Activity Report 2019

Project-Team Stars

Spatio-Temporal Activity Recognition Systems

RESEARCH CENTER
Sophia Antipolis - Méditerranée

THEME
Vision, perception and multimedia interpretation

Table of contents

Project-Team Stars

Creation of the Team: 2012 January 01, updated into Project-Team: 2013 January 01

Keywords:

Computer Science and Digital Science:

A2.1.9. - Synchronous languages
A2.1.11. - Proof languages
A2.3.3. - Real-time systems
A2.4.2. - Model-checking
A2.4.3. - Proofs
A2.5. - Software engineering
A3.2.1. - Knowledge bases
A3.3.2. - Data mining
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.5. - Bayesian methods
A3.4.6. - Neural networks
A4.7. - Access control
A5.1. - Human-Computer Interaction
A5.3.2. - Sparse modeling and image representation
A5.3.3. - Pattern recognition
A5.4.1. - Object recognition
A5.4.2. - Activity recognition
A5.4.3. - Content retrieval
A5.4.5. - Object tracking and motion analysis
A9.1. - Knowledge
A9.2. - Machine learning
A9.3. - Signal analysis

Other Research Topics and Application Domains:

B1.2.2. - Cognitive science
B2.1. - Well being
B7.1.1. - Pedestrian traffic and crowds
B8.1. - Smart building/home
B8.4. - Security and personal assistance

1. Team, Visitors, External Collaborators

Research Scientists

François Brémond [Team leader, Inria, Senior Researcher, HDR]
Sabine Moisan [Inria, Researcher, HDR]
Jean-Paul Rigault [University Côte d'Azur, Emeritus]
Monique Thonnat [Inria, Senior Researcher, HDR]
Antitza Dantcheva [Inria, Researcher]

Philippe Robert [Inria, CoBTeK, Senior Researcher]

Faculty Members

Elisabetta de Maria [University Côte d'Azur, Associate Professor]
Fréderic Précioso [University Côte d'Azur, Associate Professor, until Sep 2019]

Post-Doctoral Fellows

Michal Balazia [University Côte d'Azur, from Oct 2019]
Abhijit Das [Inria, until Apr 2019]
S L Happy [Inria, until Aug 2019]
Alexandra König [Inria, CoBTeK, until Nov 2019]
Ujjwal Ujjwal [Inria, from Dec 2019]

PhD Students

Ujjwal Ujjwal [VEDECOM, until Nov 2019]
Hao Chen [ESI, CIFRE fellowship, from May 2019]
Rui Dai [University Côte d'Azur, from Oct 2019]
Srijan Das [University Côte d'Azur]
Juan Diego Gonzales Zuniga [Kontron, CIFRE fellowship, from April 2018]
Jen Cheng Hou [Inria]
Thibaud L'Yvonnet [Inria]
Yaohui Wang [Inria]

Technical staff

Sébastien Gilabert [Inria, Engineer]
Rachid Guerchouche [Inria, Engineer]
Soumik Mallick [Inria, Engineer, until Mar 2019]
Minh Khue Phan Tran [Inria, Engineer, until Jul 2019, granted by Bpifrance Financement S.A.]
Duc Minh Tran [Inria, Engineer]

Interns and Apprentices

Di Yang [Ekinnox, from Mar 2019 until Aug 2019]
Jérémie Bertrand [Inria, from Jun 2019 until Aug 2019]
Snehashis Majhi [Inria, from Sep 2019]
Rodrigo Ignacio Rivera Galvez [Inria, until Mar 2019]
Sagar Sagar [Inria, from Feb 2019 until May 2019]
Saurav Sharma [Inria, from Jul 2019 until Oct 2019]
Vikas Thamizharasan [Inria, until Jan 2019]

Visiting Scientists

Nagi Ould Taled Aly [PhD, University of Nouakchott, Mauritania, until Mar 2019]
David Anghelone [Thales, from Sep 2019]
Seungryul Baek [PhD, Imperial College London, from Jun 2019 until Aug 2019]
Abdorrahim Bahrami [PhD, University of Ottawa, from Jun 2019 until Sep 2019]
Gaelle Darrot [Univ Côte d'Azur, from Oct 2019 until Nov 2019]
Carole Hanocq [Univ Côte d'Azur, from Oct 2019 until Nov 2019]
Audrey Sayaque [Univ Côte d'Azur, from Oct 2019 until Nov 2019]

External Collaborators

Baptiste Fosty [Ekinnox, until Feb 2019]
Daniel Gaffé [University Côte d'Azur]
Benjamin Renoust [Median Technologies, from Nov 2019]
Annie Ressouche [University Côte d'Azur, retired]
Ines Sarray [Univ Côte d'Azur]
Jean-Yves Tigli [University Côte d'Azur, until Jan 2019]
Piotr Tadeusz Bilinski [Oxford University, until Nov 2019]
Carlos Caetano [Federal University of Minas Gerais, Brazil, until Oct 2019]

2. Overall Objectives

2.1. Presentation

The **STARS (Spatio-Temporal Activity Recognition Systems)** team focuses on the design of cognitive vision systems for Activity Recognition. More precisely, we are interested in the real-time semantic interpretation of dynamic scenes observed by video cameras and other sensors. We study long-term spatio-temporal activities performed by agents such as human beings, animals or vehicles in the physical world. The major issue in semantic interpretation of dynamic scenes is to bridge the gap between the subjective interpretation of data and the objective measures provided by sensors. To address this problem Stars develops new techniques in the field of computer vision, machine learning and cognitive systems for physical object detection, activity understanding, activity learning, vision system design and evaluation. We focus on two principal application domains: visual surveillance and healthcare monitoring.

2.1.1. Research Themes

Stars is focused on the design of cognitive systems for Activity Recognition. We aim at endowing cognitive systems with perceptual capabilities to reason about an observed environment, to provide a variety of services to people living in this environment while preserving their privacy. In today world, a huge amount of new sensors and new hardware devices are currently available, addressing potentially new needs of the modern society. However the lack of automated processes (with no human interaction) able to extract a meaningful and accurate information (i.e. a correct understanding of the situation) has often generated frustrations among the society and especially among older people. Therefore, Stars objective is to propose novel autonomous systems for the **real-time semantic interpretation of dynamic scenes** observed by sensors. We study long-term spatio-temporal activities performed by several interacting agents such as human beings, animals and vehicles in the physical world. Such systems also raise fundamental software engineering problems to specify them as well as to adapt them at run time.

We propose new techniques at the frontier between computer vision, knowledge engineering, machine learning and software engineering. The major challenge in semantic interpretation of dynamic scenes is to bridge the gap between the task dependent interpretation of data and the flood of measures provided by sensors. The problems we address range from physical object detection, activity understanding, activity learning to vision system design and evaluation. The two principal classes of human activities we focus on, are assistance to older adults and video analytic.

A typical example of a complex activity is shown in Figure 1 and Figure 2 for a homecare application. In this example, the duration of the monitoring of an older person apartment could last several months. The activities involve interactions between the observed person and several pieces of equipment. The application goal is to recognize the everyday activities at home through formal activity models (as shown in Figure 3) and data captured by a network of sensors embedded in the apartment. Here typical services include an objective assessment of the frailty level of the observed person to be able to provide a more personalized care and to monitor the effectiveness of a prescribed therapy. The assessment of the frailty level is performed by an Activity Recognition System which transmits a textual report (containing only meta-data) to the general practitioner who follows the older person. Thanks to the recognized activities, the quality of life of the observed people can thus be improved and their personal information can be preserved.

The ultimate goal is for cognitive systems to perceive and understand their environment to be able to provide appropriate services to a potential user. An important step is to propose a computational representation of people activities to adapt these services to them. Up to now, the most effective sensors have been video cameras due to the rich information they can provide on the observed environment. These sensors are currently perceived as intrusive ones. A key issue is to capture the pertinent raw data for adapting the services to the people while preserving their privacy. We plan to study different solutions including of course the local processing of the data without transmission of images and the utilization of new compact sensors developed for interaction (also called RGB-Depth sensors, an example being the Kinect) or networks of small non visual sensors.

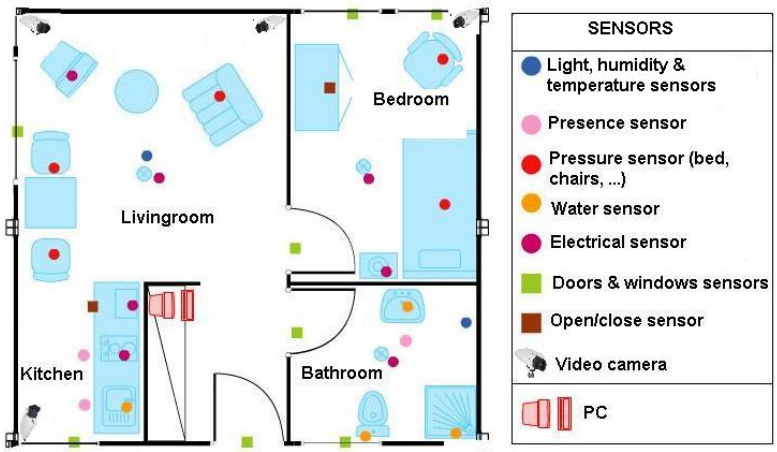

Figure 1. Homecare monitoring: the set of sensors embedded in an apartment

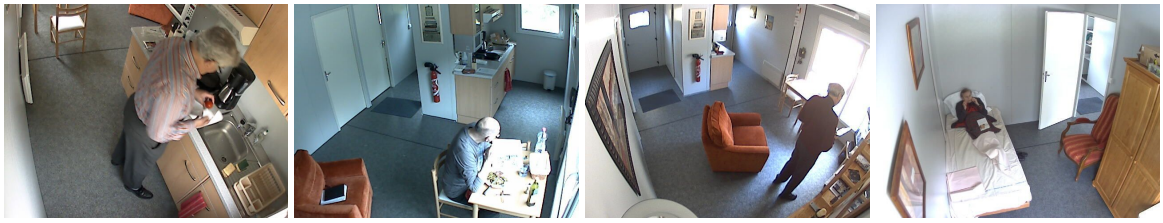

Figure 2. Homecare monitoring: the different views of the apartment captured by 4 video cameras

Activity (*PrepareMeal*,
PhysicalObjects((p : Person), (z : Zone), (eq : Equipment))
Components((s_inside : InsideKitchen(p, z))
(s_close : CloseToCountertop(p, eq))
(s_stand : PersonStandingInKitchen(p, z)))
Constraints((z->Name = Kitchen)
(eq->Name = Countertop)
(s_close->Duration >= 100)
(s_stand->Duration >= 100))
Annotation(AText("prepare meal")
AType("not urgent")))

Figure 3. Homecare monitoring: example of an activity model describing a scenario related to the preparation of a meal with a high-level language

2.1.2. International and Industrial Cooperation

Our work has been applied in the context of more than 10 European projects such as COFRIEND, ADVISOR, SERKET, CARETAKER, VANAHEIM, SUPPORT, DEM@CARE, VICOMO, EIT Health. We had or have industrial collaborations in several domains: *transportation* (CCI Airport Toulouse Blagnac, SNCF, Inrets, Alstom, Ratp, Toyota, GTT (Italy), Turin GTT (Italy)), *banking* (Crédit Agricole Bank Corporation, Eurotelis and Ciel), *security* (Thales R&T FR, Thales Security Syst, EADS, Sagem, Bertin, Alcatel, Keeneo), *multimedia* (Thales Communications), *civil engineering* (Centre Scientifique et Technique du Bâtiment (CSTB)), *computer industry* (BULL), *software industry* (AKKA), *hardware industry* (ST-Microelectronics) and *health industry* (Philips, Link Care Services, Vistek).

We have international cooperations with research centers such as Reading University (UK), ENSI Tunis (Tunisia), Idiap (Switzerland), Multitel (Belgium), National Cheng Kung University, National Taiwan University (Taiwan), MICA (Vietnam), IPAL, I2R (Singapore), University of Southern California, University of South Florida, University of Maryland (USA).

2.1.2.1. Industrial Contracts

- *Toyota:* (Action Recognition System):
  This project run from the 1st of August 2013 up to 2023. It aimed at detecting critical situations in the daily life of older adults living home alone. The system is intended to work with a Partner Robot (to send real-time information to the robot for assisted living) to better interact with older adults. The funding was 106 Keuros for the 1st period and more for the following years.
- *Gemalto:* This contract is a CIFRE PhD grant and runs from September 2018 until September 2021 within the French national initiative SafeCity. The main goal is to analyze faces and events in the invisible spectrum (i.e., low energy infrared waves, as well as ultraviolet waves). In this context models will be developed to efficiently extract identity, as well as event - information. This models will be employed in a school environment, with a goal of pseudo-anonymized identification, as well as event-detection. Expected challenges have to do with limited colorimetry and lower contrasts.
- *Kontron:* This contract is a CIFRE PhD grant and runs from April 2018 until April 2021 to embed CNN based people tracker within a video-camera.
- *ESI:* This contract is a CIFRE PhD grant and runs from September 2018 until March 2022 to develop a novel Re-Identification algorithm which can be easily set-up with low interaction.

3. Research Program

3.1. Introduction

Stars follows three main research directions: perception for activity recognition, action recognition and semantic activity recognition. **These three research directions are organized following the workflow of activity recognition systems**: First, *the perception* and *the action recognition* directions provide new techniques to extract powerful features, whereas *the semantic activity recognition* research direction provides new paradigms to match these features with concrete video analytic and healthcare applications.

Transversely, we consider a *new research axis in machine learning*, combining a priori knowledge and learning techniques, to set up the various models of an activity recognition system. A major objective is to automate model building or model enrichment at the perception level and at the understanding level.

3.2. Perception for Activity Recognition

Participants: François Brémond, Antitza Dantcheva, Sabine Moisan, Monique Thonnat.

Activity Recognition, Scene Understanding, Machine Learning, Computer Vision, Cognitive Vision Systems, Software Engineering

3.2.1. Introduction

Our main goal in perception is to develop vision algorithms able to address the large variety of conditions characterizing real world scenes in terms of sensor conditions, hardware requirements, lighting conditions, physical objects, and application objectives. We have also several issues related to perception which combine machine learning and perception techniques: learning people appearance, parameters for system control and shape statistics.

3.2.2. Appearance Models and People Tracking

An important issue is to detect in real-time physical objects from perceptual features and predefined 3D models. It requires finding a good balance between efficient methods and precise spatio-temporal models. Many improvements and analysis need to be performed in order to tackle the large range of people detection scenarios.

Appearance models. In particular, we study the temporal variation of the features characterizing the appearance of a human. This task could be achieved by clustering potential candidates depending on their position and their reliability. This task can provide any people tracking algorithms with reliable features allowing for instance to (1) better track people or their body parts during occlusion, or to (2) model people appearance for re-identification purposes in mono and multi-camera networks, which is still an open issue. The underlying challenge of the person re-identification problem arises from significant differences in illumination, pose and camera parameters. The re-identification approaches have two aspects: (1) establishing correspondences between body parts and (2) generating signatures that are invariant to different color responses. As we have already several descriptors which are color invariant, we now focus more on aligning two people detection and on finding their corresponding body parts. Having detected body parts, the approach can handle pose variations. Further, different body parts might have different influence on finding the correct match among a whole gallery dataset. Thus, the re-identification approaches have to search for matching strategies. As the results of the re-identification are always given as the ranking list, re-identification focuses on learning to rank. "Learning to rank" is a type of machine learning problem, in which the goal is to automatically construct a ranking model from a training data.

Therefore, we work on information fusion to handle perceptual features coming from various sensors (several cameras covering a large scale area or heterogeneous sensors capturing more or less precise and rich information). New 3D RGB-D sensors are also investigated, to help in getting an accurate segmentation for specific scene conditions.

Long term tracking. For activity recognition we need robust and coherent object tracking over long periods of time (often several hours in video surveillance and several days in healthcare). To guarantee the long term coherence of tracked objects, spatio-temporal reasoning is required. Modeling and managing the uncertainty of these processes is also an open issue. In Stars we propose to add a reasoning layer to a classical Bayesian framework modeling the uncertainty of the tracked objects. This reasoning layer can take into account the a priori knowledge of the scene for outlier elimination and long-term coherency checking.

Controlling system parameters. Another research direction is to manage a library of video processing programs. We are building a perception library by selecting robust algorithms for feature extraction, by insuring they work efficiently with real time constraints and by formalizing their conditions of use within a program supervision model. In the case of video cameras, at least two problems are still open: robust image segmentation and meaningful feature extraction. For these issues, we are developing new learning techniques.

3.3. Action Recognition

Participants: François Brémond, Antitza Dantcheva, Monique Thonnat.

Machine Learning, Computer Vision, Cognitive Vision Systems

3.3.1. Introduction

Due to the recent development of high processing units, such as GPU, this is now possible to extract meaningful features directly from videos (e.g. video volume) to recognize reliably short actions. Action Recognition benefits also greatly from the huge progress made recently in Machine Learning (e.g. Deep Learning), especially for the study of human behavior. For instance, Action Recognition enables to measure objectively the behavior of humans by extracting powerful features characterizing their everyday activities, their emotion, eating habits and lifestyle, by learning models from a large number of data from a variety of sensors, to improve and optimize for example, the quality of life of people suffering from behavior disorders. However, Smart Homes and Partner Robots have been well advertised but remain laboratory prototypes, due to the poor capability of automated systems to perceive and reason about their environment. A hard problem is for an automated system to cope 24/7 with the variety and complexity of the real world. Another challenge is to extract people fine gestures and subtle facial expressions to better analyze behavior disorders, such as anxiety or apathy. Taking advantage of what is currently studied for self-driving cars or smart retails, there is a large avenue to design ambitious approaches for the healthcare domain. In particular, the advance made with Deep Learning algorithms has already enabled to recognize complex activities, such as cooking interactions with instruments, and from this analysis to differentiate healthy people from the ones suffering from dementia.

To address these issues, we propose to tackle several challenges:

3.3.2. Action recognition in the wild

The current Deep Learning techniques are mostly developed to work on few clipped videos, which have been recorded with students performing a limited set of predefined actions in front of a camera with high resolution. However, real life scenarios include actions performed in a spontaneous manner by older people (including people interactions with their environment or with other people), from different viewpoints, with varying framerate, partially occluded by furniture at different locations within an apartment depicted through long untrimmed videos. Therefore, a new dedicated dataset should be collected in a real-world setting to become a public benchmark video dataset and to design novel algorithms for ADL activity recognition. A special attention should be taken to anonymize the videos.

3.3.3. Attention mechanisms for action recognition

Activities of Daily Living (ADL) and video-surveillance activities are different from internet activities (e.g. Sports, Movies, YouTube), as they may have very similar context (e.g. same background kitchen) with high intra-variation (different people performing the same action in different manners), but in the same time low inter-variation, similar ways to perform two different actions (e.g. eating and drinking a glass of water). Consequently, fine-grained actions are badly recognized. So, we will design novel attention mechanisms for action recognition, for the algorithm being able to focus on a discriminative part of the person conducting the action. For instance, we will study attention algorithms, which could focus on the most appropriate body parts (e.g. full body, right hand). In particular, we plan to design a soft mechanism, learning the attention weights directly on the feature map of a 3DconvNet, a powerful convolutional network, which takes as input a batch of videos.

3.3.4. Action detection for untrimmed videos

Many approaches have been proposed to solve the problem of action recognition in short clipped 2D videos, which achieved impressive results with hand-crafted and deep features. However, these approaches cannot address real life situations, where cameras provide online and continuous video streams in applications such as robotics, video surveillance, and smart-homes. Here comes the importance of action detection to help recognizing and localizing each action happening in long videos. Action detection can be defined as the ability to localize starting and ending of each human action happening in the video, in addition to recognizing each action label. There have been few action detection algorithms designed for untrimmed videos, which are based on either sliding window, temporal pooling or frame-based labeling. However, their performance is too low to address real-word datasets. A first task consists in benchmarking the already published approaches to study their limitations on novel untrimmed video datasets, recorded following real-world settings. A second task

could be to propose a new mechanism to improve either 1) the temporal pooling directly from the 3DconvNet architecture using for instance Temporal Convolution Networks (TCNs) or 2) frame-based labeling with a clustering technique (e.g. using Fisher Vectors) to discover the sub-activities of interest.

3.3.5. View invariant action recognition

The performance of current approaches strongly relies on the used camera angle: enforcing that the camera angle used in testing is the same (or extremely close to) as the camera angle used in training, is necessary for the approach performs well. On the contrary, the performance drops when a different camera view-point is used. Therefore, we aim at improving the performance of action recognition algorithms by relying on 3D human pose information. For the extraction of the 3D pose information, several open-source algorithms can be used, such as openpose or videopose3D (from CMU or Facebook research, https://github.com/CMU-Perceptual-Computing-Lab/openpose. Also, other algorithms extracting 3d meshes can be used. To generate extra views, Generative Adversial Network (GAN) can be used together with the 3D human pose information to complete the training dataset from the missing view.

3.3.6. Uncertainty and action recognition

Another challenge is to combine the short-term actions recognized by powerful Deep Learning techniques with long-term activities defined by constraint-based descriptions and linked to user interest. To realize this objective, we have to compute the uncertainty (i.e. likelihood or confidence), with which the short-term actions are inferred. This research direction is linked to the next one, to Semantic Activity Recognition.

3.4. Semantic Activity Recognition

Participants: François Brémond, Sabine Moisan, Monique Thonnat.

Activity Recognition, Scene Understanding, Computer Vision

3.4.1. Introduction

Semantic activity recognition is a complex process where information is abstracted through four levels: signal (e.g. pixel, sound), perceptual features, physical objects and activities. The signal and the feature levels are characterized by strong noise, ambiguous, corrupted and missing data. The whole process of scene understanding consists in analyzing this information to bring forth pertinent insight of the scene and its dynamics while handling the low level noise. Moreover, to obtain a semantic abstraction, building activity models is a crucial point. A still open issue consists in determining whether these models should be given a priori or learned. Another challenge consists in organizing this knowledge in order to capitalize experience, share it with others and update it along with experimentation. To face this challenge, tools in knowledge engineering such as machine learning or ontology are needed.

Thus we work along the following research axes: high level understanding (to recognize the activities of physical objects based on high level activity models), learning (how to learn the models needed for activity recognition) and activity recognition and discrete event systems.

3.4.2. High Level Understanding

A challenging research axis is to recognize subjective activities of physical objects (i.e. human beings, animals, vehicles) based on a priori models and objective perceptual measures (e.g. robust and coherent object tracks).

To reach this goal, we have defined original activity recognition algorithms and activity models. Activity recognition algorithms include the computation of spatio-temporal relationships between physical objects. All the possible relationships may correspond to activities of interest and all have to be explored in an efficient way. The variety of these activities, generally called video events, is huge and depends on their spatial and temporal granularity, on the number of physical objects involved in the events, and on the event complexity (number of components constituting the event).

Concerning the modeling of activities, we are working towards two directions: the uncertainty management for representing probability distributions and knowledge acquisition facilities based on ontological engineering techniques. For the first direction, we are investigating classical statistical techniques and logical approaches. For the second direction, we built a language for video event modeling and a visual concept ontology (including color, texture and spatial concepts) to be extended with temporal concepts (motion, trajectories, events ...) and other perceptual concepts (physiological sensor concepts ...).

3.4.3. Learning for Activity Recognition

Given the difficulty of building an activity recognition system with a priori knowledge for a new application, we study how machine learning techniques can automate building or completing models at the perception level and at the understanding level.

At the understanding level, we are learning primitive event detectors. This can be done for example by learning visual concept detectors using SVMs (Support Vector Machines) with perceptual feature samples. An open question is how far can we go in weakly supervised learning for each type of perceptual concept (i.e. leveraging the human annotation task). A second direction is to learn typical composite event models for frequent activities using trajectory clustering or data mining techniques. We name composite event a particular combination of several primitive events.

3.4.4. Activity Recognition and Discrete Event Systems

The previous research axes are unavoidable to cope with the semantic interpretations. However they tend to let aside the pure event driven aspects of scenario recognition. These aspects have been studied for a long time at a theoretical level and led to methods and tools that may bring extra value to activity recognition, the most important being the possibility of formal analysis, verification and validation.

We have thus started to specify a formal model to define, analyze, simulate, and prove scenarios. This model deals with both absolute time (to be realistic and efficient in the analysis phase) and logical time (to benefit from well-known mathematical models providing re-usability, easy extension, and verification). Our purpose is to offer a generic tool to express and recognize activities associated with a concrete language to specify activities in the form of a set of scenarios with temporal constraints. The theoretical foundations and the tools being shared with Software Engineering aspects.

The results of the research performed in perception and semantic activity recognition (first and second research directions) produce new techniques for scene understanding and contribute to specify the needs for new software architectures (third research direction).

4. Highlights of the Year

4.1. Highlights of the Year

4.1.1. People detection

People detection is a very challenging topic, where many top-level research groups are competing and have already proposed impressive approaches (e.g. Faster-RCNN, SSD, YOLO). Yet, we were able to design a novel algorithm able to better balance the speed and accuracy trade-off on the most challenging pedestrian detection benchmarks (e.g. Caltech and Citypersons).

4.1.2. Person Re-Identification

Person Re-Identification is a very challenging task, where current Computer Vision algorithms manage to obtain better results than humans. By proposing a simple and elegant technique, based on Spatial-Channel Partitions, we have obtained the best performance compared to the State-of-the-art approaches on the most popular benchmark datasets (e.g. Market-1501, CUHK03 and MARS).

4.1.3. Action recognition

This year, we have proposed several action recognition approaches able to outperform the State-of-the-art algorithms and get nearly maximal performance on most of ADL benchmark video datasets (e.g. Northwestern-UCLA Multiview Action 3D, NTUTU-RGB and DAHLIA). We have also released a novel ADL benchmark video dataset, which is more challenging, as it has been collected within real-world settings.

4.1.4. Awards

Antitza Dantcheva and Abhijit Das recceived a Best Poster Award at the 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019) in Lille, France (the flagship face analysis conference) for the paper: "Robust remote heart rate estimation from face utilizing spatial-temporal attention" [28].

5. New Software and Platforms

5.1. SUP

Scene Understanding Platform

KEYWORDS: Activity recognition - 3D - Dynamic scene

FUNCTIONAL DESCRIPTION: SUP is a software platform for perceiving, analyzing and interpreting a 3D dynamic scene observed through a network of sensors. It encompasses algorithms allowing for the modeling of interesting activities for users to enable their recognition in real-world applications requiring high-throughput.

- Participants: Etienne Corvée, François Brémond, Hung Nguyen and Vasanth Bathrinarayanan
- Partners: CEA - CHU Nice - USC Californie - Université de Hamburg - I2R
- Contact: François Brémond
- URL: https://team.inria.fr/stars/software

5.2. VISEVAL

FUNCTIONAL DESCRIPTION: ViSEval is a software dedicated to the evaluation and visualization of video processing algorithm outputs. The evaluation of video processing algorithm results is an important step in video analysis research. In video processing, we identify 4 different tasks to evaluate: detection, classification and tracking of physical objects of interest and event recognition.

- Participants: Bernard Boulay and François Brémond
- Contact: François Brémond
- URL: http://www-sop.inria.fr/teams/pulsar/EvaluationTool/ViSEvAl_Description.html

6. New Results

6.1. Introduction

This year Stars has proposed new results related to its three main research axes: (i) perception for activity recognition, (ii) action recognition and (iii) semantic activity recognition.

6.1.1. Perception for Activity Recognition

Participants: François Brémond, Juan Diego Gonzales Zuniga, Abhijit Das, Antitza Dantcheva, Ujjwal Ujjwal, Srijan Das, David Anghelone, Monique Thonnat.

The new results for perception for activity recognition are:

- Handling the Speed-Accuracy Trade-off in Deep Learning based Pedestrian Detection (see 6.2)
- Deep Learning applied on Embedded Systems for People Tracking (see 6.3)
- Partition and Reunion: A Two-Branch Neural Network for Vehicle Re- identification (see 6.4)
- Improving Face Sketch Recognition via Adversarial Sketch-Photo Transformation (see 6.5)
- Impact and Detection of Facial Beautification in Face Recognition: An Overview (see 6.6)
- Computer Vision and Deep Learning applied to Facial analysis in the invisible spectra (see 6.7)

6.1.2. Action Recognition

Participants: François Brémond, Juan Diego Gonzales Zuniga, Abhijit Das, Antitza Dantcheva, Ujjwal Ujjwal, Srijan Das, Monique Thonnat.

The new results for action recognition are:

- ImaGINator: Conditional Spatio-Temporal GAN for Video Generation (see 6.8)
- Characterizing the State of Apathy with Facial Expression and Motion Analysis (see 6.9)
- Dual-threshold Based Local Patch Construction Method for Manifold Approximation And Its Application to Facial Expression Analysis (see 6.10)
- A Weakly Supervised Learning Technique for Classifying Facial Expressions (see 6.11)
- Robust Remote Heart Rate Estimation from Face Utilizing Spatial- temporal Attention (see 6.12)
- Quantified Analysis for Epileptic Seizure Videos (see 6.13)
- Toyota Smarthome: Real-World Activities of Daily Living (see 6.15)
- Looking deeper into Time for Activities of Daily Living Recognition (see 6.15.1)
- Self-Attention Temporal Convolutional Network for Long-Term Daily Living Activity Detection (see 6.16)

6.1.3. Semantic Activity Recognition

Participants: François Brémond, Elisabetta de Maria, Antitza Dantcheva, Srijan Das, Abhijit Das, Daniel Gaffé, Thibaud L'Yvonnet, Sabine Moisan, Jean-Paul Rigault, Annie Ressouche, Ines Sarray, Yaohui Wang, S L Happy, Alexandra König, Philippe Robert, Monique Thonnat.

For this research axis, the contributions are:

- DeepSpa Project (see 6.17)
- Store Connect and Solitaria (see 6.18)
- Synchronous Approach to Activity Recognition (see 6.19)
- Probabilistic Activity Modeling (see 6.20)

6.2. Handling the Speed-Accuracy Trade-off in Deep Learning based Pedestrian Detection

Participants: François Brémond, Ujjwal Ujjwal.

Pedestrian detection is a specific instance of the more general problem of object detection. Pedestrian detection plays a fundamental role in many modern applications involving but not limited to *autonomous vehicles* and *surveillance systems*. These applications as many others are safety-critical. This implies that the cost of not correctly detecting a pedestrian is very high. At the same time applications such as the ones mentioned before, are expected to be real-time. This implies that a pedestrian be detected with minimum time delay. The subject of our recent work has been to design a pedestrian detector which is capable of detecting pedestrians with a high accuracy and high speed – two traits which are known to be difficult to achieve simultaneously.

Most of the pedestrian detectors in computer vision are derived from general-category object detectors. We reflect upon its implication in terms of speed and accuracy below.

6.2.1. Speed-Accuracy Trade-off

Speed and accuracy of object detectors are mutually trade-off factors. Emphasis on higher accuracy usually entails intensive computations which sacrifice the detection speed. On the other hand, emphasis on higher detection speed usually leads to simpler computations which sacrifice the detection accuracy.

We have recently been able to balance this trade-off by identifying that the means of computations on anchors are a major source of the speed-accuracy trade-off. Anchors are hypothetical bounding boxes and are reminiscent of sliding windows used in earlier works on object detection. There are two distinct means of processing anchors – *feature pooling* and *feature probing*. We have recently demonstrated that feature pooling is a costlier strategy than feature probing in terms of computational cost. However, in contrast, feature pooling is a more precise means to process anchors.

We leverage this difference in our approach by utilizing feature pooling throughout in our system. However, in order to gain in terms of run-time performance, we reduce the number of anchors to be processed. This reduction does allow us to process a small number of relevant anchors with high precision.

The block diagram of our proposed approach is shown in figure 4.

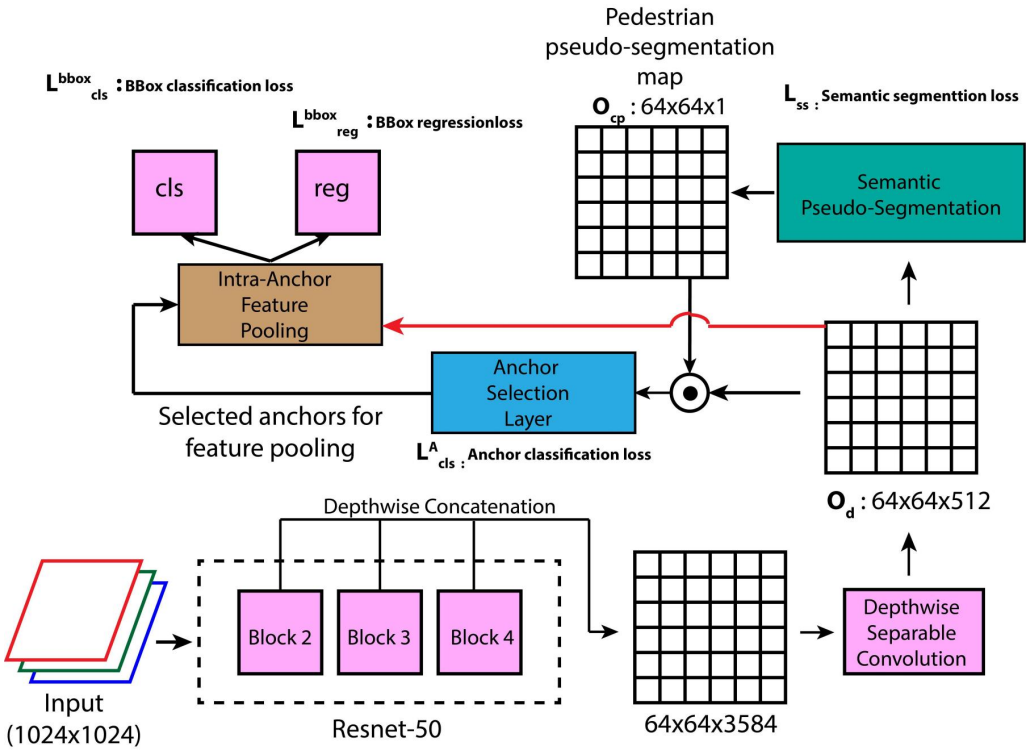

Figure 4. The block diagram of our proposed approach

We fuse the feature maps of multiple layers in order to improve the feature diversity. An increased feature diversity assists in learning from a range of hierarchical features generated by a convolutional neural network, often abbreviated as CNN. A depth-wise separable convolutional layer then further processes the fused feature

map in order to reduce the number of feature dimensions. One of the prime novelties in our work is the use of pseudo-semantic segmentation. Pseudo-semantic segmentation allows one to obtain a rough estimate of the localization of pedestrians in the form of a heatmap. This step is important, as it provides us with a basis to select a small set of anchors instead of processing all the tiling anchors on the feature map. An anchor classification layer uses anchor-specific kernel sizes to classify a given anchor as positive or negative. A positive or negative anchor is characterized by the overlap between the anchor and the ground truth bounding box during training. This overlap is measured in terms of the well known intersection-over-union (IoU) metric in computer vision. The positive anchors are then pooled from, followed by classification and regression to obtain the final detection.

6.2.2. Results

Method	**Stages**	**LAMR**		**Speed**
		caltech-reasonable (test) (w/o CP pre-training) (CP pre-trained)	**citypersons (val) (trained only on CP)**	
Faster-RCNN	2	12.10	15.4	7
SSD	1	17.78 (16.36)	19.69	48
YOLOv2	1	21.62 (20.83)	NA	60
RPN-BF	2	9.6 (NA)	NA	7
MS-CNN	2	10.0 (NA)	NA	8
SDS-RCNN	2	7.6 (NA)	NA	5
ALF-Net	1	4.5 (NA)	12.0	20
Rep-Loss	2	5.0 (4.0)	13.2	-
Ours	1.5	**4.76 (3.99)**	**8.12**	**32**

Figure 5. Performance comparison of the proposed method with other methods for caltech-reasonable test set and citypersons validation set. The speed figures are in frames per second.

Figure 5 summarizes the performance of the proposed approach vis-à-vis other approaches. The proposed approach provides significant improvements over other approaches in terms of both speed and accuracy. From figure 5 it is clear that we benefit from initial training on the citypersons data set. Moreover, we obtain the state-of-art performance on the citypersons data set, improving the existing best performing techniques by nearly 4 LAMR points.

6.3. Deep Learning applied on Embedded Systems for People Tracking

Participants: Juan Diego Gonzales Zuniga, Ujjwal Ujjwal, François Brémond, Serge Tissot [Kontron].

Our work objective is two-fold: a) Perform tracking of multiple people in videos, which is an instance of Multiple Object Tracking (MOT) problem, and b) optimize this tracking on embedded and open source hardware platforms such as OpenVINO and ROCm.

People tracking is a challenging and relevant problem since it needs multiple additional modules to perform the data association between nodes. In addition, state-of-the-art solutions require intensive memory allocation and power consumption which are not available on embedded hardware. Most architectures either require great amounts of memory or large computing time to achieve a state-of-the-art performance, these results are mostly achieved with dedicated hardware at data centers.

6.3.1. Online Joint Detection and Tracking

In people tracking, we are questioning the main paradigm that is tracking-by-detection which heavily relies on the performance of the underlying detection method. This requires access to a highly accurate and robust people detector. On the other hand, few frameworks attempt detect and track people jointly. Our intent is to perform people tracking *online* and *jointly with detection*.

We are trying to determinate a manner in which a single model can both perform detection and tracking simultaneously. Along these lines, we experimented with a variation of I3D on the Posetrack data set that takes an input of 8 frames in order to create heatmaps along multiple frames as seen in Figure 6. Giving that the data of Posetrack or MOT cannot train a network as I3D, we are doing the pretraining with the synthetic JTA-Dataset.

This work is inspired by the less common methods of traking-by-tracks and tracking-by-tracklets. Both [40] and [41] generate multi-frame bounding box tuplet proposals and extract detection scores and features with a CNN and LSTM, respectively. Recent researches improve object detection by applying optical flow to propagate scores between frames.

Another method we implemented is by using the detections of previous frames as proposal for the data association, it only uses the IOU between two objects as a distance metric. This approach is simple and efficient assuming the objects do not move drastically. An improved method increases the performance by using a siamese network to conserve identity across frames and predictions for death and birth of tracks.

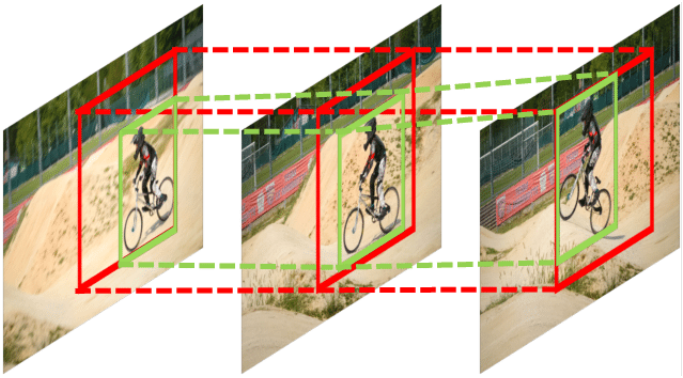

Figure 6. People tracking by tubelets

6.3.2. OpenVINO and ROCm

Regarding embedded hardware, we focus on enlarging both implementation and experimentation of two specific frameworks; OpenVINO and ROCm.

OpenVINO allows us to transfer deep learning models into Myriad and KeemBay chips, taking advantage of their capacity to compute multiple operations without the need of much power consumption. We have thoroughly tested their power consumption under different scenarios as well as implemented many qualitative algorithms with these two platforms, Figure 7 shows the Watt consumption and frame rate of the most popular backbone networks, making it viable to use on embedded applications with a reasonable 25FPS.

For ROCm, we have used the approach of [38] to optimize the compiler execution for a variety of CNN features and filters using a substitute GPU with similar computation capability as Nvidia but still remaining a low branch consumption around 15 Watts.

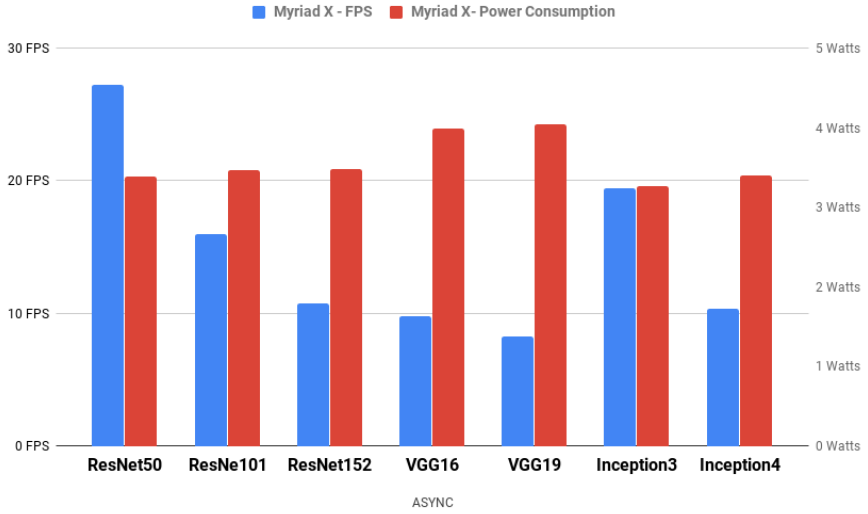

Figure 7. Power Consumption vs Frame rate

6.4. Partition and Reunion: A Two-Branch Neural Network for Vehicle Re-identification

Participants: Hao Chen, Benoit Lagadec, François Brémond.

The smart city vision raises the prospect that cities will become more intelligent in various fields, such as more sustainable environment and a better quality of life for residents. As a key component of smart cities, intelligent transportation system highlights the importance of vehicle re-identification (Re-ID). However, as compared to the rapid progress on person Re-ID, vehicle Re-ID advances at a relatively slow pace. Some previous state-of-the-art approaches strongly rely on extra annotation, like attributes (vehicle color and type) and key-points (wheels and lamps). Recent work on person Re-ID shows that extracting more local features can achieve a better performance without considering extra annotation. In this work, we propose an end-to-end trainable two-branch Partition and Reunion Network (PRN) for the challenging vehicle Re-ID task. Utilizing only identity labels, our proposed method outperforms existing state-of-the-art methods on four vehicle Re-ID benchmark datasets, including VeRi-776, VehicleID, VRIC and CityFlow-ReID by a large margin. The general architecture of our proposed method is represented in the Figure 8.

6.4.1. Learning Discriminative and Generalizable Representations by Spatial-Channel Partition for Person Re-Identification

In Person Re-Identification (Re-ID) task, combining local and global features is a common strategy to overcome missing key parts and misalignment on models based only on global features. Using this combination, neural networks yield impressive performance in Re-ID task. Previous part-based models mainly focus on spatial partition strategies. Recently, operations on channel information, such as Group Normalization and Channel Attention, have brought significant progress to various visual tasks. However, channel partition has not drawn much attention in Person Re-ID. We conduct a study to exploit the potential of channel partition in Re-ID task [32]. Based on this study, we propose an end-to-end Spatial and Channel partition Representation network (SCR) in order to better exploit both spatial and channel information. Experiments conducted on three mainstream image-based evaluation protocols including Market-1501, DukeMTMC-ReID and CUHK03 and

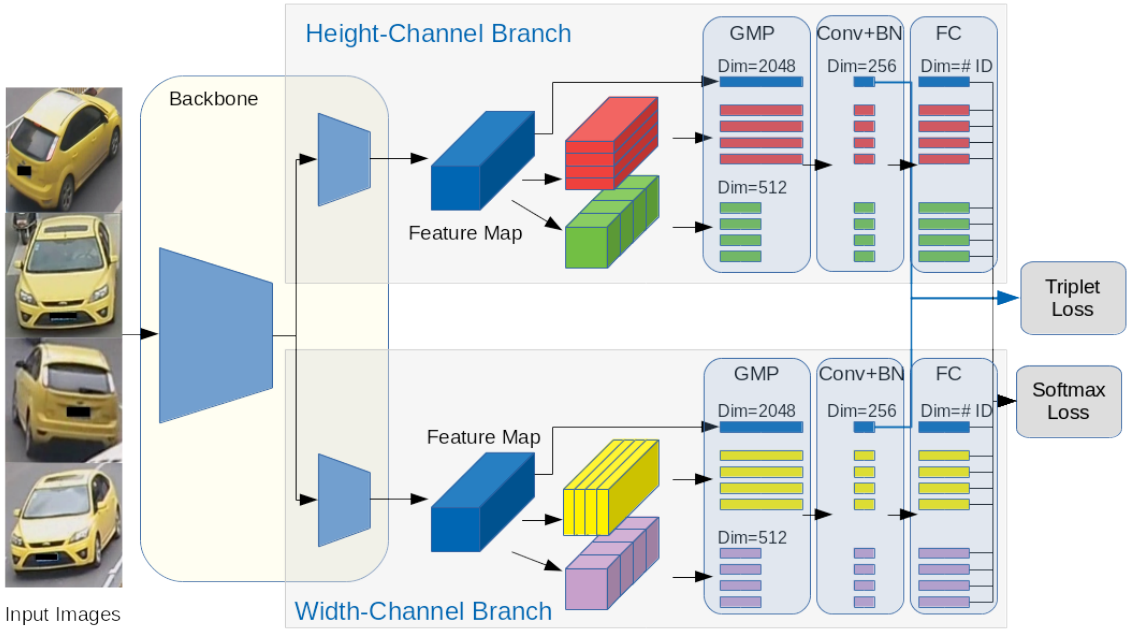

*Figure 8. General architecture of our proposed model. In this work, a ResNet-50 is used as our backbone network. Layers after conv4_1 in Resnet-50 are duplicated to split our network into 2 independent branches. GMP refers to Global Max Pooling. Conv refers to 1*1 convolutional layer, which aims to unify dimensions of global and local feature vectors. FC refers to fully connected layer. BN refers to Batch Normalization layer. In the test phase, all the feature vectors (Dim=256) after Batch Normalization layer are concatenated together as an appearance signature (Dim=256*18).*

one video-based evaluation protocol MARS validate the performance of our model, which outperforms previous state-of-the-art in both single and cross domain Re-ID tasks. The general architecture of our proposed method is represented in the Figure 9.

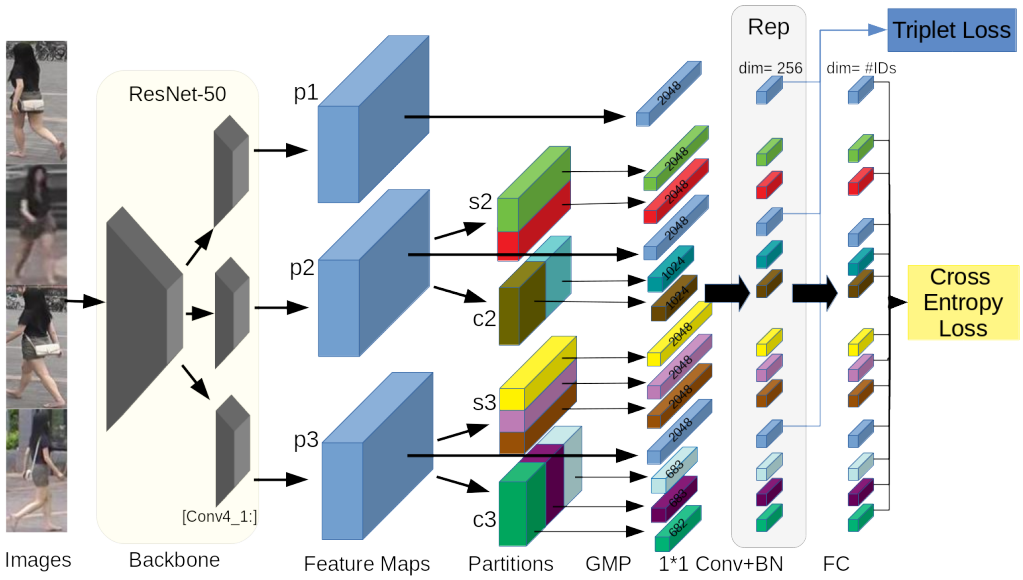

*Figure 9. Spatial and Channel Partition Representation network. For the backbone network, we duplicate layers after conv4_1 into 3 identical but independent branches that generate 3 feature maps "p1", "p2" and "p3". Then, multiple spatial-channel partitions are conducted on the feature maps. "s2" and "c2" refer to 2 spatial parts and 2 channel groups. "s3" and "c3" refer to 3 spatial parts and 3 channel groups. After global max pooling (GMP), dimensions of global (dim = 2048) and local (dim = 2048, 1024*2 and 683*2+682) features are unified by 1*1 convolution (1*1 Conv) and batch normalization (BN) to 256. Then, fully connected layers (FC) give identity predictions of input images. All the dimension unified feature vectors (dim = 256) are aggregated together as appearance representation (Rep) for testing.*

6.5. Improving Face Sketch Recognition via Adversarial Sketch-Photo Transformation

Participants: Antitza Dantcheva, Shikang Yu [Chinese Academy of Sciences], Hu Han [Chinese Academy of Sciences], Shiguang Shan [Chinese Academy of Sciences], Xilin Chen [Chinese Academy of Sciences].

participants

Face sketch-photo transformation has broad applications in forensics, law enforcement, and digital entertainment, particular for face recognition systems that are designed for photo-to-photo matching. While there are a number of methods for face photo-to-sketch transformation, studies on sketch-to-photo transformation remain limited. In this work, we proposed a novel conditional CycleGAN for face sketch-to-photo transformation. Specifically, we leveraged the advantages of CycleGAN and conditional GANs and designed a feature-level loss to assure the high quality of the generated face photos from sketches. The generated face photos were used, as a replacement of face sketches, and particularly for face identification against a gallery set of mugshot photos. Experimental results on the public-domain database CUFSF showed that the proposed approach was able to generate realistic photos from sketches, and the generated photos were instrumental in improving the

sketch identification accuracy against a large gallery set. This work has been presented at the IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019) [30].

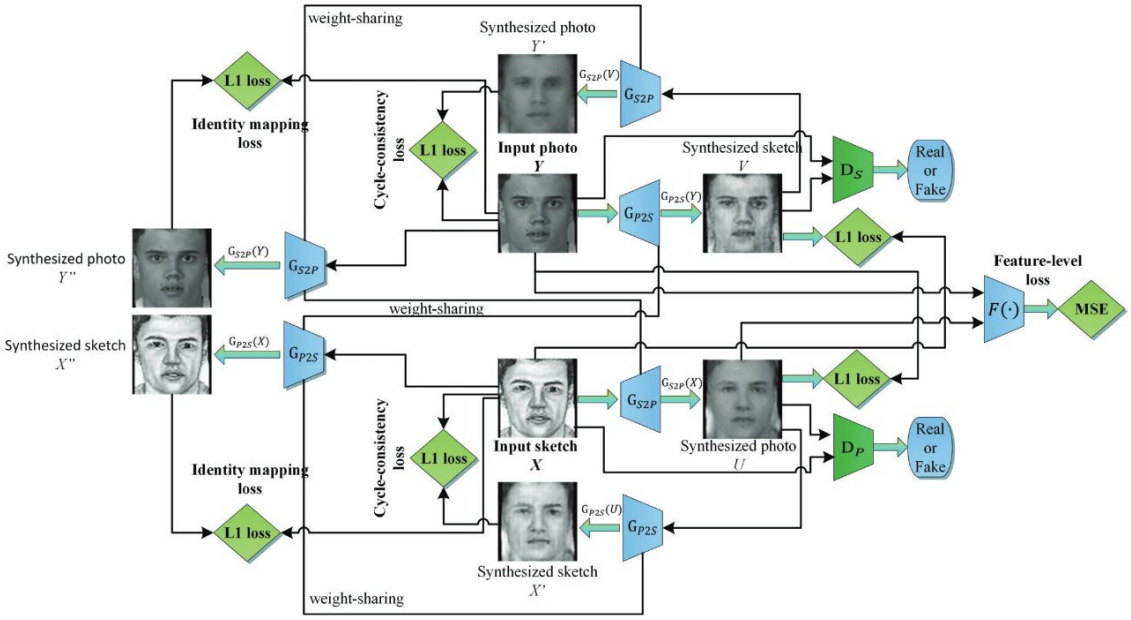

Figure 10. Overview of the proposed GAN for sketch-to-photo transformation using feature-level loss.

6.6. Impact and Detection of Facial Beautification in Face Recognition: An Overview

Participants: Antitza Dantcheva, Christian Rathgeb [Hochschule Darmstadt], Christoph Busch [Hochschule Darmstadt].

Facial beautification induced by plastic surgery, cosmetics or retouching has the ability to substantially alter the appearance of face images. Such types of beautification can negatively affect the accuracy of face recognition systems. In this work, a conceptual categorisation of beautification was presented, relevant scenarios with respect to face recognition were discussed, and related publications were revisited. Additionally, technical considerations and trade-offs of the surveyed methods were summarized along with open issues and challenges in the field. This survey is targeted to provide a comprehensive point of reference for biometric researchers and practitioners working in the field of face recognition, who aim at tackling challenges caused by facial beautification. This work was published in IEEE Access [18].

6.7. Computer Vision and Deep Learning applied to Facial analysis in the invisible spectra

Participants: David Anghelone, Antitza Dantcheva.

The goal of our work is to analyze faces, as well as recognize events in the invisible spectra. In the last few years, face analysis has been a highly active area and has attracted a lot of interest from the scientific community. Limitations encountered in the visible spectrum such as illumination-restriction have the ability to be overcome in the infrared spectrum. We explored the state-of-the-Art of facial analysis in the invisible spectrum including low energy infrared waves, as well as ultraviolet waves. In this context we have captured images in each spectra and intend to process the data. We aim at designing a model, which extracts biometric features. The key challenges are the processing of contours, shape, etc. This subject is within the framework of the national project *SafeCity*: Security of Smart Cities.

6.8. ImaGINator: Conditional Spatio-Temporal GAN for Video Generation

Participants: Yaohui Wang, Antitza Dantcheva, Piotr Bilinski [University of Warsaw], François Brémond.

keywords: GANs, Video Generation

Generating human videos based on single images entails the challenging simultaneous generation of realistic and visual appealing appearance and motion. In this context, we propose a novel conditional GAN architecture, namely ImaGINator [35] (see Figure 11), which given a single image, a condition (label of a facial expression or action) and noise, decomposes appearance and motion in both latent and high level feature spaces, generating realistic videos. This is achieved by (i) a novel spatio-temporal fusion scheme, which generates dynamic motion, while retaining appearance throughout the full video sequence by transmitting appearance (originating from the single image) through all layers of the network. In addition, we propose (ii) a novel transposed (1+2)D convolution, factorizing the transposed 3D convolutional filters into separate transposed temporal and spatial components, which yields significantly gains in video quality and speed. We extensively evaluate our approach on the facial expression datasets MUG and UvA-NEMO, as well as on the action datasets NATOPS and Weizmann. We show that our approach achieves significantly better quantitative and qualitative results than the state-of-the-art (see Table 1).

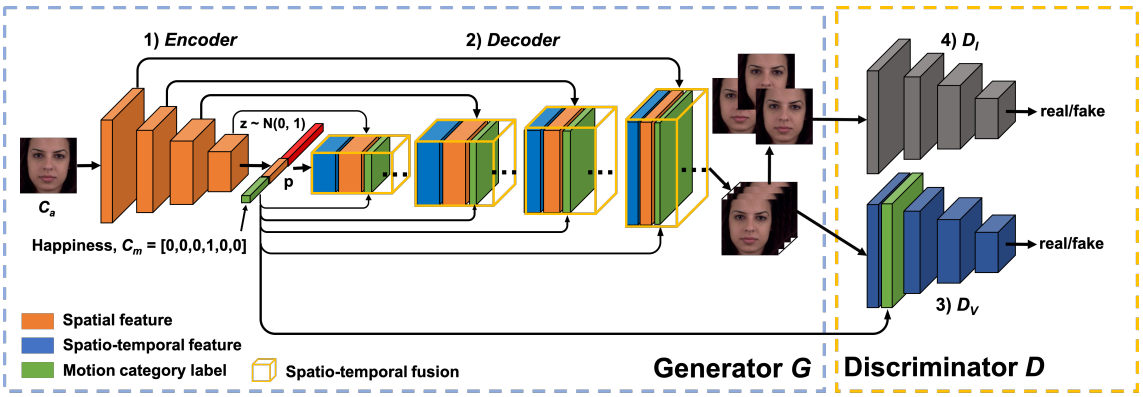

Figure 11. ***Overview of the proposed ImaGINator****. In the GeneratorG, the Encoder firstly encodes an input image c_a into a single vector p. Then, the Decoder produces a video based on a motion c_m and a random vector z. By using spatio-temporal fusion, low level spatial feature maps from the Encoder are directly concatenated into the Decoder. While D_I discriminates whether the generated images contain an authentic appearance, D_V additionally determines whether the generated videos contain an authentic motion.*

Table 1. Evaluation of VGAN, MoCoGAN and proposed ImaGINator with respect to image quality (SSIM/PSNR) and video quality (FID).

	MUG		**NATOPS**	
	SSIM/PSNR	FID	SSIM/PSNR	FID
VGAN	0.28/14.54	74.72	0.72/20.09	167.71
MoCoGAN	0.58/18.16	45.46	0.74/21.82	49.46
ImaGINator	**0.75/22.63**	**29.02**	**0.88/27.39**	**26.86**
	Weizmann		**UvA-NEMO**	
	SSIM/PSNR	FID	SSIM/PSNR	FID
VGAN	0.29/15.78	127.31	0.21/13.43	30.01
MoCoGAN	0.42/17.58	116.08	0.45/16.58	29.81
ImaGINator	**0.73/19.67**	**99.80**	**0.66/20.04**	**16.16**

6.9. Characterizing the State of Apathy with Facial Expression and Motion Analysis

Participants: S L Happy, Antitza Dantcheva, Abhijit Das, François Brémond, Radia Zeghari [Cobtek], Philippe Robert [Cobtek].

Reduced emotional response, lack of motivation, and limited social interaction comprise the major symptoms of apathy. Current methods for apathy diagnosis require the patient's presence in a clinic, and time consuming clinical interviews and questionnaires involving medical personnel, which are costly and logistically inconvenient for patients and clinical staff, hindering among other large scale diagnostics. In this work we introduced a novel machine learning framework to classify apathetic and non-apathetic patients based on analysis of facial dynamics, entailing both emotion and facial movement. Our approach catered to the challenging setting of current apathy assessment interviews, which include short video clips with wide face pose variations, very low-intensity expressions, and insignificant inter-class variations. We tested our algorithm on a dataset consisting of 90 video sequences acquired from 45 subjects and obtained an accuracy of 84% in apathy classification. Based on extensive experiments, we showed that the fusion of emotion and facial local motion produced the best feature set for apathy classification. In addition, we trained regression models to predict the clinical scores related to the mental state examination (MMSE) and the neuropsychiatric apathy inventory (NPI) using the motion and emotion features. Our results suggested that the performance can be further improved by appending the predicted clinical scores to the video-based feature representation. This work has been presented at the IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019) [25].

6.10. Dual-threshold Based Local Patch Construction Method for Manifold Approximation And Its Application to Facial Expression Analysis

Participants: S L Happy, Antitza Dantcheva, Aurobinda Routray [IIT Kharagpur].

In this paper, we propose a manifold based facial expression recognition framework which utilizes the intrinsic structure of the data distribution to accurately classify the expression categories. Specifically, we model the expressive faces as the points on linear subspaces embedded in a Grassmannian manifold, also called as expression manifold. We propose the dual-threshold based local patch (DTLP) extraction method for constructing the local subspaces, which in turn approximates the expression manifold. Further, we use the affinity of the face points from the subspaces for classifying them into different expression classes. Our method is evaluated on four publicly available databases with two well known feature extraction techniques. It is evident from the results that the proposed method efficiently models the expression manifold and improves the recognition accuracy in spite of the simplicity of the facial representatives. This work has been presented at the European Signal Processing Conference (EUSIPCO'19) [26].

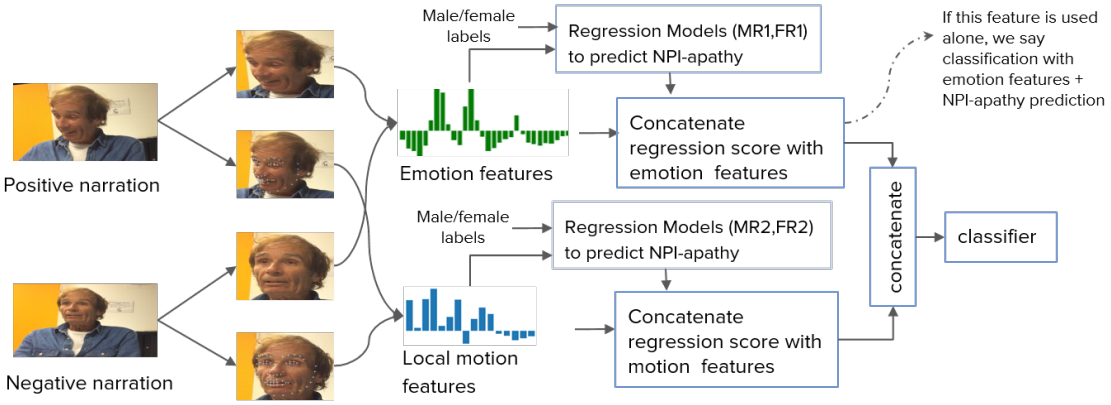

Figure 12. Overall framework for apathy detection from facial videos.

6.11. A Weakly Supervised Learning Technique for Classifying Facial Expressions

Participants: S L Happy, Antitza Dantcheva, François Brémond.

The universal hypothesis suggests that the six basic emotions: anger, disgust, fear, happiness, sadness, and surprise, are being expressed by similar facial expressions by all humans. While existing datasets support the universal hypothesis and comprise of images and videos with discrete disjoint labels of profound emotions, real-life data contains jointly occurring emotions and expressions of different intensities. Models, which are trained using categorical one-hot vectors often over-fit and fail to recognize low or moderate expression intensities. Motivated by the above, as well as by the lack of sufficient annotated data, we propose a weakly supervised learning technique for expression classification, which leveraged the information of unannotated data. Crucial in our approach was that we first trained a convolutional neural network (CNN) with label smoothing in a supervised manner and proceeded to tune the CNN-weights with both labelled and unlabelled data simultaneously. Experiments on four datasets demonstrated large performance gains in cross-database performance, as well as showed that the proposed method achieved to learn different expression intensities, even when trained with categorical samples. This work was published in Pattern Recognition Letters [15].

6.12. Robust Remote Heart Rate Estimation from Face Utilizing Spatial-temporal Attention

Participants: Antitza Dantcheva, Abhijit Das, Xuesong Niu [Chinese Academy of Sciences], Xingyuan Zhao [Chinese Academy of Sciences], Hu Han [Chinese Academy of Sciences], Shiguang Shan [Chinese Academy of Sciences], Xilin Chen [Chinese Academy of Sciences].

We proposed an end-to-end approach for robust remote heart rate (HR) measurement gleaned from facial videos. Specifically the approach was based on remote photoplethysmography (rPPG), which constitutes a pulse triggered perceivable chromatic variation, sensed in RGB-face videos. Incidentally rPPGs can be affected in less-constrained settings. To unpin the shortcoming, the proposed algorithm utilized a spatio-temporal attention mechanism, which placed emphasis on the salient features included in rPPG-signals. In addition, we proposed an effective rPPG augmentation approach, generating multiple rPPG signals with varying HRs from a single face video. Experimental results on the public datasets VIPL-HR and MMSE-HR showed that the proposed method outperformed state-of-the-art algorithms in remote HR estimation. This work has been presented at the IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019) [28].

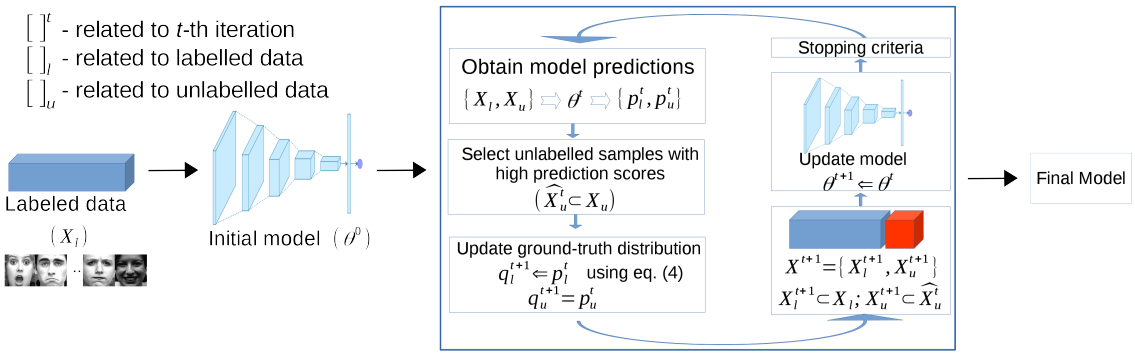

Figure 13. Workflow of the proposed method for weakly supervised learning of facial expressions.

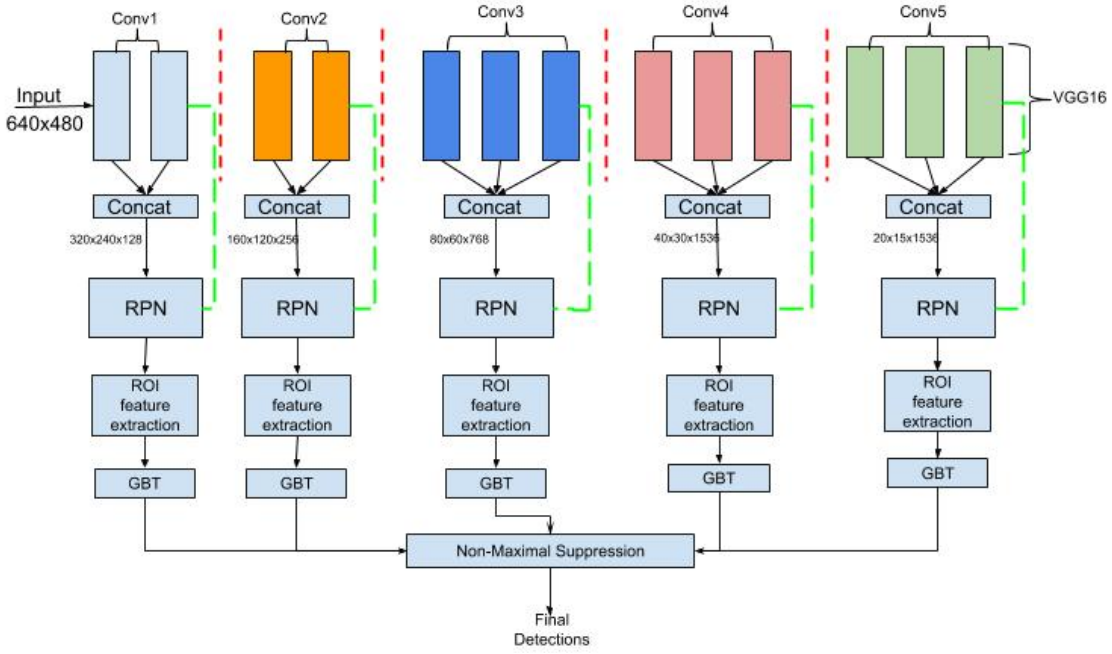

Figure 14. Overview of the proposed end-to-end trainable approach for rPPG based remote HR measurement via representation learning with spatial-temporal attention.

6.13. Quantified Analysis for Epileptic Seizure Videos

Participants: Jen-Cheng Hou, Monique Thonnat.

Epilepsy is a type of neurological disorder, affecting around 50 million people worldwide. Epilepsy's main symptoms are seizures, which are caused by abnormal neuronal activities in the brain. To determine appropriate treatments, neurologists assess manifestation of patients' behavior when seizures occur. Nevertheless, there are few objective criteria regarding the procedure, and diagnosis could be biased due to subjective evaluation. Hence it is important to quantify patients' ictal behaviors for better assessment of the disorder. In collaboration with Dr. Fabrice Bartolomei and Dr. Aileen McGonigal from Timone Hospital, Marseille, we have access to video recordings from epilepsy monitoring unit for analysis, with consent from ethics committee (IRB) and the patients involved.

6.13.1. Seizure Video Classification and Background Video Collection

In an epilepsy monitoring unit, EEG and video recording are usually collected. For patients who need brain surgery to remove lobes that produce seizures, stereo-EEG (SEEG) recordings are particularly measured. SEEG is an intrusive measurement and provides information of the seizure type. We have 86 seizure videos from 20 patients along with the corresponding SEEG conclusion (i.e. pre-frontal epilepsy, occipital epilepsy, etc.). In this study, the goal is to classify seizure videos to their seizure types. Classification was conducted by fine-tuning a pre-trained video classification model, I3D, with 10-fold cross-validation. Due to the relatively small volume of data we have and the challenging nature of our videos, the performance was not satisfactory enough. Inspired by recent semi-supervised works in leveraging large unlabeled dataset for better adaptation to certain tasks, we are collecting large volume of background videos in the epilepsy monitoring unit, in which patients' behavior are normal, such as eating, sleeping, and talking. The volume of the background video can be up to 1000 hours, which could be taken as unlabeled dataset for semi-supervised learning in our case.

6.13.2. Quantifying Rhythmic Rocking Movement with Head Tracking

In this study, six seizures from three patients with pre-frontal epilepsy were analyzed. The duration of rocking was 15-40 seconds, with marked regularity throughout each seizure. Our objective is to document time-evolving frequencies of antero-posterior rocking body movements occurring during seizures. We adopted MobileNet [39] as our backbone model for detecting head of the patient, and hence obtain the trajectories of head movement (see Figure 15). After smoothing the trajectories and find the valid peaks corresponding to the antero-posterior movement, we compute the time-evolving movement frequency for each seizure video. Whereas the rocking frequency varied substantially between patients and seizures (0.3-1Hz), coefficient of variation of frequency was low ($\leq 12\%$). The study report is under review for a medical journal.

6.14. Skeleton Image Representation for 3D Action Recognition

Participants: Carlos Caetano, François Brémond.

Due to the availability of large-scale skeleton datasets, 3D human action recognition has recently called the attention of computer vision community. Many works have focused on encoding skeleton data as skeleton image representations based on spatial structure of the skeleton joints, in which the temporal dynamics of the sequence is encoded as variations in columns and the spatial structure of each frame is represented as rows of a matrix. To further improve such representations, we introduce a novel skeleton image representation to be used as input of Convolutional Neural Networks (CNNs), named SkeleMotion. The proposed approach encodes the temporal dynamics by explicitly computing the magnitude and orientation values of the skeleton joints. Different temporal scales are employed to compute motion values to aggregate more temporal dynamics to the representation making it able to capture long-range joint interactions involved in actions as well as filtering noisy motion values. Experimental results demonstrate the effectiveness of the proposed representation on 3D action recognition outperforming the state-of-the-art on NTU RGB+D 120 dataset. This work has been published in AVSS 2019 [31].

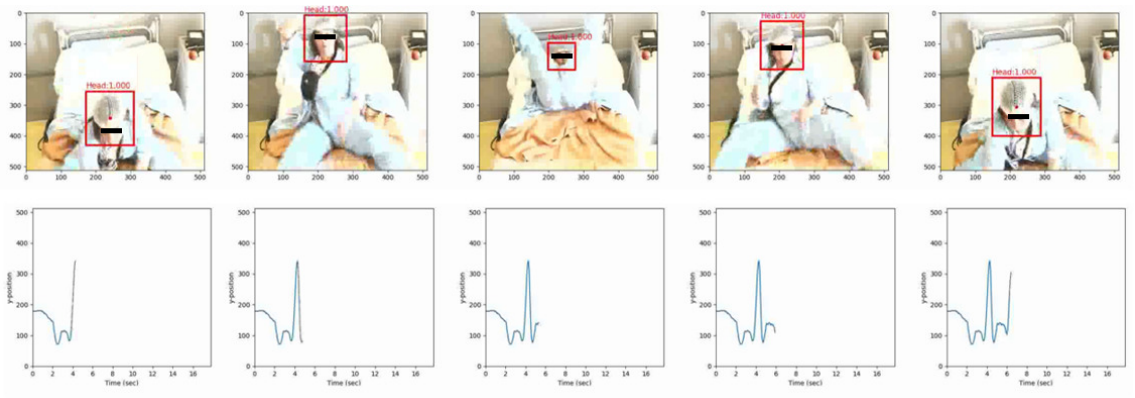

Figure 15. The first row demonstrates the image samples of the antero-posterior movement. The second row shows the position of the head through time in the vertical direction.

In another work, we have explore how to better represent motion information in a video. The temporal component of videos provides an important clue for activity recognition, as a number of activities can be reliably recognized based on the motion information. In view of that, this work proposes a novel temporal stream for two-stream convolutional networks based on images computed from the optical flow magnitude and orientation, named Magnitude-Orientation Stream (MOS), to learn the motion in a better and richer manner. Our method applies simple non-linear transformations on the vertical and horizontal components of the optical flow to generate input images for the temporal stream. Moreover, we also employ depth information to use as a weighting scheme on the magnitude information to compensate the distance of the subjects performing the activity to the camera. Experimental results, carried on two well-known datasets (UCF101 and NTU), demonstrate that using our proposed temporal stream as input to existing neural network architectures can improve their performance for activity recognition. Results demonstrate that our temporal stream provides complementary information able to improve the classical two-stream methods, indicating the suitability of our approach to be used as a temporal video representation. two-stream convolutional networks, spatiotemporal information, optical flow, depth information. This work has been published in the Journal of Visual Communication and Image Representation [14].

6.15. Toyota Smarthome: Real-World Activities of Daily Living

Participants: Srijan Das, Rui Dai, François Brémond.

The performance of deep neural networks is strongly influenced by the quantity and quality of annotated data. Most of the large activity recognition datasets consist of data sourced from the Web, which does not reflect challenges that exist in activities of daily living. In this work, we introduce a large real-world video dataset for activities of daily living: Toyota Smarthome. The dataset consists of 16K RGB+D clips of 31 activity classes, performed by seniors in a smarthome. Unlike previous datasets, videos were fully unscripted. As a result, the dataset poses several challenges: high intra-class variation, high class imbalance, simple and composite activities, and activities with similar motion and variable duration. Activities were annotated with both coarse and fine-grained labels. These characteristics differentiate Toyota Smarthome from other datasets for activity recognition as illustrated in 16.

As recent activity recognition approaches fail to address the challenges posed by Toyota Smarthome, we present a novel activity recognition method with attention mechanism. We propose a pose driven spatio-temporal attention mechanism through 3D ConvNets. We show that our novel method outperforms state-of-the-art methods on benchmark datasets, as well as on the Toyota Smarthome dataset. We release the dataset

for research use at https://project.inria.fr/toyotasmarthome. This work is done in collaboration with Toyota Motors Europe and is published in ICCV 2019 [21].

Figure 16. Sample frames from Toyota Smarthome dataset: 1-7 label at the right top corner respectively correspond to camera view 1, 2, 3, 4, 5, 6 and 7 as marked in the plan of the apartment on the right. Image from camera view (1) Drink from can, (2) Drink from bottle, (3) Drink form glass and (4) Drink from cup are all fine grained activities with a coarse label drink. Image from camera view (5) Watch TV and (6) Insert tea bag show activities with large source-to-camera distance and occlusion. Images with camera view (7) Enter illustrate the RGB image and the provided 3D skeleton.

6.15.1. Looking deeper into Time for Activities of Daily Living Recognition

Participants: Srijan Das, Monique Thonnat, François Brémond.

In this work, we introduce a new approach for Activities of Daily Living (ADL) recognition. In order to discriminate between activities with similar appearance and motion, we focus on their temporal structure. Actions with subtle and similar motion are hard to disambiguate since long-range temporal information is hard to encode. So, we propose an end-to-end Temporal Model to incorporate long-range temporal information without losing subtle details. The temporal structure is represented globally by different temporal granularities and locally by temporal segments as illustrated in fig. 17. We also propose a two-level pose driven attention mechanism to take into account the relative importance of the segments and granularities. We validate our approach on 2 public datasets: a 3D human activity dataset (NTU-RGB+D) and a human action recognition dataset with object interaction dataset (Northwestern-UCLA Multiview Action 3D). Our Temporal Model can also be incorporated with any existing 3D CNN (including attention based) as a backbone which reveals its robustness. This work has been accepted in WACV 2020 [20].

6.16. Self-Attention Temporal Convolutional Network for Long-Term Daily Living Activity Detection

Participants: Rui Dai, François Brémond.

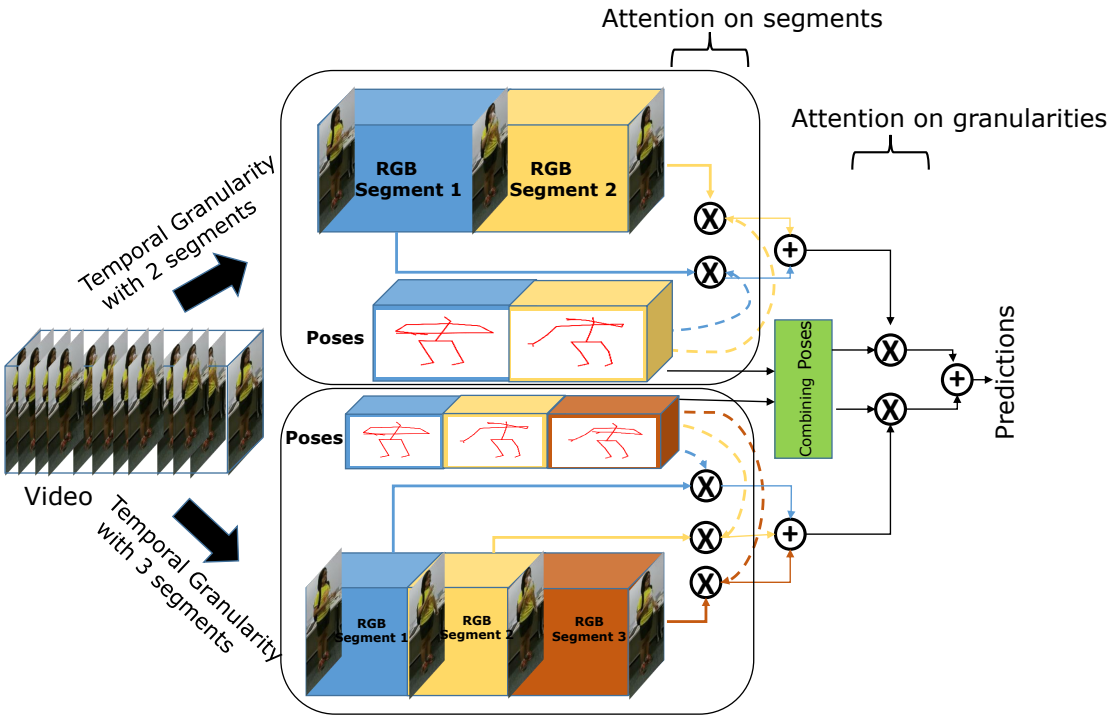

Figure 17. Framework of the proposed approach in a nutshell for two temporal granularities. The articulated poses soft-weight the temporal segments and the temporal granularities using a two-level attention mechanism.

This year, we proposed a Self-Attention - Temporal Convolutional Network (SA-TCN), which is able to capture both complex activity patterns and their dependencies within long-term untrimmed videos [34]. This attention block can also embed with other TCN-nased models. We evaluate our proposed model on DAily Home LIfe Activity Dataset (DAHLIA) and Breakfast datasets. Our proposed method achieves state-of-the-art performance on both datasets.

6.16.1. Work Flow

Given an untrimmed video, we represent each non-overlapping snippet by a visual encoding over 64 frames. This visual encoding is the input to the encoder-TCN, which is the combination of the following operations: 1D temporal convolution, batch normalization, ReLu, and max pooling. Next, we send the output of the encoder-TCN into the self-attention block to capture long-range dependencies. After that, the decoder-TCN applies the 1D convolution and up sampling to recover a feature map of the same dimension as visual encoding. Finally, the output will be sent to a fully connected layer with softmax activation to get the prediction. Fig 18 and 19 provide the structure of our model.

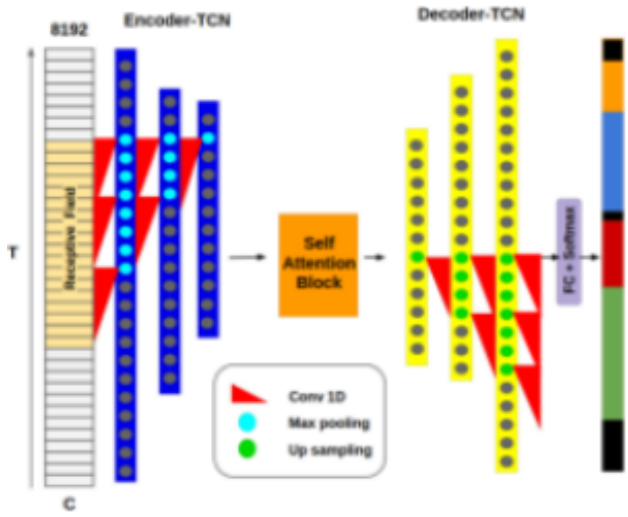

Figure 18. ***Overview****. The model contains mainly three parts: (1) visual encoding, (2) encoder-decoder structure, (3) attention block*

6.16.2. Result

We evaluated the proposed method on two daily-living activity datasets (DAHLIA, Breakfast) and achieved state-of-the-art performances. We compared with these following State-of the arts: DOHT, Negin *et al.*, GRU , ED-TCN, TCFPN.

6.17. DeepSpa Project

Participants: Alexandra König, Rachid Guerchouche, Minh Tran-Duc, Antitza Dantcheva, S L Happy, Abhijit Das.

The DeepSpa (Deep Speech Analysis, January 2019 - June 2020) project aims to deliver telecommunication-based neurocognitive assessment tools for early screening, early diagnostic and follow-up of cognitive disorders, mainly in elderly. The target is also clinical trials addressing Alzheimer's and other neurodegenerative diseases. By combining AI in speech recognition and video analysis for facial expression recognition, the proposed tools allow remote cognitive and psychological testing, thereby saving time and money.

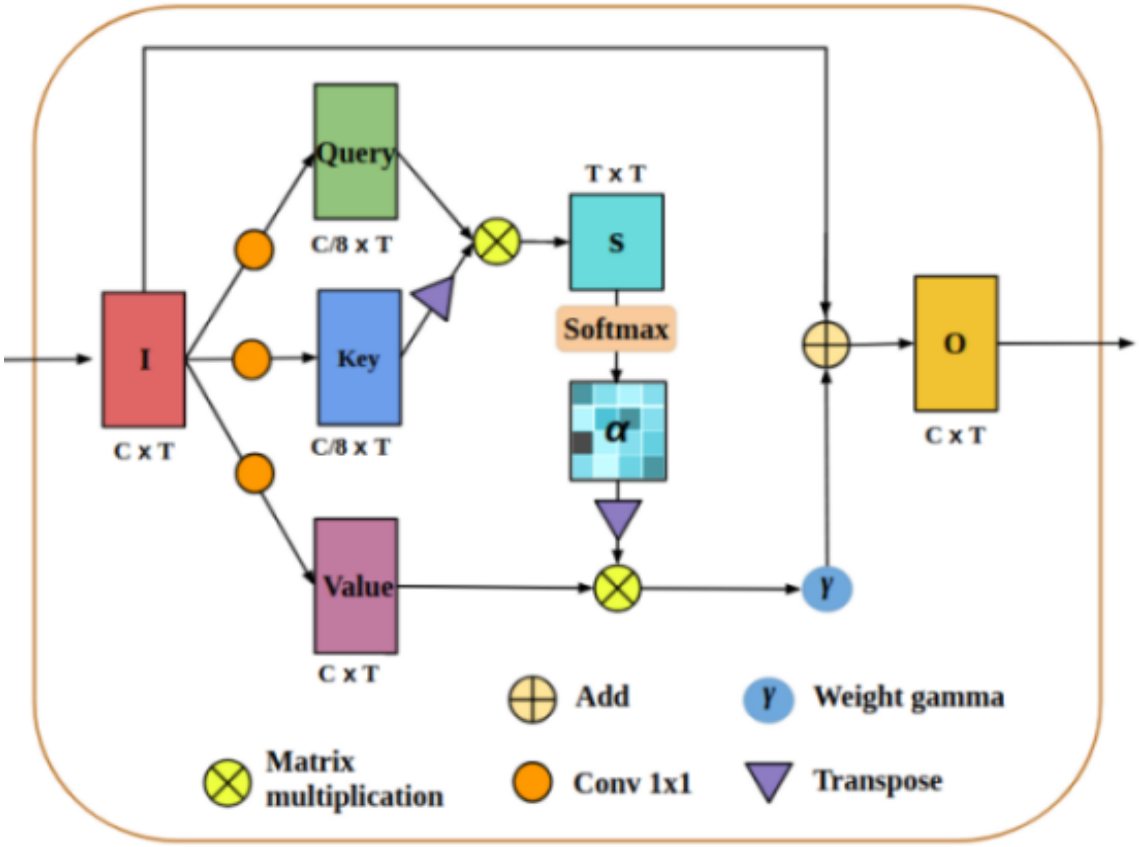

*Figure 19. **Attention block**. This figure presents the structure of attention block*

Table 2. Activity detection results on DAHLIA dataset with the average of view 1, 2 and 3. * marked methods have not been tested on DAHLIA in their original paper.

Model	FA1	F-score	IoU	mAP
DOHT	0.803	0.777	0.650	-
GRU*	0.759	0.484	0.428	0.654
ED-TCN*	0.851	0.695	0.625	0.826
Negin *et al.*	0.847	0.797	0.723	-
TCFPN*	0.910	**0.799**	0.738	**0.879**
SA-TCN	**0.921**	0.788	**0.740**	0.862

Table 3. Activity detection results on Breakfast dataset.

Model	FA1	F-Score	IoU	mAP
GRU	0.368	0.295	0.198	0.380
ED-TCN	0.461	0.462	0.348	0.478
TCFPN	**0.519**	0.453	0.362	0.466
SA-TCN	0.497	**0.494**	**0.385**	**0.480**

Table 4. Average precision of ED-TCN on DAHLIA.

Activities	**Background**	**House work**	**Working**	**Cooking**
AP	0.36	0.65	0.95	0.96
Activities	**Laying table**	**Eating**	**Clearing table**	**Wash dishes**
AP	0.90	0.97	0.80	0.97

Table 5. Combination of attention block with other TCN-based model: TCFPN. (Evaluated on DAHLIA dataset)

Model	**FA1**	**F-score**	**IoU**	**mAP**
TCFPN	0.910	**0.799**	0.738	0.879
SA-TCFPN	**0.917**	**0.799**	**0.748**	**0.894**

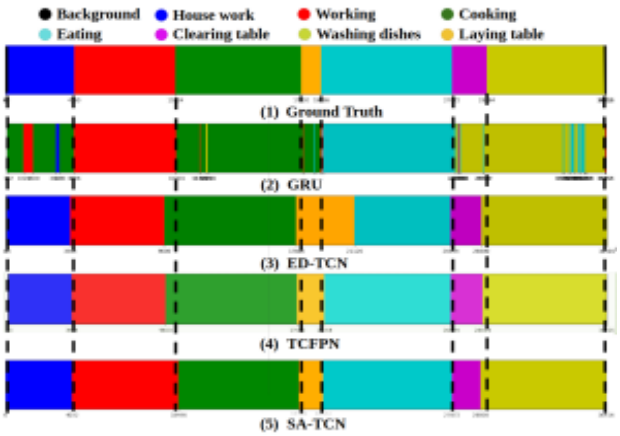

Figure 20. ***Detection visualization****. The detection visualization of video 'S01A2K1' in DAHLIA: (1) ground truth, (2) GRU, (3) ED-TCN, (4) TCFPN and (5) SA-TCN.*

The partners of the project are:

- Inria: technical partner and project coordinator
- University of Maastricht: clinical partner
- Jansen & Jansen: pharma partner and business champion
- Association Innovation Alzheimer: subgranted clinical partner
- Ki-element: subgranted technical partner.

6.17.1. Project structure

The DeepSpA project is structured in two use-cases:

- Use-case 1: remote assessment through phone for early screening of cognitive disorders (University of Maastricht, Jansen & Jansen and Ki-element): using AI based speech recognition; assessments through phone are made possible. A clinical trial is currently running in Maastricht (by end 2019, 70 subjects will be included, and 50 others will be included in 2020), the goal is to study the feasibility of such phone assessment in comparison to face-to-face assessment.
- Use-case 2: remote assessment through video-conference system (telemedicine tool) (Inria, Jansen & Jansen and Association Innovation Alzheimer): Inria developed a telemedicine tool which allows complete remote assessment. AI based speech and facial expression recognition empower the cognitive assessment by providing extra features useful for clinicians.

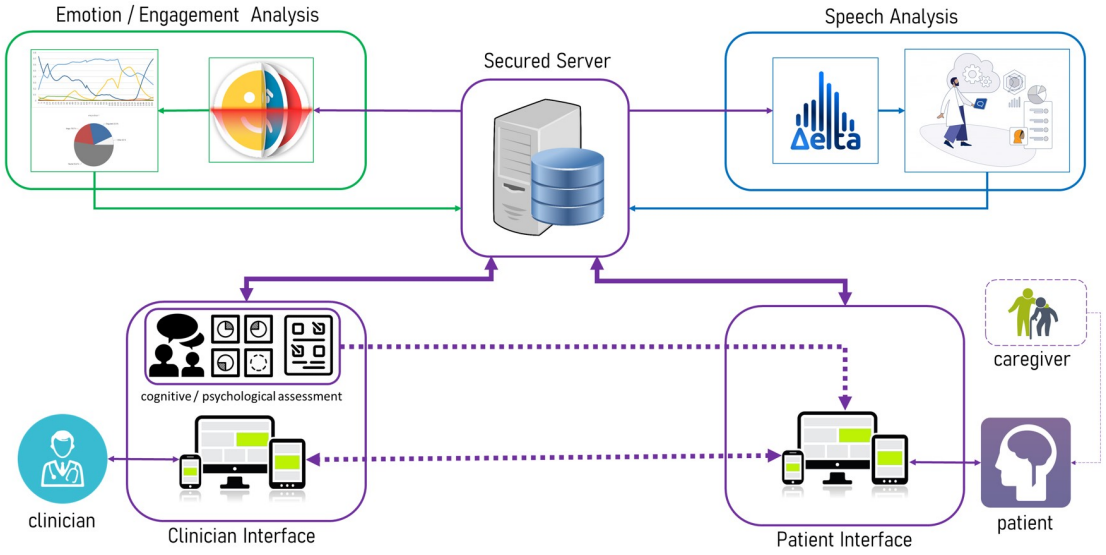

Figure 21. Global view of the telemedicine tool developed by Inria (STARS).

6.17.2. Telemedicine / Clinical Study with Digne-les-Bains

In order to evaluate the feasibility of remote assessment through the telemedicine tool, a collaboration with the city of Digne-les-Bains started in March 2019. The Hospital of Digne-les-Bains, la Maison de la Santé and the ADMR (association dealing with isolated people) are involved in a running clinical study, which aims at evaluating the feasibility of the remote assessment in two different setups:

- Clinical setup: a fixed place where the participant will undergo the telemedicine session: clinic, hospital, pharmacy, health centres ....
- Mobile Units: a mobile unit goes to the subjects home, the telemedicine session is done inside the mobile unit (e.g., van).

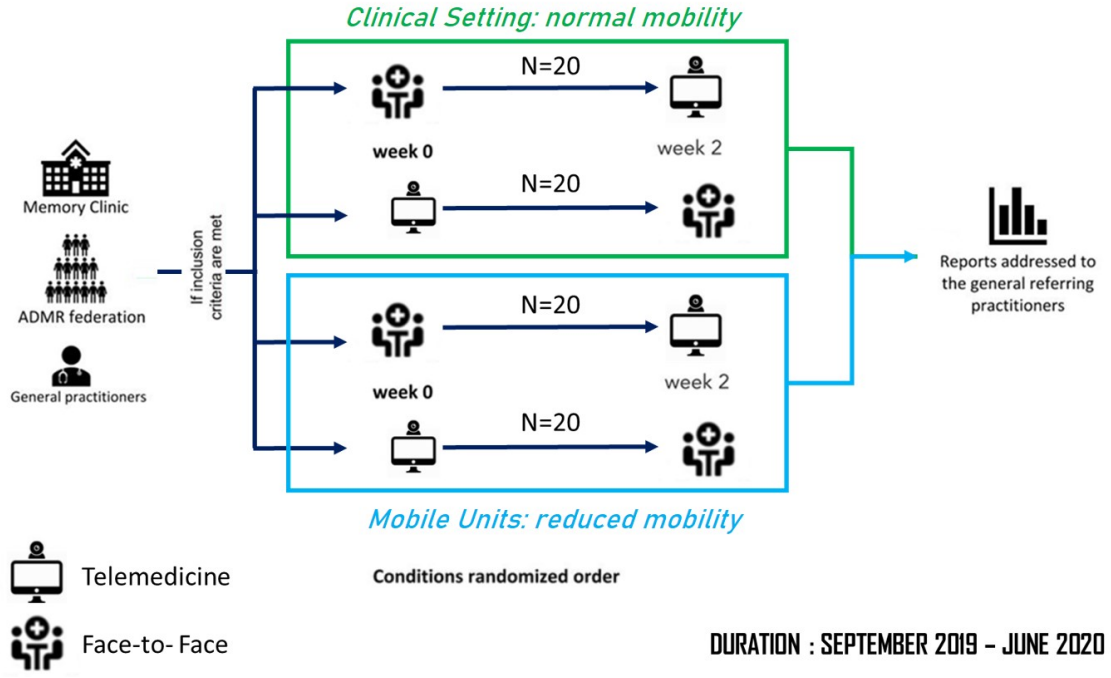

Figure 22. Global view of the clinical study with Digne-les-Bains.

We already started including subjects in the clinical setup case. By end 2019, we expect to include about 15 subjects, 25 extra subjects will be included during 2020. Mobile units setup will be tested during 2020.

First results and observations already showed that the telemedicine tool allows full assessments. Clinicians and patients showed strong interest and appreciation of such tool.

6.17.3. Facial expressions recognition and engagement evaluation in the telemedicine tool

The STARS team is doing research on facial expressions analysis, which could be integrated as part of the vision module of the telemedicine tool [25].

Notable software related to this research is the provided API on the cloud, which allows sending video files and retrieving emotions, gaze direction, facial movements and head direction (implemented by S L Happy).

6.18. Store Connect and Solitaria

Participants: Sébastien Gilabert, Minh Khue Phan Tran, François Brémond.

Store-Connect was a consortium aiming at detecting and positioning people in a supermarket. Several technologies were explore such as computing and merging trajectories obtained from the mobile phone of customers and from video cameras. In a second step, the goal was to detect all the 'stop' events of the customers while shopping in the store.

6.18.1. SupICP

We have developed with the SED team, SupICP, a platform for integrating all plugins developed by STARS team. Our main contribution is the Ontology Language Plugin. With this plugin, we can use contextual and knowledge information inside scenarios designed for video event recognition. Currently, we are improving this plugin for combining the Ontology Language with Deep Learning technology towards "Action recognition based on Deep Learning and Ontology Language".

We have also installed this software at the Institute Claude Pompidou, in order to conduct clinical trials, and to work with medical scientists.

6.18.2. Solitaria

The aim of this project is to combine data extracted from domestic sensors and from video cameras, and to implement this plugin into SupICP to monitor older people at home.

6.19. Synchronous Approach to Activity Recognition

Participants: Daniel Gaffé, Sabine Moisan, Annie Ressouche, Jean-Paul Rigault, Ines Sarray.

Activity Recognition aims at recognizing and understanding sequences of actions and movements of mobile objects (human beings, animals or artifacts), that follow the predefined model of an activity. We propose to describe activities as a series of actions, triggered and driven by environmental events.

This year we mainly refined the ADeL description language, the semantics of some of its instructions and their compilation into equation systems. We also improved the recognition engine and the synchronizer to better handle the synchronous/asynchronous transformation.

Work remains to be done to complete a full framework to generate generic recognition systems and automatic tools to interface with static and dynamic analysis tools, such as model checkers or performance monitors.

6.19.1. Activity Description Language

The ADeL language was designed to describe various activities, it provides two different (and equivalent) formats: graphical and textual. This year we started to describe use case examples in the medical domain: serious games and exercises for patients having cognitive problems, such as Alzheimer or autistic persons. This kind of games are used to test patients and to evaluate their behavior and interactions. These use cases lead us to improve the language and part of its semantics. An example of the graphical format describing a simple exercise activity is given in figure 23.

Work remains to be done to improve the usability of the language by our end-users.

6.19.2. Synchronizer

Using the synchronous paradigm makes time manipulation easy thanks to determinism and synchronous parallellism; moreover, tools exist to support formal verification. However, the sensor environment is asynchronous and it is thus necessary to transform asynchronous events given by sensors into synchronous logical instants. It is a difficult problem that does not have an exact and complete solution. We introduced a component called "synchronizer" between the environment sensors and the recognition engine. The synchronizer is responsible for filtering the sensor data, grouping them into logical instants, and sending these instants to the recognition engine.

We specified a generic algorithm, based on *awaited* events, i.e. the events which may trigger transitions to a next state. These events are provided by each automaton in each state. This algorithm is parametrized by heuristics to adapt to different situations. There are two main points of variation in the synchronizer where heuristics can be applied: when processing data coming from the sensors (to collect and combine raw data) and when building logical instants (to decide on the end of instants and to manage preemptions).

This year we finished to implement a first version of the synchronizer (for one single activity to recognize), we defined different heuristics, and we tested the synchronizer algorithm on some uses cases with these heuristics [11].

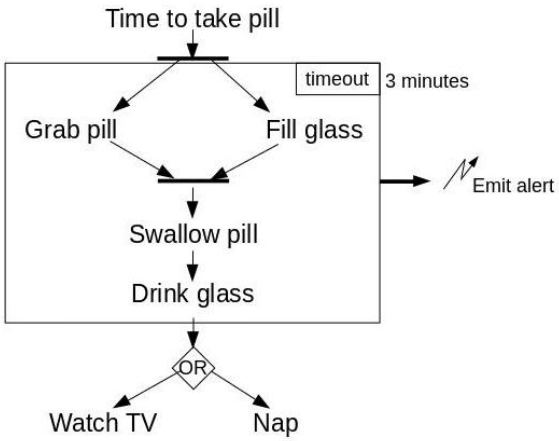

Figure 23. Example of a simple activity (patient should take a pill at a given time) including a parallel and a timeout instructions.

6.20. Probabilistic Activity Modeling

Participants: Elisabetta de Maria, Sabine Moisan, Jean-Paul Rigault, Thibaud L'Yvonnet.

Serious games constitute a domain in which real-time activity recognition is particularly relevant: the expected behavior is well identified and it is possible to rely on different sensors (biometric and external) while playing the game. We focus on games to help in diagnosis and treatment of patients.

We developed a formal approach to model such activities, taking into account possible variations in human behavior. All the scenarios of an activity are not equivalent: some are typical (thus frequent) while others seldom happen. We propose to quantify the likelihood of these variations by associating probabilities with the key actions of the activity description. We rely on a formal model based on probabilistic discrete-time Markov chains (DTMCs). We used the PRISM framework and its model checking facilities to express and check interesting temporal logic properties (PCTL).

As a use case, we considered a serious game to analyze the behavior of Alzheimer patients. We encoded this game as a DTMC in PRISM and we defined several meaningful PCTL properties that are then automatically tested thanks to the PRISM model checker. Two kinds of properties may be defined: those to verify the model and those oriented toward the medical domain. The latter may give indications to a practitioner regarding a patient's behavior. These properties include the use of PRISM "rewards" to quantify the performance of patients.

We expect that such a modeling approach could provide doctors with new indications for interpreting patients' performance and we identified three medically interesting outcomes for this approach. First, to evaluate a new patient before the first diagnosis of doctors, we can compare her game performance to a reference model representing a "healthy" behavior. Second, to monitor known patients, a customized model can be created according to their first results, and, over time, their health improvement or deterioration could be monitored. Finally, to pre-select a cohort of patients, we can use a reference model to determine, in a fast way, whether a new group of patients belongs to this specific category.

This year we first addressed the model definition and its suitability to check behavioral properties of interest [24]. Indeed, this is mandatory before envisioning any clinical study.

The next step will be to validate our approach as well as to test its scalability on three other serious games selected with the help of clinicians. We wrote a medical protocol to be submitted to CERNI proposing clinical experimentations with patients. This protocol will be a collaboration with the ICP institute, member of the CoBTEX laboratory. The new games will be modeled in PRISM and different configurations (for example for Mild, Moderate or Severe Alzheimer) will be set up with the participation of clinicians. Then, several groups of patients will play these games and their results will be recorded to calibrate our initial models.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts and Grants with Industry

Stars team has currently several experiences in technological transfer towards industrials, which have permitted to exploit research result:

7.1.1. Ekinnox

is a spin-off project of the Stars team which aims at improving the rehabilitation process for caregivers and patients. Thanks to a computer vision based system (camera combined with algorithms detecting human motion), Ekinnox provides a simple and efficient tool to quantify and visualize the performance of patients (e.g. gait parameters computation such as side-by-side video comparison, automatic sequencing of video or 3D display) during their rehabilitation process. This company was created at the beginning of 2017.

7.1.2. Toyota

is working with Stars on action recognition software to be integrated on their robot platform. This project aims at detecting critical situations in the daily life of older adults alone at home. This will require not only recognition of ADLs but also an evaluation of the way and timing in which they are being carried out. The system we want to develop is intended to help them and their relatives to feel more comfortable because they know that potential dangerous situations will be detected and reported to caregivers if necessary. The system is intended to work with a Partner Robot - HSR - (to send real-time information to the robot) to better interact with the older adult.

7.1.3. Vedecom

is interested in developing algorithms for people detection for self-driving cars. Among many challenges in pedestrian detection, the ones of interest are a) Scale- handling, b) Occlusion-handling and c) Cross-dataset generalization. Each of the aforementioned challenges is critical to enable modern applications like self-driving vehicles become safe enough for active deployment. To improve the performance of contemporary pedestrian detectors, one of our first major idea is to use multiple layers of a CNN simultaneously. Towards this, we proposed a new pedestrian detection system called Multiple-RPN. Another recent work is adding pseudo-segmentation information to pedestrian detection. The proposed features of our system perform close to the best performing detectors today.

7.1.4. Kontron

has a collaboration with Stars, which runs from April 2018 until April 2021 to embed CNN based people tracker within a video-camera. Their system uses Intel VPU modules, such as Myriad X (MA2485), based on OpenVino library.

7.1.5. The company ESI

(European System Integration) has a collaboration with Stars, which runs from September 2018 until March 2022 to develop a novel Re-Identification algorithm which can be easily set-up with low interaction for video-surveillance applications. ESI provides software solutions for remote monitoring stations, remote assistance, video surveillance, and call centers. It was created in 1999 and ESI is a leader in the French remote monitoring market. Nowadays, ensuring the safety of goods and people is a major problem. For this reason, surveillance technologies are attracting growing interest and their objectives are constantly evolving: it is now a question of automating surveillance systems and helping video surveillance operators in order to limit interventions and staff. One of the current difficulties is the human processing of video, as the multiplication of video streams makes it difficult to understand meaningful events. It is therefore necessary to give video surveillance operators suitable tools to assist them with tasks that can be automated. The integration of video analytics modules will allow surveillance technologies to gain in efficiency and precision. In recent times, deep learning techniques have been made possible by the advent of GPU processors, which offer significant processing possibilities. This leads to the development of automatic video processing.

7.1.6. Fantastic Sourcing

is a French SME specialized in micro-electronics, it develops e-health technologies. Fantastic Sourcing is collaborating with Stars through the UCA Solitaria project, by providing their Nodeus system. Nodeus is a IoT (Internet of Things) system for home support for the elderly, which consists of a set of small sensors (without video cameras) to collect precious data on the habits of isolated people. Solitaria project performs a multi-sensor activity analysis for monitoring and safety of older and isolated people. With the increase of the ageing population in Europe and in the rest of the world, keeping elderly people at home, in their usual environment, as long as possible, becomes a priority and a challenge of modern society. A system for monitoring activities and alerting in case of danger, in permanent connection with a device (an application on a phone, a surveillance system ...) to warn relatives (family, neighbours, friends ...) of isolated people still living in their natural environment could save lives and avoid incidents that cause or worsen the loss of autonomy. In this R&D project, we propose to study a solution allowing the use of a set of innovative heterogeneous sensors in order to: 1) detect emergencies (falls, crises, etc.) and call relatives (neighbours, family, etc.); 2) detect, over short or longer predefined periods, behavioural changes in the elderly through an intelligent analysis of data from sensors.

7.1.7. Nively

is a French SME specialized in e-health technologies, it develops position and activity monitoring of activities of daily living platforms based on video technology. Nively's mission is to use technological tools to put people back at the center of their interests, with their emotions, identity and behavior. Nively is collaborating with Stars through the UCA Solitaria project, by providing their MentorAge system. This software allows the monitoring of elderly people in nursing homes in order to detect all the abnormal events in the lives of residents (falls, runaways, strolls, etc.). Nively's technology is based on RGBD video sensors (Kinects type) and a software platform for event detection and data visualization. Nively is also in charge of Software distribution for the ANR Activis project. This project is based on an objective quantification of the atypical behaviors on which the diagnosis of autism is based, with medical (diagnostic assistance and evaluation of therapeutic programs) and computer scientific (by allowing a more objective description of atypical behaviors in autism) objectives. This quantification requires video analysis of the behavior of people with autism. In particular, we propose to explore the issues related to the analysis of ocular movement, gestures and posture to characterize the behavior of a child with autism. Thus, Nively will add autistic behavior analysis software to its product range.

More bilateral Grants with industries is available at: http://www-sop.inria.fr/members/Francois.Bremond/topicsText/researchProjec

8. Partnerships and Cooperations

8.1. Regional Initiatives

See CoBTek, Nice Hospital, FRIS

8.2. National Initiatives

See Vedecom

8.2.1. ANR

8.2.1.1. ENVISION

Program: ANR JCJC

Project acronym: ENVISION

Project title: Computer Vision for Automated Holistic Analysis of Humans

Duration: October 2017-September 2020.

Coordinator: Antitza Dantcheva (STARS)

Abstract: The main objective of ENVISION is to develop the computer vision and theoretical foundations of efficient biometric systems that analyze appearance and dynamics of both face and body, towards recognition of identity, gender, age, as well as mental and social states of humans in the presence of operational randomness and data uncertainty. Such dynamics - which will include facial expressions, visual focus of attention, hand and body movement, and others, constitute a new class of tools that have the potential to allow for successful holistic analysis of humans, beneficial in two key settings: (a) biometric identification in the presence of difficult operational settings that cause traditional traits to fail, (b) early detection of frailty symptoms for health care.

8.2.2. FUI

8.2.2.1. Visionum

Program: FUI

Project acronym: Visionum

Project title: Visonium.

Duration: January 2015- December 2018.

Coordinator: Groupe Genious

Other partners: Inria (Stars), StreetLab, Fondation Ophtalmologique Rothschild, Fondation Hospitalière Sainte-Marie.

Abstract: This French project from Industry Minister aims at designing a platform to re-educate at home people with visual impairment.

8.2.2.2. StoreConnect

Program: FUI

Project acronym: StoreConect.

Project title: StoreConnect.

Duration: September 2016 - June 2019.

Coordinator: Ubudu (Paris).

Other partners: Inria (Stars), STIME (groupe Les Mousquetaires Paris), Smile (Paris), Thevolys (Dijon).

Abstract: StoreConnect is a FUI project started in 2016 and ended in 2019. The goal is to improve the shopping experience for customers inside supermarkets by adding new sensors such as cameras, beacons and RFID. By gathering data from all the sensors and combining them, it is possible to improve the way to communicate between shops and customers in a personalized way. StoreConnect acts as a middleware platform between the sensors and the shops to process the data and extract interesting knowledge organized via ontologies.

8.2.2.3. ReMinAry

Program: FUI

Project acronym: ReMinAry.

Project title: ReMinAry.

Duration: September 2016 - June 2020.

Coordinator: GENIOUS Systèmes,

Other partners: Inria (Stars), MENSIA technologies, Institut du Cerveau et de la Moelle épinière, la Pitié-Salpêtrière hospital.

Abstract: This project is based on the use of motor imagery (MI), a cognitive process consisting of the mental representation of an action without concomitant movement production. This technique consists in imagining a movement without realizing it, which entails an activation of the brain circuits identical to those activated during the real movement. By starting rehabilitation before the end of immobilization, a patient operated on after a trauma will gain rehabilitation time and function after immobilization is over. The project therefore consists in designing therapeutic video games to encourage the patient to re-educate in a playful, autonomous and active way in a phase where the patient is usually passive. The objective will be to measure the usability and the efficiency of the re-educative approach, through clinical trials centered on two pathologies with immobilization: post-traumatic (surgery of the shoulder) and neurodegenerative (amyotrophic lateral sclerosis).

8.3. European Initiatives

8.3.1. Collaborations in European Programs, Except FP7 & H2020

See EIT Health.

8.4. International Initiatives

8.4.1. Inria International Labs

- $EASafEE$: Associated team (2018-2020) Safe and Easy Environment for Alzheimer disease and related disorders. Inria Stars, National Taipei University of Technology Taiwan and CoBTeK team. The objective of SafEE is to develop an automated home support system, using information and communication technologies (ICT), to support the loss of autonomy and to improve the quality of life of the elderly population.
- $FER4HM$: Inria International Lab (2017-2020) Facial Expression Recognition for Health Monitoring. Coordinator: François Brémond, Antitza Dantcheva. Other partners: Chinese Academy of Sciences (CAS) (China). FER4HM aims to investigate computer vision methods for facial expression recognition in patients with Alzheimer's disease. Most importantly though, the project seeks to be part of a paradigm shift in current health care, efficiently and cost-effectively finding objective measures to (a) assess different therapy treatments, as well to (b) enable automated human-computer interaction in remote scale health care-frameworks.

8.4.1.1. Other IIL projects

- RESPECT

  Program: ANR PRCI (French-German, ANR-DFG)

  Project acronym: RESPECT

  Project title: Reliable, secure and privacy preserving multi-biometric person authentication

  Duration: April 2019-March 2023.

  Coordinator: Antitza Dantcheva (STARS)

Abstract: In spite of the numerous advantages of biometric recognition systems over traditional authentication systems based on PINs or passwords, these systems are vulnerable to external attacks and can leak data. Presentations attacks (PAs) – impostors who manipulate biometric samples to masquerade as other people – pose serious threats to security. Privacy concerns involve the use of personal and sensitive biometric information, as classified by the GDPR, for purposes other than those intended. Multi-biometric systems, explored extensively as a means of improving recognition reliability, also offer potential to improve PA detection (PAD) generalisation. Multi-biometric systems offer natural protection against spoofing since an impostor is less likely to succeed in fooling multiple systems simultaneously. For the same reason, previously unseen PAs are less likely to fool multi-biometric systems protected by PAD. RESPECT, a Franco-German collaborative project, explores the potential of using multi-biometrics as a means to defend against diverse PAs and improve generalisation while still preserving privacy. Central to this idea is the use of (i) biometric characteristics that can be captured easily and reliably using ubiquitous smart devices and, (ii) biometric characteristics which facilitate computationally manageable privacy preserving, homomorphic encryption.

The research focuses on characteristics readily captured with consumer-grade microphones and video cameras, specifically face, iris and voice. Further advances beyond the current state of the art involve the consideration of dynamic characteristics, namely utterance verification and lip dynamics. The core research objective is to determine which combination of biometrics characteristics gives the best biometric authentication reliability and PAD generalisation while remaining compatible with computationally efficient privacy preserving BTP schemes.

- $VIdeoSeizureAnalysis$: Inserm-Inria PhD grant (October 2018- September 2021). Partners: Prof F Bartolomei Inserm UMR 1106 La Timone Hospital Marseille and M Thonnat DR Inria Stars Sophia Antipolis. The objective of the PhD thesis entitled Quantified video analysis of seizure semiology in epilepsy is to provide new automated and objective analysis and interpretation of recorded videos of patients during epilepsy seizures.

8.4.2. Inria Associate Teams Not Involved in an Inria International Labs

8.4.2.1. SafEE (Safe & Easy Environment)

Title: SafEE (Safe Easy Environment) investigates technologies for the evaluation, stimulation and intervention for Alzheimer patients. The SafEE project aims at improving the safety, autonomy and quality of life of older people at risk or suffering from Alzheimer.

International Partner (Institution - Laboratory - Researcher):

National Taipei University of Technology Taipei (Taiwan) - Dept. of Electrical Engineering - Chao-Cheng Wu

Start year: 2018

See also: https://project.inria.fr/safee2/

SafEE (Safe Easy Environment) investigates technologies for the evaluation, stimulation and intervention for Alzheimer patients. The SafEE project aims at improving the safety, autonomy and quality of life of older people at risk or suffering from Alzheimer's disease and related disorders. More specifically the SafEE project : 1) focuses on specific clinical targets in three domains: behavior, motricity and cognition 2) merges assessment and non pharmacological help/intervention and 3) proposes easy ICT device solutions for the end users. In this project, experimental studies will be conducted both in France (at Hospital and Nursery Home) and in Taiwan.

8.4.2.2. Declared Inria International Partners

See Taiwan, China

8.4.3. Participation in Other International Programs

8.4.3.1. International Initiatives

FER4HM

Title: Facial expression recognition with application in health monitoring

International Partner (Institution - Laboratory - Researcher):

Institute of Computing Technology (ICT) of the Chinese Academy of Sciences (CAS) - Prof. Hu HAN

Duration: 2017 - 2019

Start year: 2017

See also: https://project.inria.fr/fer4hm/

The proposed research aims to provide computer vision methods for facial expression recognition in patients with Alzheimer's disease. Most importantly though, the work seeks to be part of a paradigm shift in current healthcare, in efficiently and cost effectively finding objective measures to (a) assess different therapy treatments, as well as to (b) enable automated human-computer interaction in remote large-scale healthcare- frameworks. Recognizing expressions in severely demented Alzheimer's disease (AD) patients is essential, since such patients have lost a substantial amount of their cognitive capacity, and some even their verbal communication ability (e.g., aphasia). This leaves patients dependent on clinical staff to assess their verbal and non-verbal language, in order to communicate important messages, as of discomfort associated to potential complications of the AD. Such assessment classically requires the patients' presence in a clinic, and time consuming examination involving medical personnel. Thus, expression monitoring is costly and logistically inconvenient for patients and clinical staff, which hinders among others large-scale monitoring. Approaches need to cater to the challenging settings of current medical recordings, which include continuous pose variations, occlusions, camera-movements, camera-artifacts, as well as changing illumination. Additionally and importantly, the (elderly) patients exhibit generally less profound facial activities and expressions in a range of intensities and predominantly occurring in combinations (e.g., talking and smiling). Both, Inria-STARS and CAS-ICT have already initiated research activities related to the here proposed topic. While both sides have studied facial expression recognition, CAS-ICT has explored additionally the use of heart rate monitoring sensed from a webcam in this context.

8.5. International Research Visitors

8.5.1. Visits of International Scientists

- Wael Abd-Almageed from the Information Sciences Institute of the University of Southern California (USC) Viterbi School of Engineering visited in January 2019.
- Timur Luguev from the Intelligent Systems Group of Fraunhofer Institute for Integrated Circuits, Germany visited STARS in March 2019.
- Alan Aboudib from College de France visited STARS in July 2019.
- Julien Pettre from Inria Rennes (Team Rainbow) visited STARS in July 2019.
- Radu Horaud from Inria Grenoble (Team Perception) visited STARS in September 2019.
- Marcos Zuniga from Universidad Tecnica Federico Santa Maria, Chile visited STARS in 2019.
- Chao-Cheng Hu from the National Taipei University of Technology, Taiwan visited STARS in October 2019.

8.5.2. Internships

Several students from India, China, South Korea

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events: Organisation

9.1.1.1. General Chair, Scientific Chair

- Elisabetta De Maria was General chair of the international conference CsBio 2019 (10th International Conference on Computational Systems-Biology and Bioinformatics), Nice, France.

9.1.1.2. Member of the Organizing Committees

- Antitza Dantcheva, Abhijit Das and François Brémond organized the special session on human health monitoring based on computer vision at the 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG'19).
- Antitza Dantcheva co-organized the Robust Tattoo Detection and Retrieval Competition (RTDRC 2019) associated to the 10th IEEE International Conferencce on Biometrics: Theory, Applications and Systems (BTAS 2019).
- Antitza Dantcheva co-organized the special session on The Future of Biometrics beyond Recognition and Anti-Spoofing at the 12th IAPR International Conference on Biometrics (ICB'19).

9.1.1.3. Chair of Conference Program Committees

- François Brémond was part of the Organization Committee for the workshop on "Crowd analysis and applications: simulations meet video analytics"
- François Brémond was part of the AVSS'19 Organization Committee
- François Brémond was a Session Chair of AVSS - 17th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Taipei, Taiwan, 18-21 September 2019.
- Elisabetta De Maria was Program chair of the international conference BIOINFORMATICS 2019 (10th International Conference on Bioinformatics Models, Methods, and Algorithms), which is part of BIOSTEC 2019 (12th International Joint Conference on Biomedical Engineering Systems and Technologies), Prague, Czech Republic.
- Elisabetta De Maria was Program chair of the international conference CsBio 2019 (10th International Conference on Computational Systems-Biology and Bioinformatics), Nice, France.
- Antitza Dantcheva was program Co-chair at the International Conference of the Biometrics Special Interest Group (BIOSIG) 2019, Darmstadt, Germany.

9.1.1.4. Member of the Conference Program Committees

- Monique Thonnat was program committee member of the conference ICPRAM 2020.
- Elisabetta De Maria was member of the program committee of the ICML Workshop on Computational Biology 2019, Long Beach, CA, USA.

9.1.1.5. Reviewer

- François Brémond was reviewer for the conferences: CVPR2019-20, ICCV2019, ICPRS-19, ICDP19, WACV 2020, AVSS19.
- François Brémond was reviewer for the Journal: IEEE International Conference on Systems, Man, and Cybernetics (SMC), IET Computer Vision, IEEE Transactions on Circuits and Systems for Video Technology
- Antitza Dantcheva was reviewer for IEEE Transactions on Information Forensics and Security (TIFS), IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM), IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), Pattern Recognition and Neurocomputing.
- Srijan Das was reviewer for KCST 2019, ICAML 2019, AVSS 2019 and WACV 2020.

9.1.2. Journal

9.1.2.1. Member of the Editorial Boards

- François Brémond was handling editor of the international journal "Machine Vision and Application".
- François Brémond was handling editor for the MDPI special issue sensors Deep Learning for multi-sensor fusion.
- Elisabetta De Maria was guest editor of the JBCB journal (Journal of Bioinformatics and Computational Biology) for the review process of selected papers for the special issue on the conference BIOINFORMATICS 2019.
- Antitza Dantcheva has been in the Editorial Board of the Journal Multimedia Tools and Applications (MTAP) since 2017.

9.1.3. Invited Talks

- François Brémond was invited at the Data Science Riviera to give the talk “Cross domain residual transfer learning for person re-identification” in April 2019.
- François Brémond was invited to give a keynote speech on people tracking at the T4S Workshop in AVSS Taipei, 21 September 2019.
- François Brémond was invited in HealthCare and AI CITEDI at Tijuana, 23 August 2019.
- Monique Thonnat was invited as Speaker on AI and Health session to the CEO Forum at VivaTech, Paris, 16 May 2019.
- Monique Thonnat was invited to give a talk entitled Where to Focus on for Human Action Recognition? at the Franco-Mexican workshop on AI, Mexico, 27-29 August 2019.
- Monique Thonnat was invited by TEC Monterey for to give a talk on AI for Daily Living Activity recognition from videos at the AI Hub Launch, Guadalajara, 27 November 2019.
- Elisabetta De Maria was invited at *School of Electrical Engineering and Computer Science*, University of Ottawa, Canada in June 2019. Title of the talk: Parameter Learning for Spiking Neural Networks Modelled as Timed Automata.
- Elisabetta De Maria was invited at *Center of Modeling, Simulation, and Interactions (MSI)*, Université Côte d’Azur, France in September 2019. Title of the talk: Parameter Learning for Spiking Neural Networks Modelled as Timed Automata.
- Srijan Das was invited at the Data Science Riviera to give the talk “Spatio-temporal attention mechanism for Activities of Daily Living” in November.
- S L Happy was invited at the Data Science Riviera to give the talk “Apathy diagnosis by analyzing facial dynamics in videos” in April.
- Abhijit Das was invited at the Data Science Riviera to give the talk “Robust face analysis employing machine learning techniques for remote heart rate estimation and towards unbiased attribute analysis” in January.

9.1.4. Leadership within the Scientific Community

- François Brémond was part of the Evaluation Committee for new Inria team creation, Chorale.
- François Brémond was part of the Evaluation Committee for PINZ Axel research application for Austrian Science Fund.
- Monique Thonnat is member of the scientific board of ENPC, Ecole Nationale des Ponts et Chaussées since June 2008.
- Jean-Paul Rigault is an ISO C++ expert and the head of the French delegation at the ISO C++ standardization committee.

- Antitza Dantcheva serves in the Technical Activities Committee of the IEEE Biometrics Council since 2017
- Antitza Dantcheva serves in the EURASIP Biomedical Image & Signal Analytics (BISA) SAT 2018-2021
- Antitza Dantcheva is member of the European Reference Network for Critical Infrastructure Protection (ERNCIP), Thematic Group Extended Virtual Fencing - use of biometric and video technologies, since 2017
- Antitza Dantcheva is member of the European Association for Biometrics, since 2018

9.1.5. Scientific Expertise

Elisabetta De Maria was facilitator of the brainstorming of the strategic axis "Humain-Biologie" during the meeting of the I3S Laboratory, Fréjus, France.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Stars Team members (e.g. François Brémond) gave the class Master Data Science M2 in *Computer Vision and Deep Learning* from December 10, 2019 to February 25, 2020.

http://www-sop.inria.fr/members/Francois.Bremond/MSclass/deepLearningWinterSchool/index.html

9.2.2. Supervision

- PhD: U. Ujjwal, "Pedestrian detection to dynamically populate the map of a crossroad" [12], Thèses, Université Côte d'Azur, November 2019.
- PhD: I. Sarray, "Conception de systèmes de reconnaissance d'activités humaines" [11], Thèses, Université Côte d'Azur, March 2019.

  PhD in progress: Srijan Das, "Action recognition of daily living activities from RGB-D videos", 2017, PhD codirected 50% François Brémond and Monique Thonnat.

  PhD in progress: Yaohui Wang, 2017, "Automated holistic human analysis", Antitza Dantcheva.

  PhD in progress: Jen-Cheng Hou: "Quantified video analysis of seizure semiology in epilepsy", 2018, PhD codirected 50% Monique Thonnat and Prof. Fabrice Bartolomei, PU-PH AMU/Inserm.

  PhD in progress: Juan Diego Gonzales Zuniga, 2018, "People Tracking using Deep Learning algorithms on embedded hardware", 70% François Brémond and 30% Serge Tissot (Fellowship CIFRE - Kontron).

  PhD in progress: Thibaud L'Yvonnet, "Relations between human behaviour models and brain models - Application to serious games", 2018, Sabine Moisan and Elisabetta De Maria.

  PhD in progress: Hao Chen, 2019, "People Re-identification using Deep Learning methods", 70% François Brémond and 30% Benoit Lagadec (Fellowship CIFRE - ESI).

  PhD in progress: Rui Dai, 2019, "Action Detection for Untrimmed Videos based on Deep Neural Networks", François Brémond.

  Abdorrahim Bahrami, "Modelling and verifying dynamical properties of biological neural networks in Coq", Elisabetta De Maria.

Srijan Das was mentor for the Emerging Technology Business Incubator (ETBI) Led by NIT Rourkela, a platform envisaged transforming the start-up ecosystem of the region.

Srijan Das mentored for B.E.N.J.I. in GirlScript Summer of Code 2019 edition.

9.2.3. Juries

- François Brémond was jury member of Tenure Track Selection: committee member for permanent position, COS informatique, Lyon 2 University, 6 May 2019
- François Brémond was jury member for habilitation, Anthony Fleury, Lille University, 20 February 2019
- François Brémond was jury member for habilitation, Stefan Duffner, Lyon University, 5 April 2019
- François Brémond was jury member of the mid-term review for 6 PhDs - Nicolas Girard (May 3rd, 2019), Melissa Sanabria (May 21st, 2019), Lucas Pascal (July 11, 2019), Claire Labit (October 8, 2019), Renato Baptista (January 17 and August 28, 2019), Magali PAYNE (December 3, 2019).
- François Brémond was jury member of the following PhD theses:

  PhD, Jennifer Vandoni, Université de Paris Saclay, Saclay, 14 May 2019.

  PhD, Amr Alyafi, Grenoble Institute of Technology, 27 May 2019.

  PhD, Cristiano Massaroni, La Sapienza University in Rome, 13 December 2019.
- Monique Thonnat was reviewer for the PhD defense of Florent Lefevre, University of Lorraine, 4 December 2019
- Monique Thonnat was president for the PhD defense of Danny Francis, Sorbonne Université, 12 December 2019
- Jean-Paul Rigault was president of the HDR jury of Julien Deantoni.
- Elisabetta De Maria was member of the Ph.D. proposal defence jury of Abdorrahim Bahrami, University of Ottawa, Canada. Title of the thesis: Verifying Dynamic Properties of Neural Networks in Coq.
- Antitza Dantcheva was reviewer for the PhD defense of Mohamed Abdul Cader, Queensland University of Technology, Australia.

9.3. Popularization

9.3.1. Articles and contents

- François Brémond was interviewed for *Web Interview* on Facial recognition: limits and challenges for society, 16 Oct 2019.
- François Brémond was interviewed for *Graphical novel* about the challenges of AI, 29 Oct 2019.
- Antitza Dantcheva was interviewed for an article in *Science et Vie* on facial analysis in October 2019.
- Antitza Dantcheva was interviewed for an article in *Charlie Hebdo* on facial recognition in February 2019.

9.3.2. Interventions

- François Brémond, Thibaud L'Yvonnet, David Anghelone and Sandrine Boute represented STARS on the 19 October at the *La fête de la science* in Palais des Congrès d'Antibes Juan-les-Pins.
- STARS presented demos for Unlimitech Sport, Lyon, 18-21 September 2019.

10. Bibliography

Major publications by the team in recent years

[1] P. BILINSKI, F. BREMOND. *Video Covariance Matrix Logarithm for Human Action Recognition in Videos*, in "IJCAI 2015 - 24th International Joint Conference on Artificial Intelligence (IJCAI)", Buenos Aires, Argentina, July 2015, https://hal.inria.fr/hal-01216849

[2] S. BĄK, G. CHARPIAT, E. CORVEE, F. BREMOND, M. THONNAT. *Learning to match appearances by correlations in a covariance metric space*, in "European Conference on Computer Vision", Springer, 2012, pp. 806–820

[3] C. F. CRISPIM-JUNIOR, V. BUSO, K. AVGERINAKIS, G. MEDITSKOS, A. BRIASSOULI, J. BENOIS-PINEAU, Y. KOMPATSIARIS, F. BREMOND. *Semantic Event Fusion of Different Visual Modality Concepts for Activity Recognition*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2016, vol. 38, pp. 1598 - 1611 [*DOI :* 10.1109/TPAMI.2016.2537323], https://hal.inria.fr/hal-01399025

[4] M. KAÂNICHE, F. BREMOND. *Gesture Recognition by Learning Local Motion Signatures*, in "CVPR 2010 : IEEE Conference on Computer Vision and Pattern Recognition", San Franscico, CA, United States, IEEE Computer Society Press, June 2010, https://hal.inria.fr/inria-00486110

[5] M. KAÂNICHE, F. BREMOND. *Recognizing Gestures by Learning Local Motion Signatures of HOG Descriptors*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2012, https://hal.inria.fr/hal-00696371

[6] S. MOISAN. *Knowledge Representation for Program Reuse*, in "European Conference on Artificial Intelligence (ECAI)", Lyon, France, July 2002, pp. 240-244

[7] S. MOISAN, A. RESSOUCHE, J.-P. RIGAULT. *Blocks, a Component Framework with Checking Facilities for Knowledge-Based Systems*, in "Informatica, Special Issue on Component Based Software Development", November 2001, vol. 25, n^{o} 4, pp. 501-507

[8] A. RESSOUCHE, D. GAFFÉ. *Compilation Modulaire d'un Langage Synchrone*, in "Revue des sciences et technologies de l'information, série Théorie et Science Informatique", June 2011, vol. 4, n^{o} 30, pp. 441-471, http://hal.inria.fr/inria-00524499/en

[9] M. THONNAT, S. MOISAN. *What Can Program Supervision Do for Software Re-use?*, in "IEE Proceedings - Software Special Issue on Knowledge Modelling for Software Components Reuse", 2000, vol. 147, n^{o} 5

[10] V. VU, F. BREMOND, M. THONNAT. *Automatic Video Interpretation: A Novel Algorithm based for Temporal Scenario Recognition*, in "The Eighteenth International Joint Conference on Artificial Intelligence (IJCAI'03)", 9-15 September 2003

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] I. SARRAY. *Design of human activity recognition systems*, Jean-Paul Bodeveix, Professeur, Université Paul Sabatier Toulouse ; Frédéric Boulanger, Professeur, Centrale SUPELEC Paris ; Frédéric Mallet, Professeur, Université Côte d'Azur ; Térésa Colombi, Directrice, LUDOTIC, March 2019, https://hal.inria.fr/tel-02145417

[12] U. UJJWAL. *Handling the speed-accuracy trade-off in deep-learning based pedestrian detection systems*, Inria Sophia Antipolis - Méditerranée ; Universite cote d'Azur, November 2019, https://hal.inria.fr/tel-02416418

Articles in International Peer-Reviewed Journals

[13] F. BREMOND, R. TRICHET. *How to train your dragon: best practices in pedestrian classifier training*, in "IEEE Access", January 2019, pp. 1-1 [*DOI :* 10.1109/ACCESS.2019.2891950], https://hal.archives-ouvertes.fr/hal-02422526

[14] C. CAETANO, V. DE MELO, F. BREMOND, J. A. DOS SANTOS, W. ROBSON SCHWARTZ. *Magnitude-Orientation Stream Network and Depth Information applied to Activity Recognition*, in "Journal of Visual Communication and Image Representation", August 2019, https://hal.archives-ouvertes.fr/hal-02422536

[15] S. L. HAPPY, A. DANTCHEVA, F. BREMOND. *A Weakly Supervised Learning Technique for Classifying Facial Expressions*, in "Pattern Recognition Letters", December 2019 [*DOI :* 10.1016/J.PATREC.2019.08.025], https://hal.inria.fr/hal-02381439

[16] V. MANERA, S. ABRAHAMS, L. AGÜERA-ORTIZ, F. BREMOND, R. DAVID, K. FAIRCHILD, A. GROS, C. HANON, M. HUSAIN, A. KÖNIG, P. LOCKWOOD, M. PINO, R. RADAKOVIC, G. ROBERT, A. SLACHEVSKY, F. STELLA, A. TRIBOUILLARD, P. D. TRIMARCHI, F. VERHEY, J. YESAVAGE, R. ZEGHARI, P. ROBERT. *Recommendations for the Nonpharmacological Treatment of Apathy in Brain Disorders*, in "American Journal of Geriatric Psychiatry", August 2019 [*DOI :* 10.1016/J.JAGP.2019.07.014], https://hal.archives-ouvertes.fr/hal-02339088

[17] F. NEGIN, F. BREMOND. *An Unsupervised Framework for Online Spatiotemporal Detection of Activities of Daily Living by Hierarchical Activity Models*, in "Sensors", October 2019, vol. 19, n^{o} 19, 4237 p. [*DOI :* 10.3390/S19194237], https://hal.archives-ouvertes.fr/hal-02422522

[18] C. RATHGEB, A. DANTCHEVA, C. BUSCH. *Impact and Detection of Facial Beautification in Face Recognition: An Overview*, in "IEEE Access", December 2019 [*DOI :* 10.1109/ACCESS.2019.DOI], https://hal.inria.fr/hal-02378939

[19] J. TRÖGER, N. LINZ, A. KÖNIG, P. ROBERT, J. ALEXANDERSSON, J. PETER, J. KRAY. *Exploitation vs. exploration—computational temporal and semantic analysis explains semantic verbal fluency impairment in Alzheimer's disease*, in "Neuropsychologia", August 2019, vol. 131, pp. 53-61 [*DOI :* 10.1016/J.NEUROPSYCHOLOGIA.2019.05.007], https://hal.archives-ouvertes.fr/hal-02339134

International Conferences with Proceedings

[20] S. DAS, A. CHAUDHARY, F. BREMOND, M. THONNAT. *Where to Focus on for Human Action Recognition?*, in "WACV 2019 - IEEE Winter Conference on Applications of Computer Vision", Waikoloa Village, Hawaii, United States, January 2019, pp. 1-10, https://hal.inria.fr/hal-01927432

[21] S. DAS, R. DAI, M. F. KOPERSKI, L. MINCIULLO, L. GARATTONI, F. BREMOND, G. FRANCESCA. *Toyota Smarthome: Real-World Activities of Daily Living*, in "ICCV 2019 -17th International Conference on Computer Vision", Seoul, South Korea, October 2019, https://hal.inria.fr/hal-02366687

[22] S. DAS, M. THONNAT, F. BREMOND. *Looking deeper into Time for Activities of Daily Living Recognition*, in "WACV 2020 - IEEE Winter Conference on Applications of Computer Vision", Snowmass village, Colorado, United States, March 2020, https://hal.inria.fr/hal-02368366

[23] S. DAS, M. THONNAT, K. SAKHALKAR, M. F. KOPERSKI, F. BREMOND, G. FRANCESCA. *A New Hybrid Architecture for Human Activity Recognition from RGB-D videos*, in "MMM 2019 - 25th International Conference on MultiMedia Modeling", Thessaloniki, Greece, Lecture Notes in Computer Science, Springer,

January 2019, vol. 11296, pp. 493-505 [*DOI :* 10.1007/978-3-030-05716-9_40], https://hal.inria.fr/hal-01896061

[24] E. DE MARIA, T. L'YVONNET, S. MOISAN, J.-P. RIGAULT. *Probabilistic Activity Recognition For Serious Games With Applications In Medicine*, in "ICFEM 2019 - FTSCS workshop", Shenzhen, China, November 2019, https://hal.inria.fr/hal-02341600

[25] S. L. HAPPY, A. DANTCHEVA, A. DAS, R. ZEGHARI, P. ROBERT, F. BREMOND. *Characterizing the State of Apathy with Facial Expression and Motion Analysis*, in "FG 2019 - 14th IEEE International Conference on Automatic Face and Gesture Recognition", Lille, France, May 2019, https://hal.inria.fr/hal-02379341

[26] S. L. HAPPY, A. DANTCHEVA, A. ROUTRAY. *Dual-threshold Based Local Patch Construction Method for Manifold Approximation And Its Application to Facial Expression Analysis*, in "EUSIPCO 2019 - 27th European Signal Processing Conference", A Coruna, Spain, September 2019, https://hal.inria.fr/hal-02378985

[27] F. M. KHAN, F. BREMOND. *Cross domain Residual Transfer Learning for Person Re-identification*, in "WACV 2019 - IEEE's and the PAMI-TC's premier meeting on applications of computer vision", Waikoloa Village, Hawaii, United States, January 2019, https://hal.inria.fr/hal-01947523

[28] X. NIU, X. ZHAO, H. HAN, A. DAS, A. DANTCHEVA, S. SHAN, X. CHEN. *Robust Remote Heart Rate Estimation from Face Utilizing Spatial-temporal Attention*, in "FG 2019 - 14th IEEE International Conference on Automatic Face and Gesture Recognition", Lille, France, May 2019, https://hal.inria.fr/hal-02381138

[29] U. UJJWAL, A. DZIRI, B. LEROY, F. BREMOND. *A One-and-Half Stage Pedestrian Detector*, in "WACV 2020 - IEEE Winter Conference on Applications of Computer Vision", Snowmass Village, United States, March 2020, https://hal.inria.fr/hal-02363756

[30] S. YU, H. HAN, S. SHAN, A. DANTCHEVA, X. CHEN. *Improving Face Sketch Recognition via Adversarial Sketch-Photo Transformation*, in "FG 2019 - 14th IEEE International Conference on Automatic Face and Gesture Recognition", Lille, France, May 2019, https://hal.inria.fr/hal-02381115

Conferences without Proceedings

[31] C. CAETANO, J. SENA, F. BREMOND, J. A. DOS SANTOS, W. ROBSON SCHWARTZ. *SkeleMotion: A New Representation of Skeleton Joint Sequences Based on Motion Information for 3D Action Recognition*, in "AVSS 2019 - 16th IEEE International Conference on Advanced Video and Signal-based Surveillance", Taipei, Taiwan, September 2019, https://hal.archives-ouvertes.fr/hal-02422551

[32] H. CHEN, B. LAGADEC, F. BREMOND. *Partition and Reunion: A Two-Branch Neural Network for Vehicle Re-identification*, in "CVPR Workshops 2019", Long Beach, United States, June 2019, https://hal.archives-ouvertes.fr/hal-02353527

[33] H. CHEN, B. LAGADEC, F. BREMOND. *Learning Discriminative and Generalizable Representations by Spatial-Channel Partition for Person Re-Identification*, in "WACV 2020 - IEEE Winter Conference on Applications of Computer Vision", Snowmass Village, United States, March 2020, https://hal.archives-ouvertes.fr/hal-02374246

[34] R. DAI, L. MINCIULLO, L. GARATTONI, G. FRANCESCA, F. BREMOND. *Self-Attention Temporal Convolutional Network for Long-Term Daily Living Activity Detection*, in "AVSS 2019 - 16th IEEE International

Conference on Advanced Video and Signal-Based Surveillance (AVSS)", Taipei, Taiwan, September 2019, https://hal.archives-ouvertes.fr/hal-02357161

[35] Y. WANG, P. BILINSKI, F. BREMOND, A. DANTCHEVA. *ImaGINator: Conditional Spatio-Temporal GAN for Video Generation*, in "WACV 2020 - Winter Conference on Applications of Computer Vision", Snowmass Village, United States, March 2020, https://hal.archives-ouvertes.fr/hal-02368319

Other Publications

[36] A. KÖNIG, V. NARAYAN, P. AALTEN, I. H. RAMAKERS, N. LINZ, J. TRÖGER, P. ROBERT. *Novel Digitalized Markers for Screening and Disease Trajectory Tracking in Clinical Trials*, July 2019, vol. 15, n^{o} 7, P158 p. , AAIC 2019 - Alzheimer's Association International Conference, Poster [*DOI :* 10.1016/J.JALZ.2019.06.4324], https://hal.archives-ouvertes.fr/hal-02339170

[37] R. ZEGHARI, P. ROBERT, V. MANERA, M. LORENZI, A. KÖNIG. *Towards a Multidimensional Assessment of Apathy in Neurocognitive Disorders*, July 2019, vol. 15, n^{o} 7, 569 p. , AAIC 2019 - Alzheimer's Association International Conference, Poster [*DOI :* 10.1016/J.JALZ.2019.06.4514], https://hal.archives-ouvertes.fr/hal-02339152

References in notes

[38] T. CHEN, T. MOREAU, Z. JIANG, L. ZHENG, E. YAN, M. COWAN, H. SHEN, L. WANG, Y. HU, L. CEZE, C. GUESTRIN, A. KRISHNAMURTHY. *TVM: An Automated End-to-end Optimizing Compiler for Deep Learning*, in "Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation", Berkeley, CA, USA, OSDI'18, USENIX Association, 2018, pp. 579–594, http://dl.acm.org/citation.cfm?id=3291168.3291211

[39] A. G. HOWARD, M. ZHU, B. CHEN, D. KALENICHENKO, W. WANG, T. WEYAND, M. ANDREETTO, H. ADAM. *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*, in "CoRR", 2017, vol. abs/1704.04861, http://arxiv.org/abs/1704.04861

[40] K. KANG, H. LI, T. XIAO, W. OUYANG, J. YAN, X. LIU, X. WANG. *Object Detection in Videos with Tubelet Proposal Networks*, 07 2017, pp. 889-897

[41] K. KANG, W. OUYANG, H. LI, X. WANG. *Object Detection from Video Tubelets with Convolutional Neural Networks*, 06 2016, pp. 817-825