ACMCV'2022

9th Annual Catalan Meeting
on Computer Vision

MASTER THESIS PRESENTATIONS | SCHEDULE

14:00 – 14:45

ALEX MARTIN MARTÍNEZ: "Lightweight neural network architectures for on-device object detection"

- **Supervisors:** Maria Vanrell Martorell (UAB)
- **Abstract:**
  The task of object detection in computer vision gets much interest in the industry since its applications can lead to saving time, costs and reducing human errors in many scenarios. The deployment of these models is mostly made using cloud systems, which can lead to security and latency problems. The alternative to the cloud is the usage of edge AI, which is the deployment of AI applications throughout physical devices without the necessity of a connection to the internet. This approach comes at the expense of restricting the set of models suitable for the final application, which vary depending on the technical specifications of the edge device being used to perform this type of deployment. In this work, we use the IMX500 device, which has a memory limitation of 8 MB and a maximum input resolution of VGA. In addition, it requires the model to be previously converted to a binary file through Sony's Digital Signal Processing tool, which imposes additional restrictions on the set of operations that can convert. The current lightweight state-of-the-art detectors are not sufficiently compact to be able to run on this device. Moreover, they are generally developed to work with smaller resolutions than the maximum input resolution the IMX500 can handle, which works against the on-device performance of the model in specific scenarios such as small object detection, which is the use case we are particularly interested in. To this end, in this work we first adapt a state-of-the-art lightweight object detector, the YOLONano, to fit into the IMX500 and then analyze how the model performance vary with the change of the input resolution, number of parameters and the different conversions steps. After carrying out these experiments using PASCAL VOC dataset for training and benchmarking, we obtain the two best performing models for 480x640 and for 288x384 input resolution. We then fine-tune these models on the Holoselecta dataset for the task of small object detection and observe an advantage in using a higher resolution.

YU PANG: "Electricity bill key information detection and recognition based on improved YOLOv5"

- **Supervisor:** Oriol Ramos (UAB)
- **Abstract:**
  With the prosperity of modern industry, commerce, and daily economic activities, more and more bills need to be sorted, manually entered and manually retrieved, which not only wastes time, but also is prone to errors. Therefore, with the help of the practice and application of artificial intelligence technology, the automatic recognition of bills is an effective way to improve work efficiency. In this work, we will perform intelligent

recognition of key information of electricity bills. We propose an implementation route that is different from the commonly used OCR method, that is, to adopt the YOLOv5 algorithm, which is currently cutting-edge in terms of both speed and accuracy, as SOTA to complete the detection of field targets, and on this basis, we incorporate two attention mechanisms to enhance its detection capability for small targets and targets with large aspect ratios. Relevant image annotation and data verification are also carried out, and then combined with the CRNN text recognition algorithm to perform text recognition on the detected key field information. The text detection model in this paper achieves an accuracy rate of 90.8%, a recall rate of 92.1%, and a text recognition model accuracy rate of 96%. After testing, the system we implemented can help users to quickly extract key information from electricity bill images.

14:45 – 15:30

FRANCESC NET BARNÉS: "Fruit detection and tracking in RGB-D videos"

- **Supervisor:** Jordi Gené Mola (UdL) & Ramon Morros (UPC)
- **Abstract:**
  In recent years, Precision Agriculture has become a useful option to improve agricultural productivity: increase production and reduce labor time. Thanks to the arise of Deep Learning and the large amount of data currently available, video fruit tracking algorithms are gaining importance for yield prediction and yield mapping. Subsequently, more precise solutions involving fruit-counting methods are being implemented with respect to the classical methods launched in previous years. In this work, an apple counting model based on Deep Learning techniques using object detection and tracking algorithms is proposed in order to face problems such as fruit occlusions. In addition, a modification of the state-of-the-art methods is proposed by including multi-modal (color, depth and IR) videos obtained with an RGB-D camera. Finally, a comparative study of the tested methods is carried out, concluding that the combination of YOLOv5x with ByteTrack outperformed other methods in terms of accuracy and inference speed (MOTA=0.6817, 15.31 frames/second), and that the use of multimodal data (depth and IR) did not add a significant value for tracking fruits in videos.

JUAN ANTONIO RODRÍGUEZ: "Text to Figure Generation"

- **Supervisors:** David Vázquez (UAB) & Pau Rodriguez (UAB)
- **Abstract:**
  Synthetic image generation has recently experienced significant improvements in domains such as natural image or art generation. Multi-modal conditioning mechanisms in image generation have shown high potential for creating complex compositions while delivering impressive realism and quality. However, the problem of figure and diagram generation remains unexplored. A challenging aspect of generating figures and diagrams is effectively rendering readable texts within the images. To alleviate this problem, we present OCR-VQGAN, an image encoder, and decoder that leverages OCR pre-trained features to optimize a text perceptual loss, encouraging the architecture to preserve high-fidelity text and diagram structure. Furthermore, we investigate recent text-to-image generation methods to create figures that match text descriptions in research papers. Unfortunately, existing datasets of figures do not offer enough data to train text-to-figure generation methods. To overcome this problem, we introduce the

Paper2Fig100k dataset, with over 100k images of figures and texts from research papers. Images of figures show architecture diagrams and methodologies of research papers available at arXiv.org from fields like artificial intelligence and computer vision. Figures usually include text and discrete objects, e.g., boxes in a diagram, with lines and arrows that connect them. Furthermore, we include detailed text descriptions of figures extracted from the paper, which allows for text-conditional figure generation. We show the effectiveness of OCR-VQGAN by conducting several experiments on the task of figure and text-in-the-wild reconstruction. We explore the qualitative and quantitative impact of weighting different perceptual metrics in the image encoder loss. Finally, we present results on the task of text-to-figure generation, using both autoregressive and diffusion-based approaches.

ERIC HENRIKSSON MARTÍ: “Lightweight Monocular 3D Vehicle Detection in Calibrated and Uncalibrated Scenarios”

- **Supervisors:** Dimosthenis Karatzas (UAB)
- **Abstract:**
  This paper explores the 3D vehicle detection problem based exclusively on monocular images, placing emphasis on creating the simplest possible yet functional models for the task. The main concept behind the CenterNet system is taken as a foundation for this purpose, involving the detection of vehicle centers and the regression of other parameters that provide further location and pose information. Several lightweight architectures with U-Net-like feature extractors are tested to implement this concept in a simplified manner, which use only a fraction of the parameters that constitute the original CenterNet model. A case with calibrated cameras is first addressed, where several model variants capable of predicting the location and orientation of 3D bounding boxes enclosing vehicles are designed for the KITTI dataset. The feasibility of estimating this same information when dealing with uncalibrated cameras is later explored through the simplification of the 3D detection problem into 2D. Images from the Ko-Per traffic monitoring dataset are first transformed into their bird's-eye view (BEV) versions using homographies, and vehicles are then detected as rotated bounding boxes in a 2D plane.

15:30 – 16:15

IGNACIO GALVE: “3D Apple segmentation and measurement in large unstructured point clouds”

- **Supervisors:** Javier Ruiz Hidalgo (UPC) & Veronica Vilaplana (UPC)
- **Abstract:**
  In recent years, there has been a rapid increase in the number of publications related to the field of agriculture. The advancement in deep learning for image and point cloud processing for mapping orchards or quality control of cattle has led to various models being adapted to each specific problem with highly accurate results. In this work, we propose an end-to-end trainable model for instance segmentation and in-field size estimation. The presented model works with large unstructured RGB point clouds obtained with a Structure from Motion algorithm. The dataset employed for training and testing is from two apple orchards captured in different stages of growth, representing different colors and sizes. Two main approaches are presented to carry out the size estimation: 1) a subnetwork with a point encoder followed by a multilayer perceptron

(MLP), and 2) an MLP fed with the features per instance generated by the segmentation model. Benchmarking with mean absolute error, the results obtained are precise, with an error of 4mm. However, all the trained models failed to establish a good correlation between the predicted sizes and the ground truth indicated by a low coefficient of determination (R squared).

<u>DAVID SERRANO LOZANO:</u> "Uncertainty as a Proxy of the Generalization Error for Marine Species Identification"

- **Supervisors:** David Masip (UOC)
- **Abstract:**
  Marine Protected Areas monitoring is a must to understand ecological processes and assess whether management aims are fulfilled. One of the best ways of doing it is by using a Remotely Operated Underwater Vehicle to collect images. However, the main drawback is the large amount of data that has to be annotated by specialists. In this thesis, we propose to go one step further and use a deep learning system to maximize the system's performance while reducing the human workload. The algorithm reports, in addition to the deterministic decision, uncertainty estimations to identify potential misclassifications. However, evaluating the model doubtfulness is not trivial and, therefore, we test several well-known and a novel metric which evaluates the quality of the estimations regarding its ranking. Furthermore, we propose a systematic method to reduce the workload from non-annotated datasets, using uncertainty as a proxy of the generalization error and automatically labelling with the model the most certain samples.

<u>BRIAN GUANG JUN DU:</u> "A visual and Semantic Two-Stream Architecture for Visual Sentiment Recognition"

- **Supervisors:** Àgata Lapedriza Garcia (UOC)
- **Abstract:**
  This thesis presents a two-stream architecture for Visual Sentiment Recognition by introducing semantic information to visual-based methods such as convolutional neural networks (CNN). The objective of this architecture is to explore external knowledge through the use of Visual Question Answering (VQA) together with language-based transformers (semantic stream), and how this information can help improve basic CNNs (visual stream) to archive state-of-the-art performance. The semantic stream adopts a VQA algorithm to extract semantic features through a series of questions; that is comparable to image captioning coupled with other features such as locations, facial expressions, and more. An extensive set of experiments showed that external knowledge is able to improve the basic visual methods by bridging the affective gap between pixel-level images and the underlying sentiment, achieving equal or better results when compared to the state-of-the-art deep learning models for Visual Sentiment Recognition.

<u>MERT YAZAN:</u> "Automatic dataset update via instance relevance assessment"

- **Supervisors:** Joost van de Weijer (UAB)
- **Abstract:**
  Every Machine Learning application is at risk of suffering from data drift. It is a phenomenon that can happen in many forms, and sometimes it can even be impossible to foresee. The high variability of the problem makes it particularly hard to come up with standardized ways to handle it. As a result, even though it is a very important problem

that affects the whole AI community, research done in this area is very scarce. With this project, a solution that can detect different types of drift in images and solve them by using as little data as possible while working completely unsupervised is proposed. We created a simulation environment that evaluates the performance of the solution before it gets deployed in production. A new type of drift which is called semantic shift is introduced, where the shift is between intra-class distributions. We are also proposing a new way to simulate drift by creating subsets from the dataset with different distributions, and a new approach to detect it by using batches. Furthermore, since we are proposing to work unsupervised, we investigated redundancy removal methods to limit the number of samples that will be used to solve drift to reduce the annotation cost.

KEVIN MARTÍN FERNÁNDEZ: "Automatic detection of fiber-cement roofs in aerial images"

- **Supervisors:** Àgata Lapedriza Garcia (UOC)
- **Abstract:**
  Asbestos is a fiber cement harmful to health. For this reason, almost all countries have laws to eliminate this material. Unfortunately, asbestos detection is a challenging task. The current procedures for identifying asbestos require human exploration, which is costly and slow. This has motivated the interest of governments and companies to develop automatic tools that can helps to detect and classify these types of materials that are dangerous to the population. This paper explores multiple computer vision techniques based on deep learning for the automatic detection of asbestos in aerial images. Concretely, we trained and tested implementations of Mask-RCNN, ResNet and Embedding spaces, and we used data augmentation and weighted sampling to overcome the data scarcity for training. The obtained results are 86.55% average of correct asbestos detection in the testing set, which shows the potential of the computer vision deep learning techniques for the automatic detection of asbestos in aerial images.

16:15 – 17:00

LAIA ALBORS ZUMEL: "A method to minimize human interaction in labelling images for flora and fauna visual identification"

- **Supervisors:** Ferran Marqués Acosta (UPC)
- **Abstract:**
  Rainforests are crucial for the environment and for the great biodiversity they contain, so it is urgent to preserve this ecosystem, increasing the knowledge we have of its fauna and flora. For this, we need to be able to quickly and easily identify the species that live there. Current image classification systems have proven very successful in this identification task, even with several classes, but they are very data-hungry. The main problem is the lack of labeled images and the high cost of labeling images at the species level, even more for rainforest species. In this work, we propose a method to help experts in this labeling task by creating a deep learning network that maps images into a feature space that clusters them by species (and other taxonomic levels), so that only one image per cluster needs to be tagged. The model is trained and evaluated on a subset of the PlantCLEF 2022 dataset.

JOSÉ MANUEL LÓPEZ CAMUÑAS: "Non-verbal Synchrony Analysis in Dyadic interactions"

- **Supervisors:** Àgata Lapedriza Garcia (UOC)
- **Abstract:**
  In the context of dyadic interactions, synchrony refers to a state of harmony in verbal and non-verbal communication signals of the two interactants. Synchronic communication naturally occurs when the interactants share common cultural norms, particularly in contexts of cooperation. Thus, automatically characterizing synchrony during a dyadic interaction is essential to richly understand the nature of the interaction. The goal of this project is to analyze the synchrony on different social signals and behaviors in dyadic interactions. We propose a novel method to quantitative measure the non-verbal synchrony in the interaction. Additionally, we study the impact of the different signals in our features, the relationship between the interactants and the nature of the interaction.

SERGIO MONTOYA DE PACO: "Fine 4D Neural Models from Uncalibrated Videos"

- **Supervisors:** Antonio Agudo (UPF)
- **Abstract:**
  Current NR-SfM methods are limited in not providing a fine and dense reconstruction with images recorded from a single monocular camera. In this work, we take advantage of the latest state-of-the-art research in NeRF and non-rigid body priors to improve the results of NR-SfM works in long sequences with large deformation. NeRF-based NR-SfM is a recent idea that has not been explored in-depth so far, which we decide to explore due to its scene representation potential. We propose to separate deformation into coarse and fine deformations, which are more adequate to represent articulated objects. Previous methods only consider a Linear Blend Skinning model or an as-rigid-as-possible assumption, therefore being very limited in the deformations they can model. Introducing fine elastic deformation refinements over the coarse estimation allows the network to correctly refine the coarse shape, modeling finer details of the geometry. We obtain better qualitative and quantitative results by following this approach compared to previous work. Our work also shows better view synthesis in rendered images from the learned Neural Radiance Field.

PAU TORRAS COLOMA: "End-to-End Symbolic Music Recognition"

- **Supervisors:** Alicia Fornés (UAB) & Sanket Biswas (CVC-UAB)
- **Abstract:**
  The field of Optical Music Recognition has seen a surge in performance thanks to the most recent advances in Computer Vision and Deep Learning as a whole. Nevertheless, the application of these advances in real-case scenarios for domains other than typeset music scores is rather complicated, as there is very limited amounts of usable musical data. In this work we propose tackling OMR as a single-step process from image to notation reconstruction, with the aim of avoiding intermediate targets and using existing transcriptions as output, provided that the image contents and the transcriptions can be aligned. We propose a notation format based on a tree-like scheme that can be inferred using sequence-to-sequence models, which have already shown to be performant in similar image transcription tasks. This tree-like nature can also be exploited by increasing the level of abstraction in the sequence left to right, inducing the model to focus on

actual primitives on the score first and then organising these primitives into higher-order compounds.

<u>VÍCTOR UBIETO NOGALES:</u> "A novel Learning Database for Sign Language Animation Synthesis from Quaternion estimations"

- **Supervisors:** Coloma Ballester (UPF) & Gloria Haro (UPF)
- **Abstract:**
  Avatar synthesis is one of the most important and challenging tasks when it comes to sign language synthesis. The scarce, useful data in the field and its multidisciplinary requirements, make this task very challenging and without many solutions. In this project we will present a system to generate sign animations from sign language videos. In particular, a novel automatic system to generate a dataset will be proposed and implemented. This approach is divided in two steps, a pose estimation, and a method to create the animations in a BVH (Biovision Hierarchy) file. Unlike the others, the presented approach estimates the rotations of each joint in quaternions instead of estimating the 3D absolute position of the joints, which are directly used to represent virtual animations. Finally, an evaluation of different state of the art approaches of generating realistic sign language animations will also be presented and compared to the presented method, both qualitatively and quantitatively.

17:00 – 17:30

<u>MARCOS MELGAR SEGOVIA:</u> "Dynamic Audio NeRF for 4D Facial Avatar Reconstruction"

- **Supervisors:** Gloria Haro (UPF)
- **Abstract:**
  During the last years, advances in computer science lead towards the possibility of capturing and synthesize 3D dynamic scenes. In particular, there an extensive research exists in the areas of human head avatars capture, with applications in Virtual Reality (VR), film and video-game industry, visual effects (VFX), chat-bots interaction and communication. Lower prices in GPUs and advances in deep learning had made neural rendering more accessible. Although, current state of the art is capable of synthesizing high quality images, they still cannot achieve photorealistic results. In this work, we present a model based on NeRF, we implemented a new loss that measures the lip synchronisation between rendered mouth image features and audio features. We also add a new facial expression database to the training stage in order to generalise for unseen expressions.

<u>MARCOS V CONDE OSORIO:</u> "Perceptual Image Enhancement for Smartphone Real-Time Applications"

- **Supervisors:** Javier Vazquez Corral (UAB)
- **Abstract:**
  Recent advances in camera designs and imaging pipelines allow us to capture high-quality images using smartphones. However, due to the small size and lens limitations of the smartphone cameras, we commonly find artifacts or degradation in the processed images. The most common unpleasant effects are noise artifacts, diffraction artifacts,

blur, and HDR overexposure. Deep learning methods for image restoration can successfully remove these artifacts. However, most approaches are not suitable for real-time applications on mobile devices due to their heavy computation and memory requirements. In this work, we propose LPIENet, a lightweight network for perceptual image enhancement, with the focus on deploying it on smartphones. Our experiments show that, with much fewer parameters and operations, our model can deal with the mentioned artifacts and achieve competitive performance compared with state-of-the-art methods on standard benchmarks. Moreover, to prove the efficiency and reliability of our approach, we deployed the model directly on commercial smartphones and evaluated its performance. Our model can process 2K resolution images under 1s without specific optimization for the mobile devices.

JOAN FONTANALS MARTÍNEZ: **"2-phase cross-modal search. Combining dual encoders with Vision and Language Transformers"**

- **Supervisors:** Dimosthenis Karatzas (UAB) & Lluís Gómez Bigorda (UAB)
- **Abstract:**

In the recent years, a lot of attention in the field of Information Retrieval has been put on the capacity to perform search between different modalities and specially between images and their associated texts. The most recent improvements on various multi-modal tasks involving images and textual information have been achieved by Vision-and-Language models, where the model needs to interact with both modalities at the same time in order to extract knowledge from the modality interactions, showing impressive results. However, the usage of these models to perform retrieval cannot scale since it is not possible to compute offline an index where to perform fast queries. In order to overcome these limitations, we propose to combine the training of two independent models for image and text and to combine its usage with the most advanced Vision-and-Language models for re-ranking in order to maintain the qualitative results while cutting down the processing time.