 – 15:30

ADVAIT DIXIT

Incorporating Anisotropic Surface Characteristics into NeRF for Cultural Heritage Preservation

Supervisors: Joost van de Weijer (UAB)

Abstract:
The process of digitizing cultural artefacts and historical sites has seen remarkable advancements, enabling the creation of highly detailed 3D representations. Traditional techniques include image-based photogrammetry or laser-based scans. However, in recent years, Neural Radiance Fields (NeRF) have emerged as a cutting-edge technique of photorealistic 3D digitization and lightning representation (non-Lambertian). Furthermore, the material-based version Ref-NeRF improves the modelling of highly reflective materials. This research incorporates in Ref-NeRF a more complex orientation modelling by including surface orientation tangents, thereby enhancing its capability to represent materials with preferred reflective

orientations. Through this anisotropic upgrade, this work improves the capacity for more faithful artefact digitization, ensuring the preservation of intricate surface details. This offers a promising avenue for safeguarding and sharing our cultural legacy in an immersive and technologically advanced manner.

KYRYL DUBOVETSKYI

Self-training OCR for Shipping Container Code Reading

Supervisors: Dimosthenis Karatzas (UAB), Marçal Rusiñol (UAB)

Abstract:
Shipping containers need to be tracked. Often it is done with optical character recognition (OCR) systems that read container numbers. Currently, OCR systems are frequently implemented using machine learning models. Such models can benefit from unlabeled data. We present research on how to improve a machine learning model for container number reading, using container images without ground truth labels. This research is focuses on self-training, that is a machine learning training procedure in which a model itself generates labels from which it learns, and associated problems like selecting pseudo-labels for following training iterations, or the impact of pseudo-labels distribution on training quality. As a result, under some circumstances, studied methods outperformed the baseline training procedure: regular training with the initial data set.

ANA HARRIS MARTÍNEZ

Generation of Synthetic Longitudinal Magnetic Resonance Images of Subjects with Multiple Sclerosis: Assessment of Brain Atrophy and their Clinical Impact

Supervisors: Gerard Martí Juan (UPF)

Abstract:
Multiple sclerosis (MS) is a chronic autoimmune disease characterized by central nervous system inflammation, demyelination, neurodegeneration and great heterogeneity. Differentiating between normal aging and MS-related brain atrophy is crucial in MS research. We propose a novel approach using generative models to simulate healthy aging in people with MS by generating follow-up magnetic resonance imaging (MRI) from baseline T1-weighted images. We compared the performance of two generative models, a conditional generative adversarial network model (cGAN) and a conditional latent diffusion model (cLDM), in addition to REPLICA, a model based on a supervised random forest approach. Our training set consists of 351 pairs of images from healthy control (HC), while the test set comprises 562 pairs of images from people with MS. Clinical relevance is evaluated using SIENA. The results indicate that cGAN outperforms REPLICA in simulating normal aging based on PBVC values. Our findings demonstrate the potential of generative models in advancing research on MS-related brain atrophy.

IGOR UGARTE MOLINET

Self-supervised Learning of Multimodal Representations in Food Recipes

Supervisors: Coloma Ballester (UPF), Gloria Haro (UPF)

Abstract:
The increasing social interest in gastronomy has filled the Internet with food information in a variety of formats, such as recipe videos. Nevertheless, most user-uploaded video data is weakly ordered and structured. Furthermore, a video recipe is a combination of multi-modal signals (audio, video and text), so information can be found split between modalities that lie in different spaces. The lack of labeling and the multi-modal nature of video signals make the automatic comparison of food recipes hard to perform. Here we introduce a self-supervised perspective to perform multi-modal temporal alignment in non-labeled recipe videos from YouCookII dataset.

This approach can be seen as a generalization of the Temporal Cycle Consistency for multi-modal signals. It will help finding common information between text, audio and video signals. We evaluate our methodology using Kendall's Tau correlation coefficient and demonstrate its effectiveness in aligning multi-modal cooking videos, with promise for improvement with better data quality. Our work would help with future downstream tasks such as recipe or cooking action retrieval, ingredient identification, and cooking instruction understanding.

MIQUEL ROMERO BLANCH confidential

Using LiDAR for 3D Object Detection in the context of Video Surveillance

ALEX CARRILLO ALZA confidential

Automatic Data Drift Detection on Edge Device Deployments

15:30 – 16:15

JOSEP BRAVO BRAVO

Dynamic Object Manipulation and Enhanced Embeddings for Advancing 3D Scene Understanding in Robotics

Supervisors: Bogdan Raducanu (UAB)

Abstract:
Accurately understanding and interpreting 3D scenes in robotic environments is a critical task for various applications, such as object manipulation and scene navigation. This paper presents an improved methodology for 3D scene understanding using as reference the CLIP-Fields framework, which integrates recent advancements in vision and language models for robotic applications. Our key contributions include: the adoption of a larger closed-set vocabulary, efficient memory usage changes in the training pipeline, analysis of various pre-trained models with altering inputs, improved query retrieval in the inference pipeline, and the creation of a dynamic point cloud to handle scene changes. Overall, the proposed contributions to the original framework enhance both textual and visual embedding leading to an improved accuracy in object localization and recognition, for several query types, when compared to the original CLIP-Fields framework, as confirmed by our experiments. Furthermore, the methodology was deployed and evaluated on a robotic platform, showcasing its real-world applicability. While the pipeline excels in textual, visual, semantic, and some relating queries, challenges persist in handling object localization for complex relational queries. Overall, the improved framework showcases promising results and sets the foundation for further advancements in 3D scene understanding for robotics.

MARCOS MUÑOZ GONZÁLEZ

Symmetry Detection in Artificial Neural Networks: From Filters to Perception

Supervisors: Alejandro Pàrraga (UAB)

Abstract:
Symmetry is an important characteristic for many disciplines, ranging from biology to the arts. Recent artificial neural networks research has shown promising results in symmetry detection, mirroring human capabilities. How neural networks perform this computation is still, however, an open research question. The purpose of this work is twofold: (I) To explore the capabilities of neural network filters in symmetry detection and (II) to expand previous results on symmetry

detection using artificial neural networks by computing the activation differences obtained from symmetric and non-symmetric dot-array stimuli. To this end, we designed an interface for analyzing filters and developed a fast, customizable dot-array symmetric stimuli generator similar to those used in human experiments. Our results show that filters capable of finding symmetric axes may exist in artificial neural networks. Additionally, we demonstrate the effect of several factors on symmetry perception in neural networks through experiments ranging from neural network architecture to stimuli variations, including color and luminance.

MICHELL VARGAS

AI-Enabled Cloud Detection in Satellite Imaginery for RISC-V Platforms

Supervisors: Daniel Ponsa (UAB), Felipe Lumbreras (UAB)

Abstract:
This master's thesis revolves around the selection of the CloudSen12 database as the foundation for training a range of AI models specialized in cloud classification and segmentation. The primary emphasis is on refining cloud detection accuracy. Following this, an optimization process is conducted on these trained models, leading to their successful integration into emulators of the RISC-V architecture. The study presents a comprehensive strategy, highlighting elevated cloud analysis capabilities within satellite imagery, coupled with seamless deployment on RISC-V-oriented satellite systems.

ADRIÀ MOLINA RODRÍGUEZ

Bridging Cross-Modal Alignment for OCR-Free Content Retrieval in Scanned Historical

Supervisors: Josep Llados Canet (UAB), Oriol Ramos (UAB)

Abstract:
In this work, we address the limitations of current approaches to document retrieval by incorporating vision-based topic extraction. While previous methods have primarily focused on visual elements or relied on optical character recognition (OCR) for text extraction, we propose a paradigm shift by directly incorporating vision into the topic space. We demonstrate that recognizing all visual elements within a document is unnecessary for identifying its underlying topic. Visual cues such as icons, writing style, and font can serve as sufficient indicators. By leveraging ranking loss functions and convolutional neural networks (CNNs), we learn complex topological representations that mimic the behavior of text representations. Our approach aims to eliminate the need for OCR and its associated challenges, including efficiency, performance, data-hunger, and expensive annotation. Furthermore, we highlight the significance of incorporating vision in historical documentation, where visually antiquated documents contain valuable cues. Our research contributes to the understanding of topic extraction from a vision perspective and offers insights into annotation-cheap document retrieval systems.

GUILLEM MARTÍNEZ SÁNCHEZ confidential

Semantic Segmentation for Memory-Constrained Edge Devices

16:15 – 17:00

ALBERT BARREIRO DÍAZ

Preserving the Past: Enhancing 3D Geometry recovery for Cultural Heritage

Supervisors: Gloria Haro (UPF)

Abstract:
This master's thesis explores the digital preservation of cultural heritage, focusing on monuments captured in challenging conditions. Neural Radiance Fields (NeRF) is found as state of the art for 3D reconstruction techniques generating new views by optimizing a neural network. However, NeRF has its limitations, especially in handling inconsistent lighting and obstacles in the scenes. To tackle these challenges, adaptations like "NeRF in the wild" and "Ha-NeRF" have emerged. Inspired by these advancements, this thesis uses "Ha-Nerf" as a baseline to push for improved performance. Experimental strategies include refining Ha-NeRF's appearance encoder, incorporating an lpips loss term for clearer reconstructions, integrating GANs for enhanced image quality, improving Ha-NeRF's masks and introducing a penalization term in the l2 based on the gradient of the ground truth images.

ALEXANDER TEMPELAAR SÁNCHEZ

Fish Monitoring System with Stereo Camera on Edge-Computing Device

Supervisors: Antonio Agudo (UPF)

Abstract:
Underwater images obtained by the Deep Vision Subsea Unit give a great amount of visual information about what is entering the fishing net it is attached to. The image acquisition system is composed of two stereo cameras and a controlled illumination setup providing a homogeneous background colour. In this paper, a pipeline is proposed to process the stereo images and provide an estimation of the number of fish instances that appear in them in addition of their species and an average size. The proposed pipeline runs on the underwater system, an edge-computing device, and reduces the necessary data to be send from the system to the vessel's bridge through an acoustic modem with very small bandwidth, allowing the fishers to know whatis entering the net and stop in an early stage if the catch is not of their interest. The three fish species that have been used to implement the algorithm are Herring, Mackerel and Bluewhiting.

SERGI MASIP CABEZA

Continual Learning of Diffusion Models with Data-free Distillation

Supervisors: Ernest Valveny Llobet (UAB), Pau Rodriguez (UAB)

Abstract:
Diffusion Models (DMs) achieve state-of-the-art performance in many generative modelling tasks such as Image Synthesis. However, their training demands substantial computational resources. Training them using continual learning (CL) would enhance efficiency and allow reusing already trained models by allowing incremental task learning while retaining prior knowledge. A CL strategy that has proven successful for other models, Generative Replay, involves interlacing samples generated from a copy of the model trained on previous tasks with samples from the current task. However, DMs have high generation times, and the sampled images may not be perfect and contain noise, causing a degradation in the denoising capabilities throughout the continual learning process. To overcome both issues, I introduce a novel approach based on distillation to transfer the knowledge from previous tasks effectively. This approach, data-free distillation of diffusion models, does not require ground truth data and

distils the entire reverse process of the DM. Experiments with an independent and identically distributed stream of data show that the students achieve a level of performance comparable to that of the teacher. Furthermore, when applied in the continual learning setting, my approach exhibits a remarkable improvement over naive generative replay applied to the DM while reducing the required number of generation steps by up to fivefold (x5). Finally, I show that DMs can be used as effective generators in generative replay for classifiers.

ANNA OLIVERAS TOUS

Large Language Models for Document Visual Question Answering

Supervisors: Dimosthenis Karatzas (UAB)

Abstract:
This study introduces innovative methods for Document Visual Question Answering (DocVQA) through the utilization of Large Language Models (LLMs). Our approach involves fine-tuning the Flan-T5 model on the DocVQA dataset with diverse context types, revealing the effectiveness of incorporating spatial information. By utilizing both the document's textual content and the corresponding bounding box locations of words, we achieve the best performance, reaching an ANLS score of 0.76, using only the text modality. Furthermore, we attempt to incorporate word recognition in the language model itself. To this purpose, we present a multimodal DocVQA pipeline and establish a pre-training task aimed at aligning the visual features of cropped word images with the LLM space. This approach enables the LLM to effectively understand and process visual information. Finally, we explore two novel methods for performing DocVQA by utilizing the visual embeddings of words. These approaches represent initial steps toward developing a comprehensive and robust solution for addressing this challenging task in an end-to-end manner.

JÚLIA ARIADNA BLANCO ARNAUS

Estimation of 3D Shape and Volume of Fire Plumes from Multiple Views

Supervisors: Josep R. Casas (UPC), Montse Pardàs (UPC)

Abstract:
An open-ended topic in wildfire analysis is obtaining and interpreting the characteristics of fires to reduce the risks. In particular, capturing the 3D spatial structure of smoke plumes and their temporal evolution can be a valuable reference to mitigate the impact of wildfires. Nowadays, 3D reconstruction methods tend to rely on textures, which presents a challenge when reconstructing surfaces that are semi-transparent or without many key-points, which is the case of smoke. This paper showcases two strategies to deal with this problem depending on the resources and number of views available: a classic shape-from-silhouette and a deep learning approach. We introduce a set of experiments to assess both solutions qualitatively and quantitatively with real and synthetic images. Results show how the deep-learning approach can outperform classic shape from silhouette without the need for a precise camera calibration stage.

AYAN BANERJEE

Enhancing Document Layout Analysis using Transformer-Based Semi-Supervised Learning and Graph-Based Knowledge Distillation

Supervisors: Josep Llados Canet (UAB), Sanket Biswas (CVC-UAB)

Abstract:
Instance-level segmentation of documents consists of assigning a class-aware and instance-aware label to each pixel of the image. It is a key step in document parsing for their

understanding. In this thesis, we present a unified transformer encoder-decoder architecture for end-to-end instance segmentation of complex layouts in document images based on only visual features. The method adapts a contrastive training with a mixed query selection for anchor initialization in the decoder. Later on, it performs a dot product between the obtained query embedding and the pixel embedding map (coming from the encoder) for semantic reasoning. To generalize the method, we have proposed a co-occurrence-guided semi-supervised strategy where the architecture can learn about the novel classes through the support set in one-stage training. However, the use of transformers is not feasible in real-time scenarios, due to the large number of model parameters. To optimize the number of model parameters without compromising the performance, we present a graph-based knowledge distillation strategy that reduces the number of model parameters from 223M to 44M. Extensive experimentation on competitive benchmarks like PubLayNet, PRIMA, Historical Japanese (HJ), TableBank, and DocLayNet demonstrate that our model with SwinL backbone achieves better segmentation performance than the existing state-of-the-art approaches with an average precision of 93.72, 54.39, 84.65, 98.04, and 76.85 respectively. The code is made publicly available at: github.com/ayanban011/SwinDocSegmenter.

17:00 – 17:45

ALVARO FRANCESC BUDRIA FERNANDEZ

3D Human Avatars with Accurate Geometry from Monocular Video

Supervisors: Francesc Moreno-Noguer (UPC)

Abstract:
Enabling the reconstruction of detailed 3D avatars from readily available "in-the-wild" videos, like those captured on smartphones, holds immense potential across various domains, including augmented reality (AR), virtual reality (VR), human-computer interaction, robotics, and entertainment. Traditional methods rely on costly calibrated multi-view systems, which are impractical for widespread use, while current techniques based on templates or explicit mesh representations struggle with generalization and temporal coherence. Addressing the growing needs of applications like the Metaverse demands more pragmatic, robust, and broadly applicable solutions. Recent advances in neural fields have paved the way for techniques like InstantAvatar, which reconstructs animatable avatars from monocular videos. However, these methods primarily focus on view synthesis, lacking accurate 3D geometric representation. This thesis introduces a novel approach that enhances InstantAvatar by incorporating geometric awareness through a volumetric rendering formulation (VolSDF). This allows for precise body shape representation and fine-grained surface details, including clothing wrinkles. Additionally, our proposed regularization scheme ensures geometric consistency, removing artifacts and enhancing surface smoothness. The presented method achieves competitive geometry reconstruction and rendering quality in a few minutes of training time. We evaluate our method on synthetic and real-world videos, demonstrating its robustness even under challenging pose changes and showing the effectiveness of the proposed regularization scheme.

RAZVAN-FLORIN APATEAN

Improving Zero-shot Composed Image Retrieval for Images with Multiple Objects

Supervisors: Lluís Gómez Bigorda (UAB)

Abstract:
In this project, we study the task of composed image retrieval (CIR), where the input query is specified in the form of an image plus some text that describes desired modifications to the input

image. Our specific focus lies in the Zero-Shot setting, where we aim to construct a CIR model without requiring labeled data for training. The current state-of-the-art approach for this task is based on pre-trained CLIP image/text encoders and a mapping network that translates images into a single text token. We show that this approach has some limitations when the input image contains more than one object and propose a modification to solve this problem. Our solution involves splitting the input image into a set of crops, each corresponding to a separate object, and then consolidating their respective text tokens into a unified query.

JIA QIANG YE ZHU

In-context Learning for Robotics Control with Feedback Loops

Supervisors: David Vázquez (UAB)

Abstract:
This paper introduces a pioneering research study in the field of robotics, large language models, and computer vision. The study proposes an innovative system designed to address the perception, control, and complexity limitations that are currently prevalent in robotics technologies. The system is architecturally designed with several critical components: a visual perception model for environmental interpretation, an object identification and localization module, and a large language model for translating high-level commands into low-level instructions for the robotic unit. A distinctive feature of the system is the integration of a feedback loop mechanism. This mechanism operates by sending low-level instructions to the robot multiple times per second and iteratively validating task completion, thereby significantly enhancing the system's operational efficiency and reliability. A significant innovation of this system is the closed-loop feedback mechanism, which allows the robot to dynamically react to changes in the positions of objects in its environment. This feature enables the robot to adjust its actions in real-time and successfully complete tasks even in the face of unexpected changes. This research can represent a significant advancement in Robotics, offering a comprehensive solution to some of the most pressing challenges in the field. The findings of this study have broad implications for the future of Robotics, paving the way for more reliable, efficient, and intelligent autonomous systems that can adapt to dynamic environments.