MobiNetVideo Newsletters

#5 - June 2020

TEC2017-88169-R MobiNetVideo (2018-2020)

Visual Analysis for Practical Deployment of Cooperative Mobile Camera Networks

http://www-vpu.eps.uam.es/MobiNetVideo/

New Registered Observer: Nokia Bell Labs Spain

Nokia Bell Labs Spain joined as Registered Observer in May 2020.

Nokia Bell Labs is one of the main global centers of research and innovation in telecommunications technologies, responsible for countless breakthroughs that have shaped the networking and communications industry. Bell Labs scientists have been recognized with 9 Nobel Prizes throughout the institution's more than 90-year history, for its contribution to society with technologies that deeply influence the future of our way of working, living and communicating,

Since its creation in 2016 the research team of Bell Labs in Nokia Spain contributes to this goal with its focus on the application of immersive media to human communication.

Nokia creates the technology to connect the world. Only Nokia offers a comprehensive portfolio of network equipment, software, services and licensing opportunities across the globe. With our commitment to innovation, driven by the award-winning Nokia Bell Labs, we are a leader in the development and deployment of 5G networks. Our communications service provider customers support more than 6.1 billion subscriptions with our radio networks, and our enterprise customers have deployed over 1,000 industrial networks worldwide. Adhering to the highest ethical standards, we transform how people live, work and communicate.

Fifth semester progress report

During this semester, research activities have been progressing properly. Although the *pandemia* has limited some activities (mainly short research stays), the project Team has been in contact via teleconferencing and mail, limiting the consequences of not direct contact. Nevertheless, in May 20202 we have decided to extend WP3 till September (D3v2 has been rescheduled accordingly) as well as delaying WP4 start till September. We will evaluate in the next months the possibility of a project extension request.

Two papers were published, and several ones are under preparation.
Deliverable D5v2 "Results Report" was published March 2020. "Content Sets" page at the project web site, has been upgraded to "Public Resources" containing both links to the pages with, as detailed in the next section, two contest sets (currently still just available on-demand) and software (available at github.con/vpulab)

Fifth semester results

Content Sets

P365LLds: A Places365 Lifelogging version Dataset. (available on-demand; to be available on-line soon)

The task of scene recognition has been classically evaluated using still images representing scenes. In the context of the MobiNetVideo project, we have created a new dataset that extrapolates Places365's classes to lifelogging/egocentric videos. The dataset is made up of 450 videos recorded with smartphones, go-pro and handheld cameras. Videos have been obtained by downloading YouTube videos licensed as Creative Commons. For each scene class in Places365, we include between one (90% of the classes) and four videos. The average length of the videos is 638 frames (around twenty-one seconds) and the median length is 600 frames per video (around twenty seconds). In overall, the dataset is approximately 34.1 GB large.

USSds: A Unified Semantic Segmentation Dataset (available on-demand; to be available on-line soon)

There is a large variety of semantic datasets. However, not all of them have the same semantic classes, and the appearance of shared classes substantially differ. The USSds represents a data integration effort to create a unified semantic dataset which—by enlarging the number of classes and the diversity of the shared classes, aims to provide a more generic benchmark for training and evaluation. The merged datasets have been relabelled to a common set of 293 semantic labels distributed into a total of 145,555 training images and 7,614 validation images. The datasets agglutinated to compose the USSds dataset are:

- COCO-Stuff, COCO-Stuff Dataset.
- Cityscapes, Cityscapes Dataset.

- ADE20K, ADE20K Dataset.
- TASKONOMY, TASKONOMY Dataset.
- Mapillary, Mapillary Dataset.

Software

Pytorch Implementation of Semantic-Aware Scene Recognition https://github.com/vpulab/Semantic-Aware-Scene-Recognition

Official Pytorch Implementation of Semantic-Aware Scene Recognition by Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós and Álvaro García-Martín (Elsevier Pattern Recognition).

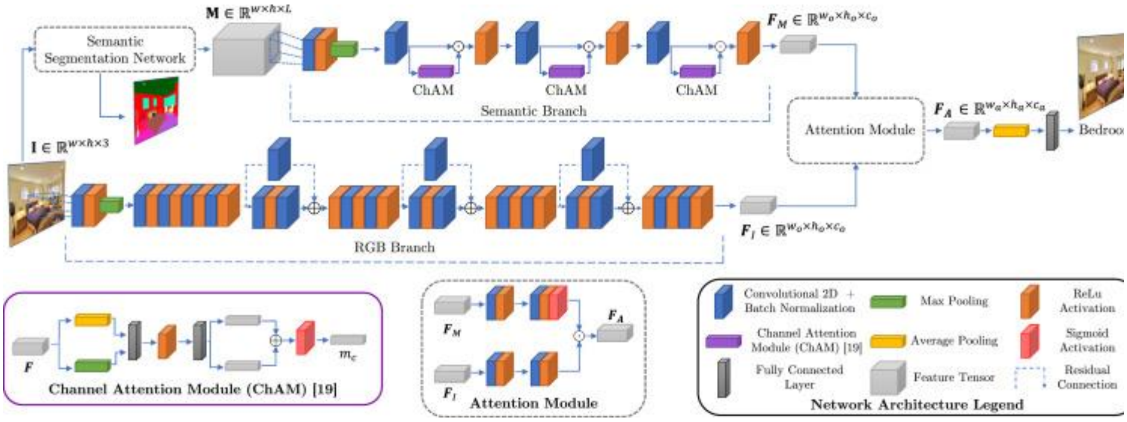
Support for Pytorch and Caffe CNNs Visualization https://github.com/MEscuderoVinolo/MobiNet-Video-CNN-Visualization

This repository contains the implementation of a visualization tool for convolutional neuronal networks.

Journals

Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, Álvaro García-Martín, **Semantic-aware scene recognition**, Pattern Recognition, Volume 102, 2020, 107256, ISSN 0031-3203 (DOI 10.1016/j.patcog.2020.107256).

Scene recognition is currently one of the top-challenging research fields in computer vision. This may be due to the ambiguity between classes: images of several scene classes may share similar objects, which causes confusion among them. The problem is aggravated when images of a particular scene class are notably different. Convolutional Neural Networks (CNNs) have significantly boosted performance in scene recognition, albeit it is still far below from other recognition tasks (e.g., object or image recognition). In this paper, we describe a novel approach for scene recognition based on an end-to-end multi-modal CNN that combines image and context information by means of an attention module. Context information, in the shape of a semantic segmentation, is used to gate features extracted from the RGB image by leveraging on information encoded in the semantic representation: the set of scene objects and stuff, and their relative locations. This gating process reinforces the learning of indicative scene content and enhances scene disambiguation by refocusing the receptive fields of the CNN towards them. Experimental results on three publicly available datasets show that the proposed approach outperforms every other state-of-the-art method while significantly reducing the number of network parameters. All the code and data used along this paper is available at: https://github.com/vpulab/Semantic-Aware-Scene-Recognition.


$\mathbf{M} \in \mathbb{R}^{w \times h \times L}$
Semantic Segmentation Network
$\boldsymbol{F_M} \in \mathbb{R}^{w_o \times h_o \times c_o}$
ChAM
ChAM
ChAM
Semantic Branch
Attention Module
$\boldsymbol{F_A} \in \mathbb{R}^{w_a \times h_a \times c_a}$
Bedroo
$\mathbf{I} \in \mathbb{R}^{w \times h \times 3}$
RGB Branch
$\boldsymbol{F_I} \in \mathbb{R}^{w_o \times h_o \times c_o}$
F
m_c
Channel Attention Module (ChAM) [19]
F_M
F_A
F_I
Attention Module
Convolutional 2D + Batch Normalization
Max Pooling
ReLu Activation
Channel Attention Module (ChAM) [19]
Average Pooling
Sigmoid Activation
Fully Connected Layer
Feature Tensor
Residual Connection
Network Architecture Legend

Conferences

Paula Moral, Álvaro García-Martín, José M. Martínez, **Vehicle Re-Identification in Multi-Camera scenarios based on Ensembling Deep Learning Features**, Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR2020), Seattle, Washington, USA, Jun. 2020, in press.

Vehicle re-identification (ReID) across multiple cameras is one of the principal issues in Intelligent Transportation System (ITS). The main challenge that vehicle ReID presents is the large intra-class and small inter-class variability of vehicles appearance, followed by illumination changes, different viewpoints and scales, lack of labelled data and camera resolution. To address these problems, we present a vehicle ReID system that combines different ReID models, including appearance and orientation deep learning features. Additionally, for results refinement re-ranking and a post-processing step taking into account the vehicle trajectory information provided by the CityFlow-ReID dataset are applied.

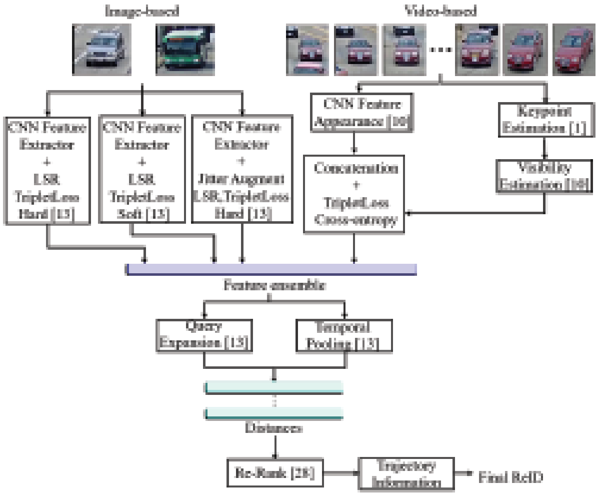


Image-based
Video-based
CNN Feature Extractor + LSR TripletLoss Hard [13]
CNN Feature Extractor + LSR TripletLoss Soft [13]
CNN Feature Extractor + Jitter Augment LSR TripletLoss Hard [13]
CNN Feature Appearance [10]
Keypoint Estimation [1]
Concatenation + TripletLoss Cross-entropy
Visibility Estimation [18]
Feature ensemble
Query Expansion [13]
Temporal Pooling [13]
Distances
Re-Rank [28]
Trajectory Information
Final ReID

Master thesis

Real-Time Target Tracking to Position a Mobile Device, Awet H. Gebreiwot (advisors: Jesús Bescós, Álvaro García-Martín), Master Thesis, Erasmus Mundus Image Processing and Computer Vision Master Program), Univ. Autónoma de Madrid, Jun. 2020.

Abstract: Real-time visual object tracking is an open problem in the field of computer vision, with multiple applications in the industry, such as autonomous vehicles, human-machine interaction, intelligent cinematography, automated surveillance, and autonomous social navigation. The challenge of tracking a target of interest is critical to all of these applications. Recently, tracking algorithms that use Siamese neural networks trained offline on large-scale datasets of image pairs have achieved the best performance exceeding real-time speed on multiple benchmarks. Results show that siamese approaches can be applied to enhance the tracking capabilities by learning deeper features of the object's appearance. This thesis aims to study, evaluate, and enhance one of the state-of-the-art tracking algorithms called SiamMask, in order to position a mobile camera to precisely follow a target. SiamMask utilizes the power of siamese networks and supervised learning approaches to solve the problem of an arbitrary object tracking in real-time speed. However, its practical applications are limited due to failures encountered during testing. In order to improve the robustness of the tracker and make it applicable for the intended real-world application, two improvements have been proposed, each addressing a different aspect of the tracking task. The first one is a data augmentation strategy to consider both motion-blur and low-resolution during training. It aims to increase the robustness of the tracker against a motion-blurred and low-resolution frames during inference. The second improvement is a target template update strategy that utilizes both the initial ground truth template and a supplementary updatable template, which considers the score of the predicted target for an efficient template update strategy by avoiding template updates during severe occlusion. All the proposed improvements were extensively evaluated and achieved state-of-the-art performance in the VOT2018 and VOT2019 benchmarks. Comparable results on VOT2018 and VOT2019 benchmarks show that the proposed approach has significantly improved the tracker's performance with respect to the original SiamMask results. A tracking system that automatically follows the presenter (Lecturer) and that creates a far more visually appealing live-video presentation is developed as a real-world application.

Incorporating Depth in Egocentric Perception, Andrija Gajic (advisor: Marcos Escudero Viñolo), Master Thesis, Erasmus Mundus Image Processing and Computer Vision Master Program), Univ. Autónoma de Madrid, Jun. 2020.

Abstract: Lately, the performance of several computer vision tasks on color images has been boosted by using deep learning methods—specially by the advent of convolutional neural networks (CNN), and by the increasing availability of large, annotated datasets. In contrast, the research done in deep learning using depth maps is scarser. However, depth data can add some very valuable information to the color data, since it implicitly contains semantic information, by knowing that pixels of an object will typically contain similar depth ranges. Also, depth data can be very useful for handling some of the biggest drawbacks of color-based models, such as illumination changes. Lately, interest on exploring likewise depth data is increasing as the depth sensors are being massively incorporated into everyday systems, making depth information accessible to the community. With the advent of depth information, the goal of this Master's thesis is to enhance existing color architectures by incorporating depth data and to understand whether such complementary information can help color-based egocentric computer vision scenarios—semantic segmentation and scene recognition. Augmented Virtuality systems incorporate specific real objects (most typically, the user) into the virtual surroundings. In order to do so, the exact pixels corresponding to the objects of interest must be discriminated. For this purpose, semantic segmentation is used. In this task, the contribution of this Master's thesis is threefold: i) we create a semi-synthetic egocentric dataset composed of 7,826 realistic images and associated pixel-wise labels of arms including different demographic factors; ii) building upon the ThunderNet architecture, we implement a deep learning semantic segmentation algorithm able to perform beyond real-time requirements (16 ms for 720x720 images). Even though the inference time was severely improved, the performance of proposed model remained competitive with the state of the art methods, resulting in mIoU scores of 57.5% in GTEA Gaze+ dataset and 50.3% in THU-Read dataset. This is, to the best of our knowledge, the first real-time deep network designed for binary segmentation of arms for Mixed Reality; iii) We explore the opportunity of synthesizing artificial depth information in order to train the networks with the depth extracted from real sensors. Scene recognition is currently one of the top-challenging research fields in computer vision, mostly due to the ambiguity between scene classes. The task of a scene recognition system is to, given an image, determine to which scene class it belongs. A system for RGB-D scene recognition is designed. We show that using depth maps can further improve the results, since the depth possesses additional cues, not very likely to be learnt from color data. We define a two-stage learning architecture consisting of three branches—color, depth and semantic, fused in the end using

attention mechanisms. Each branch is firstly maximized in terms of precision on its own. In this case, we show that the proper encoding is crucial for the depth branch and that HHA (Horizontal disparity, Height, Angle) representation leads to the best results. Moreover, we show that proper pre-training makes a great difference when fine-tuning to small datasets. After all branches have been optimized, weights inside them are frozen and different attention modules are trained and evaluated. In the end, using Hadamard combination proved to be the most prolific. Finally, we reached performances very comparable to the current state of the art methods, resulting in a 60.0% Top@1 precision in the SUN RGB-D dataset. We also provide an extensive quantitative and qualitative evaluation of our model.

People detection in omnidirectional cameras: development of a deep learning architecture based on a spatial grid of classifiers, Enrique Sepúlveda Jorcano (advisor: Pablo Carballeira López), Master Thesis, Erasmus Mundus Image Processing and Computer Vision Master Program), Univ. Autónoma de Madrid, Jun. 2020.

Abstract: Deep learning has arisen as one of the best tools to use in computer vision. With the potential of adapting to almost any problem, deep learning CNN (Convolutional Neural Networks) are highly effective in multiple tasks for computer vision. The application of deep learning in traditional cameras has improved the performance in detection tasks with respect to the state-of-the-art methods that do not use deep learning. Images from traditional cameras are perspective projections of the real world with almost no distortion. This allows us to train networks that can learn the general appearance of an object and assume that they are spatial-invariant between frames. However, the use of omnidirectional cameras has increased in the last years thanks to the advantage they offer with respect of traditional cameras: a wider Field of View. While in conventional cameras it usually does not go further than 60o, in omnidirectional cameras it can reach values of 160o. With a single omnidirectional camera, therefore, we are able to cover a wider area than with a traditional camera, reducing costs of deployment. But this comes at a cost, and it's that they introduce a great distortion, making the objects change their appearance depending on their position in the image. This makes some of the existing deep learning techniques to highly reduce their performance, and it's necessary the use of new techniques. The proposed method in this thesis uses CNNs to extract characteristics from omnidirectional images to detect objects using a spatial-aware grid of classifiers. The original idea was to use a HOG features and grid of SVM classifiers SVM where each of the classifier will learn from a region of the image. In this way we avoid the problem of using one single classifier that has learnt general spatial-invariant characteristics. In [1] the goal was to create a robust and real-time detector using (1) HOG to extract a feature vector from the image and (2) a grid of SVM (Support Vector Machine) classifiers that would predict the position of the objects. In our work we propose to substitute this two-step architecture with a CNN that could be used end-to-end. We have tested Alexnet, Resnet18 and Resnet50. Part of this project has been the study of data augmentation techniques to improve the performance of the system. An example of this is the creation of synthetic images where people are added from random frames to others. The best performance has been obtained with Resnet50.

Desarrollo de aplicaciones móviles de clasificación y detección de objetos a partir de redes convolucionales ligeras **(Development of mobile application for objtect classification and detection based on light convolutional networks)**, Paulo C. Casa Robles (advisor: Pablo Carballeira López), Trabajo Fin de Máster (Master Thesis), Master en Ingeniería de Telecomunicación, Univ. Autónoma de Madrid, Jun. 2020.

Abstract: In recent years, the artificial vision has evolved rapidly due to the development of convolutional neural networks, and this is demonstrated by the numerous publications and research of the scientific community in several of the possible applications that it may have, such as: the classification, detection, segmentation, or recognition of certain objects. In addition, on the other hand, it is also known the high demand that mobile applications have at this moment. Therefore, this Final Master Project is presented with the intention of uniting both subjects, specifically with the development of mobile applications in the tasks of classification and object detection. However, integrating machine vision tasks into a mobile device is a complex problem of great interest. This is solved with light convolutional networks since they have certain characteristics, among which stand out, the memory efficiency and model precision, being necessary properties that an application demands on a mobile device for its correct operation. For this, it has been necessary to explore the different alternatives that the state of the art offers us to incorporate pre-trained models of light convolutional networks into mobile devices. One of the deep learning platforms that is currently in constant development and that will have a fundamental role in the integration of these models will be TensorFlow Lite. Thus, this TFM presents the necessary techniques and configurations to convert the models into TensorFlow Lite format and later allow the insertion of the models into mobile devices with Android operating system using developer tools. Finally, this document contributes a comparative study of the classification and detection tasks, providing conclusions of the different integrated convolutional networks performances in terms of efficiency and computational complexity.

Graduate thesis

Clasificación de imágenes con redes neuronales profundas mediante conjuntos de entrenamiento reducidos y aprendizaje "few-shot" (**Deep learning based image classification using reduced training sets and few-shot training**), Guillermo E. Torres Alonso (advisor: Miguel Ángel García), Trabaja Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jun. 2020.

Abstract: This Bachelor Thesis has consisted in the implementation and experimental evaluation of a deep neural network that allows the supervised classification of images through reduced training sets. This network was described in a paper published by F. Sung, Y. Yang y L. Zhang, T. Xiang in 2018 in one of the most prestigious international conferences on Computer Vision and Pattern Recognition (IEEE CVPR 2018). Differently to conventional training, which uses millions of labelled images, "few-shot" learning aims at classifying images from reduced training sets. In this work, we have implemented the neural network described in the reference paper and we have evaluated its performance with respect to a set of images acquired over the UAM campus.

Desarrollo de un marco de trabajo para segmentación semántica en bases de datos de imágenes urbanas (**Development of a framework for semantic segmentation in urban images' databases**), Javier González Cabrero (advisor: Pablo Carballeira López), Trabaja Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jun. 2020.

Abstract: Navigation through unstructured environments is a basic ability of intelligent creatures, and therefore, is of fundamental interest in the study and development of artificial intelligence. The capacity to automate it, being able to navigate without the use of maps, only with images through urban environments is being the target of incipient works. The automatic navigation systems are based on systems trained with images of the urban environment, of the city. The segregation of fixed and mobile elements can be useful to improve the training process of these systems, making their learning based on the appearance of fixed elements, and not on mobile elements that can distort the learning process, and therefore, operation. In this context, the semantic segmentation of objects could help to improve the location and guidance system. The aim of the work is to develop a framework for semantic segmentation in urban image databases. Taking advantage of the availability of Google Street View, it is used as a database for the implementation of the work because of its worldwide coverage and photographic content, showing images of different locations with different camera characteristics such as field of view, heading, pitch, etc. This makes it possible to cover the 360o with images from a single location, showing different points of view of the location. As 360o information is available, this framework includes the reprojection of semantic masks in the sphere to add the redundant information of the different points of view. In view of this, the following hypothesis arises: could the use of the different points of view help to improve the semantic segmentation obtained from a single point of view? To assess and respond to this, a framework for evaluating different reprojection and aggregation algorithms is established. The proposed aggregation methods are simple. This work gathers a set of preliminary conclusions on a limited set of data, showing that in most cases the results obtained by the simple aggregation methods implemented do not exceed those obtained by direct segmentation.

The MobiNetVideo (TEC2017-88169-R) project is supported by

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD