ACMCV'2020

7th Annual Catalan Meeting
on Computer Vision

MASTER THESIS PRESENTATIONS | SCHEDULE

8:30 – 9:15

AGUSTINA POSE: "**Automatic reseed area estimation from aerial images "**

- **Supervisors:** Felipe Lumbreras & Daniel Ponsa
- **Abstract:**
  The goal of this work is to create a set of computer vision and deep learning algorithms for automatically identifying reseed areas in crop fields, by using multispectral drone images as input. The research developed in this thesis is split in two different sections: automatic segmentation of objective vegetation (discriminating weeds and any kind of spurious vegetation) and automatic estimation of crop rows. Both problems were tackled by training a U-net, which yielded very good results. In order to overcome the lack of data and ground truth, a series of computer vision algorithms were also developed for performing an exhaustive process of data augmentation. By superimposing the obtained results (identification of vegetation / crop lines) we can identify the zones along crop rows with and without objective vegetation, being the latter the zones that should be reseeded.

JOAN FRANCESC SERRACANT LORENZO: "**Spatial-Temporal Graph Convolutional Networks for Nonverbal Language in Entrepreneurial Pitching Sessions"**

- **Supervisor:** Coloma Ballester
- **Abstract:**
  Nonverbal language plays a role when entrepreneurs pitch their business ideas to potential investors. Some authors have proposed that several crucial clues from nonverbal communication can affect on accessing early stage investments and even long-term firm survival. The aim of this project is to apply recent supervised deep learning strategies to this field of investigation. Our proposal leverages from a recent work [1] where the authors gathered a new dataset from entrepreneurial pitching sessions recorded on video and found a set of nonverbal language characteristics that strongly correlate to investment outcome and start-up performance in early stages. We use their dataset to extract skeletal information of speakers and learn patterns in human body poses that correlate with those nonverbal language characteristics. Inspired by the state of the art techniques in human action recognition, we propose a method that uses Spatial-temporal Graph Convolutional Networks, adapted to perform a regression instead of classification, given the nature of our data. The obtained results exhibit good performance with a mean average error of1.7%. Due to the lack of reference results on our custom dataset, we implemented other traditional machine learning techniques with hand-crafted features to compare our results with.

ROGER CASALS VILARDELL: "Conditional GAN for Synthetic Text Dataset Generation"

- **Supervisor:** Dimosthenis Karatzas
- **Abstract:**
  This project proposes a conditioned generative model to deal with the large data needs required to train a deep neural network for text recognition. We present a generative adversarial network (GAN) that defines the appropriate geometric corrections to apply to synthetic text in order to compose it with a background image, as well as making appearance modifications to make the composite look realistic. The geometric consistency is achieved through the utilization of a spatial transformer network (STN) while a second generator is responsible for the seamless integration of foreground and background, with both modules being connected such that end-to-end training can be conducted without supervision. The proposed GAN is evaluated on the task of synthesizing license plate numbers onto cars on real-life scenarios that are used to train a better text recognition model.

9:15 – 10:00

SERGIO CASAS PASTOR: "Domain Adaptation for object detection and semantic segmentation in autonomous driving"

- **Supervisor:** Antonio López
- **Abstract:**
  Capturing road-annotated data in real environments for autonomous driving is a process that wastes many resources. Collecting data in a virtual environment is a possible solution to overcome this limitation. However, data are not properly transferred between different domains due to variance in scene, objects and camera conditions. The work presented in this document uses two domain adaptation approaches, Cycle GAN and MUNIT, to adapt unlabeled unpaired images from a source domain to a target domain using CARLA simulator. To evaluate the performance of the adaptation, two autonomous driving sub-tasks are applied to the transformed images: road object detection and semantic segmentation. Results show that using adapted data has a significant performance increase in comparison to source domain data in semantic segmentation. Moreover, the adapted models sometimes segment certain elements of images more precise than the target domain model. Consequently, we obtain a good baseline to use data from the simulation in autonomous driving tasks without the need of obtaining annotated data in a real environment. The next stage of the project would include the use of domain adaptation for an end-to-end autonomous driving model.

YIXIONG YANG: "Deep Single-Image Relighting Based on Intrinsic Decomposition"

- **Supervisors:** Maria Vanrell & Hassan A. Sial
- **Abstract:**
  Scene relighting from a single image pursuits to solve the problem of generating a new image of the same scene under a different objective light provided as input. In this project, the aim is to explore different image-to-image neural networks and some physical properties to be used to solve single-image relighting. We explore how the network training is affected by 3 physical constraints: (a) the intrinsic decomposition of shading and reflectance of the input image; (b) the reflectance consistency of an image set of the same scene under different light conditions; and (c) the explicit estimation of

the input light properties. All the proposed architectures have been trained on a new version of the synthetic SID dataset and the quantitative evaluation does not allow us to state clear conclusions yet, since we need further analysis of the results and test on further datasets. However, from a qualitative point of view our approach is showing very promising results, thus it seems that adding physical constraints makes the performance of the networks more superior than a single encoder-decoder architecture.

KEYAO LI: "**Solving a classification task by spiking neural networks**"

- **Supervisors:** Xavier Otazu
- **Abstract:**
  In this thesis we try to solve a simple classification task using a biologically plausible computational mechanism. We use a spiking neural network (using NEST simulator software) whose learning process is the biological mechanism called Spike Time Dependent Plasticity (STDP). This mechanism changes the connections between nodes of the network depending on the temporal synchronization of spikes. Input data is encoded, similarly to biological cells, using Gaussian receptive fields. We applied this architecture to MNIST dataset, obtaining high accuracy classification results.

STANISLAW KAJETAN MORAWSKI: "Unsupervised Shakespeare Style Captioning"

- **Supervisors:** Dimosthenis Karatzas & Ali Biten
- **Abstract:**
  Most of the deep learning attempts at captioning result in factual, neutral and boring captions. This may be valuable in certain scenarios (like historical archives), but in the world of entertainment is a big limitation. In this work we have explored image captioning in poetic and dignified English, mimicking the written language of William Shakespeare. Using SemStyle as a baseline we have tried to improve upon the framework using a different style. Also we have trained the model to mimic the modern equivalent of Shakespearean style. Furthermore we have examined the differences in performance of our model when built on top of different CNN feature extractors.

10:30 – 11:15

MARIA VILA ABAD: "Instance segmentation annotation on construction site images"

- **Supervisors:** Luís Herranz
- **Abstract:**
  Construction is one of the world's largest but also most inefficient and wasteful industries. One of its biggest problems is that though we design digitally, we still construct manually. This causes the constructed structure to significantly diverge from the designed model. Digitally monitoring the evolution of construction sites can help closing the gap between virtual and real buildings as it allows us to quickly discover errors and deviations from the virtual model. Part of this monitoring consists on segmenting instances of building elements in images of construction sites to control their accurate matching with the virtual design. Deep Convolutional Neural Networks have been proven to succeed in this task but they need large annotated data to learn from. Manually labelling an instance segmentation dataset is costly and time-consuming and may not be viable in many occasions. In this thesis, we present two different procedures to obtain instance segmentation annotations at a reduced human supervision cost. The

first approach tackles the problem by using 3D information from the virtual building model and the construction site to automatically find correspondences between them in order to obtain instance segmentation labels. The second approach consists on a Weakly Semi-Supervised method that learns to predict instance segmentation masks from image-level labels (which are cheaper to annotate) and a small number of instance segmentation masks. We obtain promising results that could lead to significant savings in annotation costs.

SARA CELA ALFONSO: "A deep program learning model for code induction from screenshot images of GUIs"

- **Supervisors:** Pau Riba & Josep Lladós
- **Abstract:**
  This work proposes a deep learning based inductive programming method for graphic program synthesis. Automatic code generation is a useful way to assist developers, allowing them to design in a higher abstraction level, focusing their efforts on tasks that really require cleverness and creativity. In particular, in the case of web development, the automatic code generation is able to translate a graphical web design to a code that, when it is parsed, synthesises the original visual input. Graphical User Interface (GUI) generation for mobile apps is a particular scenario where interfaces are usually and structurally similar. The main contribution of this work is an attention-based model that generates a syntactically valid textual description of a screenshot of a GUI using a Domain Specific Language (DSL). This textual output, a valid program, is parsed and translated to an XML code ready to be rendered. Using as baseline the pix2code model [?], we propose an improvement consisting in an added attention mechanism, as well as better adapting the DSL vocabulary to a machine translation task. We provide a quantitative and qualitative experimental evaluation overcoming the state of the art in terms of standard metrics. Additionally, for the sake of a proof of concept, we have developed a web application that given an input GUI image captured from an Android device, processes it with the proposed model, and generates a DSL compliant output and the corresponding compiler's output in XML, ready to be rendered

YAEL TUDELA: "Use of Two-stage Object Detection Architectures for Real-time Polyp Detection and Classification in Colonoscopy Sequences"

- **Supervisors:** Jorge Bernal
- **Abstract:**
  Colorectal cancer is one of the main causes of deaths worldwide. Early detection and histological class prediction of its precursor lesion, the polyp, is key to reduce its mortality and to improve the efficiency of the procedure. In this work we study how two-stage object detection architectures can be used to provide an accurate polyp detection and classification on video sequences while keeping a trade-off between performance and resource consumption that allows the use of such a system in the exploration room. We use Faster R-CNN and Mask R-CNN as base architectures and we propose several cost efficient improvements aiming to alleviate some of the drawbacks that they present. Our experiments show promising results and indicate a benefit associated with the use of more different polyps (even if we only have a few shots of each of them) rather than including additional video sequences.

NILAI SALLENT: "Object detection of unsorted parts for industrial applications"

- **Supervisors:** Josep Ramon Casas
- **Abstract:**
  Automated bin-picking is a core problem in robotics. The project presents an alternative method for those cases where no rigid 3D shape of the object can be matched with the point cloud to extract correlations for grasping. Thus, the method utilised is an object detection model that employs the information of both grayscale images and the scene's depth to extract the matches. The results of these experiments have been compared with the standard object detection on RGB images for contrast. In summary, the information that adds the scene's depth improves the mAP metric by 1.5% while maintaining the IoU precision of the bounding boxes.

MARC ORÓS: "SepNet: Managing dataset bias to reduce object hallucination in captioning"

- **Supervisors:** Dimosthenis Karatzas
- **Abstract:**
  Most machine learning methods are susceptible to biases in training data, which can hinder their performance in certain situations. Image captioning is a research area where human introduced biases are important and have significant effects. We describe our implementation of a system that can be used to generate proxy training data with reduced bias by separating the foreground and background in order to use different representations and perform object replacements, as well as a model that utilises this separated image data to perform image captioning. We also evaluate multiple variants of our model on the MS-COCO dataset to evaluate its performance both in the general case as well as situations where the dataset's bias could make typical captioning models perform worse.

11:15 – 12:00

RICHARD SEGOVIA: "Automatic waste recycling using hyperspectral cameras and neural networks"

- **Supervisors:** Ramon Baldrich
- **Abstract:**
  In this work, we present the development of a novel solution for hyperspectral waste segmentation based on a modified U-NET with a ResNet18 backbone. This project is part of a collaboration between PICVISA, a company that specializes in residue (waste, garbage) sorting systems, and the Computer Vision Center. PICVISA provides the image dataset and groundtruth acquired with different hyperspectral cameras. The main goal of the project is to extend existing technologies on RGB images to nonlimited band images. For this, we first explored various methods to correctly normalize the images of the dataset in order to feed them to a neural network. Next, we modified a network that was originally designed to work with RGB images (three channels) to process hyperspectral images (multiple channels) as well. These modifications include the usage of 1×1 convolutions (network-in-network), wider decoder layers, and wider encoder layers. Finally, we found it necessary to group some classes that are variants of the same material because the confusion matrix showed that the model struggled to differentiate these classes. We conclude that the additional information that hyperspectral images

provide helps to improve the segmentation. Our results showed that models trained with hyperspectral images perform better than the ones trained with RGB images.

<u>OSCAR MAÑAS:</u> "Self-Supervised Visual Representation Learning for Remote Sensing"

- **Supervisors:** Xavier Giró
- **Abstract:**
  With the creation of large-scale annotated datasets such as the ImageNet, fully-supervised machine learning methods have become the standard for solving computer vision tasks. These methods require large amounts of annotated data, which is usually obtained with crowdsourcing tools or social media tags. However, these approaches do not scale for specialized domains, such as medical or satellite imaging, where annotations must be provided by experts at a prohibitive cost. Recently, self-supervised learning has emerged as an alternative for obtaining transferable visual representations from unlabeled data. Models based on these representations match the performance of fully-supervised models while only requiring a small fraction of the annotations. In this project, we aim to explore the application of self-supervised learning methods in the remote sensing domain.

<u>LAURA MORA:</u> "Brain Tumor Segmentation using 3D-CNNs with Uncertainty Estimation"

- **Supervisors:** Verónica Vilaplana
- **Abstract:**
  Automation of brain tumors in 3D magnetic resonance images (MRIs) is key to assess the diagnostic and treatment of the disease. In recent years, convolutional neural networks (CNNs) have shown improved results in the task. However, high memory consumption is still a problem in 3D-CNNs. Moreover, most methods do not include uncertainty information, which is especially critical in medical diagnosis. This work studies 3D encoder-decoder architectures, trained with patching techniques to reduce memory consumption and decrease the effect of unbalanced data. We also introduce voxel-wise uncertainty information, both epistemic and aleatoric using test-time dropout and data-augmentation respectively.

<u>DIEGO ALEJANDRO VELAZQUEZ:</u> "Logo Detection With No Priors"

- **Supervisors:** Jordi Gonzàlez
- **Abstract:**
  This master thesis is based on DETR, a work presented by FAIR, which treats the object detection problem as a set prediction problem directly, resulting in an end-to-end fully differentiable pipeline. This approach does not require any domain specific prior to be fed into the model, and its performance is comparable to current state of the art methods. We evaluate the effectiveness of this new approach on the task of logo detection and work on improving its results on the detection of small objects, while keeping the model simple and fully differentiable. We obtain good results when compared to a strong Faster R-CNN baseline.