ACMCV'2021

8th Annual Catalan Meeting
on Computer Vision

MASTER THESIS PRESENTATIONS | SCHEDULE

8:00 – 8:45

ADITYA SANGRAM SINGH RANA: "Event Detection in Football using Graph Convolutional Networks 3.1"

- **Supervisors:** Francesc Moreno-Noguer & Antonio Rubio Romano
- **Abstract:**

The massive growth of data collection in sports has opened numerous avenues for professional teams and media houses to gain insights from this data. The data collected includes per frame player and ball trajectories, and event annotations such as passes, fouls, cards, goals etc. Graph Convolutional Networks (GCNs) have recently been employed to process this highly unstructured tracking data which can be otherwise difficult to model because of lack of clarity on how to order players in a sequence and how to handle missing objects of interest. In this thesis, we focus on the goal of automatic event detection from football videos. We show how to model the players and the ball in each frame of the video sequence as a graph, and present the results for different types of graph convolutional layers and losses that can be used to model the temporal context present around each action.

8:45 – 9:30

CLARA GARCIA MOLL: "Beyond one-meter resolution: A comparative study on super-resolution methods"

- **Supervisors:** Felipe Lumbreras Ruiz & Javier Marin Tur
- **Abstract:**

Super Resolution (SR) offers a great opportunity to improve certain remote sensing applications since these techniques deal with increasing the resolution providing much more details in the images. Moreover, further techniques were developed due to the recent important breakthrough in deep convolutional neural networks (CNNs). In this work, we conduct a comparative study of the most recent SR methods on High Resolution satellite images. Specifically, we compare them at different scale factors, one-meter resolution being our starting point. Among the different methods we assess a contestant clearly stands above the rest qualitatively and quantitatively (PSNR, SSIM, SWD, FID). Additionally, aiming at improving the PSF and by doing so easing the resolving task, we incorporate a deblurring layer as a pre-stage. The resolved images yielded by the model when using such a layer outperform the original ones. A full detail of the experimental setup and data preparation are provided. It is our hope that this study provides a better insight on the latest advances on Super Resolution when using satellite imagery.

POL ALBACAR FERNANDEZ: "Switching furniture in room images using Generative Models"

- **Supervisors:** Javier Ruiz Hidalgo
- **Abstract:**
  This thesis aims to explore possible ways of automatically placing the desired furniture seen on a single image into room scenes using Generative Adversarial Networks (GANs). GANs have made great success in synthesizing high-quality images; however, how to manage the synthesis process of these models and customize the output image is much less explored. It has been found that modulating the input latent space of the generator can modify factors in the generated image, but such manipulation changes the entire image. In this work, two approaches are presented, both using Deep Learning Generative Models to automatically control locally the generated scenes. The first approach is based on taking advantage of the Visual Features learned from solving the task of image generation. The second approach is based on a novel generative architecture called ADGAN, where the core idea is to embed room elements into a latent space as independent codes and then achieve control of these elements by mixing their latent spaces to feed the generator through a block called Decompose Component Encoder. Visual results and comparisons between methods are presented.

ELOI BOVÉ CANALS: "SynchroSem: Loosely coupled multi-modal semantic mapping for synchronized Lidar-Visual-Inertial Systems"

- **Supervisors:** Josep R. Casas & Bharath Sankaran
- **Abstract:**
  We present SynchroSem: a loosely coupled multi-modal semantic SLAM approach for synchronized Lidar-Visual-Inertial (LVI) systems. Our system integrates pose measurements from two different tightly coupled odometry systems, Visual-Inertial (VI) and Lidar-Inertial (LI). These are loosely coupled by means of a pose graph representation that enables a robust late-fusion of pose measurements, being resilient to geometric and visual degeneracy. The developed multi-modal semantic feature

extraction allows the global optimization system to reduce long term operation errors via loop closure, achieving State-of-the-Art performance on public LVI datasets.Our method has also been deployed on a custom robotics platform. We developed a simple and reproducible hardware synchronization system using commercial-off-the-shelf hardware, and it is made publicly available at https://halops.github.io/SynchroSem.

9:30 – 10:15

MAR FERRER FERRER: "Fruit size estimation using Multitask Deep Neural Networks"

- **Supervisor:** Javier Ruiz Hidalgo & Jordi Gené Mola
- **Abstract:**

The measurement of fruit size is of great interest to estimate the ripeness of the crop and predict the harvest resources in advance. Non-invasive estimation of fruit size remains a challenging task that has to deal with occlusions, which may be caused by the foliage or shadows. This work proposes a novel technique for in-field apple measurement based on Deep Neural Networks (DNNs). The proposed framework has been trained with RGB-D data and consists of an end-to-end multitask architecture that performs the following tasks: 1) detect and segment every fruit from its surroundings; 2) estimate the diameter of each of the detected fruits. The network has been trained to perform instance segmentation and amodal segmentation, which when combined, allow us to see the relation between the occlusion percentage of the apple and the error of the diameter estimation. Our presented model is based on the Mask R-CNN architecture, which was extended in order to achieve all of the required tasks. This methodology was tested with a total of 2491 apples at different stages of growth, with diameters varying from 27 mm to 95 mm. We obtained a F1-score for instance segmentation of 0.78 and a mean absolute error of the diameter estimation of 6.8 mm. These state-of-the-art results show the potential of Deep Learning for fruit size estimation tasks.

CARMEN GARCÍA NOGUEIRAS: "Segmentation and mosaicking of intraoperative fetoscopic video for fetal surgeries"

- **Supervisors:** Mario Ceresa
- **Abstract:**

Twin-to-twin transfusion syndrome (TTTS) is a pregnancy condition affecting identical twins. In these cases, the blood flow among the twins is uneven. The most common treatment consists in photocoagulating the abnormal connections (anastomoses). During this intervention, a fetoscopic camera inspects the placenta's surface. This device has a limited field of view (FoV) and manoeuvrability. In addition, the quality of the obtained images is poor due to the lower resolution of the cameras and the amniotic

fluid turbidity. To tackle this, we propose a mosaicking algorithm supported by a deep semantic segmentation of the placenta. Finally, we present qualitative and quantitative results and discuss their suitability within the frame of ASTRA objectives.

JOSEP BRUGUÉS I PUJOLRÀS: “**Scene Text Visual Question Answering and Visual Question Generation: A multilingual approach**”

- **Supervisors:** Lluís Gómez i Bigorda & Dimosthenis Karatzas
- **Abstract:**
  During the last decade, Visual Question Answering (VQA) and Visual Question Generation (VQG) deep learning architectures have become trending topics in the Computer Vision community. Such models have a big potential for many types of applications, but lack the ability to perform well on more than one language at a time due to the lack of bilingual and multilingual data and the use of monolingual word embeddings in training. In this work, we hypothesise the possibility to obtain bilingual and multilingual VQA and VQG models. In that regard, we use already established models that use monolingual word embeddings as part of their pipeline and substitute them for FastText and BPEmb multilingual word embeddings that have been aligned to English. We employ the EST-VQA dataset in Chinese and English, and the ST-VQA dataset, which has been translated from English to other languages. On the one hand, we demonstrate that it is possible to obtain bilingual and multilingual VQA models with a minimal loss in performance in languages not used during training. On the other hand, we show that we can generate questions in multiple languages with a single VQA model.

IAN PAU RIERA SMOLINSKA: “Pedestrian Detection from 3D Geometry in High Density Point Clouds”

- **Supervisors:** Josep R. Casas & Santiago Royo
- **Abstract:**
  With the irruption of autonomous driving in recent years, pedestrian detection has gained momentum. Despite the maturity of 2D detection on RGB images, there has been a tendency to add 3D sensors, such as Light Detection and Ranging (LiDAR), to complement data in situations where the 2D approach fails due to environment conditions. 3D detection datasets are still under construction and not as mature as for 2D object detection, and most of the state-of-the-art architectures rely on a single dataset: Kitti. Besides, the high computational cost of point cloud processing, caused most of the approaches to either exploit the 2D detection and back-project it to the point cloud, or to reduce its size by means of grouping points into voxels. The aim of this project is to explore how the characteristics of the input data can affect the performance of 3D detection, using solely the geometric information. To that end, we exploit PointRCNN, a point-based architecture that performs object detection directly over the raw point cloud data. We have annotated a pedestrian-oriented dataset captured with a L3CAM sensor from Beamagine, that provides a high density point cloud. The sensor also provides a synchronous RGB capture that helps in the annotation process. In this project we

compare the detection results obtained using PointRCNN on the Kitti and Beamagine datasets.

10:15 – 11:00

AITOR SÁNCHEZ ABELLÁN: "Exploiting new modalities with CLIP"

- **Supervisors:** Pau Riba & Pau Rodriguez
- **Abstract:**
  Typical vision datasets are labor intensive and costly to create while teaching only a narrow set of visual concepts. CLIP aims to solve this problem with a model which efficiently learns from natural language supervision using image-text pairs. In this work we extend CLIP to operate using new modalities such as audio or video. Our proposal is generalized to work with n streams of data and makes use of the VGG-Sound audio-visual dataset to train these new branches. New modality retrieval such as text to video or audio to image is incorporated. Moreover, the expected improvement on retrieval performance after the alignment of the new latent space is tested and analyzed.

ÒSCAR LORENTE COROMINAS: "Multi-view 3D People Reconstruction combining Parametric and Non-parametric models"

- **Supervisors:** Xavier Giró-i-Nieto & Francesc Moreno-Noguer
- **Abstract:**
  3D reconstruction of human bodies from multiple images has been a long-standing problem in computer vision. It is typically addressed using statistical models of the human body, which describe the geometry by a small number of parameters encoding 3D pose and shape. Non-parametric representations are alternatives that gain expressiveness for cloth capture, but have difficulties in recovering reasonable 3D human shapes when camera views are too sparse. In this dissertation, we aim to leverage the advantages of parametric and non-parametric models by extending the parametric Skinned Multi-Person Linear Model (SMPL) with Implicit Differentiable Renderer (IDR), an architecture that implicitly represents the geometry as a zero level-set of a neural network. The neural surface of IDR is typically initialized as a sphere, which allows rendering objects of all types. However, our work focuses on the reconstruction of human bodies, so we explore the contribution of parametric 3D human models such as SMPL as priors. The evaluation has been performed on a subset of the Renderpeople dataset, using as metrics for 3D reconstruction the Chamfer-L1 and point-to-surface distances, as well as PSNR for the corresponding renderings. The obtained results confirm that in scenarios where the camera views are too sparse, using an SMPL model as a prior improves 3D reconstruction and accelerates convergence. Finally, we propose a strategy

based on an attention mechanism for IDR to improve the results on the head of the person, where the original IDR pipeline struggles to achieve a detailed reconstruction.

<u>SIDDHANT BHAMBRI:</u> "Interpretability of Deep Neural Networks In Continual Learning Settings"

- **Supervisors:** Bogdan Raducanu
- **Abstract:**
  Deep learning (DL) using deep neural networks (DNNs) and artificial intelligence (AI) is making our life easy. Today AI is playing crucial roles in one of the most important domains of intelligence which are autonomous driving and medical imaging. We can see in the coming future a very bright and rigorous use of deep learning in almost all sort of domains. However, the black-box nature of DNNs has become one of the primary obstacles for their wide acceptance in mission-critical applications such as medical diagnosis and therapy. Due to the huge potential of deep learning, interpreting neural networks has recently attracted much research attention. In this work we are going to investigate and study the deep aspects of these black boxes (artificial deep neural networks) in the domain of continual learning (CL) settings and try to visualise and analyse the different aspects of the functioning and behaviour of these architectures.

<u>ANTONI RODRÍGUEZ VILLEGAS:</u> "End-to-end License Plate detection and recognition"

- **Supervisors:** Marçal Rossinyol & Dimosthenis Karatzas
- **Abstract:**
  Automatic Number Plate Recognition (ANPR) systems are an active research topic due to their many applications and the present availability of data and high computational power.Current ANPR systems consist of two main steps: detection and recognition. The quality of the detections affect greatly that of the recognition, since the image is cropped based on what is believed to be the area of interest. Also, these systems tend to be computationally exhaustive, forcing them to be deployed in dedicated and expensive hardware.In this work, we aim to develop an efficient and lightweight end-to-end trainable model that is able to detect and read a license plate in a single step, eliminating detection-recognition dependencies. For this purpose, we created a neural network architecture with independent branches for each task that can be trained end-to-end. Finally, we quantised and tested this model in order to be deployed in on-the-edge devices.

11:00 – 11:45

KHANH NGUYEN VAN: "Multi-modal Image Captioning in Wikipedia"

- **Supervisors:** Ali Furkan Biten, Andres Mafla Delgado & Dimosthenis Karatzas
- **Abstract:**
  Typical image captioning systems operate solely on a set of visual object features and its relationships, whereas the contextual information or prior knowledge in the world are still totally ignored. Such incorporation, by any chance, are crucial as they can serve as a plentiful source of valuable information for the model to exploit in order to produce higher-quality descriptions. With this in mind, we aim to build a captioning model that is able to interpret the scene with contextual information such as Named Entities taken into account. More specifically, in this work we focus on the problem of generating captions for images contained in articles. We propose a novel model that extend the classic sequence-to-sequence attention model in image captioning: Show, Attend and Tell in two aspects: (1) Enriching the model's input with data from different modalities and analyzing their semantic correlation to improve the generative capability and (2) Employing a hard-attention mechanism to adaptively copy words from the source text via a pointer network, which allows the network to handle of out-of-vocabulary words. Furthermore, we introduce Wiki Dataset, a dataset for the multi-modal image captioning task and report experimental results with our model applied to it.

GERMAN BARQUERO GARCÍA: "Behavior forecasting for social interactions: a multimodal skeleton-based approach"

- **Supervisors:** Xavier Baró, Sergio Escalera & Cristina Palmero
- **Abstract:**
  Many works focus on predicting the motion or trajectory of individuals engaged in a particular action, which intends to reduce the inherent stochasticity of the future. We open a new horizon by aiming at forecasting human behavior in dyadic interactions. In such scenarios, the ability to anticipate human behavior implies an implicit knowledge of the underlying mechanisms of communication involving cognitive, affective, and behavioral perspectives. This knowledge is key for many applications in robotics, medicine and psychology. In this work, we introduce an extended version of the UDIVA dataset which contains automatically extracted face, body and hands landmark annotations for 145 dyadic sessions among 134 participants. We use it to deeply analyze the current limitations of interaction forecasting, most of them derived from the multimodal nature of the future and the huge dimensionality attached to human behavior. In parallel, we propose a multimodal recurrent model based on the popular seq2seq model, which serves as a baseline for future research on this topic. Finally, we present an ablation study to discuss the effects of leveraging multimodal data such as audio and participants metadata.

VERNON STANLEY ALBAYEROS DUARTE: "StyleArm: a style-transferring robot arm"

- **Supervisors:** Fernando Vilariño
- **Abstract:**
  As the paradigm of human-computer interaction shifts to increasingly intelligent systems that require no direct user inputs to provide services, the way we interact with our machines is still primarily "active interactions", as the computer requires some sort of direct user input to provide its content. In this master's thesis, we create a physical proof of concept that implements computer vision algorithms to aid in interaction, in the form of a self-built robotic arm with a camera that is able to interface with a Raspberry Pi, a small computer. The proof of concept utilizes a server-client connection with a higher-powered machine capable of relaying images treated by a fast Style Transfer GAN in near real-time, switching the context of the style transfer depending on the facial expression of the user. The goal of this robot is to provide the user with a way to interact with computer vision processes they might find useful, like Style Transfer or image stitching for story-telling and publication on social media accounts. Initially, this project was meant to optimize a style transfer GAN for use on a Raspberry Pi and provide a completely autonomous project, but we were unable to produce results that could be used in real-time.

DHANANJAY NAHATA: "Applications of Continual Learning Settings in Medical Imaging"

- **Supervisors:** Joost Van De Weijer
- **Abstract:**
  The application of Artificial Intelligence (AI) has tremendously impacted our lives, which has led to rapid demand and usage of technologies in various domains such as classification, object detection (OD) and recognition, etc. Deep neural networks (DNNs) had already achieved a significant milestone in tackling the classification problems in classifying a certain task, where it generally learns from well-defined phase, thus acquiring knowledge of that task only on which it is trained, but these architectures face the issue of the catastrophic forgetting, where it fails to extend the knowledge of the previous task when trained on a new task. Incremental learning (IL) tries to address this issue by accommodating the knowledge of the previous tasks and current tasks continually, without training the model from the scratch for different tasks. In this work, we are trying to address classification problem task incremental learning on the medical images of breast cancer and rectal colon using various state-of-the-art methods, where we define certain tasks and the network tries to learn each task sequentially.