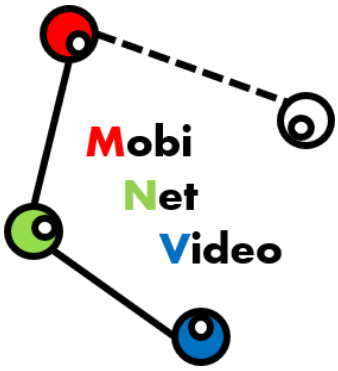

TEC2017-88169-R MobiNetVideo (2018-2020)

Visual Analysis for Practical Deployment of Cooperative Mobile Camera Networks

D5 v2

Results Report

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by

AUTHORS LIST

José M. Martínez	josem.martinez@uam.es
Pablo Carballeira López	Pablo.carballeira@uam.es
Juan Carlos San Miguel Avedillo	juancarlos.sanmiguel@uam.es
Marcos Escudero Viñolo	marcos.escudero@uam.es
Álvaro García-Martín	alvaro.garcia@uam.es
Jesús Bescós Cano	j.bescos@uam.es

HISTORY

Version	Date	Editor	Description
0.1	06/06/2019	José M. Martínez	First Draft version
0.2	12/06/2019	José M. Martínez	Structure for Contributions and draft for several sections
0.3		Pablo Carballeira	Contributions
0.4		Juan Carlos San Miguel	Contributions
0.5		Marcos Escudero	Contributions
0.6		Álvaro García-Martín	Contributions
0.7		José M. Martínez	Working Draft edition
0.7		Marcos Escudero	Working Draft comments
0.8		Pablo Carballeira	Working Draft comments
0.9		Álvaro García-Martín	Working Draft comments
0.10		Juan Carlos San Miguel	Working Draft comments
0.11	31/07/2019	José M. Martínez	Final Working Draft – Editorial checking
1.0	31/07/2019		First version
1.1	10/03/2020	José M. Martínez	Structure for Contributions and draft for several sections
1.2	13/03/2020	Marcos Escudero	Contributions
1.3	15/03/2020	Juan Carlos San Miguel	Contributions
1.4	20/03/2020	Álvaro García-Martín	Contributions

1.5	22/03/2020	Pablo Carballeira López	Contributions
1.5	24/03/2020	Jesús Bescós Cano	Contributions
1.6	30/03/2020	José M. Martínez	Final Working Draft – Editorial checking
2.0	30/07/2020		Second version

CONTENTS:

1. Introduction

This *report* recapitulates the results obtained within the MobiNetVideo project. The results are announced as they are obtained at the Web site (http://www-vpu.eps.uam.es/MobiNetVideo).

1.1. Document structure

This document contains the following chapters:

- Chapter 1: Introduction to this document
- Chapter 2: Publications
- Chapter 3: Project Documents
- Chapter 4: Public Resources: Content Sets and Software
- Chapter 5: Workshops and Seminars
- Chapter 6: Overview of project progress
- Chapter 7: Main achievements of the project

2. Publications

2.1. Journals

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

2.2. Book Chapters

2.3. Conferences

[2] Zhu P. et al., "VisDrone-VDT2018: The Vision Meets Drone Video Detection and Tracking Challenge Results", in Computer Vision – ECCV 2018 Workshops. ECCV 2018, L. Leal-Taixé, S. Roth (eds.) Lecture Notes in Computer Science, Vol. 11133. Springer, Cham, 2019, pp. 496-518. (ISBN 978-3-030-11020-8) (DOI 10.1007/978-3-030-11021-5_29)

[3] Elena Luna, Paula Moral, Juan C. SanMiguel, Álvaro García-Martín, José M. Martínez, "VPULab participation at AI City Challenge 2019", Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR2019), Long Beach, CA, USA, Jun. 2019, in press.

2.4. PhD Thesis

2.5. Master Thesis

[4] Mejora del rendimiento de redes convolucionales entrenadas para el reconocimiento de escena mediante el uso de información sobre los objetos comunes a éstas (Refocusing a scene recognition ConvNet by incorporating object priors), Raúl García Jiménez, (advisor: Marcos Escudero Viñolo), Trabajo Fin de Máster (Master Thesis), Master en Ingeniería de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2018.

[5] Object detection and association in multiview scenarios based on Deep Learning, Paula Moral de Eusebio (advisor: Álvaro García-Martín), Trabajo Fin de Máster (Master Thesis), Master en Investigación e Innovación en TIC – Programa Internacional de Múltiple Titulación IPCV (Image Processing and Computer Vision Master Program), Univ. Autónoma de Madrid, Jul. 2019.

[6] Learning how to modify training rates in scene-recognition, Miguel Basarte Mena, (advisor: Marcos Escudero Viñolo), Trabajo Fin de Máster (Master Thesis), Master en Investigación e Innovación en TIC – Programa Internacional de Múltiple Titulación IPCV (Image Processing and Computer Vision), Univ. Autónoma de Madrid, Sep. 2019.

2.6. Graduate Thesis

[7] Detección de objetos en imágenes urbanas de Google Street View (Object detection in urban images from Google Street View), Paula Guerra Toni (advisor: Pablo Carballeira), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2019.

[8] Adaptación de un sistema de detección de personas en cámaras omnidireccionales a descriptores Deep Learning (Adaptation of a system for people detection with omnidirectional cameras to Deep Learning descriptors), Nicolás García Crespo (advisor: Pablo Carballeira), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2019.

[9] Detección jerárquica de grupos de personas con CNNs (Hierarchical detection of groups of people using CNNs), Antonio Campoy Cordero (advisor: Álvaro García-Martín), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2019.

[10] Re-identificación de personas (People re-identification), Daniel Sáez García (advisor: Álvaro García-Martín), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2019.

[11] Reconocimiento de escenas exteriores mediante redes neuronales profundas entrenadas con la base de datos places (Scene recognition using Deep Neural Networks trained with the Places database), Santiago Vicente Moñivar (advisor: Miguel Ángel García), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Oct. 2019.

3. Project Documents

3.1. First 19 months Deliverables

3.1.1. D1.1 “System Infrastructure” version 1 (December 2018 - delayed July 2019)

This deliverable describes the infrastructure available at VPULab for the development of the tasks proposed within the MobiNetVideo project (TEC2017-88169-R. Describing both the available hardware for video sequences adquisition (“Adquisition infrastructure”) and the hardware and software environments for data processing and analysis (“Processing infrastructure”).

3.1.2. D1.2 “Camera simulation” version 1 (December 2018 - delayed July 2019)

This deliverable describes the work related with the task T.1.2 Cameras network simulation which supports other tasks for generating test data. We focus on the simulator “Multi-camera System Simulator (MSS)” to describe its structure and the developed features within the context of this project. Moreover, we also integrated other developments for the MSS simulator.

3.1.3. D1.3 “Evaluation datasets” version 1 (December 2018 - delayed July 2019)

This deliverable describes the work related with the task T.1.3 “Generation of datasets. Support to other tasks for generating test data and defining evaluation methodologies”. It includes the selection of appropriate datasets (sequences and associated ground-truth) and the ones generated within the project. Additionally, the evaluation methodologies used for the evaluation for the project algorithms and systems is also described.

3.1.4. D2 “Feasibility studies: algorithms and findings” ((September 2018 - rescheduled March 2019 - delayed July 2019)

This deliverable describes the work related with tasks T.2.1 People tracking for active vision, T.2.2 Object detection for collision detection, T.2.3 Scene categorization for lifelogging and T.2.4 Multi-target tracking for UAV monitoring.

All these WP2 tasks aim at performing a study of current technologies for applications related to heterogeneous camera networks where camera mobility plays a key role. Such studies are performed on public datasets. The main objective is the identification of suitable state-of-the-art video analysis tools (e.g., segmentation, tracking and detection),

by the implementation and evaluation of their performance in single mobile cameras for use as a baseline for comparison with the achievements to be developed within WP3 and WP4.

The technologies covered by the deliverable are:

- People tracking for active vision
- Object detection for collision detection
- Scene categorization for lifelogging
- Multi-target tracking for UAV monitoring

3.1.5. D3 "Technologies for mobile camera networks" version 1 (June 2019 – rescheduled July 2019 – rescheduled September 2019)

This deliverable was rescheduled to September 2019 in July.

This deliverable will describe the new technologies developed within WP3 for applications related to heterogeneous camera networks where camera mobility plays a key role, covering different areas grouped in WP3 tasks:

- T.3.1 Scene Recognition
- T3.2 Semantic Segmentation
- T3.3 Multi-view matching
- T.3.4 Cooperative detection and tracking

3.1.6. D5 "Results Report" version 1 (September 2018 - rescheduled December 2018 - cancelled December 2018)

3.1.7. D5 "Results Report" version 1 (July 2019)

The first version of this document presenting the results achieved during the first 19 months of the project.

3.1.8. D1.1 "System Infrastructure" version 2 (December 2019 -not required-)

3.1.9. D1.2 "Camera simulation" version 2 (December 2019 -not required-)

3.1.10. D1.3 "Evaluation datasets" version 2 (December 2019 -not required-)

3.1.11. D5 "Results Report" version 2 (March 2020)

This document, updating the results of the Project till March 2020.

3.2. Future Deliverables

3.2.1. D3 “Technologies for mobile camera networks” version 2 (June 2020)

3.2.2. D1.1 “System Infrastructure” version 3 (December 2020, if required)

3.2.3. D1.2 “Camera simulation” version 3 (December 2020)

3.2.4. D1.3 “Evaluation datasets” version 3 (December 2020, if required)

3.2.5. D4 “Deployment and application scenarios” (December 2020)

3.2.6. D5 “Results Report” version 3 (December 2020)

3.3. Technical Reports

4. Public Resources

4.1. 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.

4.2. Software

- Pytorch Implementation of Semantic-Aware Scene Recognition [1]
  https://github.com/vpulab/Semantic-Aware-Scene-Recognition.
- Support for Pytorch and Caffe CNNs Visualization [6]
  https://github.com/MEscuderoVinolo/MobiNet-Video-CNN-Visualization.

5. Workshops and Seminars

6. Overview of project progress

The provisional announcement of project approval arrived in the first days of 2018, but the official project approval was not published till mid-June, allowing to start the official opening of the project at the university. Whilst waiting for the notification of definitive approval, we have been working in the lines described in the proposal, but without being able to work with financial support.

During the first year (in July and September), the workplan suffered several adjustments:

- D2 and D5v1, as well as the first Newsletter, were delayed
- M1.1 and M1.2 were delayed.
- WP2 tasks were extended till M15 and all milestones and deliverables have been scheduled for the end of WP activities.
- With respect to WP5, the first developers' workshop has been cancelled and D5v1 (this document) rescheduled to September 2018), was finally cancelled, being the first version scheduled June 2019.

Vaelsys joined as Registered Observer in October 2018. Vaelsys is a Spanish company devoted mainly intelligent video analysis. Vaelsys mission is to allow businesses from diverse sectors to take advantage of the opportunities offered by the wealth of data generated daily by video cameras.

During the third semester, the first set of deliverables (D1.1 "System Infrastructure" version 1, D1.2 "Camera simulation" version 1, D1.3 "Evaluation datasets" version 1, D2 "Feasibility studies: algorithms and findings", and D5 "Results Report" version 1 -tis document-) as well as Newsletter#3 were finally rescheduled and published July 2019. Deliverable D3 "Technologies for mobile camera networks" version 1, has been delayed from July 2019 to September 2019.

From July 2019 till March 2020, Deliverable D3 "Technologies for mobile camera networks" version 1, was published (September 2019), as well as MobiNetVideo Newsletter #4 (December 2019). The work has focused on WP3 and it was decided to delay WP4 kick-off until July 2020, once WP3 is completed.

Considering this Results Report, we will work on potential adjustments to the workplan and project schedule during July-September.

7. Main Achievements of the Project

7.1. First 18 months main achievements

7.1.1. WP1: Infrastructure and datasets

- T1.1: Infrastructure update and maintenance
  - Acquisition and configuration of new hardware for GPU-based processing of image and video signals (Deliverable D1.1)
  - Acquisition and configuration of new hardware for research activities on "People Tracking for Active Vision" of WP2 and WP3 (Deliverable D1.1)
- T1.2: Cameras network simulation
  - Integration of previous research on simulation of camera networks using the Unity3D game engine. Generation of documentation for the integrated project. (Deliverable D1.2)
  - Design of an API using Python and Matlab. On-going Implementation and testing.
- T1.3: Generation of datasets
  - Development of a tool for the extraction of urban images through predefined routes, using the Google Directions and Street View APIs [7].
  - Compilation of existing datasets for multi-target tracking from UAV-based video sequences. (Deliverable D1.3)
  - Compilation of existing datasets for scene recognition in still images. (Deliverable D1.3)
  - Aggregation of existing datasets for semantic segmentation, creating a unified dataset. (Deliverable D1.3)
  - Creation of a new dataset datasets for scene recognition in life-logging videos. (Deliverable D1.3)

7.1.2. WP.2: Feasibility studies

- T2.1: People tracking for active vision
  - Adaptation of a people detection system for omnidirectional cameras to deep learning descriptors [8].
- T2.2: Object detection for collision detection
  - Evaluation of state-of-the-art in people detection methods with a wearable camera. (Deliverable D2.1).
  - Evaluation of state-of-the-art in people re-identification methods [10].
  - Evaluation of state-of-the-art in car re-identification methods [5].
  - Development of a people detection approach in presence of groups with CNNs [9].
- T2.3: Scene categorization for life-logging
  - Evaluation of state-of-the-art object detectors in urban images obtained with the tool developed in T1.3 [7].

- Evaluation of state-of-the-art scene recognition methods in still images [4].
- Evaluation of state-of-the-art semantic segmentation methods in life-logging videos compiled in T1.3. A Graphical User Interface has been designed to ease results' understanding. (Deliverable D2.3).
- Evaluation of state-of-the-art scene recognition methods in life-logging videos compiled in T1.3. A Graphical User Interface has been designed to ease results' understanding. (Deliverable D2.3).
- Adaptation of a convolutional neural network visualization tool for accounting scene recognition learning process. (Deliverable D2.3).

- T2.4: Multi-target tracking for UAV monitoring
  - Generation of a framework for evaluating algorithms multi-target tracking (developed in Python or Matlab) on existing datasets or custom datasets and using standard evaluation protocols (Deliverable D2.1).
  - Evaluation of selected state-of-the-art approaches using the VisDrone dataset and CarPK dataset captured from drones (Deliverable D2.1).

7.1.3. WP.3: Enabling technologies

- T3.1: Scene identification
  - Development of a framework for scene recognition in Python which combines semantic and appearance deep features by means of an attention module in an end-to-end multi-modal CNN.
  - Implementation of evaluation protocols to estimate scene recognition performance.
- T3.2: Semantic segmentation
  - Development of a framework for semantic segmentation in Python to allow the learning of the unified semantic segmentation dataset.
- T3.3: Multi-view matching
  - Development of a framework for multi-camera object re-identification
  - Implementation of evaluation protocols to estimate object re-identification performance
  - Participation [3][5] in the AICity challenge 2019 (CVPR workshop Challenge Track 2: City-Scale Multi-Camera Vehicle Re-Identification) https://www.aicitychallenge.org/2019-data-sets/.
- T3.4: Cooperative detection and tracking
  - Development of a framework for single-camera Multiple-Object-Tracking in Python which employs spatial and appearance (deep) features.
  - Implementation of evaluation protocols to estimate single camera multiple-object tracking performance.
  - Participation[2] in the VisDrone challenge 2018 (ECCV workshop) http://aiskyeye.com/.
  - Development of a framework for multi-camera Multiple-Object-Tracking in Python which employs spatial and appearance (deep) features.
  - Implementation of evaluation protocols to estimate multicamera multiple-object tracking performance

- Participation [3] in the AICity challenge 2019 (CVPR workshop) https://www.aicitychallenge.org/2019-data-sets/.

7.1.4. WP.4: Deployment of applications

No activity was planned during the first 18 months of the project in this workpackage tasks.

- T4.1: Case studies: implementation and testing
- T4.2: Real-time and network issues
- T4.3 Sensor issues

7.1.5. WP.5: Management and dissemination

- T5.1: Management
  - During the first 18 month, several workplan updates have been done due to the disarrangements caused by the delay in the project kick-off and the resolution of the FPI scholarship.
- T5.2: Dissemination
  - The web page was created and it is being updated timely.
  - Three MobiNetVideo Newsletters were published.
  - As expected, the number of results ready for publication during this period is limited. A detailed plan will be done early September.
- T5.3 Workshop organization
  - The 2018 Developers Workshop was cancelled. Currently, we are evaluating the planning of such workshops for the rest of the project.

7.2. Main achievements during months 19 to 27

7.2.1. WP1: Infrastructure and datasets

- T1.1: Infrastructure update and maintenance
  - Design and Development of a virtualization platform to enable the research and prototyping algorithms with shared computation and storage resources. It also allows concurrent access and remote working by multiple users. On-going documentation and testing.
- T1.2: Cameras network simulation
  - Update of the simulator backbone software (Unity 3D) to latest version.
  - Update of the Python API to Python version 3.7. Completed testing and examples of usage.
  - Addition of new features: cameras of the scene, PTZ cameras, continuous/on-demand running modes and semantic maps for camera feeds. On-going documentation and testing.
- T1.3: Generation of datasets
  - Creation of the **P365LLds** dataset for scene recognition in life-logging videos.
  - Creation of the **USSds** dataset to increase the generality and diversity in the training and evaluation of semantic segmentation methods.

7.2.2. WP.2: Feasibility studies

- T2.1: People tracking for active vision
  - Development of a people detection system for omnidirectional cameras based on an end-to-end deep learning architecture (ongoing work)
  - Adaptation of a state-of-the-art real-time tracker to enhance its robustness. Participation into the Visual Object Tracking 2020 Challenge, specifically into the short-term real-time challenge.
- T2.2: Object detection for collision detection
  - Development of an algorithm for automatic anticipating accidents or collisions in board first-person view camera videos (ongoing work).
- T2.3: Scene categorization for life-logging
  - Statistical analysis of the increase in shared semantic class divergence provided by the **USSds** semantic segmentation dataset (ongoing work).
  - Evaluation of state-of-the-art semantic segmentation urban images obtained with the tool developed in T1.3 [7] (ongoing work).
  - Design, development and limited evaluation of a transfer learning method that moderately conserves previous learning in a CNN trained for scene recognition [6].
- T2.4: Multi-target tracking for UAV monitoring

7.2.3. WP.3: Enabling technologies

- T3.1: Scene identification
  - Design, development and evaluation of a scene-recognition method that takes advantage of the fact that the learning of scenes is inherent to the learning of the objects they include to improve scene-recognition by relying on semantic-segmentation with moderate increasing of the CNN capacity [1].
  - Design and development of computer vision methods for the temporal arrangement and clustering of video sequences based on scene-recognition results (ongoing work).
- T3.2: Semantic segmentation
  - Design, development and (limited) evaluation of self-taught and self-paced methods for training severe class-unbalanced datasets as the **USSds** semantic segmentation dataset (ongoing work).
- T3.3: Multi-view matching
  - Work towards participating in the AICity challenge 2020 https://www.aicitychallenge.org/
- T3.4: Cooperative detection and tracking
  - Development of an algorithm for cooperative detection of moving targets in multicamera scenarios based on Deep Features and Mixup learning. On-going documentation and testing.
  - Integration of cooperative detection for target tracking with multiple overlapped cameras. On-going documentation and testing.

7.2.4. WP.4: Deployment of applications

WP4 was planned for starting at month 24, but in order to concentrate resources in WP3, we have decided to delay WP4 kick-off till the planned end of WP3 (M30 – June 2020). Therefore, W4 will start July 2020 (M31).

- T4.1: Case studies: implementation and testing
- T4.2: Real-time and network issues
- T4.3 Sensor issues

7.2.5. WP.5: Management and dissemination

- T5.1: Management
  - During months 19-27 (June 2019-March 2020), there were no major management issues besides the delay in the kick-off of WP4 and the very limited mobility (research stays and conferences).
  - Taking into account the project progress, before summer we will evaluate the possibility of asking for an extension in order to be able to fulfil properly WP4 activities as well as further dissemination of the project.
- T5.2: Dissemination
  - The web page was updated timely.
  - The fourth MobiNetVideo Newsletter was published December 2020.
- T5.3 Workshop organization
  - Developers Workshops are cancelled. Currently, we are evaluating the planning of a final workshop for the disseminating project results to EPOs and industry.