Activity Report 2017

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:

A2.1.8. - 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. Personnel

Research Scientists

François Brémond [Team leader, Inria, Senior Researcher, HDR]
Sabine Moisan [Inria, Researcher, HDR]
Annie Ressouche [Inria, Researcher]
Jean Paul Rigault [Professor Emeritus, Univ de Nice - Sophia Antipolis]
Monique Thonnat [Inria, Senior Researcher, from Jul 2017, HDR]

Faculty Member

Frédéric Precioso [Univ de Nice - Sophia Antipolis, Associate Professor, from Mar 2017 until Aug 2017]

Post-Doctoral Fellows

Antitza Dantcheva [Inria, until Feb 2017]
Naveed Ejaz [ERCIM, from May 2017 until Jun 2017]
Alexandra Konig [Inria, from Oct 2017]
Minh Khue Phan Tran [Centre Hospitalier Universitaire de Nice, until Apr 2017]

PhD Students

Srijan Das [Inria, from Aug 2017]
Juan Diego Gonzales Zuniga [Inria, from Sep 2017 until Nov 2017]
Auriane Gros [Centre Hospitalier Universitaire de Nice, until Sep 2017]
Michal Koperski [Inria, until Nov 2017]
Amir Nazemi [Inria, from Jul 2017]
Farhood Negin [Inria]
Thi Lan Anh Nguyen [Inria]
Ines Sarray [Inria]
Ujjwal Ujjwal [VEDCOM, granted by CIFRE]
Yaohui Wang [Inria, from Dec 2017]

Technical staff

Abdelrahman Gaber Abubakr [Inria, from Nov 2017]
Julien Badie [Inria, granted by BPIFRANCE FINANCEMENT SA]
Manikandan Bakthavatchalam [Inria, until Jan 2017]
Carlos Fernando Crispim Junior [Inria, until Aug 2017, granted by BPIFRANCE FINANCEMENT SA]
Antitza Dantcheva [Inria, from Mar 2017]
Anais Ducoffe [Inria, until Jun 2017]
Furqan Muhammad Khan [Inria]
Michal Koperski [Inria, from Dec 2017]
Soumik Mallick [Inria, from Nov 2017]
Valeria Manera [Inria, until Aug 2017]
Thanh Hung Nguyen [Inria]
Javier Ortiz [Inria, until Jun 2017]
Minh Khue Phan Tran [Inria, from Nov 2017]
Rémi Trichet [Inria]
Seongro Yoon [Inria, until Oct 2017]
Matias Marin [Inria, until May 2017]

Interns

Abdelrahman Gaber Abubakr [Inria, from Jun 2017 until Sep 2017]
Nagi Aly [University Abdel Malek Essadi, Marocco, from Oct 2017 until Nov 2017]
Salwa Baabou [Inria, from Apr 2017 until Sep 2017]
Killian Barrere [Ecole normale supérieure de Rennes, from May 2017 until Jul 2017]
Florent Bartoccioni [Inria, from May 2017 until Aug 2017]
Emmanuelle Bineau [Centre Hospitalier Universitaire de Nice, from Oct 2017 until Nov 2017]
Barbara Bonnel [UNS, UFR Médecine, from Sep 2017]
Yu-Feng Chen [Inria, from Feb 2017 until Aug 2017]
Srijan Das [Inria, until May 2017]
Chandraja Dharmana [Inria, until Jun 2017]
Cédric Girard-Riboulleau [Inria, until Jun 2017]
Abhishek Goel [Inria, from Aug 2017]
Dorine Havyarimana [Inria, from Apr 2017 until Sep 2017]
Shaira Kansal [Inria, until Jul 2017]

Kartik Kartik [Inria, until Jul 2017]
Kuan-Ru Lee [Inria, from Aug 2017]
Aimen Neffati [CESI, Valbonne France, from Sep 2017 until Oct 2017]
Rahul Pandey [Inria, until May 2017]
Aurore Rainouard [UNS, UFR Médecine, from Sep 2017]
Francesco Verrini [Agence Erasmus+ France, from Jun 2017]
Jules Diez [CESI, Sophia Antipolis, France, from Jun 2017]

Administrative Assistant
Laurence Briffa [Inria]

Visiting Scientists
Adlen Kerboua [University of 20 aout 1955, until Oct 2017]
Xue Le [University of Cornwell, USA, from Feb 2017 until Jul 2017]
Hassan Loulou [ESTACA, Paris, until Jan 2017]
Behzad Mirmahboub [University of Technology, ISFAMA, Iran, from Apr 2017 until Jul 2017]
Salwa Baabou [Guest PhD, from Sep 2017]

External Collaborators
Abdelrahman Gaber Abubakr [Univ de Bourgogne, from Sep 2017 until Oct 2017]
Carlos Fernando Crispim Junior [Univ Lumière Lyon, from Sep 2017]
Daniel Gaffé [Univ de Nice - Sophia Antipolis]
Sébastien Gilabert [Automat, Ajaccio, from Oct 2017]
Philippe Robert [Centre hospitalier universitaire de Nice]
Guillaume Sacco [Centre hospitalier universitaire de Nice]
Jean Yves Tigli [Univ de Nice - Sophia Antipolis]
Slawomir Bak [Disney Research, from May 2017]
Piotr Tadeusz Bilinski [Oxford University]
Frederic Precioso [Univ de Nice - Sophia Antipolis, from Sep 2017]

2. Overall Objectives

2.1. Presentation

2.1.1. Research Themes

STARS (Spatio-Temporal Activity Recognition Systems) 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.

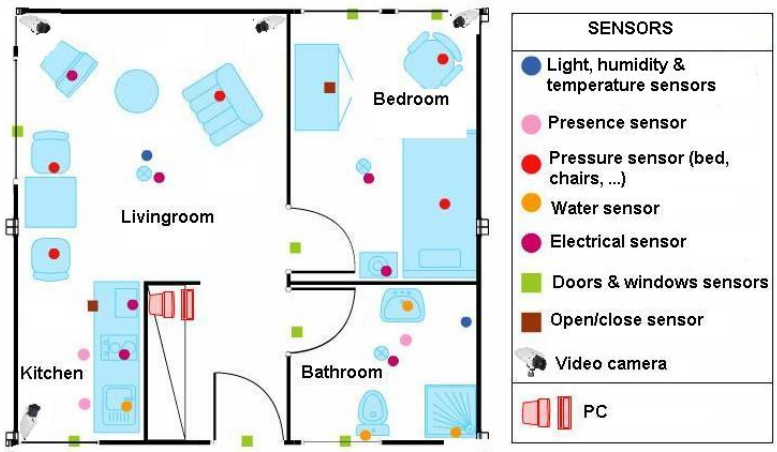


Figure 1. Homecare monitoring: the set of sensors embedded in an apartment

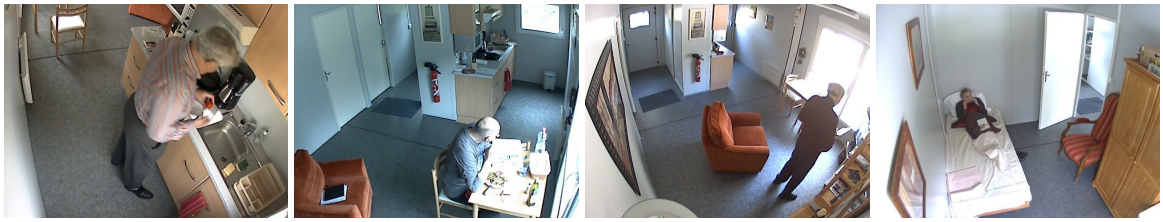


Figure 2. Homecare monitoring: the different views of the apartment captured by 4 video cameras

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

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

for interaction (also called RGB-Depth sensors, an example being the Kinect) or networks of small non visual sensors.

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. We had or have industrial collaborations in several domains: *transportation* (CCI Airport Toulouse Blagnac, SNCF, Inrets, Alstom, Ratp, 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* (Multitel (Belgium), Thales Communications, Idiap (Switzerland)), *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), 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).

3. Research Program

3.1. Introduction

Stars follows three main research directions: perception for activity recognition, semantic activity recognition, and software engineering for activity recognition. **These three research directions are interleaved**: *the software engineering* research direction provides new methodologies for building safe activity recognition systems and *the perception* and *the semantic activity recognition* directions provide new activity recognition techniques which are designed and validated for concrete video analytic and healthcare applications. Conversely, these concrete systems raise new software issues that enrich the software engineering research direction.

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, Sabine Moisan, Monique Thonnat.

Computer Vision; Cognitive Systems; Learning; Activity Recognition.

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 videosurveillance 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. Semantic Activity Recognition

Participants: François Brémond, Sabine Moisan, Monique Thonnat.

Activity Recognition, Scene Understanding, Computer Vision

3.3.1. 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.

3.3.2. 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 ...).

3.3.3. 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.

3.3.4. 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, they will be detailed in section 3.4.

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).

3.4. Software Engineering for Activity Recognition

Participants: Sabine Moisan, Annie Ressouche, Jean-Paul Rigault, François Brémond.

Software Engineering, Generic Components, Knowledge-based Systems, Software Component Platform, Object-oriented Frameworks, Software Reuse, Model-driven Engineering

The aim of this research axis is to build general solutions and tools to develop systems dedicated to activity recognition. For this, we rely on state-of-the art Software Engineering practices to ensure both sound design and easy use, providing genericity, modularity, adaptability, reusability, extensibility, dependability, and maintainability.

This research requires theoretical studies combined with validation based on concrete experiments conducted in Stars. We work on the following three research axes: *models* (adapted to the activity recognition domain), *platform architecture* (to cope with deployment constraints and run time adaptation), and *system verification* (to generate dependable systems). For all these tasks we follow state of the art Software Engineering practices and, if needed, we attempt to set up new ones.

3.4.1. Platform Architecture for Activity Recognition

In the former project teams Orion and Pulsar, we have developed two platforms, one (VSIP), a library of real-time video understanding modules and another one, LAMA [12], a software platform enabling to design not only knowledge bases, but also inference engines, and additional tools. LAMA offers toolkits to build and to adapt all the software elements that compose a knowledge-based system.

Figure 4 presents our conceptual vision for the architecture of an activity recognition platform. It consists of three levels:

- The **Component Level**, the lowest one, offers software components providing elementary operations and data for perception, understanding, and learning.
  - *Perception components* contain algorithms for sensor management, image and signal analysis, image and video processing (segmentation, tracking...), etc.
  - *Understanding components* provide the building blocks for Knowledge-based Systems: knowledge representation and management, elements for controlling inference engine strategies, etc.
  - *Learning components* implement different learning strategies, such as Support Vector Machines (SVM), Case-based Learning (CBL), clustering, etc.

  An Activity Recognition system is likely to pick components from these three packages. Hence, tools must be provided to configure (select, assemble), simulate, verify the resulting component combination. Other support tools may help to generate task or application dedicated languages or graphic interfaces.
- The **Task Level**, the middle one, contains executable realizations of individual tasks that will collaborate in a particular final application. Of course, the code of these tasks is built on top of the components from the previous level. We have already identified several of these important tasks: Object Recognition, Tracking, Scenario Recognition... In the future, other tasks will probably enrich this level.

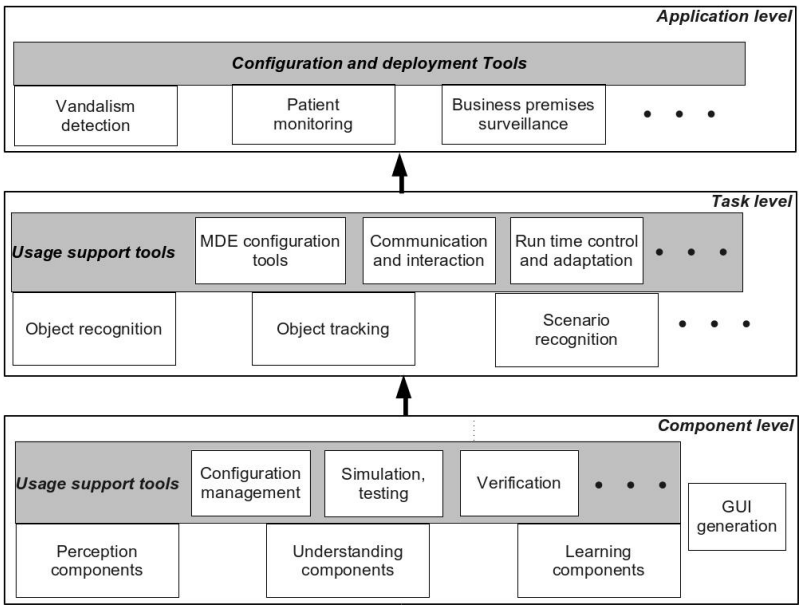


Figure 4. Global Architecture of an Activity Recognition The gray areas contain software engineering support modules whereas the other modules correspond to software components (at Task and Component levels) or to generated systems (at Application level).

For these tasks to nicely collaborate, communication and interaction facilities are needed. We shall also add MDE-enhanced tools for configuration and run-time adaptation.

- The **Application Level** integrates several of these tasks to build a system for a particular type of application, e.g., vandalism detection, patient monitoring, aircraft loading/unloading surveillance, etc.. Each system is parameterized to adapt to its local environment (number, type, location of sensors, scene geometry, visual parameters, number of objects of interest...). Thus configuration and deployment facilities are required.

The philosophy of this architecture is to offer at each level a balance between the widest possible genericity and the maximum effective reusability, in particular at the code level.

To cope with real application requirements, we shall also investigate distributed architecture, real time implementation, and user interfaces.

Concerning implementation issues, we shall use when possible existing open standard tools such as NuSMV for model-checking, Eclipse for graphic interfaces or model engineering support, Alloy for constraint representation and SAT solving for verification, etc. Note that, in Figure 4, some of the boxes can be naturally adapted from SUP existing elements (many perception and understanding components, program supervision, scenario recognition...) whereas others are to be developed, completely or partially (learning components, most support and configuration tools).

3.4.2. Discrete Event Models of Activities

As mentioned in the previous section (3.3) we have started to specify a formal model of scenario dealing with both absolute time and logical time. Our scenario and time models as well as the platform verification tools rely on a formal basis, namely the synchronous paradigm. To recognize scenarios, we consider activity

descriptions as synchronous reactive systems and we apply general modeling methods to express scenario behavior.

Activity recognition systems usually exhibit many safeness issues. From the software engineering point of view we only consider software security. Our previous work on verification and validation has to be pursued; in particular, we need to test its scalability and to develop associated tools. Model-checking is an appealing technique since it can be automatized and helps to produce a code that has been formally proved. Our verification method follows a compositional approach, a well-known way to cope with scalability problems in model-checking.

Moreover, recognizing real scenarios is not a purely deterministic process. Sensor performance, precision of image analysis, scenario descriptions may induce various kinds of uncertainty. While taking into account this uncertainty, we should still keep our model of time deterministic, modular, and formally verifiable. To formally describe probabilistic timed systems, the most popular approach involves probabilistic extension of timed automata. New model checking techniques can be used as verification means, but relying on model checking techniques is not sufficient. Model checking is a powerful tool to prove decidable properties but introducing uncertainty may lead to infinite state or even undecidable properties. Thus model checking validation has to be completed with non exhaustive methods such as abstract interpretation.

3.4.3. Model-Driven Engineering for Configuration and Control and Control of Video Surveillance systems

Model-driven engineering techniques can support the configuration and dynamic adaptation of video surveillance systems designed with our SUP activity recognition platform. The challenge is to cope with the many—functional as well as nonfunctional—causes of variability both in the video application specification and in the concrete SUP implementation. We have used *feature models* to define two models: a generic model of video surveillance applications and a model of configuration for SUP components and chains. Both of them express variability factors. Ultimately, we wish to automatically generate a SUP component assembly from an application specification, using models to represent transformations [43]. Our models are enriched with intra- and inter-models constraints. Inter-models constraints specify models to represent transformations. Feature models are appropriate to describe variants; they are simple enough for video surveillance experts to express their requirements. Yet, they are powerful enough to be liable to static analysis [85]. In particular, the constraints can be analyzed as a SAT problem.

An additional challenge is to manage the possible run-time changes of implementation due to context variations (e.g., lighting conditions, changes in the reference scene, etc.). Video surveillance systems have to dynamically adapt to a changing environment. The use of models at run-time is a solution. We are defining adaptation rules corresponding to the dependency constraints between specification elements in one model and software variants in the other [42], [111], [98].

4. Application Domains

4.1. Introduction

While in our research the focus is to develop techniques, models and platforms that are generic and reusable, we also make effort in the development of real applications. The motivation is twofold. The first is to validate the new ideas and approaches we introduce. The second is to demonstrate how to build working systems for real applications of various domains based on the techniques and tools developed. Indeed, Stars focuses on two main domains: **video analytic** and **healthcare monitoring**.

4.2. Video Analytics

Stars has long-lasting experience in video analytics for security, an area of paramount importance due to the high need for public security, accompanied by the availability of a vast amount of cameras. Related applications range from public safety, smart retail, crowd management to maintenance of secure sites. One current challenge of video analytics is extracting the meaningful information from big video data in an efficient manner, with the goal of high-level semantic descriptions of the activities occurring in the area under surveillance.

Video analytics covers research directions including tracking, object detection, object classification, behavior recognition, pose estimation, and semantic scene modeling. Deep neural networks have brought to the fore great progress in the area of large-scale video analytics and specifically in object recognition and action recognition. However, many challenges remain open, mainly in the context of real-life surveillance scenarios and specifically concerning the robustness, adaptability, and scalability. Prominent open challenges comprise people detection and tracking in crowded environments, as well as detecting abnormal activities, which is often associated with processing of very large video data in the presence of noise and occlusions.

Stars has accumulated a well selected network of industrial partners ranging from end-users, integrators and software editors to provide data, objectives, evaluation and funding.

4.3. Healthcare Monitoring

Since 2011, we have initiated a strategic partnership (called CobTek) with Nice hospital [62], [112] (CHU Nice, Prof P. Robert) to start ambitious research activities dedicated to healthcare monitoring and to assistive technologies. These new studies address the analysis of more complex spatio-temporal activities (e.g. complex interactions, long term activities).

4.3.1. Research

To achieve this objective, several topics need to be tackled. These topics can be summarized within two points: finer activity description and longitudinal experimentation. Finer activity description is needed for instance, to discriminate the activities (e.g. sitting, walking, eating) of Alzheimer patients from the ones of healthy older people. It is essential to be able to pre-diagnose dementia and to provide a better and more specialized care. Longer analysis is required when people monitoring aims at measuring the evolution of patient behavioral disorders. Setting up such long experimentation with dementia people has never been tried before but is necessary to have real-world validation. This is one of the challenge of the European FP7 project Dem@Care where several patient homes should be monitored over several months.

For this domain, a goal for Stars is to allow people with dementia to continue living in a self-sufficient manner in their own homes or residential centers, away from a hospital, as well as to allow clinicians and caregivers remotely provide effective care and management. For all this to become possible, comprehensive monitoring of the daily life of the person with dementia is deemed necessary, since caregivers and clinicians will need a comprehensive view of the person's daily activities, behavioral patterns, lifestyle, as well as changes in them, indicating the progression of their condition.

4.3.2. Ethical and Acceptability Issues

The development and ultimate use of novel assistive technologies by a vulnerable user group such as individuals with dementia, and the assessment methodologies planned by Stars are not free of ethical, or even legal concerns, even if many studies have shown how these Information and Communication Technologies (ICT) can be useful and well accepted by older people with or without impairments. Thus one goal of Stars team is to design the right technologies that can provide the appropriate information to the medical carers while preserving people privacy. Moreover, Stars will pay particular attention to ethical, acceptability, legal and privacy concerns that may arise, addressing them in a professional way following the corresponding established EU and national laws and regulations, especially when outside France. Now, Stars can benefit from the support of the COERLE (Comité Opérationnel d'Evaluation des Risques Légaux et Ethiques) to help it to respect ethical policies in its applications.

As presented in 3.1, Stars aims at designing cognitive vision systems with perceptual capabilities to monitor efficiently people activities. As a matter of fact, vision sensors can be seen as intrusive ones, even if no images are acquired or transmitted (only meta-data describing activities need to be collected). Therefore new communication paradigms and other sensors (e.g. accelerometers, RFID, and new sensors to come in the future) are also envisaged to provide the most appropriate services to the observed people, while preserving their privacy. To better understand ethical issues, Stars members are already involved in several ethical organizations. For instance, F. Brémond has been a member of the ODEGAM - "Commission Ethique et Droit" (a local association in Nice area for ethical issues related to older people) from 2010 to 2011 and a member of the French scientific council for the national seminar on "La maladie d'Alzheimer et les nouvelles technologies - Enjeux éthiques et questions de société" in 2011. This council has in particular proposed a chart and guidelines for conducting researches with dementia patients.

For addressing the acceptability issues, focus groups and HMI (Human Machine Interaction) experts, will be consulted on the most adequate range of mechanisms to interact and display information to older people.

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

Antitza Dantcheva has been awarded with the prestigious ANR Jeunes chercheuses / Jeunes chercheurs grant and is Principal Investigator of the project "ENVISION (see 9.2.1.3). Computer Vision for Automated Holistic Analysis of Humans" 2017–2020.

Antitza Dantcheva has received the Best Paper Award (Runner Up) at the 3rd IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2017) for the work "Spoofing Faces Using Makeup: An Investigative Study" , co-authored by Cunjian Chen, Thomas Swearingen and Arun Ross from the Michigan State University, USA.

BEST PAPER AWARD:

[26]
C. CHEN, A. DANTCHEVA, T. SWEARINGEN, A. ROSS. *Spoofing Faces Using Makeup: An Investigative Study*, in "IEEE International Conference on Identity, Security and Behavior Analysis 2017", New Delhi, India, February 2017, https://hal.archives-ouvertes.fr/hal-01430020

6. New Software and Platforms

6.1. EGMM-BGS

KEYWORD: 2D
FUNCTIONAL DESCRIPTION: This alogorithm allows to distinguish between the mobile pixels (except shadows) and pixels belonging to the background of the image.

- Participants: Anh Tuan Nghiem, François Brémond and Vasanth Bathrinarayanan
- Contact: François Brémond

6.2. MTS

Multi camera Tracking System
KEYWORD: Vision

FUNCTIONAL DESCRIPTION: This tool allows to find an appearence of interest in a following system with multi cameras.

- Participants: François Brémond and Slawomir Bak
- Contact: François Brémond

6.3. PALGate

KEYWORDS: Health - Home care - Handicap

- Contact: David Daney

6.4. PrintFoot Tracker

KEYWORD: Video analysis
FUNCTIONAL DESCRIPTION: Following of mobile object moving from single camera video streams.

- Participants: Duc Phu Chau, François Brémond and Monique Thonnat
- Contact: François Brémond

6.5. Proof Of Concept Néosensys (Poc-NS)

KEYWORD: Video analysis
FUNCTIONAL DESCRIPTION: This software is composed of 3 applications dedicated to show the techniques that will be applied by Néosensys Stars start-up. The software PoC-NS is a graphical interface allowing switching between these 3 applications. These applications are dedicated to help videosurveillance opertators in stores, in the fight against theft. There are the following: 1. Auto-side swith: allows to swith from a camera to another one by a single translation moving (left-riht) in a set of cameras in parallel. 2. Re-identification: Based on EGMM-BGS and PrintFoot Tracker software (both registed at APP), this application allows to find a person in several camera registrations, during a specific time, by clicking once on the person in a video. 3. Assisted following: allows (by hand) to follow a person in a camera network, with the feature of an automatic switch from a camera to another one when the person moves in a controlled area.

- Participants: Annunziato Polimeni, Bernard Boulay, François Brémond, Julien Gueytat, Slawomir Bak, Sofia Zaidenberg and Yves Pichon
- Partner: Neosensys
- Contact: François Brémond

6.6. py_ad

py action detection
FUNCTIONAL DESCRIPTION: Action Detection framework Allows user to detect action in video stream. It uses model trained in py_ar.

- Participants: François Brémond and Michal Koperski
- Contact: Michal Koperski

6.7. py_ar

py action recognition
FUNCTIONAL DESCRIPTION: Action Recognition training/evaluation framework. It allows user do define action recognition experiment (on clipped videos). Train, test model, save the results and print the statistics.

- Participants: François Brémond and Michal Koperski
- Contact: Michal Koperski

6.8. py_sup_reader

FUNCTIONAL DESCRIPTION: This is a library which allows to read video saved in SUP format in Python.

- Participant: Michal Koperski
- Contact: Michal Koperski

6.9. py_tra3d

py trajectories 3d
KEYWORD: Videos
SCIENTIFIC DESCRIPTION: New video descriptor which fuses trajectory information with 3D information from depth sensor.
FUNCTIONAL DESCRIPTION: 3D Trajectories descriptor. Compute 3D trajectories descriptor proposed in http://hal.inria.fr/docs/01/05/49/49/PDF/koperski-icip.pdf.

- Participants: François Brémond and Michal Koperski
- Contact: Michal Koperski

6.10. 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, Thanh 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

6.11. sup_ad

sup action detection
SCIENTIFIC DESCRIPTION: This software introduces the framework for online/real-time action recognition using state-of-the-art features and sliding window technique.
FUNCTIONAL DESCRIPTION: SUP Action Detection Plugin Plugin for SUP platform which performs action detection using sliding window and Bag of Words. It uses an input data model trained in py_ar project.

- Participants: François Brémond and Michal Koperski
- Contact: Michal Koperski

6.12. 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

6.13. bomotech

KEYWORDS: 3D - Video analysis - Kinect - 2D
FUNCTIONAL DESCRIPTION: Software dedicated to walking analysis using a Kinect deep camera.

- Authors: Melaine Gautier and Baptiste Fosty
- Partner: Mélaine Gautier
- Contact: Melaine Gautier

6.14. BMC_1

- Authors: Anaïs Ducoffe, Julien Badie, Manikandan Bakthavatchalam, Vasanth Bathrinarayanan, Anh Tuan Nghiem, Duc Phu Chau, Slawomir Bak, Ghada Bahloul and Nicolas Chleq
- Contact: François Brémond

6.15. CLEM

FUNCTIONAL DESCRIPTION: The Clem Toolkit is a set of tools devoted to design, simulate, verify and generate code for LE programs. LE is a synchronous language supporting a modular compilation. It also supports automata possibly designed with a dedicated graphical editor and implicit Mealy machine definition.

- Participants: Annie Ressouche and Daniel Gaffé
- Contact: Annie Ressouche
- URL: http://www-sop.inria.fr/teams/pulsar/projects/Clem/

6.16. Person Manual Tracking in a Static Camera Network (PMT-SCN)

- Participants: Anaïs Ducoffe, Annunziato Polimeni, Bernard Boulay, Julien Gueytat and Sofia Zaidenberg
- Partner: Neosensys
- Contact: Anaïs Ducoffe

6.17. sup_ad_ont

SUP Activity detection with ontologies
KEYWORD: Activity recognition
FUNCTIONAL DESCRIPTION: SUP plugin for activity detection, with manually defined ontologies.

- Participants: François Brémond, Michal Koperski and Dario Dotti
- Contact: Michal Koperski

7. New Results

7.1. Introduction

This year Stars has proposed new results related to its three main research axes : perception for activity recognition, semantic activity recognition and software engineering for activity recognition.

7.1.1. Perception for Activity Recognition

Participants: François Brémond, Etienne Corvée, Antitza Dancheva, Furqan Muhammad Khan, Michal Koperski, Thi Lan Anh Nguyen, Javier Ortiz, Remi Trichet, Ujjwal Ujjwal, Srijan Das, Monique Thonnat.

The new results for perception for activity recognition are:

- Pedestrian detection: Training set optimization (see 7.2)
- Pedestrian Detection Using Deep Learning (see 7.3)
- Deep Learning applied on Embedded Systems for people detection (see 7.4)
- Facial Analysis (see 7.5)
- Multi-Object Tracking using Multi-Channel Part Appearance Representation (see 7.6)
- Tracklets Pre-Processing for Signature Computation in the Context of Multi-Shot Person Re-Identification (see 7.7)
- Multi-shot Person Re-identification in surveillance videos (see 7.8)
- Person Re-Identification using Pose-Driven Body Parts (see 7.9)
- Human Action Recognition in Videos with Local Representation (see 7.10)
- Action Detection in Untrimmed Videos (see 7.11)
- RGB-D based Action Recognition using CNNsf (see 7.12)
- Recognizing Human Actions Using RGB Sport Videos From the Web (see 7.13)

7.1.2. Semantic Activity Recognition

Participants: Carlos Fernando Crispim Junior, Kartik Kartik, Farhood Negin, Thanh Hung Nguyen, Kuan-Ru Lee, Antitza Dantcheva, Auriane Gros, Alexandra Koening, Guillaume Sacco, Philippe Robert, François Brémond, Monique Thonnat.

For this research axis, the contributions are :

- Event Recognition Based on Depth Image (see 7.14)
- Recognition of Daily Activities by Embedding Visual Features within a Semantic Language (see 7.15)
- Cognitive Assessment Using Gesture Recognition (see 7.16)
- Geometric and Visual Features Fusion for Action Recognition (see 7.17)
- Probabilistic Logic for Activity Recognition (see 7.18)
- Recognizing Retracing of Steps Using Walk Comparison (see 7.19)
- Safe & Easy Environment for Alzheimer Disease and related disorders (see 7.20)
- Early detection of cognitive disorders such as dementia on the basis of speech analysis ELEMENT (see 7.21)
- Serious Exergames for Cognitive Stimulation (see 7.22)

7.1.3. Software Engineering for Activity Recognition

Participants: Sabine Moisan, Annie Ressouche, Jean-Paul Rigault, Ines Sarray, Thanh Hung Nguyen, Daniel Gaffé, Julien Badie, Anais Ducoffe, Dorine Havyarimana, Cedric Girard-Riboulleau, François Brémond, Minh Khue Phan Tran, Philippe Robert.

The contributions for this research axis are:

- Defining an activity description language for end-users and its semantics (see 7.23)
- The Clem Workflow (see 7.24)
- Study of Temporal Properties of Neuronal Archetypes (see 7.25)
- Maintaining the engagement of older adults with dementia while interacting with serious game(see 7.26)
- Application of deep learning on healthcare (see 7.27)
- Brick & Mortar Cookies (see 7.28)

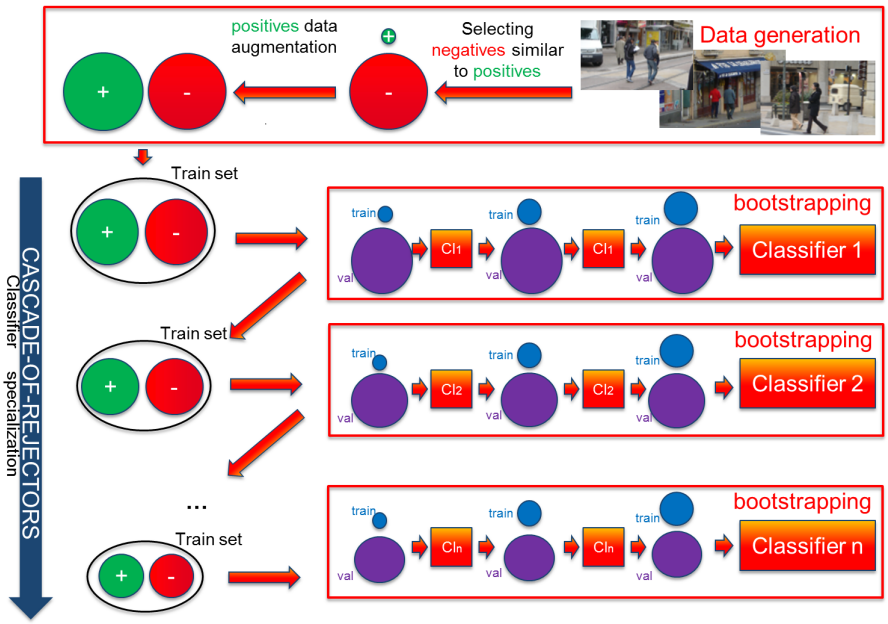


Figure 5. Training pipeline. The initial training set generation selects data while balancing negative and positive sample cardinalities. A cascade of classifiers is then trained on it, each independent classifier being learnt through bootstrapping. balanced positive and negative sets is sought all along the cascade. Each circle surface is proportional to the set's cardinality that it represents.

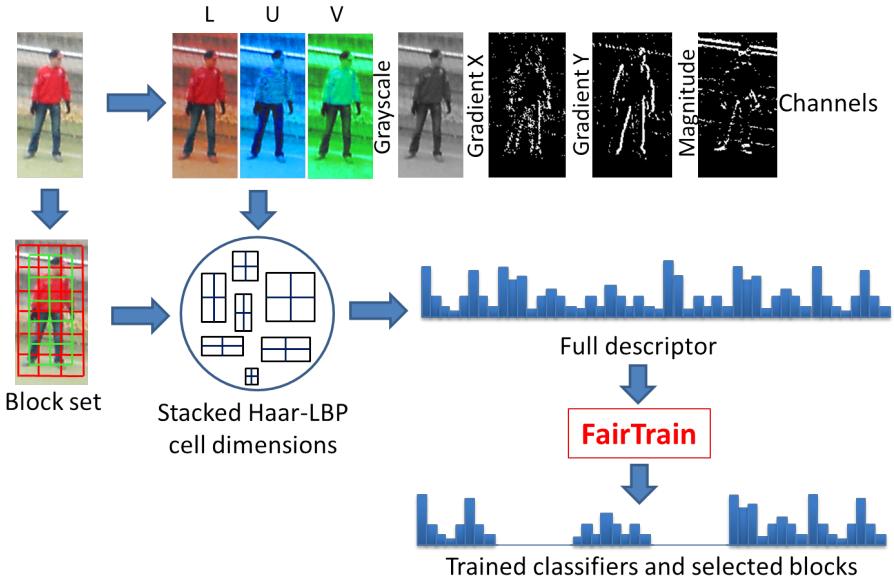


Figure 6. LBP Channel features pipeline.

7.2. Pedestrian Detection: Training Set Optimization

Participants: Remi Trichet, Javier Ortiz.

keywords: computer vision, pedestrian detection, classifier training, data selection, data generation, data weighting, feature extraction

This year's work builds on the near real-time pedestrian detector introduced last year. Let's recall that this detector novelty mainly focusses on our training set generation protocol, named *FairTrain* [39]. The methodology, illustrated in figure 5, decomposes in two distinct parts: The initial training set generation and the classifier training. The initial training set generation carefully selects data from a set of images while balancing negative and positive sample cardinalities. We then train a cascade of 1 to n classifiers. This cascade could consist of a cascade-of-rejectors [57], [143], [48], [118], [122], a soft cascade [99], or both. In addition, each independent classifier is learnt through bootstrapping [59], [69] to improve performance. One key aspect is to seek balanced positive and negative sets at all time. Hence, all along the cascade, the minority class is oversampled to create balanced positive and negative sets. See [39] for details.

This year's improvement on this framework is two-fold: refined experimentation and Local Binary Pattern (LBP) channel descriptor.

In many aspects, the construction of a training set remains similar to what it was at the birth of the domain, some related problems are not well studied, and sometimes still tackled empirically. This work studies the pedestrian classifier training conditions. More than a survey of existing training techniques, our experimentation highlights impactful parameters, potential new research directions, and combination dilemmas. They allowed us to better understand and parametrized our pipeline. Second, we introduce a 12-valued filter representation based on LBP. Indeed, various improvements now allow for this texture feature to provide a very discriminative, yet compact descriptor. This new LBP-based channel descriptor outperforms channel features [65] while requiring a fraction of the original LBP memory footprint. Uniform patterns [100] and Haar-based LBP [56] are employed to shrink the filter dimension in accordance to our needs. Also, cell stacking and new filter combination restriction based on proposal window coverage are successfully

applied. Finally, a more reliable feature selection technique is introduced to construct a lower dimension final descriptor without harming its discriminability. Experiments on the Inria and Caltech-USA datasets, respectively presented in tables 1 and 2 validate these progresses.
In the light of these results, combining the *FairTrain* data selection pipeline with CNN features appears like the obvious next step.

Table 1. Comparison with the state-of-the-art on the Inria dataset. Near real-time methods are separated from others. Ours is in bold. Deep learning techniques are in red. Computation times (CPU/GPU) are calculated according to 640×480 resolution frames. The used metric is the log-average miss rate (the lower the better). Best viewed in color.

Evaluation method	Log-average miss rate	Speed(CPU/*GPU*)
HoG [59]	46%	0.5fps
HoG-LBP [127]	39%	Not provided
MultiFeatures [129]	36%	$<$ 1fps
FeatSynth [45]	31%	$<$ 1fps
MultiFeatures+CSS [123]	25%	No
Channel Features [65]	21%	0.5fps
FPDW [64]	21%	2-5fps
DPM [70]	20%	$<$ 1fps
RF local experts [95]	15.4%	3fps
PCA-CNN [81]	14.24%	$<$ 0.1fps
CrossTalk cascades [66]	18.98%	30-60fps
VeryFast [46]	18%	8/*135fps*
WordChannels [57]	17%	0.5/*8fps*
SSD [92]	15%	*56fps*
LBP-Channels full	**14.3%**	**0.5/ *7.5fps***
LBP-Channels selected	**13.6%**	**0.7/ *10fps***
FRCNN [110]	13%	*7fps*
RPN+PF [140]	7%	*6fps*

7.3. Detection of Pedestrians Using Deep Learning

Participants: Ujwal Ujwal, Frederic Precioso, Nagi Aly, François Brémond.

keywords: Deep learning, CNN

7.3.1. Introduction

The problem of pedestrian detection shares many important characteristics with the scenario of general object detection, its applications have much more practical and widespread ramifications. This includes areas such as surveillance, monitoring and autonomous vehicles. Traditional approaches such as HoG-based detection [59], [60] and Deformable Parts Model(DPM) [71], [102] based detection have been reasonably successful. However in the wake of recent interests in autonomous vehicles where the need of safety is utmost, it is pertinent to expect a much higher degree of performance from pedestrian detection systems.

The advent and popularity of deep learning beckons us to investigate it in search for such a high-performance system. Deep learning has been very successful in object detection problems of a more general taste as reflected by a large number of very successful systems. In our work, we focus upon investigating deep learning for designing high-performance pedestrian detection systems.

This work has been done in collaboration with Bertrand Leroy (VEDECOM)

Table 2. Comparison with the state-of-the-art on the Caltech dataset. Near real-time methods are separated from others. Ours is in bold. Deep learning techniques are in red. Computation times (CPU/GPU) are calculated according to 640×480 resolution frames. The used metric is the log-average miss rate (the lower the better). Best viewed in color.

Evaluation method	Log-average miss rate	Speed(CPU/*GPU*)
HoG [59]	69%	0.5fps
DPM [70]	63.26%	$<$ 1fps
FeatSynth [45]	60.16%	$<$ 1fps
MultiFeatures+CSS [123]	60.89%	No
FPDW [64]	57.4%	2-5fps
Channel Features [64]	56.34%	0.5fps
Roerei [46]	48.35%	1 fps
MOCO [54]	45.5%	$<$ 1fps
JointDeep [103]	39.32%	$<$ 1fps
SquaresChnFtrs [47]	34.8%	$<$ 1fps
InformedHaar [137]	34.6%	$<$ 0.63fps
Spatial pooling [104]	29.2%	$<$ 1fps
Checkboards [138]	24.4%	$<$ 1fps
FRCNN [110]	56%	*7fps*
CrossTalk cascades [66]	53.88%	30-60fps
WordChannels [57]	42.3%	0.5/*8fps*
LBP-Channels full	**39.1%**	**0.5/ *7.5fps***
LBP-Channels selected	**35.9%**	**0.7/ *10fps***
SSD [92]	34%	*56fps*
RPN+PF [140]	10%	*6fps*

7.3.2. State-of-the-art investigations

This year, we continued our investigations into the state-of-the-art deep learning based systems which have been proposed or have been applied to pedestrian detection. Deep learning has yet been without much theoretical foundations. The vastly practical and experimental playground of deep learning makes it very important to investigate existing systems [75], [92], [140] through thorough experiments in order to better comprehend their behavior in a vast variety of scenarios where pedestrian detection might be desired. Our investigations offered us insights such as the following :

1. **Performance limitations of current systems**: We were able to conclude a number of important scenarios where present state-of-art systems stutter in their detection performance. This primarily includes the following instances :
   - *Small-scale People*: This refers to people who are far away from the camera; thus appearing small in size. We also refer to such instances as far-range people, who are often missed.
   - *Occluded People*: People in urban environments are often occluded or semi-occluded by various entities such as lamp-posts and other vehicles to name a few.
   - *Seated People*: Very often people who are either in a sitting position or riding a vehicle are often missed. This effect is much more pronounced in coupling with the previously mentioned case of small-scale people.
2. **Suboptimal usage of CNN architectures**: Convolution Neural Network (CNN) architectures are the backbone of deep learning based object detection systems. CNNs are hierarchical layers of neurons (e.g - as in multi-layer perceptrons (MLP)) albeit with more involved operations. We observe that the lower layers of a CNN are only implicitly utilized by extracting features from the last convolutional layer. We consider this to be suboptimal owing to our observations during our experiments that lower layers of a CNN indeed detect some important features which may prove useful with respect to scenarios such as small-scale people and occluded people.

Outcome

Our investigations have enabled us to focus upon some important aspects of our problem and have thus narrowed our focus. This allows us to focus upon relevant portions of system design. We expect this to induce more productivity in our future work.
Following these state-of-art studies we plan to coalesce our findings in a review paper which we aim to submit shortly to a journal.

7.3.3. Detection of small-scale people

As mentioned before, our state-of-the-art studies enabled us to identify that CNN architectures might being used in a suboptimal way. To take this investigation further, we worked upon the design of a better system which can make use of all the hierarchies of a CNN. We are correcting some implementation issues with the aforementioned system, although in our first experiments it did provide us with a miss-rate of 13.98% as against the state-of-art miss-rate of 9%. Miss-rate refers to the number of pedestrian instances which were not detected (thus *false negative*). Hence a lower miss-rate gestures at a better performing system. In our first experiments although we have a miss-rate roughly 5% worse than the state-of-art, but we find it encouraging given that in our experiments we used a much smaller CNN. A smaller CNN gestures at a lower capacity for feature extraction. We believe that by employing a better-performing CNN, much better results may be warranted.

Outcome

Our work in this problem is currently moving ahead of our first experiments where we demonstrated the validity of our conclusions that suboptimal usage of CNN architectures might be a possibility in existing systems. We are currently focused upon our second set of experiments which involve employing a better CNN and conducting more exhaustive investigations into the system performance and behavior.

7.4. Deep Learning applied on Embedded Systems for people detection

Participants: Juan Diego Gonzales Zuniga, Ujjwal Ujjwal, François Brémond.

keywords: Deep learning, CNN, Embedded Systems

7.4.1. Introduction

One of the problems with people detection is the amount of resources it takes for quality results. Most architectures either require big memory or large computing time to achieve a state-of-the-art position, these results are mostly achieved with dedicated hardware at data centers. The applications for an embedded hardware with these capabilities are limitless: automotive, security and surveillance, augmented reality and healthcare just to name a few. But the state-of-the-art architectures are mostly focused on accuracy than resources consumption [74] [75] [140].

The popularity of deep learning invites us to explore high-performance algorithms. In our work, we have to consider improving the systems' accuracy and reducing resources for a real-time application on people detection. This will lead towards new and efficient deep learning solutions.

7.4.2. State-of-the-art investigations

Deep learning lacks a strong theoretical background and a significant part of the knowledge is by investigating existing systems [75] [92] [140]. In order to better grasp the behavior in a different range of scenarios, we started our investigation to comprehend the nature of deep learning by diving into architectures that multi-task different activities. The combination of detection and segmentation shed light on the mutual improvements for people detection as seen in [67] and [50].

Another key part of our investigation was also to experiment with low time consuming architectures such as [109], an architecture that takes less time than [75] [92] [140] but still competitive and fairly flexible.

Our investigations offered us insights such as the following :

1. **Performance limitations of current systems**: We were able to conclude a number of important scenarios where present state-of-the-art systems stutter in their detection performance. This primarily includes the following instances :
   - *Loss function*: This refers to the feedback that will be reinserted for training. Different and more complex loss functions have different results. In other words, it is not the quantity of samples to train but more so the way they are trained.
   - *Usage of filters*: It is commonly used among deep learning architectures to have a small filter size, this improves the field view of an image but also increases the number of parameters to control.
   - *Time-Computation*: Most architectures with high-performance double the work to refine their precedent results, the accuracy of the solution is undeniable but the cost of computation and the memory resources also get affected.
2. **Suboptimal usage of CNN architectures**: (see section 7.3).

7.4.3. Outcome

Our investigations have showed us to focus more on the quality of training and not so much on the quantity. This allows us to focus upon relevant portions of system design. We expect this to give us clues on how to increase accuracy without compromising the resources.
Following these state-of-the-art studies we plan to coalesce our findings in a review paper which we aim to submit to a journal shortly.

This work has been done in collaboration with Serge Tissot (Kontron).

7.5. Facial Analysis

Participants: Antitza Dantcheva, Hung Thanh Nguyen, Philippe Robert, François Brémond.

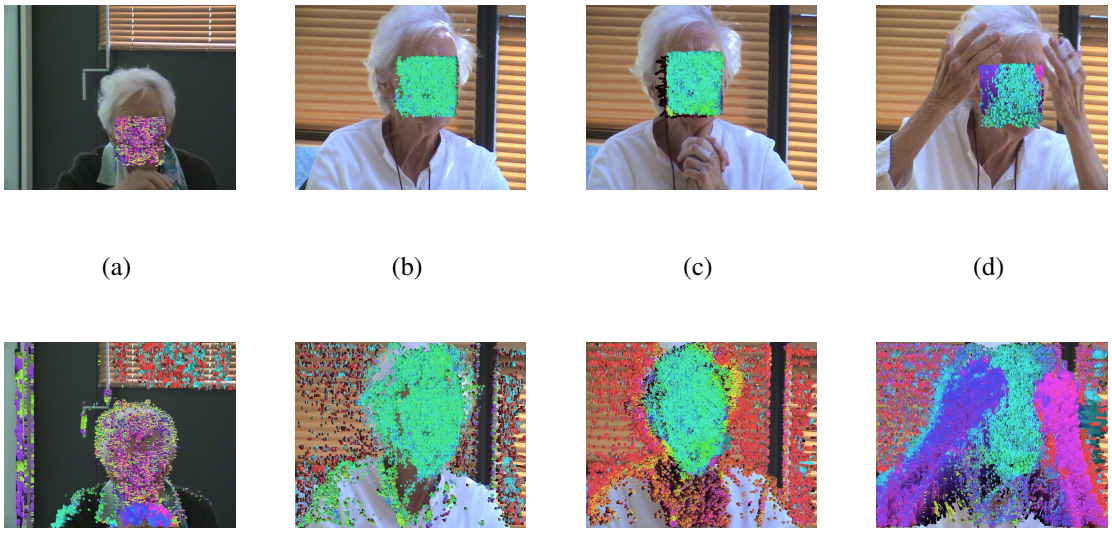
keywords: automated healthcare, healthcare monitoring, expression recognition, gender estimation, soft biometrics, biometrics, visual attributes

7.5.1. Automated Healthcare: Facial-expression-analysis for Alzheimer's patients in musical mnemotherapy

This work was done in collaboration with Piotr Bilinski (Univ. Oxford, UK), Jean-Claude Broutart (GSF Noisiez, France), Arun Ross (MSU, USA), Cunjian Chen (MSU, USA), Thomas Swearingen (MSU, USA), Ester Gonzalez-Sosa (UAM, Spain) and Julian Fierrez (UAM, Spain), Ruben Vera-Rodriguez (UAM, Spain), Jean-Luc Dugelay (Eurecom, France)

Recognizing expressions in patients with major neurocognitive disorders and specifically Alzheimer's disease (AD) 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 the 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. In this work, we present a novel approach for automated recognition of facial activities and expressions of severely demented patients, where we distinguish between four activity and expression states, namely *talking*, *singing*, *neutral* and *smiling*. Our approach caters to the challenging setting of current medical recordings of music-therapy sessions, which include continuous pose variations, occlusions, camera-movements, camera-artifacts, as well as changing illumination. An additional important challenge that we tackle has to do with the fact that the (elderly) patients exhibit generally less profound facial activities and expressions, which furthermore occur in combinations (e.g., talking and smiling).

Our proposed approach is based on the extension of the Improved Fisher Vectors (IFV) for videos, representing a video-sequence using both, local, as well as the related spatio-temporal features. We test our algorithm on a dataset of over 229 video sequences, acquired from 10 AD patients. We obtain the best results in *personalized facial expression and activity recognition*, where we train the proposed algorithm on video sequences related to each patient, individually. The results are promising and they have sparked substantial interest in the medical community. We believe that the proposed approach can play a key part in assessment of different therapy treatments, as well as in remote large-scale healthcare-frameworks.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 7. Example images of facial dense trajectories in video sequences related to the facial expressions and activities (a) neutral, (b) smile, (c) talking and (d) singing of a patient in the acquired dataset of AD patients.

Facial expression and activity recognition [37]

We report the average MCA in Table 3 of the proposed algorithm and 2 further variations thereof. Specifically, we investigate as a first variation (a) the performance without face detection. The rational is that head and hands movement might contain potentially useful information in expression recognition (see Fig. 7). However, we observe that, due to a vast amount of camera-artifacts (*i.e.*, the static background containing a seemingly considerable amount of motion), the analysis of the full-frames reduces the recognition accuracy. Further, we report (b) the performance of the original IFV scheme. Our proposed algorithm significantly outperforms the original IFV-scheme from 50.83% to 53.99% (without face detection), and from 58.4% to 63.37% (with face detection). Additionally, we note that face detection significantly improves the performance of our proposed algorithm, namely from 53.99% to 63.37%.

Table 3. Average Mean Class Accuracy (MCA).

IFV, no face detection	50.83%
IFV, face detection	58.4%
Spatio-temporal IFV, no face detection	53.99%
Spatio-temporal IFV, face detection	63.37%

7.5.2. Can a smile reveal your gender?

Automated gender estimation has numerous applications including video surveillance, human computer-interaction, anonymous customized advertisement and image retrieval. Most commonly, the underlying algorithms analyze facial appearance for clues of gender.

Deviating from such algorithms in [61] we proposed a novel method for gender estimation, exploiting dynamic features gleaned from smiles and show that (a) facial dynamics incorporate gender clues, and (b) that while for adults appearance features are more accurate than dynamic features, for subjects under 18 years old facial dynamics outperform appearance features. While it is known that sexual dimorphism concerning facial appearance is not pronounced in infants and teenagers, it is interesting to see that facial dynamics provide already related clues.

The obtained results suggest that smile-dynamic include pertinent and complementary to appearance gender information. Such an approach is instrumental in cases of (a) omitted appearance-information (*e.g.* low resolution due to poor acquisition), (b) gender spoofing (*e.g.* makeup-based face alteration), as well as can be utilized to (c) improve the performance of appearance-based algorithms, since it provides complementary information.

7.5.3. Vulnerabilities of Facial Recognition Systems

Makeup can be used to alter the facial appearance of a person. Previous studies have established the potential of using makeup to obfuscate the identity of an individual with respect to an automated face matcher. We analyzed [26] the potential of using makeup for spoofing an identity, where an individual attempts to impersonate another person's facial appearance (see Fig. 8). In this regard, we first assembled a set of face images downloaded from the internet where individuals use facial cosmetics to impersonate celebrities. We next determined the impact of this alteration on two different face matchers. Experiments suggest that automated face matchers are vulnerable to makeup-induced spoofing and that the success of spoofing is impacted by the appearance of the impersonator's face and the target face being spoofed (see Fig. 9). Further, an identification experiment was conducted to show that the spoofed faces are successfully matched at better ranks after the application of makeup. To the best of our knowledge, this was the first work that systematically studied the impact of makeup-induced face spoofing on automated face recognition.

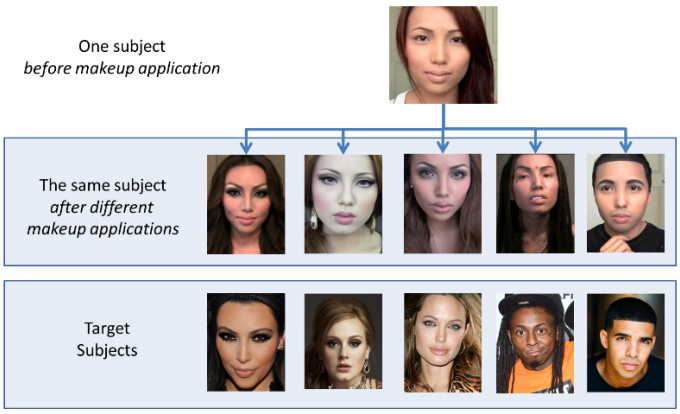
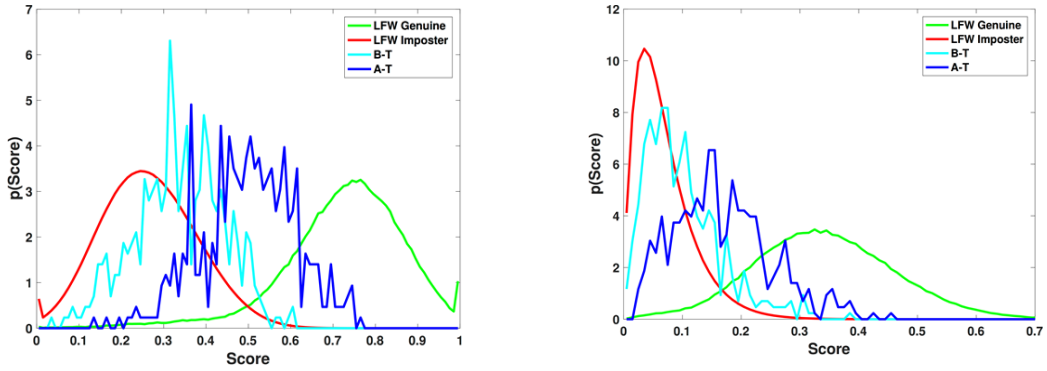


Figure 8. The subject on the top attempts to resemble identities in the bottom row (labeled "Target Subjects") through the use of makeup. The result of these attempts can be seen in the top row. Dataset available under http://antitza.com/makeup-datasets.html.



(a) (b)

Figure 9. Normalized histogram of similarity scores from the B-T subset (the image before makeup is matched against the target image), A-T subset (the image after makeup is matched against the target image), and the LFW dataset for the (a) COTS and (b) VGG face matchers.

7.6. Multi-Object Tracking using Multi-Channel Part Appearance Representation

Participants: Thi Lan Anh Nguyen, Furqan Muhammad Khan, Farhood Negin, François Brémond.

Keywords: Tracklet fusion, Multi-object tracking, Appearance Representation

Multi-object tracking (MOT) has been one of the fundamental problems in computer vision, essential for lots of applications (e.g home-care, house-care, security systems, etc.). The main objective of MOT is to estimate the states of multiple objects while identifying these objects under appearance and motion variation in time. This problem becomes very challenging due to frequent occlusion by background or other objects, object pose as well as illumination variation, etc.

Depending on the time of data association process, tracking algorithms can be categorized into 2 types: short-term and long-term tracking. Short-term trackers [108], [115] associate object detections in current frame with the most matching object trajectories in the past. These methods are able to perform online processing based on frame-to-frame association and therefore, could be applied in real-time applications. In general, short-term trackers use bipartite matching methods for short-term data association where Hungarian algorithm is the most popular method. Although these methods are computationally inexpensive, object identification could fail due to inaccurate detections (false alarms) and only short-term occlusions can be handled. Long-term trackers [139], [113] can overcome the shortcomings of short-term trackers by extension of the bipartite matching into network flow. The direct acyclic graph in [139] was formed where vertices are object detections or short tracklets and edges are the similarity links between vertices. In [113], the track of a person forms a clique and MOT is formulated as constraint maximum weight clique graph. The data association solutions for these long-term trackers are found through minimum-cost flow algorithm. However, long-term tracking methods also have their obvious drawbacks, such as: their huge computational cost due to iterative association process to generate globally optimized tracks and their pre-requirement for entire object detection in a given video.

Recently, some proposed trackers tried to combine both short-term and long-term tracking methods in a framework to perform online object tracking. The MOT methods in [44], [121] use a frame-to-frame association to generate tracklets followed by a tracklet association process with a time buffer latency. However, their performance is limited by their object features and tracklet representation. These methods utilize basic features (e.g. 2D information, color histogram or constant velocity) applied on whole body parts and use normal Gaussian distribution to describe the object. This way of representation could lose important

information to discriminate objects and consequently, could fail to track objects in complex scene conditions (such as occlusion, low video resolution or insufficient lighting of environment).

On the other hand, multiple-shot person re-identification methods [89], [136] [31] gained high performances in matching objects from different camera views. In order to match a given person in a camera to the closest person in a gallery in another camera, these re-identification methods use efficient features and object representations. These methods are adapted to solve problems that involve pose and camera view setting variation. Since person Re-identification usually deals with identification of a person from different camera views, it is expected that using Re-id representation becomes even more effective in single-view multi-object tracking problem.

Therefore, we propose a robust online multi-object tracking method named MTSTracker which extends object representation and methods proposed for re-identification domain to address problems in MOT. While the re-identification works in offline mode, MTSTracker works in online mode. This method uses a time-window buffer to extract tracklets and associates tracklets in each time-window by using Re-identification techniques. MTSTracker integrates a short-term and long-term trackers in a comprehensive framework. The short-term tracker generates object trajectories called tracklets. Object features are computed for full and body parts, then, each tracklet is represented by a set of multi-modal feature distribution modeled by GMMs. The long-term tracker associates tracklets after mis-detections or occlusions based on learning Mahalanobis distance between GMM components. In order to learn this metric, KISSME [84] algorithm is adopted to learn feature transformations between different scenes by directly learning transformation between probability distributions. Experiments on two public datasets MOT2015 and ParkingLot show that MTSTracker performs well when compared to state-of-the-art tracking algorithms. This contribution has been published in the international conference AVSS 2017 [35].

7.7. Tracklets Pre-Processing for Signature Computation in the Context of Multi-Shot Person Re-Identification

Participants: Salwa Baabou, Furqan Muhammad Khan, Thi Lan Anh Nguyen, Thanh Hung Nguyen, François Brémond.

keywords: Person Re-Identification (Re-ID), tracklet, signature representation.

7.7.1. Tracklets pre-processing/representation

The person Re-Identification (Re-ID) system is divided into two steps: *i)* constructing a person's appearance signature by extracting feature representations which should be robust against pose variations, illumination changes and occlusions and *ii)* Establishing the correspondence/matching between feature representations of probe and gallery by learning similarity metrics or ranking functions (see Figure 10). However, appearance based person Re-ID is a challenging task due to disparities of human bodies and visual ambiguities across different cameras. Therefore, we focus on how to pre-process tracklets to compute the signature and represent it for Multi-shot person Re-ID to handle high appearance variance and occlusions.

We have worked on CHU Nice dataset. First, we used the SSD detector [91] to get the detection results. Second, inspired by the Multi-Object Tracking in [35], we extracted the tracklets of persons and then we used these tracklets to compute the signatures of individuals based on Part Appearance Mixture PAM approach [31].

Figure 11 shows a visualizaton of a sample from CHU Nice dataset of the detection results.

Figure 12 presents a visualization of a sample of consecutive frames from CHU Nice dataset of the tracking results.

Figure 13 shows a representation of some samples of tracklets of a person from CHU Nice dataset.

7.7.2. Experimental results

CHU Nice dataset

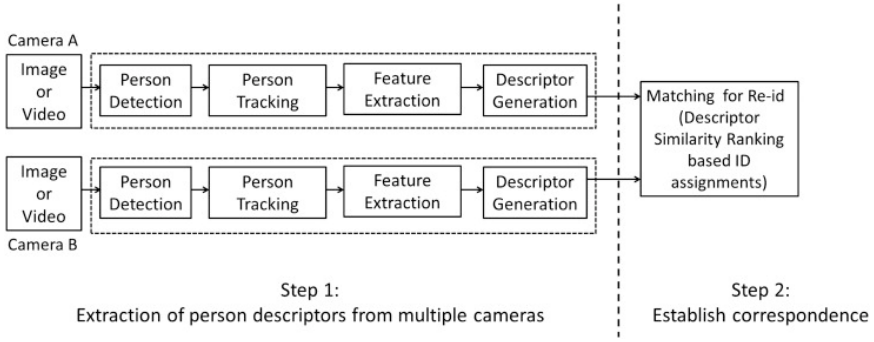
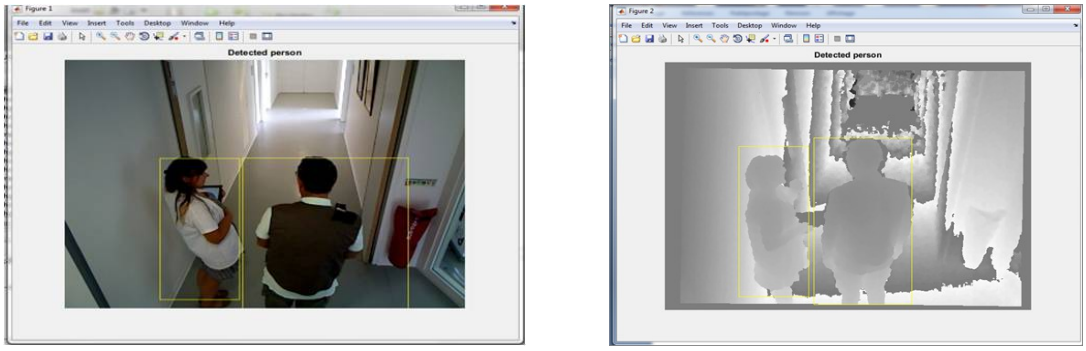


Figure 10. Re-ID Diagram



(a) (b)

Figure 11. A visualization of a sample from CHU Nice dataset of the Detection Results

(a) (b) (c) (d) (e) (f)

Figure 12. A visualization of a sample of frames from CHU Nice dataset of the Tracking Results

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure 13. A sample of tracklets representation from CHU Nice Dataset

Table 4. Top ranked Recognition rates (%) on CHU Nice dataset

Methods	Rank-1	Rank-5	Rank-10	Rank-20
LOMO+XQDA	30.7	64.6	80.3	90.3
PAM-LOMO+XQDA	38.5	69.2	84.6	100.0
PAM-LOMO+KISSME	81.8	90.9	100	100

It is an RGB-D dataset (RGB+Depth) and Collected in the hospital of Nice (CHU) in Nice, France. Most of the people recruited for this dataset were elderly people, aged 65 and above, of both genders. It contains 615 videos with 149365 frames acquired from 2 cameras: one in the corridor and another camera in the room.

Table 4 shows the recognition rate (%) at different ranks (rank-1, 5, 10, 20) of a baseline method LOMO+XQDA [89], PAM-LOMO+XQDA and PAM-LOMO+KISSME on CHU Nice dataset. From the above experiments, we notice that PAM-LOMO+KISSME achieves good performance on three datasets; 81.8% rank-1 recognition rate. This shows that our adaptation of feature descriptor LOMO [89] and metric learning KISSME [84] to PAM representation is effective.

Limitations

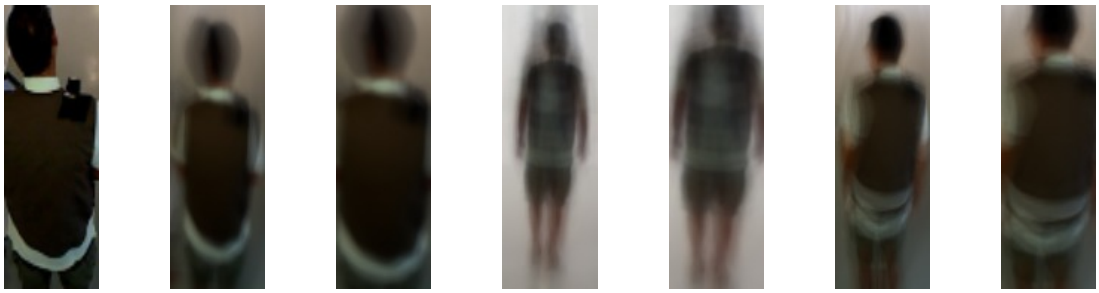
As shown in Figure 14, a visualization of a selected samples from CHU Nice dataset of PAM signature representation is presented. Indeed, we visualize each GMM component by constructing a composite image. Given appearance descriptor, we compute the likelihood of an image belonging to a model component and then by summing images of corresponding person weighted by their likelihood we generate the composite image. We can say that our signature representation is able to cater variance in person's pose and orientation as well as illumination, it deals also with occlusions and is able to reduce effect of background. However, we can notice that this PAM signature representation presents some limitations, specially on our own dataset CHU Nice, which can affect the quality of our signature representation (see Figure 5). Among these challenging problems, we can cite:

- Bad detection
- Number of frames by pose
- Number of GMM components not adequate with the number of person's pose/orientation and depends of the low-level features used.

We are actually working to improve the PAM signature representation by using the skeleton and extracting the pose machines from our dataset CHU Nice.

7.8. Multi-shot Person Re-identification in surveillance videos

Participants: Furqan Khan, Seongro Yoon, François Brémond.



(a) (b) (c) (d) (e) (f) (g)

Figure 14. A visualization of selected samples of signature representation from CHU Nice Dataset: the first image is the input image used to learn appearance model. It's followed by the composite images, one for each component of the GMM mixture. Optimal number of GMM components for each appearance model varies between persons. GMM components focus on different pose and orientation of the person.

keywords: person re-identification, appearance modeling, long term visual tracking

7.8.1. Efficient Video Summarization Using Principal Person Appearance for Video-Based Person Re-Identification

In video-based person re-identification, while most work has focused on problems of person signature representation and matching between different cameras, intra-sample variance is also a critical issue to be addressed. There are various factors that cause the intra-sample variance such as detection/tracking inconsistency, motion change and background. However, finding individual solutions for each factor is difficult and complicated. To deal with the problem collectively, we assume that it is more effective to represent a video with signatures based on a few of the most stable and representative features rather than extract from all video frames. In this work, we propose an efficient approach to summarize a video into a few of discriminative features given those challenges. Primarily, our algorithm learns principal person appearance over an entire video sequence, based on low-rank matrix recovery method. We design the optimizer considering temporal continuity of the person appearance as a constraint on the low-rank based manner. In addition, we introduce a simple but efficient method to represent a video as groups of similar frames using recovered principal appearance. Experimental results (Table 5) show that our algorithm combined with conventional matching methods outperforms state-of-the-art on publicly available datasets PRID2011 [77] and iLIDS-VID [125].

In order to get a deeper insight, Figure 15 presents some qualitative results visualizing principal appearance groups discovered by our approach without manual supervision. We compare our results with the approach of Shu *et.al.* [116] and found our results to be relatively more visually coherent and to have more groups. Details of this work can be found in our BMVC paper [36].

7.8.2. Multi-shot Person Re-identification using Part Appearance Mixture

Appearance based person re-identification in real-world video surveillance systems is a challenging problem for many reasons, including ineptness of existing low level features under significant viewpoint, illumination, or camera characteristic changes to robustly describe a person's appearance. One approach to handle appearance variability is to learn similarity metrics or ranking functions to implicitly model appearance transformation between cameras for each camera pair, or group, in the system. The alternative, that is followed in this work, is the more fundamental approach of improving appearance descriptors, called *signatures*, to cater for high appearance variance and occlusions. A novel signature representation for *multi-shot* person re-identification, called *Part Appearance Mixture* (PAM), is henceforth presented that uses multiple appearance models, each describing appearance as a probability distribution of a low-level feature for a certain portion of

Table 5. Comparison of recognition rates (%) at different ranks of various Re-ID methods on PRID and iLIDS-VID. Best results are highlighted in bold.

	PRID				**iLIDS-VID**			
Method	**r=1**	**r=5**	**r=10**	**r=20**	**r=1**	**r=5**	**r=10**	**r=20**
HOG3D+RankSVM [125]	19.4	44.9	59.3	77.2	12.1	29.3	41.5	56.3
Color+RankSVM [125]	29.7	49.4	59.3	71.1	16.4	37.3	48.5	62.6
DVR [125]	28.9	55.3	65.5	82.8	23.3	42.4	55.3	68.6
ColorLBP [78]+RankSVM	34.3	56.0	65.5	77.3	23.2	44.2	54.1	68.8
DVDL [80]	40.6	69.7	77.8	85.6	25.9	48.2	57.3	68.9
Color+LFDA [106]	43.0	73.1	82.9	90.3	28.0	55.3	70.6	88.0
AFDA [88]	43.0	72.7	84.6	91.9	37.5	62.7	73.0	81.8
DSVR [126]	40.0	71.1	84.5	92.2	39.5	61.1	71.7	81.0
MTL-LORAE [119]	-	-	-	-	43.0	60.1	70.3	85.3
STFV3D+KISSME [93]	64.1	87.3	89.9	92.0	43.8	69.3	80.0	90.0
CNN+KISSME [142]	69.9	90.6	-	98.2	48.8	75.6	-	92.6
RFA-Net+RankSVM [131]	58.2	85.8	93.4	97.9	49.3	76.8	85.3	90.0
CNN+XQDA [142]	77.3	93.5	-	99.3	53.0	81.4	-	95.1
APR+XQDA [73]	68.6	94.6	97.4	98.9	55.0	87.5	93.8	97.2
TDL [133]	56.7	80.0	87.6	93.6	56.3	87.6	95.6	98.3
RCNN [96]	70.0	90.0	95.0	97.0	58.0	84.0	91.0	96.0
PPA+Euclidean	66.6	90.1	93.5	96.7	29.6	55.7	67.6	79.7
PPA+KISSME	85.7	98.9	**99.9**	**100.0**	65.7	92.3	96.8	99.1
PPA+XQDA	**87.6**	**99.2**	99.6	99.9	**66.8**	**93.9**	**97.8**	**99.8**

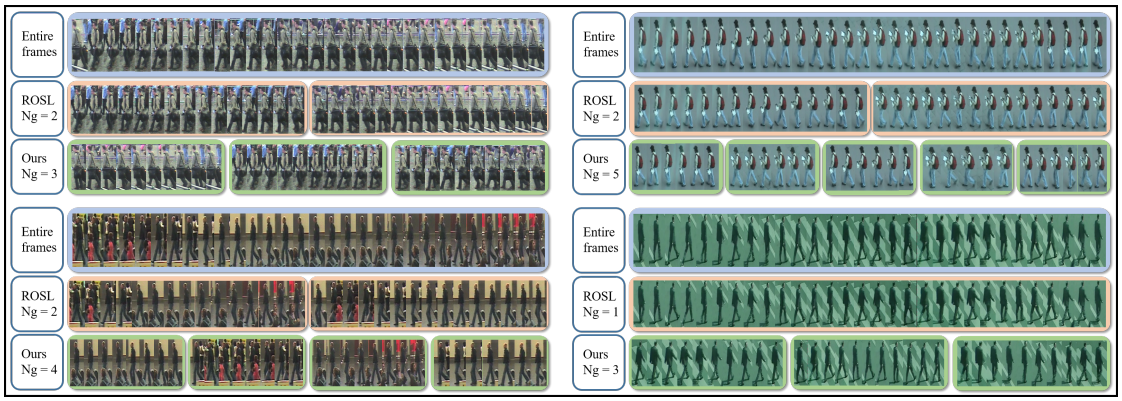


Figure 15. Visualization of principal appearance groups. Examples of our algorithm result are shown in comparison with ROSL [116] for the same process. The top and bottom on the left column are ID:218 and ID:016 of iLIDS-VID [125], and the top and bottom on the right column are ID:001 and ID:185 of PRID [77], respectively. N_g means the number of image groups.

an individual's body. It caters for high variance in a person's appearance by automatically trading compactness with variability as can be visually seen in the results presented in Figure 16.

Signature representation has probabilistic interpretation of appearance signatures that allows for application of information theoretic similarity measures. A signature is acquired over coarsely localized body regions of a person in a computationally efficient manner instead of reliance on fine parts localization. We also define a Mahalanobis based distance measure to compute similarity between two signatures. The metric is also amenable to existing metric learning methods and appearance transformation between different scenes can be learned directly using proposed signature representation. Combined with metric learning, rank-1 recognition rates of 92.5% and 79.5% are achieved on PRID2011 [77] and iLIDS-VID [125] datasets, respectively, which establish a new state-of-the-art on both the datasets. Detailed comparisons with other contemporary unsupervised and supervised re-identification methods are presented in table 6 and table 7.

Table 6. Recognition rates (%) at different ranks for unsupervised methods. Best results are highlighted in bold.

Spatiotemporal Methods								
	PRID2011				iLIDS-VID			
Method	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
HOG3D [93]	20.7	44.5	57.1	76.8	8.3	28.7	38.3	60.7
FV3D [93]	38.7	71.0	80.6	90.3	25.3	54.0	68.3	87.7
STFV3D [93]	42.1	71.9	84.4	91.6	**37.0**	**64.3**	**77.0**	**86.9**
Spatial Methods								
	PRID2011				iLIDS-VID			
Method	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
SDALF [68]	5.2	20.7	32.0	47.9	6.3	18.8	27.1	37.3
eSDC [141]	25.8	43.6	52.6	62.0	10.2	24.8	35.5	52.9
FV2D [94]	33.6	64.0	76.3	86.0	18.2	35.6	49.2	63.8
PAM-HOG	50.6	72.2	83.6	93.0	22.9	44.3	55.7	69.3
PAM-LOMO	**70.6**	**90.2**	**94.6**	**97.1**	33.3	57.8	68.5	80.5

Figure 16. Visualization of full-body appearance mixtures of HOG descriptor. For each person, first image is one of the input images used to learn appearance model. It is followed by the composite images, one for each component of the GMM. Optimal number of components for each appearance model varies between persons. (a)-(d) GMM components focus on different pose and orientation of person. (e)-(g) Transient occlusions are implicitly dealt with in appearance models as components focus on pose and orientation. (h) GMM components focus on different person alignments in the bounding box.

Table 7. Recognition rates (%) at different ranks of supervised methods. Best results are highlighted in bold.

Dictionary or Feature Learning Methods								
	PRID2011				iLIDS-VID			
Method	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
DVDL [80]	40.6	69.7	77.8	85.6	25.9	48.2	57.3	68.9
Color+LFDA [106]	43.0	73.1	82.9	90.3	28.0	55.3	70.6	88.0
AFDA [88]	43.0	72.7	84.6	91.9	37.5	62.7	73.0	81.8
MTL-LORAE [119]	-	-	-	-	43.0	60.1	70.3	85.3
RCNN [96]	70.0	90.0	95.0	97.0	58.0	84.0	91.0	96.0
Metric or Rank Learning Methods								
	PRID2011				iLIDS-VID			
Method	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
HOG3D+RankSVM [125]	19.4	44.9	59.3	77.2	12.1	29.3	41.5	56.3
Color+RankSVM [125]	29.7	49.4	59.3	71.1	16.4	37.3	48.5	62.6
ColorLBP [78]+RankSVM	34.3	56.0	65.5	77.3	23.2	44.2	54.1	68.8
DVR [125]	28.9	55.3	65.5	82.8	23.3	42.4	55.3	68.6
DSVR [126]	40.0	71.1	84.5	92.2	39.5	61.1	71.7	81.0
STFV3D+KISSME [93]	64.1	87.3	89.9	92.0	43.8	69.3	80.0	90.0
LOMO+XQDA [90]	-	-	-	-	53.0	78.5	86.9	93.4
LOMO+SBSR+XQDA [53]	-	-	-	-	68.5	87.9	93.0	96.3
RFA-Net+RankSVM [131]	58.2	85.8	93.4	97.9	49.3	76.8	85.3	90.0
CNN+KISSME [142]	69.9	90.6	-	98.2	48.8	75.6	-	92.6
CNN+XQDA [142]	77.3	93.5	-	99.3	53.0	81.4	-	95.1
PAM-HOG+KISSME	55.3	80.7	90.2	95.6	33.9	60.0	70.2	79.1
PAM-LOMO+KISSME	**92.5**	**99.3**	**100.0**	**100.0**	**79.5**	**95.1**	**97.6**	**99.1**

For further details, please refer to our paper [31].

7.9. Person Re-Identification using Pose-Driven Body Parts

Participants: Behzad Mirmahboub, Furqan Khan, François Brémond.

keywords: appearance-based person re-identification, pose estimation, human body parts, mask.

7.9.1. Introduction

Person re-identification is the problem of recognizing persons between several non-overlapped cameras. The main assumption is that the persons don't change their clothings between cameras. General approach is to extract discriminating color and texture features form images and calculate their distances as a measure of similarity. Most of the works consider whole body to extract descriptors. However, human body may be occluded or seen from different views that prevents correct matching between persons. We propose to use a reliable pose estimation algorithm to extract meaningful body parts and extract descriptor from each part separately.

7.9.2. Body Parts

"OpenPose" [52] is a state-of-the-art pose estimation algorithm that detects 15 body joints as shown in Fig. 17 (a). An example of pose estimation result on MARS dataset [142] is shown in Fig. 17 (b). We used joint positions to define 12 body parts as shown in Fig. 17 (c). Our idea is to extract image descriptor from each part and calculate their distances separately. Distance between two images can be computed by weighted average of all distances between body parts.

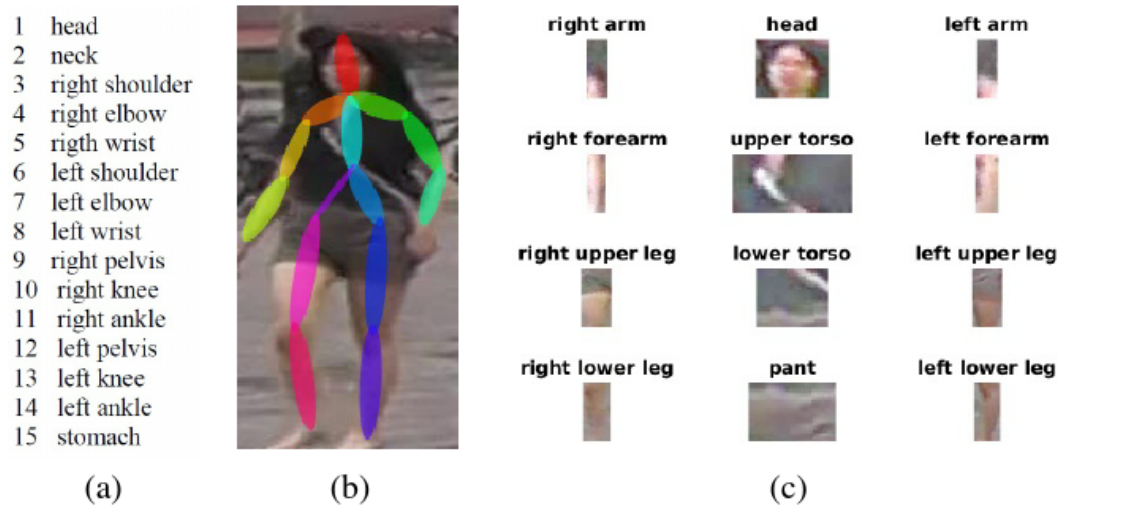


Figure 17. Human body joints and parts (a) Body joints that are detected with pose estimation algorithm (b) An example image from MARS dataset with estimated pose (c) Different body parts that we defined for feature extraction.

LOMO [90] is a famous descriptor for person re-identification that divides each image into horizontal stripes and finds the maximum bins of color and texture histograms in each stripe. We modified this code to use it on body parts.

7.9.3. Body Mask

Another challenge in person re-identification is different backgrounds between cameras. Background usually adds unnecessary information to descriptors that is not related to the person in the image, resulting in the mismatch between persons. We used the results of pose estimation algorithm to find a mask for whole body as can be seen in Fig. 18. Since the masked image has many zero-value pixels that are not related to the persons, we modified LOMO descriptor to remove the effect of those zero pixels.

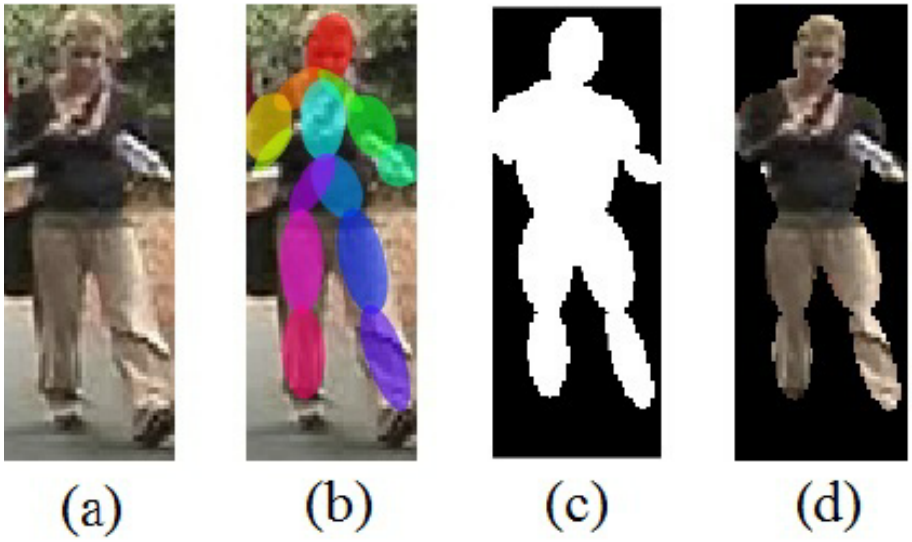


Figure 18. Generating mask from pose estimation (a) An example image from VIPeR dataset [76] (b) Result of pose estimation (c) Mask that we define based on pose (d) Masked image

7.9.4. Conclusion

Preliminary experiments show some potentials of using pose estimation for ReID, but not as accurate as global signature. One shortcoming of our work may be that we relied on LOMO descriptor that is essentially designed for the whole image. Suitable descriptor such as deep features [130] should be designed for body parts. In case of proper descriptor, part-based re-identification is promising to cope with the problem of pose and view point variations. This work can also be extended to detect mid-level features or attributes [86] (such as gender, long hair, jeans, t-shirt etc.) that are more reliable than low-level descriptors (such as gradients and histogram).

7.10. Human Action Recognition in Videos with Local Representation

Participants: Michal Koperski, François Brémond.

keywords: Computer Vision, Action Recognition, Machine Learning, Deep Learning, Artificial Intelligence

This work targets recognition of human actions in videos. Action recognition can be defined as the ability to determine whether a given action occurs in the video. This problem is complicated due to the high complexity of human actions such as appearance variation, motion pattern variation, occlusions, etc.

Recent advancements in either hand-crafted or deep-learning methods significantly improved action recognition accuracy. But there are many open questions, which keep action recognition task far from being solved.

Current state-of-the-art methods achieved satisfactory results mostly based on features, which focus on a local spatio-temporal neighborhood. But human actions are complex, thus the following question that should be answered is how to model a relationship between local features, especially in spatio-temporal context.

In previous years, we proposed 2 methods which try to answer that challenging problem. In the first method [49], we proposed to measure a pairwise relationship between features with Brownian Covariance. In the second method [83], we proposed to model spatial-layout of features with respect to person bounding box, achieving better or similar results as skeleton based methods. Our methods are generic and can improve both hand-crafted and deep-learning based methods.

Another open question is whether 3D information can improve action recognition. Currently, most of the state-of-the-art methods work on RGB data, which is missing 3D information. In addition, many methods use 3D information only to obtain body joints, which is still challenging to obtain. In our previous work, we showed that 3D information can be used not only for joints detection. We proposed [82] a novel descriptor which introduces 3D trajectories computed on RGB-D information.

In this year work we provide comprehensive study of methods proposed in the previous years, which is a part of PhD thesis [22] defended on 9th November 2017. In the evaluation part, we focus particularly on daily living actions – performed by people in their daily self-care routine. In the scope of our interest are actions like eating, drinking, cooking. Recognition of such actions is particularly important for patient monitoring systems in hospitals and nursing homes. Daily living action recognition is also a key component of assistive robots.

To evaluate the methods proposed in this work we created a large-scale dataset, which consists of 160 hours of video footage of 20 senior people. The videos were recorded in 3 different rooms by 7 RGB-D sensors. We have annotated the videos with 28 action classes. The actions in the dataset are performed in un-acted and unsupervised way, thus the dataset introduces real-world challenges, absent in many public datasets.

We proposed also new GHOG descriptor which is able to capture rough static pose information from person bounding box without need of skeleton detection. In our PhD thesis we show that fusion of GHOG with descriptors, which capture dynamic information (eg. [49], [83], [82]) leads to significant recognition accuracy improvement.

Finally, we claim that ability to process a video in real-time will be a key factor in future action recognition applications. All methods proposed in this work are ready to work in real-time. We proved our claim empirically by building a real-time action detection system, which was successfully adapted by Toyota company in their robotic systems.

We have also evaluated our methods on our Smarthomes dataset as well as on publicly available datasets: CAD60, CAD120 and MSRDailyActivity3D. Our experiments show that the methods proposed in this thesis improve state-of-the-art results.

More detail description can be found in the PhD thesis [22].

7.11. Action Detection in Untrimmed Videos

Participants: Abhishek Goel, Michal Koperski, François Brémond.

7.11.1. Problem Statement

The problem addressed in this work is *Online Action Detection in Untrimmed Videos*. The task of action detection can be broken down into two major modules, namely Action Recognition module and Temporal Localization module. Action Recognition module is responsible for assigning an action label to a trimmed video clip that is having only one action from start to the end of the clip. Temporal localization module on the other hand is responsible for deciding upon the start and end of the action present in an untrimmed video. The work has been done on the Smarthomes Dataset [22].

7.11.2. Action Detection Framework

- **Recognition Module:** The recognition module used in this work, makes use of trajectory features [72], [128] for describing the input frames. These features are clustered using a 512 centroid Gaussian Mixture Model (GMM) and encoded using Fisher Vector. Finally Fisher Vectors are then used as input to SVM classifier, which is trained in a one vs all fashion. Figure 19 gives an overview of the recognition model used in this work.

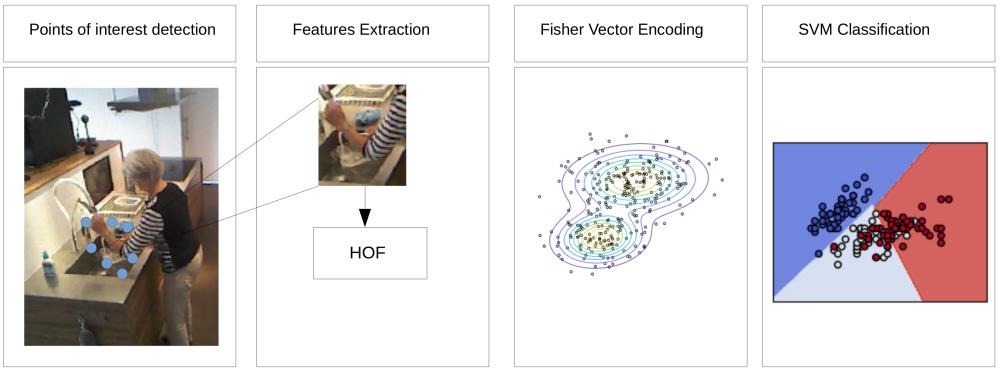


Figure 19. Steps involved in the action recognition part of the Action Detection Module. The first step is to detect the feature points in the input image. Since, in the feature detector used is Dense Trajectories [128], the feature points have been obtained using dense sampling followed by removal of points in homogeneous points. For each of the interest point, HOF descriptor is computed. These features are then clustered using GMM and encoded using fisher vectors. Finally, a SVM classifier is trained using the fisher vectors obtained for different video clips of different action class.

- **Temporal Localization Module:** The temporal localization module makes use of a sliding window architecture [72], [101], [120], [97], [134] to give candidate clips, intervals which might have action of interest, to the action recognition module to get the label for that interval.

7.11.3. Challenges

The major challenges when working with untrimmed videos in an online fashion are to identify the intervals where there are *No Action of interest* present and to identify the transition from the No Action interval to an interval containing an action of interest. In order to address these two problems, two new methods were proposed.

7.11.4. Proposed Methods

Distance Based Sliding Window

The first method, named **"Distance Based Sliding Window"** defined an actioness criterion based on the distance of a Fisher Vector from the hyperplane of a class of a trained classifier to address the problem of identifying the No Action intervals. Figure 20 gives an overview of the proposed approach.

Past and Future Windows

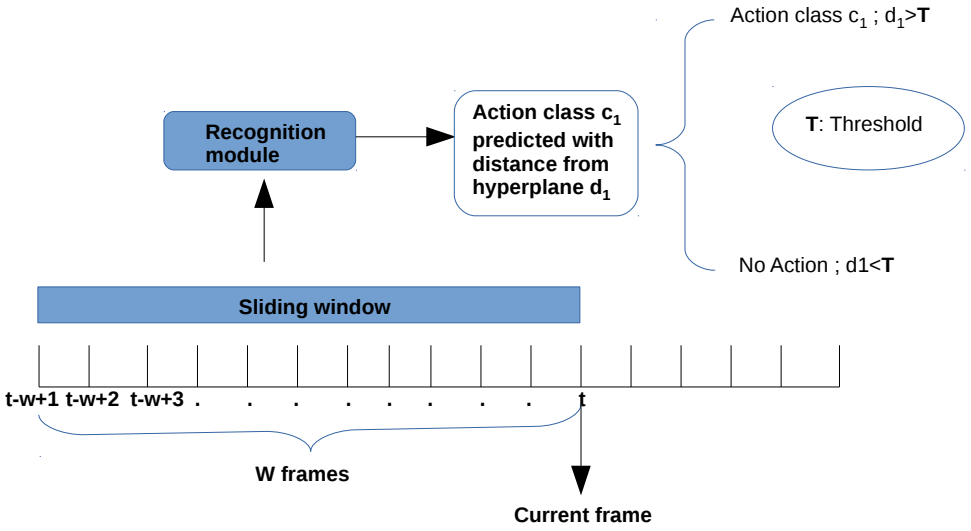


Figure 20. Approach Distance Based Sliding Window. First, a candidate, a short clip, is selected using the sliding window. This candidate is sent as an input to the action recognition system which returns the predicted class along with the distance of the fisher vector from the hyperplane of this class. Finally, if this distance is greater than a threshold T, the predicted label is that class otherwise it is No Action interval.

The second method, named **"Past and Future Windows"** addressed the second issue with a sliding window architecture which makes use of some of the future frames in order to get an action label for the current frame. The task is to perform Online Action detection in which ideally we have information only about the frames that have been seen till now and prediction for the current frame has to be done on the basis of this information. The term "future" refers to the frames which come after the frame in consideration. Since now the label is getting predicted for a frame after seeing some more frames after it, a delay is introduced in the prediction of the label. This delay is equivalent to *W* frames, where *W* is the window size. Figure 21 gives an overview of the proposed method.

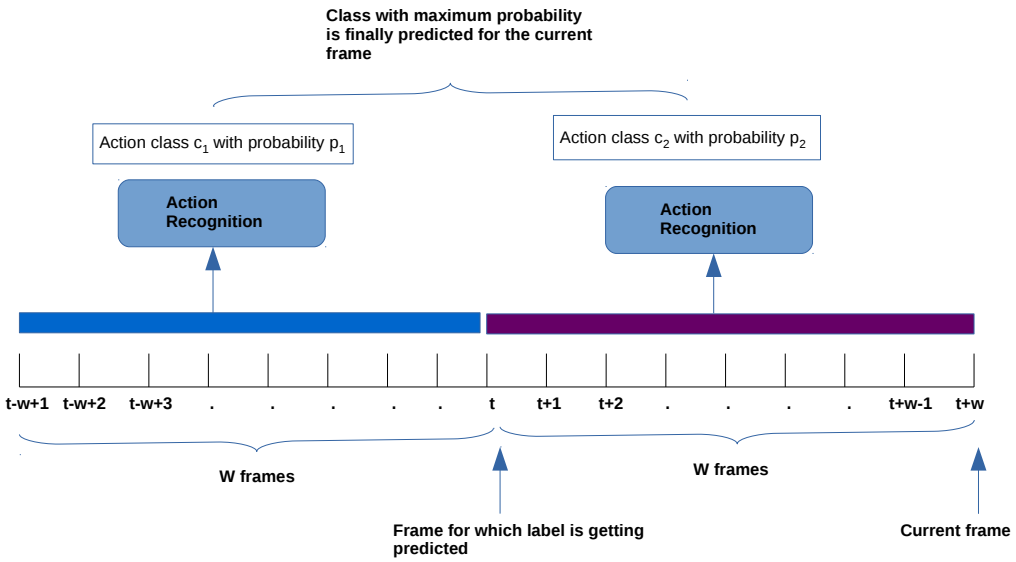


Figure 21. Past and Future windows approach. For the current frame, two temporal windows of size W frames are considered. The first window contains past frames and the second one future frames relative to the frame for which the label has to be predicted. Both the windows predict an action label with a probability with the help of an Action Recognition module. The final class label is corresponding to the class which returns the highest probability.

7.12. RGB-D based Action Recognition using CNNs

Participants: Srijan Das, Michal Koperski, François Brémond.

In the first half of the year, we focused on using the newly introduced Pose Machines [51] to extract skeletons from RGB frames. Then, our objective was to study how different skeleton extraction methods affect the performance of action recognition. For skeleton extraction from RGB data we used Pose Machines [51] and from Depth data we used Kinect sensors. Since, our final objective is to recognize actions from videos, we use an action recognition network proposed in [55]. This action network takes part patches (right hand, left hand,

upper body, full body and full image) around the joints to produce CNN features. The framework considers both the appearance flow and the optical flow so as to produce the concatenated CNN features. These features followed by a max-min aggregation are used as an input of a SVM to classify actions.
Finally, we propose a fusion of classifiers trained based on each skeleton extraction methods discussed above to improve the action recognition performance. The framework is depicted in fig. 22. We validate our approach on CAD60, CAD120 and MSRDailyActivity3D, achieving the state-of-the-art results. We chose daily living action datasets due to its application to healthcare and robotics. The proposed framework has been published in AVSS 2017 [28].

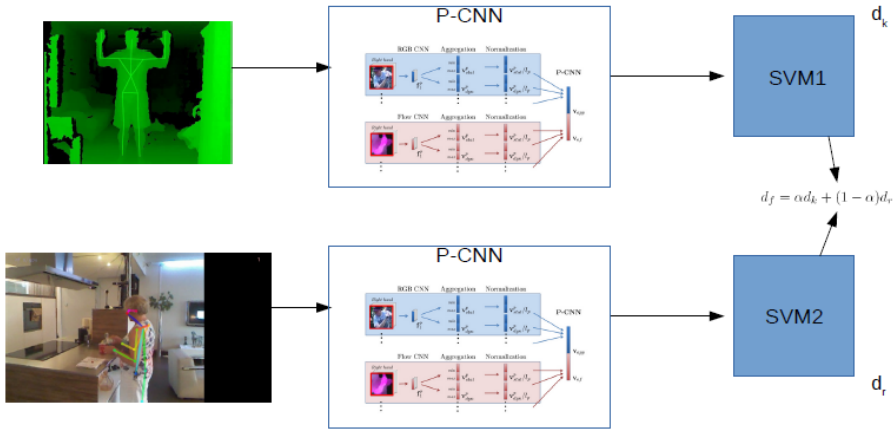


Figure 22. Framework of the proposed approach in [28]. Here, d_i represents the classifier scores for $i-th$ network.

Though our proposed approach performs well on CAD60 and MSRDailyActivity3D, it does not perform well on CAD120 because of the wrong patches detected during the skeleton extraction technique. Moreover, we also observed that sometimes noisy skeletons results in better action recognition as compared to well detected skeletons. This is because, noisy skeletons include the objects incurred in the actions since we use part patches in the action recognition network. So, now we include the object reference as well by considering extra large part patches in our modified P-CNN network for action recognition.

Now, we are focusing on how to include temporal information for action recognition. In our recent work, we discuss the limitation of not taking the temporal information into account for action recognition. Our objective is to introduce the temporal evolution of skeleton sequences with Recurrent Neural Networks (RNNs).

This work has been done in collaboration with Francesca Gianpiero (Toyota Motors Europe).

7.13. Recognizing Human Actions Using RGB Sport Videos From the Web

Participants: Amir Nazemi, François Brémond.

keywords: Action Recognition, Activity Recognition, Video Summarization, Web Sport Videos, Golf Videos.

The aim of this work is to extract sport actions from a web sport streaming video and use them for highlight detection. The sport videos which is used in this research is Golf videos. The report explains 4 steps including the data preparation, methods selection and excremental results.

7.13.1. Data Preparation

Figure 23. The output of human poses detection on one frame of Golf video dataset

Table 8. The Golf dataset.

Class names	Number of samples
Tee shot + Geometrical Features	73
Putt	70
standing	81

Table 9. The experimental results of performing two different methods on the golf dataset.

Methods	Accuracy on Golf Dataset
LSTM + Geometrical Features	91.5 %
P-CNN	97.32 %

First, from a streaming video a dataset is built. This dataset contains 3 action classes such as Tee-shot, Putt and Standing. Table 8 shows the dataset description.

7.13.2. Framework

After preparing the dataset next step is to define the solutions for the problem. Since one of the main goal of this research is to provide a general solution for sport video then we proposed a solution based on the skeleton or human poses. Our proposed framework contains human pose detection, human tracking and action recognition respectively. For human pose detection we used a recent method named open-pose [105]. For human pose tracking we used a tracking method of Inria STARS SUP framework. Finally for action recognition we did some experiments for choosing the best method.

7.13.3. Methods selection

From different methods in the field of action recognition we selected the P-CNN [55] method which is the state of the art on some data-set. Additionally for having an alternative solution which is faster than P-CNN we proposed a method based on geometrical features of human poses. We used the geometrical features in a Long Short-Term Memory (LSTM) structure to characterize the second solution.

7.13.4. Experimental Results

Table 9 shows the results of selected methods on the prepared golf dataset. As it is illustrated in the table 9 the P-CNN method works better than a method with LSTM and geometrical features.

7.14. Event Recognition Based on Depth Image

Participants: Kuan-Ru Lee, Carlos F. Crispim Junior, Yu-Feng Chen, François Brémond.

keywords: event recognition, depth image

7.14.1. Introduction

We proposed several methods to improve event recognition that are related to human and bed. Our final goal is to provide a helpful system to support the real-time observation of patients by medical personnel.

7.14.2. Experiments

The experiments were designed in several continuous steps. The first is the basic model by using time relative operator and two particular zones, the others were developed on this basic model. Theoretically, the performance of the accuracy will be increased step by step.

Basic model

The idea of the basic model is to combine the information of time and position to achieve the goal. Therefore, we create two particular zones to represent the area of the bed and the surrounding region of the bed.

$$Zone_M \textit{ before } Zone_N$$

M and N represent the bed and the surrounding area. The time relative operator was assigned as before. To distinguish the direction of the person, we can simplify the expression by switching the zones. However, the model was not able to detect the activity because the video was not completely recorded. Due to the ontology language that we had designed, what had been done by person will be initialized in this case, and we named these problems as "Frames Jump". On the other hand, our model was sensitive to the location of person. Extra detection will happen when the person who was standing next to bed, and just simply a step backward but not sitting on the bed. By integrating the above problems, we decided to develop a new method to replace the function that is used to distinguish the direction of the person.

Distance Analysis between the Person and the Bed

To recognize the events get-in-bed and get-out-bed, we compute the vertical distance between the person and a line of reference, which is horizontally passing from the center of the bed. To avoid the noise influencing the instance value of the distance, we used majority voting rules to represent the general direction of the moving person.

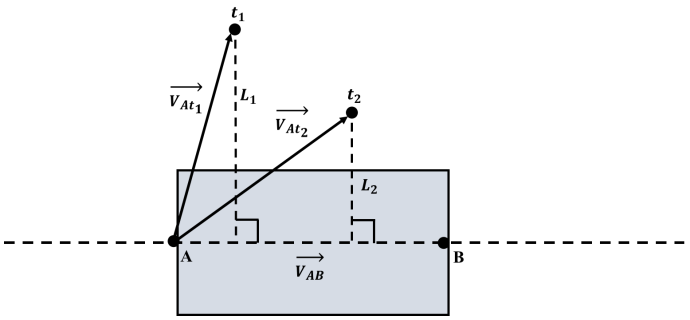


Figure 24. Schematic diagram of the proposed method

In Figure 24, the rectangle represents the zone of bed, and a horizontal line cross through the reference points A and B. The point t_i represents the position of the person at frame i while i $\in \{1, 2, ..., N\}$. To project t_i on the reference line, the distance between t_i and the reference line, L_i, can be easily calculated. The majority voting rules can provide an average value representing the final distance.

7.14.3. Frame Jump

Due to the detection process, the system will start to detect the person while he or she moves. In this case, some frames may be lost when the person is still on the bed but in the getting out processes. Based on our ontology language, we designed a new model which just focused on those events which happened surrounding the bed area.

7.14.3.1. Performance

Table 10. Contingency table of two events

	Get in Bed					Get out Bed				
Method	TP	FP	FN	Pr.	Re.	TP	FP	FN	Pr.	Re.
A	6	14	10	0.3	0.37	7	9	8	0.43	0.46
B	9	2	8	0.81	0.52	7	5	8	0.58	0.46
B + C	12	6	5	0.6	0.70	14	10	1	0.58	0.93

In the previous section, the basic model, which we labeled as method A, faced two problems, frame jump and extra detecting. Those problems lead high false positive and false negative. To solve the unnecessary extra detecting, we propose to analyse the distance between the person and the bed with the method B.

In table 10 we can notice the improvement of TP, FP, and FN. The FP value for method B is much lower than for method A. The new approach using the average filter and majority voting rule to eliminate the noise of movement, reduces the chance of misdetection and improves the whole performance. After we considered the solution of frame jump as method C, and combined it with method B, the number of FN reduced a lot. The increments of FP are caused by the overlap detection of both methods.

7.15. Recognition of Daily Activities by Embedding Visual Features within a Semantic Language

Participants: Francesco Verrini, Carlos F. Crispim Junior, Michal Koperski, François Brémond.

keywords: Activity of Daily Living, RGBD Sensors, Activity Recognition

The recognition of complex actions is still a challenging task in Computer Vision especially in daily living scenarios where problems like occlusion and limited field of view are very common. Recognition of Activity Daily Living (ADL) could improve the quality of life and supporting independent and healthy living of older or/and impaired people by using information and communication technologies at home, at the workplace and in public spaces. A method based on the development of a scenario model with semantic logic and a priori knowledge formalism is able to take into account spatio-temporal information of the scene but falls short in identifying finer events occurring in a specific area, e.g. it identifies that the person is sitting but we cannot determine whether the person is only sitting, using a laptop, or watching television. For this method the supervision of experts is also needed [58]. In this method after detection and tracking, event recognition is performed (fig 25). On the other side, action recognition through visual words [83] improves recognition of actions with low amount of motion but contextual information of the scene is not taken into account. The goal of this work is to merge the two methods trying to use the spatio-temporal information of the ontology model to improve the results of the action recognition through visual words. The actions detected with visual words are implemented as Primitive States in the scenario and then used as Components of Composite States to merge them with spatio-temporal pattern that the people display while performing activities of daily living (e.g. Person Inside Zone Sink, Person moving between zone Entry and zone Corridor). In a challenging Dataset such as SmartHome [22] where a high variance intra-class and low variance inter-class is present results for some actions improves in precision and recall thanks to spatial information, e.g. for Clean Dishes the precision with the proposed method increases form 29% to 42% thanks to the definition of a spatial zone for the sink. A drawback of this method is that True Positive can not increase being strictly dependent on machine learning pipeline, for this reason Precision improves due to the less number of False Positive.

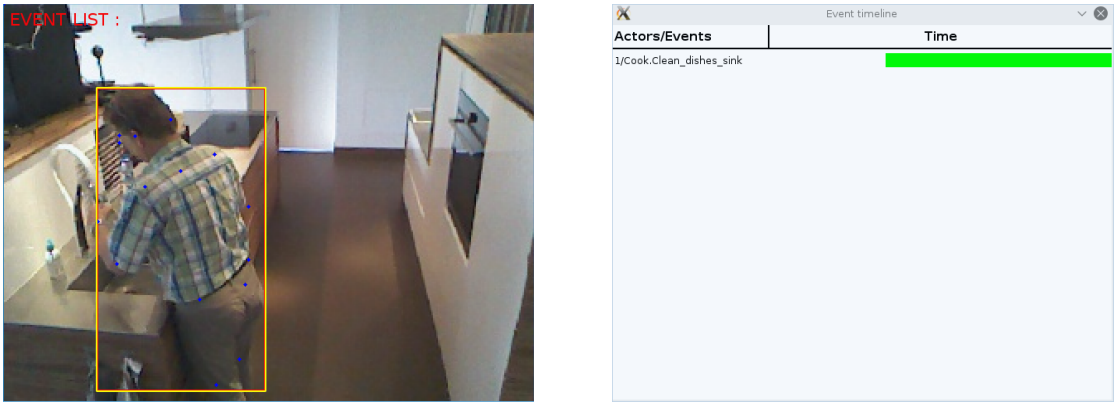


Figure 25. left:Person Performing Cook.Clean_dishes_sink, right:Timeline of the action Cook.Clean_dishes_sink

7.16. Cognitive Assessment Using Gesture Recognition

Participants: Farhood Negin, Michal Koperski, Philippe Robert, François Brémond, Pau Rodrigez, Jeremy Bourgeois.

keywords: Human computer interaction, Computer assisted diagnosis, Cybercare industry applications, Medical services, Patient monitoring, Pattern recognition.

7.16.1. The Praxis test and clinical diagnosis

Praxis test is a gesture-based diagnostic test which has been accepted as diagnostically indicative of cortical pathologies such as Alzheimer's disease. Despite being simple, this test is oftentimes skipped by the clinicians. In this study, we proposed a novel framework to investigate the potential of *static* and *dynamic* upper-body

gestures based on the Praxis test and their potential in a medical framework to automatize the test procedures for computer-assisted cognitive assessment of older adults.

In order to carry out gesture recognition as well as correctness assessment of the performances we have recollected a novel challenging RGB-D gesture video dataset [1] recorded by Kinect v2, which contains 29 specific gestures suggested by clinicians and recorded from both experts and patients performing the gesture set. Moreover, we propose a framework to learn the dynamics of upper-body gestures, considering the videos as sequences of short-term clips of gestures. Our approach first uses body part detection to extract image patches surrounding the hands and then, by means of a fine-tuned convolutional neural network (CNN) [87] model, it learns deep hand features which are then linked to a long short-term memory (LSTM) [79] to capture the temporal dependencies between video frames.

We report the results of four developed methods using different modalities. The experiments show effectiveness of our deep learning based approach in gesture recognition and performance assessment tasks. Satisfaction of clinicians from the assessment reports indicates the impact of our proposed framework corresponding to the diagnosis.

7.16.2. Proposed Method

Four methods have been applied to evaluate the dataset (Figure 26). Each path (indicated with different colors) learns its representation and performs gesture recognition independently given RGB-D stream and pose information as input.

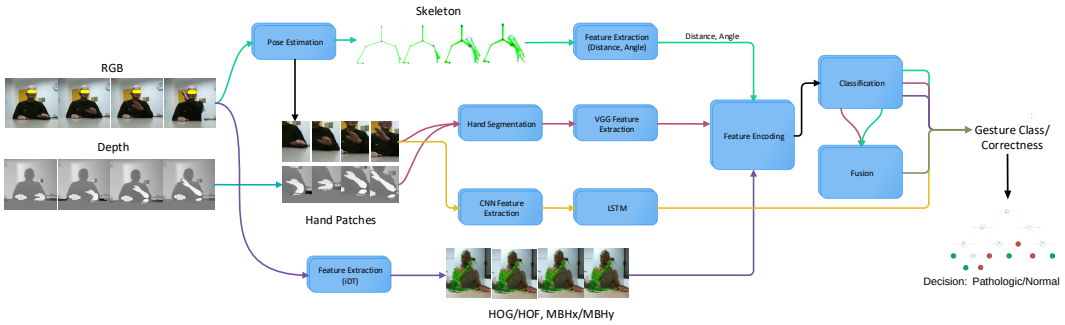


Figure 26. The data flow for the four methods applied on the Praxis dataset. Each method flow is separated by using a different color code.

Skeleton Based Method: Similar to [132] the joint angle and distance features are used to define global appearance of the poses. Prior to the classification (different from [132]), a temporal window-based method is employed to capture temporal dependencies among consecutive frames and to differentiate pose instances by notion of temporal proximity.

Multi-modal Fusion: The skeleton feature captures only global appearance of a person, while deep VGG features [117] extracted from RGB video stream acquire additional information about hand shape and dynamics of the hand motion which is important for discriminating gestures, specially the ones with similar poses. Due to sub-optimal performance of immediate concatenation of the high-dimensional features, a late fusion scheme for class probabilities is adopted.

Local Descriptor Based Method: Similar to action recognition techniques which use improved dense trajectories [128], a feature extraction step is followed by a Fisher vector-based encoding scheme.

[1] https://team.inria.fr/stars/praxis-dataset/

Deep Learning based Method: Influenced by recent advancements in representation learning methods, a convolutional neural network based representation of hands is coupled with a LSTM to effectively learn both temporal dependencies and dynamics of the hand gestures. In order to make decisions about condition of a subject (normal vs pathologic) and perform a diagnostic prediction, a decision tree is trained by taking output of gesture recognition task into account.

It should be noticed that for all the developed methods we assumed that the subjects are in a sitting position in front of the camera where only their upper-body is visible. We also assume that the gestures are already localized and the input to the system is short-term clipped videos.

		Accuracy			Correctness		
Method		Static	Dynamic	Average	Static	Dynamic	Average
Skeleton	Distance	70.04	56.99	63.51	72.04	59.93	65.98
	Angle	57.21	51.44	54.32	68.13	62.16	65.14
	Distance+Angle	61.83	55.78	58.80	70.06	61.49	65.77
Multimodal Fusion	RGB (VGG)	67.63	63.18	65.40	68.21	63.54	65.87
	RGB (VGG)+Skeleton	72.43	62.75	67.59	70.72	64.55	67.63
improved dense trajectories (iDT)	HOG/HOF	65.04	61.31	63.17	61.89	57.37	59.63
	MBHx/MBHy	70.32	75.49	72.90	55.63	72.93	64.28
Deep Learning	CNN+LSTM	**92.88**	**76.61**	**84.74**	**93.80**	**86.28**	**90.04**

Figure 27. Comparison of the obtained results for static and dynamic gestures using proposed methods in terms of accuracy of gesture classification and correctness of performance with other baseline methods (Best performances are indicated in bold).

In this work we made a stride towards non-invasive detection of cognitive disorders by means of our novel dataset and an effective deep learning pipeline that takes into account temporal variations, achieving 90% average accuracy on classifying gestures for diagnosis. The performance measurements of the applied algorithms are given in Figure 27.

We proposed a computer-assisted solution to undergo evaluation of automatic diagnosis process with help of computer vision. The evaluations of the system can be delivered to the clinicians for further assessment in decision making processes. We have collected a unique dataset from 60 subjects targeting analysis and recognition of the challenging gestures included in the Praxis test. To better evaluate the dataset we have applied different baseline methods using different modalities. Using CNN+LSTM we have shown strong evidence that complex near range gesture and upper body recognition tasks have potential to be employed in medical scenarios. In order to be practically useful, the system will be evaluated with a larger population.

7.17. Geometric and Visual Features Fusion for Action Recognition

Participants: Adlen Kerboua, Farhood Negin, François Brémond.

keywords: skeleton, geometric features, visual features, CNN.

The proposed activity recognition system consists in the fusion of two types of features: geometric features and visual features. The geometric features are computed from the 2D/3D skeleton joints that represent the articulations of the human body, which can be extracted by modern methods of pose estimation from RGB/RGB-D images, such as DeepCut pose estimation [107] or OpenPose [51]. The visual features are extracted by a convolutional neural network CNN from RGB patches representing both hands, the main steps of this method are illustrated in Figure 28.

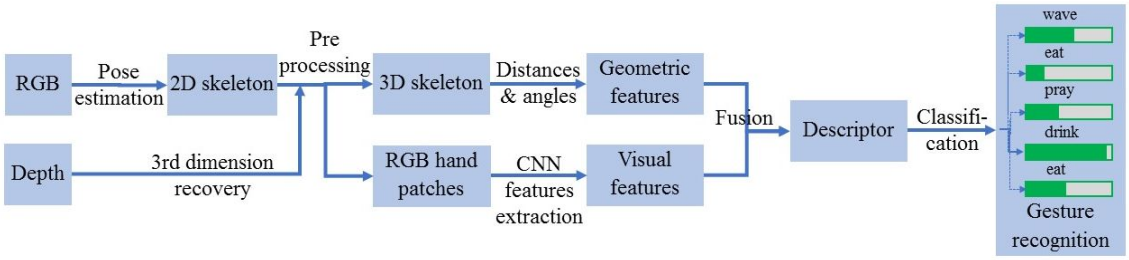


Figure 28. The data flow of our framework.

7.17.1. Geometric features

Before computing the geometrical features, we apply several preprocessing steps on the skeleton joints:

- First, we reduce noise in joints estimation by smoothing joints position over frames by applying local regression using weighted linear least squares and a 2nd degree polynomial model.
- Second, according to individual speed, similar action can be performed in different time duration by each subject resulting different number of frames, to uniform the skeleton information over frames we use cubic interpolation of the values at neighboring grid points in each respective dimension this method also permits to remove outliers joints wrongly estimated.
- Then, in order to compensate variations in body sizes which can causes intra-classes variations and confuse the classifier, we follow the method presented in [135] by imposing the same limbs (skeleton segments) lengths for skeletons of all individuals in the dataset.
- Finally, we compute rotations matrix for the first frame, then we apply same rotations for all frames in the video to remove the camera variations while keeping the over action dynamic, we also translate the skeleton to make the hip center of the first frame like the origin (figure 29).

To get the geometric features, we compute Euclidean distances and spherical coordinates (angles) between every skeleton joints pair-wise belonging to the same frame and adjacent frames.

7.17.2. Visual features

After the extraction of the skeletal joints during the precedent phase we can also extract the RGB patches representing the parts of the body most used to perform actions, which are in most cases the two hands, then we use several types of CNN (VGG16, VGG19, Resnet, ...) to extract the visual features. Different from our preliminary work of gesture recognition in the PRAXIS project, we merge those features with the geometrical one by concatenation to get the final descriptor used for classification of activities. During the test phase, a cross-subject test protocol is used to reduce the overfitting phenomenon and to obtain solutions that can be generalized.

7.17.3. Experiments

In the experimentation phase, we choose several datasets, including some properties of the STARS team, such as the PRAXIS dataset, this innovative test battery conducted on people with Alzheimer's disease is very useful to evaluate the evolution of this disease. This dataset contains more than 29 gestures divided between static and dynamic gestures, repeated several times by 58 patients, and contained a total of 3227 gestures performed in a correct manner and others in an inconsistent way. STARS also uses the most popular public datasets in the scientific community in order to evaluate the proposed methods compared to the current state of the art, such as NTU RGB+D [114] and ChaLearn [124].

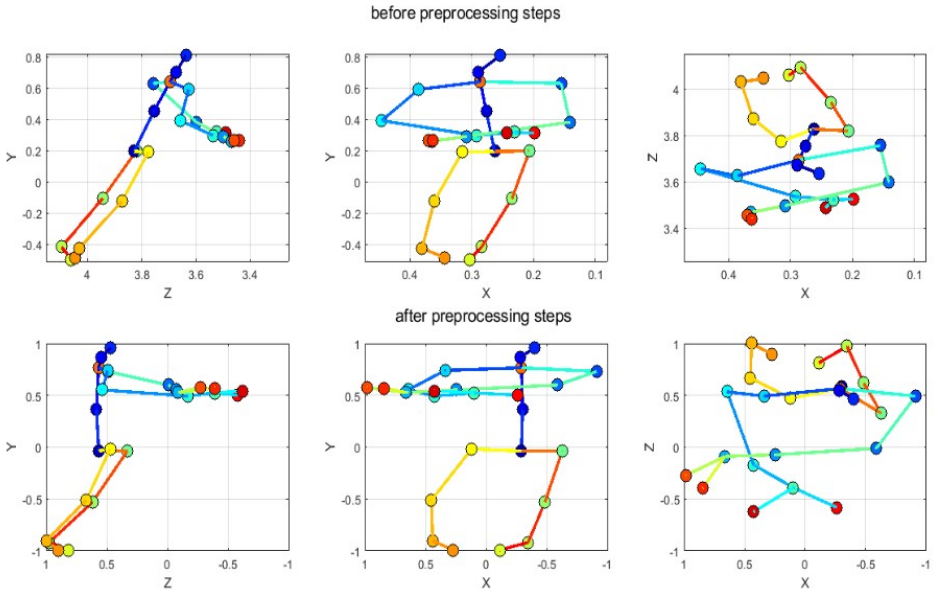


Figure 29. The data flow of our framework.

7.18. Probabilistic Logic for Activity Recognition

Participants: Carlos F. Crispim-Junior, François Brémond.

keywords: ProbLog2, Horn clauses, Probabilistic models, activity recognition, uncertainty management

In this line of investigation, we have been working on novel models to associate the robust nature of probabilistic logic to noisy observations and the knowledge representation of ontological languages. Our end goal is to design models that are capable of representing the hierarchical structure of complex events (entities, sub-events and constraints) at the same time they can handle the uncertainty of real-life settings, like noisy observations from the vision pipeline. Currently, knowledge representations underperform under noisy scenarios, while prior work in probabilistic logic has provided support either to reason about uncertainty related to entity recognition (probability of recognizing entity A) or to violation knowledge constraints (relevance of violation of constraint i to model y).

This work have been carried out in partnership with KU Leuven and the first results of this joint work have been published on the workshop entitled Assisted Computer Vision and Robotics, which was organized during the 2017 edition of the International Conference on Computer Vision. In this paper we propose BEHAVE, a person-centered pipeline for probabilistic event recognition (Fig. 30). The proposed pipeline firstly detects the set of people in a video frame, then it searches for correspondences between people in the current and previous frames (i.e., people tracking). Finally, event recognition is carried for each person using probabilistic logic models (PLMs, ProbLog2 language). PLMs represent interactions among people, home appliances and semantic regions. They also enable one to assess the probability of an event given noisy observations of the real world. BEHAVE was evaluated on the task of online (non-clipped videos) and open-set event recognition (e.g., target events plus none class) on video recordings of seniors carrying out daily tasks. Results have shown that BEHAVE improves event recognition accuracy by handling missed and partially satisfied logic models.

Future work will investigate how to extend PLMs to represent other types of relations, like temporal relations, and how to incorporate low-level information from deep architectures, like Deep Convolution Neural Networks.

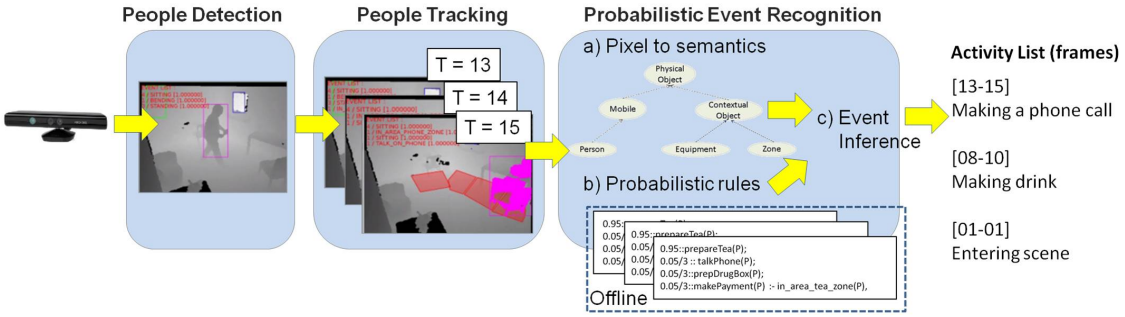


Figure 30. BEHAVE:Behavioral analysis of visual events for assisted living scenarios

7.19. Recognizing Retracing of Steps Using Walk Comparison

Participants: Kartik Kartik, Carlos F. Crispim-Junior, François Brémond.

keywords: Alzheimer, Retracing of steps, Activity Recognition

The recent advancements in the technology have paved the way to complete automation and detection of human behavior.Dementia in Alzheimer's causes memory loss and affects other mental abilities which are necessary for tasks of daily life. This discussion deals with recognizing retracing of steps in order to observe incidents or characteristics related to dementia.

7.19.1. Introduction

In recent years it has been observed a huge increase in the dementia related problems in the older population. A lot of new methodologies are being developed to recognize and observe the characteristics of people suffering from dementia. One such example is retracing the same path back and forth due to memory loss. The patient is observed walking, then stopping for a few moments or abruptly turning and then walking in the opposite direction following the same path.This discussion caters to recognition of such events using comparison between two consecutive walks. The comparison is done on the basis of position of the patient in each frame during the walk.

7.19.2. Methodology

We generate the xml1 and xml3 file by performing event recognition on SUP. SUP stands for Scene Understanding Platform which performs event and activity recognition.

In order to compute the relation between two consecutive walks that account for retracing of steps, we need to compare the positions at all points of the two walks. The information about spatial properties (2 and 3 dimensional features) are extracted from the xml1 file and accessed in python by converting the xml1 file to pandas data-frame. Similarly, the xml3 file is converted to pandas data-frame and all the frames corresponding to a walk are stored in a list.

Now, the next step is to compare every pair of frames (because they constitute a walk event) to the pair of frames prior to them.

For example g=[110, 119, 148, 178, 194, 208, 247, 295] here, 110-119 is one walk event,148-178 is another walk event and so on. So what we have to do is to compare the walk instance 147-178 to the walk instance 110-119 and walk instance 194-208 will be compared with 148-178. This comparison between the two walk instances is done on the basis of positions at each point of the walk. It has been taken into consideration to compare the positions of second walk in reverse order.

For example if frames 1325, 1350, 1367, 1397 are two walk events that correspond to retracing of footsteps, where 1325-1350 is one walk instance (going) and 1367-1397 is another walk instance (returning). We compare the positions of 1367-1397 to 1350-1325 in reverse order thus ensuring that the two walk events are a perfect candidate for step retracing.

These events are then checked for instances in which the person exits the scene and re-enters the scene following the same trajectory. Such events are of no interest from a clinical point of view hence these cases are dropped from the final list of frames that describe step retracing.

Further, to not miss the events of step retracing in a single walk,we propose to segment the walk and compare for positions using a sliding window. Each walk is compared for positions after the 20 frames of the walk. If the positions are closely related, the walk qualifies as an event of step retracing.

7.19.3. Tables

The output can be computed in the following format

	event id	event	start frame	end frame
0	0	ALLER	68967	68979
1	1	RETOUR	68989	69000
4	2	ALLER	96773	96799
5	3	RETOUR	96809	96849
8	4	ALLER	100293	100320
9	5	RETOUR	100369	100396
12	6	ALLER	121357	121369
13	7	RETOUR	121383	121396
16	8	ALLER	125693	125707
17	9	RETOUR	125718	125736
20	10	ALLER	136516	136524
21	11	RETOUR	136531	136560
24	12	ALLER	137845	137856
25	13	RETOUR	137886	137908

7.20. Safe & Easy Environment for Alzheimer Disease and related disorders

Participants: Auriane Gros, François Brémond.

As part of the SafEE project (see 9.2.1.2), we have implemented two clinical protocols, one for patients in nursing home with Alzheimer disease and the other at home for patients with a frailty syndrome. The goal was to automatically recognize emotional disorders. In nursing home, we sought to detect events specific to the quality of sleep (effective hours of sleep, awakenings at night) to provide an objective measure of sleep disorders. For this purpose, several activity analysis methods have been used to increase the accuracy of automatic recognition. The events of interest were represented by "get in bed" and "get out bed". The performance of the three models used (A; B and B + C) allowed a performance of 0.71 (F1 score). At home, we focused on anxiety-like behavior (repetitive gestures, parasitic gestures), apathy (locomotion, gestures with goals) and agitation (retracing steps, locomotion). In order to be able to identify algorithms with the greatest precision possible, clinicians have realized annotations via the Viper software. In order to annotate these events we have developed a language based on a dedicated ontology for the recognition of activity behaviors that we have deemed of interest. In order to carry out an automatic recognition of the activities of interest, files (xml1 and xml3) were generated in the Scene Understanding Platform (SUP). For the recognition of the

activity "to retrace one's steps" we have calculated the relation between two consecutive walks by comparing all the position points of these walks. The spatial properties information was extracted from the xml1 file and converted to Python Pandas data to analyze and visualize the data. The first results highlighted: -a gradual increase in social interactions during the follow-up; -an increase in gestures with goals; -a decrease in retracing one's steps. The studies were conducted on 6 patients during three months of follow-up in nursing home and three patients during six months of follow-up at home.

7.21. Early detection of cognitive disorders such as dementia on the basis of speech analysis ELEMENT

Participants: Alexandra Koenig, Antitza Dantcheva, François Brémond.

This year we have contributed to the ELEMENT activity (see 9.3.2). The goal of this activity has been directed towards automated screening for cognitive decline in non-clinical settings, resulting in faster, earlier diagnosis and intervention. Inria in collaboration with the Association Innovation Alzheimer (AIA)created the French speech corpus allowing automatic detection of dementia on the basis of speech analysis. The employed speech analysis has been augmented by the means of facial expression recognition.

Due to the rapidly ageing population, the number of people with dementia in the EU will triple by 2050. The proposed speech-based screening app supports early detection and intervention, which in turn significantly reduces cost of care and preserves quality of life. The people best placed to spot early cognitive decline are carers, social workers, and family. But there is a clear lack of affordable, usable screening apps that people without medical training can use to validate these concerns and to provide actionable data for medical professionals. The approach can also be used to track mental health and other neurological conditions. The proposed solution supplements neuropsychological assessment with sophisticated and unobtrusive natural biomarkers extracted from speech data that is collected outside of medical consultations. It provides rich information about cognitive and emotional characteristics and can be used to inform clinical judgment during consultations, saving time and money. The project will bring to the European market a new product for fully-automated, reliable, unobtrusive, self-managed screening for cognitive decline, in particular dementia, and other cognitive disorders. It will allow earlier detection and, through that, more effective interventions resulting in the reduction of overall costs associated with treatment and rehabilitation. For users is will offer the comfort of flexible usage without visiting professional physicians.

The target customer group can be characterized as individuals 60+ living either at home or in residential care facilities, as well as their families, caregivers, charities, social services, other stakeholders involved in supporting older persons. The Activity will enable the first-of-that-kind product allowing implementing sophisticated and unobtrusive neuropsychological assessment within minutes right at home or at easy reachable locations without the support of professional clinicians. The initial target markets are France and UK, with the goal to start focused marketing there at the end of 2017 - beginning 2018. The focus of initial marketing will be three-fold: (1) residential care facilities as access points for groups of users, (2) social services providing homecare, and (3) pharmacies with the modern trend of turning them from sales points to service providers. Therefore, the main method of revenue generation will be corporate subscriptions purchased for specific number of users. In the middle term (2-3 years), building on the footprint at the market for corporate clients, the company will start sales for individual clients. In this case the focus of marketing strategy will be on general practitioners as major recommenders and market agents.

Inria in collaboration with AIA created the French speech corpus allowing automatic detection of dementia on the basis of speech analysis. Until now, the corpus contains samples of 149 recorded participants from which 40 Healthy controls, 40 Major cognitive disorder and 57 minor cognitive disorder, for the rest the diagnosis in missing). Data collection is ongoing and will be coupled next year systematically with the video recording.
The following shows the list of transcribed audio files :
Semantic verbal fluency : 74/149
Phonetic fluency: 36/149
Pictures: 167/ 289

Counting backwards: 7/149
Sentence repeating : 168/ 596
Postive story: 77/140
Negative story: 76/149
Motivation: 75/149
In this year, target use case scenario and user requirements were defined. ki elements UG (haftungs-beschraenkt) was established on March 15th, 2017, with the purpose of commercializing the technologies matured and integrated within the ELEMENT project. Significant efforts have been invested in the preparation of the first public demonstration of the product's prototype Delta. In collaboration with DFKI, several research papers were published on the audio data collected in Nice (see [33], [32], [25]).

7.22. Serious exergames for Cognitive Stimulation

Participants: Guillaume Sacco, Monique Thonnat.

A serious exergame is a video game combining cognitive and physical stimulation with a positive impact on patients affect. We have worked to develop and assess X-Torp, a serious exergame which is played with a Kinect™ (see Figure 31).

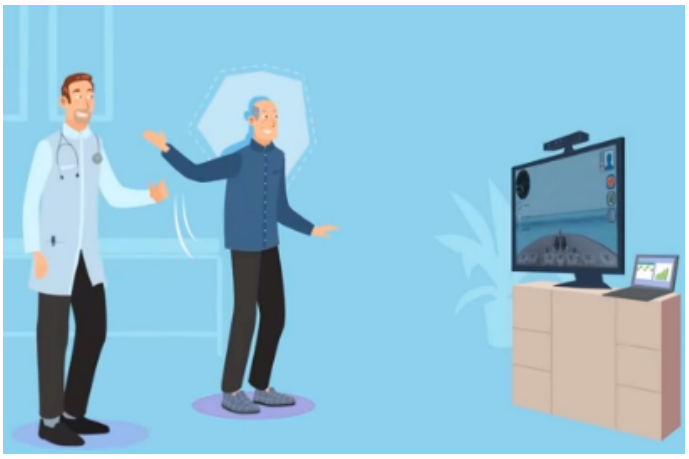


Figure 31. Illustration of the modality of use of X-Torp by a patient with a therapist

This naval battle serious exergame contains two game modes: a scenario mode and a therapist mode. The scenario mode is the core of the game. It combines exploration of an open world and mini-games. Moving in the open world (ocean with islands) corresponds to the physical exercise phase. The mini games (missions proposed on the islands) correspond to the phases of stimulation / cognitive evaluation. The game includes a system of experience points, which reflects the progression of the player in the scenario mode. The therapist mode contains direct access to the virtual versions of the neuropsychological tests proposed in the scenario mode. Therapist mode also contains a physical test that can be performed in a virtual environment.

The integration of serious exergames as a tool to manage patients with neurocognitive disorders could have a major interest in the evaluation of cognitive abilities for two reasons: the design of tests and their more ecological character. Indeed, if the serious game is regularly used by the patient (in autonomy possibly guided by an avatar but not by a therapist), we think that it can allow a much more fine and relevant evaluation of the residual cognitive functions. The first reason for this is that in patients with mild or moderate cognitive impairment, there is a variation in test performance that may be related to anxiety, but also to the variability of observable performance over the day and day to day for some patients.

The innovative aspect of our approach is both a combination of physical and cognitive activity within the same game and the use of goal-directed motor skills for physical stimulation. Indeed, the use of virtual support to guide physical activity allows a longer duration of activity and therefore a better training.

If we consider the continuum of care of patients with neurocognitive disorders at prodromal or mild stages, we find that contacts with the health care system are limited. In addition, the availability of health professionals who can offer stimulation activities remains limited compared to the massive demand generated by the large number of patients with neurocognitive disorders. The interest of our approach is that it makes it possible to free oneself largely from the healthcare professional (which intervenes only for consultations of synthesis and reassessment) and thus allows many more patient follow-up for one healthcare professional.

Thus the development of playful serious exergames, with adequate design and in sufficient numbers, available for example via online gaming platforms, would allow regular use by patients in their homes. The serious exergame would then be a tool for training cognitive functions and physical abilities. It would also be a tool for regular and objective assessment and monitoring of the cognitive and physical performance of patients. It could even be envisaged the creation of game allowing the early detection of cognitive dysfunctions directly from the home of the subjects and inviting them if necessary to consult their physician.

7.23. Activity Description Language

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 artefacts), that follow the predefined model of an activity. We propose to describe activities as a series of actions, triggered and driven by environmental events.

Due to the large range of application domains (surveillance, safety, health care ...), we propose a generic approach to design activity recognition systems that interact continously with their environment and react to its stimuli at run-time. Such recognition system must satisfy stringent requirements: dependability, real time, cost effectiveness, security and safety, correctness, completeness ... To enforce most of these properties, our approach is to base the configuration of the system as well as its execution on formal techniques. We chose the *Synchronous Approach* which provides formal bases to perform static analysis, verification and validation, but also direct implementation.

Based on the synchronous approach, we have created a new user-oriented activity description language (named ADeL) to express activities and to automatically generate recognition automata. This language relies on two formal semantics, a behavioral and an equational one. This year, we continued to work on this topic: we improved both the syntax of the ADeL language to be easier to use by non computer-scientists and its semantics to generate synchronous automata.

As the world is not synchronous and since we are working with the synchronous paradigm, we have to face the classical problem of sampling. Our systems have to deal with asynchronous events coming from the environment. This year we started to define an asynchronous/synchronous transformer component, that we call Synchronizer, to transform asynchronous sensor events into synchronous "instants".

7.23.1. Activity Description Language (ADeL)

ADeL is a (synchronous) language that allows non-computer scientists to describe activities and behaviors to be recognized. It is a modular and hierarchical language, which means that activities can be simple or composed of one or more sub-activities. ADeL has the notions of (typed) roles, events and sub-activities, flow of control... It supports parallelism, variants (choices), and repetitions. ADeL relies on a set of formally specified control and temporal operators.

We provide our language with both a graphical and textual format. We propose a graphical tool which displays several windows, mainly to declare the scene where the activity will take place (zones, roles and equipment) and to describe the activity (expected events and "story board"). However, it may be difficult to express complex activities in a purely graphical way, thus we also provide an equivalent textual form.

This year we conducted a preliminary heuristic evaluation to define the layout of the graphical tool in collaboration with our ergonomist partners from LudoTIC. An example of a scene description window for a medical application (serious game) is shown in Fig. 32. We made a first positive user test with a doctor, more extensive user tests are planned.

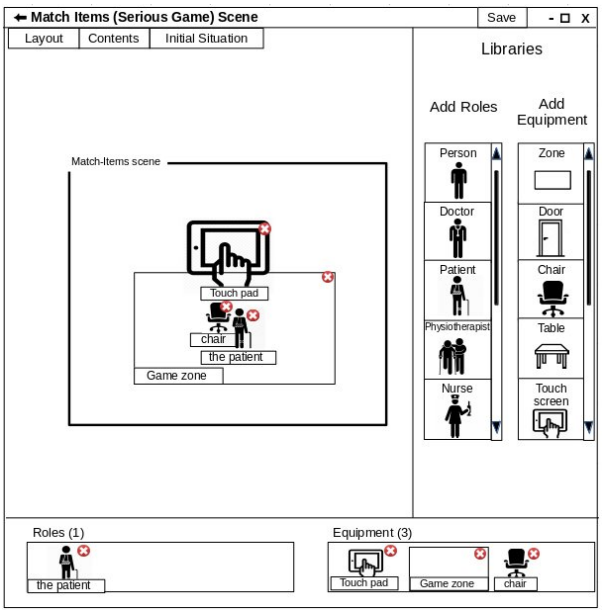


Figure 32. Graphical format of a serious game description (selection of roles and equipment)

7.23.2. Synchronizer

The main drawback of the synchronous paradigm is that the world is not synchronous in general. Thus it requires to transform asynchronous physical flows of events into a succession of discrete instants.

This year, we specified a Synchronizer component to address this issue. The Synchronizer filters physical asynchronous events, decides which ones may be considered as "simultaneous" and aggregates the latter into logical instants. The sequence of these instants constitutes the logical time of our recognition systems. In general, no exact simultaneity decision algorithm exists but several empirical strategies and heuristics may be used for determining instant boundaries. We have to take into account these parameters in the Synchronizer specification. This year we completed the UML specification of the Synchronizer and we started its first implementation.

7.23.3. Semantics

We defined two semantics for the ADeL language [34]. First, conditional rewriting rules are a classical and rather natural way to formally express the intuitive semantics. This form of *behavioral semantics* gives an abstract description and a clear interpretation of a program behavior. However it is not convenient as an implementation basis nor suitable for proofs (e.g., model-checking). Hence we also define an *equational semantics* which maps an ADeL program into a Boolean equation system representing its finite state machine. The ADeL compiler can easily translate this equation system into an efficient code not only for our runtime recognition component, but also for other tools such as model-checkers.

Since we have two different semantics, it is mandatory to establish their relationship. In fact we proved that the execution of a program based on the equational semantics also conforms to the behavioral semantics [34], [38].

7.24. The Clem Workflow

Participants: Annie Ressouche, Daniel Gaffé, Dorine Havyarimana.

Keywords: Synchronous languages, Synchronous Modeling, Model checking, Mealy machine, Logical Decision Diagram

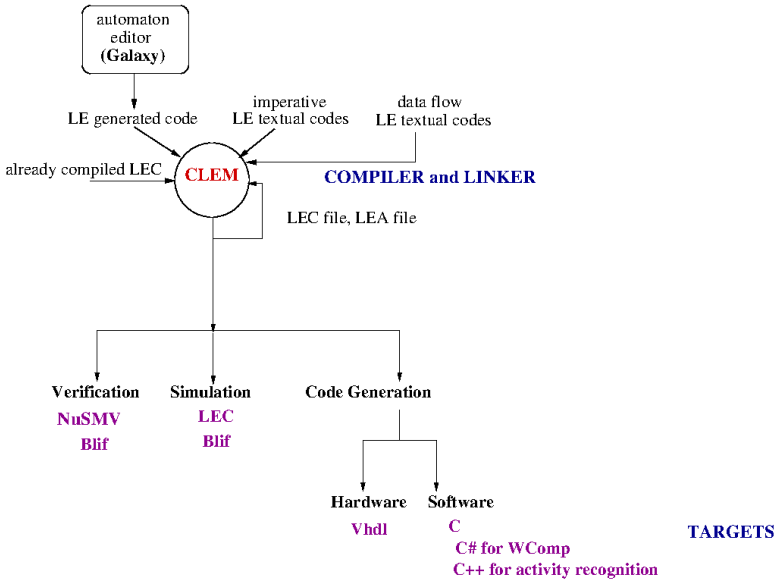


Figure 33. The Clem Toolkit

This research axis concerns the theoretical study of a synchronous language LE with modular compilation and the development of a toolkit around the language (see Figure 33) to design, simulate, verify, and generate code for programs. The novelty of the approach is the ability to manage both modularity and causality.

This year, we continued to focus on the improvement of both the LE language and compiler concerning data handling and generation of back-ends, required by other research axis of the team. We also improved the design of the simulator (developped these two last years) for LE programs which integrates the data part of the language. The simulator GUI has been designed in Qt. The simulator takes into account the values carried by signals. We implemented an external computation of data values and a communication between the compiler and the simulator through a socket mechanism. However, in this first implementation, there were no predefined types and their related methods. As a consequence, the end user must define them in external files, loaded by the simulator. To avoid the definition of such files for basic types (integer, real,...) and basic operations ($\leq$, $=$, $\geq$,...), this year we improved the simulator and the exchange mechanism to be able to integrate predefined types.

Now, we want to extend the verification side of CLEM. To this aim, this year we begin to replace the fundamental representation of Boolean values as BDD (Binary Decision Diagrams) with LDD (Logical Decision Diagrams), which allow to encode integer values in a very efficient way. During the internship of Dorine Havyarimana [41], such a LDD library has been implemented, to replace the BDD one. Then the

validation mechanism of CLEM would take into account properties of integer data. However, this is a first step and the integration of a new model checking techniques (like satisfiability modulo theories model checking) is a future work.

7.25. Study of Temporal Properties of Neuronal Archetypes

Participants: Annie Ressouche, Daniel Gaffé, Cedric Girard-Riboulleau.

Keywords: biologic archetypes, Leaky Integrate and Fire Modeling, Model Coupling, Neural Spiking networks, Synchronous Languages, Model Checking Synchronous Modeling, model-checking, lustre, temporal logic, probabilistic models, network reduction.

Last year, we began a collaboration with the I3S CNRS laboratory and Jean Dieudonné CNRS laboratory to verify temporal properties of neuronal archetypes. There exist many ways to connect two, three or more neurons together to form different graphs. We call archetypes only the graphs whose properties can be associated with specific classes of biologically relevant structures and behaviors. These archetypes are supposed to be the basis of typical instances of neuronal information processing. To model these different representative archetypes and express their temporal properties, we used a synchronous programming language dedicated to reactive systems (Lustre). Then, we generated several back ends to interface different model checkers supporting data types and automatically validate these properties. We compared the respective results, that mainly depend on the underlying abstraction methods used in model checkers [63].

This year, during the internship of Thibaud l'Yvonnet [2] we tackle the next logical step and proceed to the study of the properties of their couplings. For this purpose, we rely on Leaky Integrate and Fire neuron modeling and we use the synchronous programming language Lustre to implement the neuronal archetypes and to formalize their expected properties. Then, we exploit an associated model checker called kind2 to automatically validate these behaviors. We show that when the archetypes are coupled either these behaviors are slightly modulated or they give way to a brand new behavior. We can also observe that different archetype couplings can give rise to strictly identical behaviors. Our results show that time coding modeling is more suited than rate coding modeling for this kind of studies. These results are published in [30].

On the other hand, in the framework of Cedric Girard Riboulleau internship, we formalize Boolean Probabilistic Leaky Integrate and Fire Neural Networks as Discrete-Time Markov Chains using the language PRISM. In our models, the probability for neurons to emit spikes is driven by the difference between their membrane potential and their firing threshold. The potential value of each neuron is computed taking into account both the current input signals and the past potential values. Taking advantage of this modeling, we propose a novel algorithm which aims at reducing the number of neurons and synaptical connections of a given network. The reduction preserves the desired dynamical behavior of the network, which is formalized by means of temporal logic formulas and verified thanks to the PRISM model checker.

These results are published in [29] and detailed in [40].

7.26. Maintaining the engagement of older adults with dementia while interacting with serious game

Participants: Minh Khue Phan Tran, Philippe Robert, François Brémond.

keywords: Older adults, Dementia, Engagement, Serious game,

The contribution of Phan-Tran's thesis [23] is to provide an approach that can help older adults with dementia while playing serious games. The approach proposes a set of interaction strategies to solve a difficult situation by suitable interactions. A strategy is a rule which defines a set of interactions to transfer from a situation to another one. A situation is defined and recognized by the perception on the users characteristics (position, posture, gesture, game performance). Once an interaction strategy is chosen, the approach helps the user throughout a 3D animated avatar 34. 11 strategies are defined as well as 21 situations and 13 interactions.

[2] funded by the NeuComp project (C@UCA), in which the Stars team is involved.

Figure 34. A serious game with avatar's assistance

Three experiments have been performed with the older participants with dementia. The results shown positive impacts on the participants engagement:

- the participants can finish the game session;
- their performance while playing with the avatar's helps is similar that the one while playing with a therapist.

The proposed approach can be improved because the types of situations defined and recognized during 3 experiments are not quite so much and the difficulty level of games used in the experiment is still low. The ongoing work aims to apply the proposed approach to a more difficult game and to explore new situations.

7.27. Application of deep learning on healthcare

Participants: Thanh Hung Nguyen, François Brémond.

Healthcare standards have changed dramatically over the past 100 years. People nowadays are aware of the importance of health in their lives. The rising income has enabled them to use private healthcare services. Since the demand is growing rapidly, the physicians need tools which can help them work effeciently in terms of time and cost. Parallely, deep learning has become very popular in the recent years because of its success in computer vision. We have proposed two applications ,VISIONUM and REMINARY, as demonstration of the impact of deep learning on the healthcare.

The first application, VISIONUM (see 9.2.2.1), can detect when the patients lose their focus on the therapeutic exercise and thereby reminds them to return back to the exercise. Since the patients can correct themselves on their own, while they are doing exercises, therapists can focus more on the performance of the exercise. It also helps the therapist to keep track of multiple patients at the same time and hence, save both time and money (see Figure 35).

The second application, REMINARY (see 9.2.2.3), aims to detect the movement of people. In this application, the movement of patients is tracked and analysed. The output gives therapists an overview of the movement of their patients. This information is used by the therapist to monitor the diseases which are related to the movement and decide if there is any improvement due to the treatment. For both applications, we can also analyse emotions like happiness of patient. This information is extremely important for the design of the exercise (see Figure 36 and Figure 37).

7.28. Brick & Mortar Cookies

Participants: Julien Badie, Manikandan Bakthavatchalam, Anais Ducoffe.

Figure 35. Examples of our application VISIONUM, we detect two people in the scene, but only the person in the red bounding box is interested. This person is detected as sitting too far from the camera using depth information, he will be reminded that he should come closer so the exercise can continue.

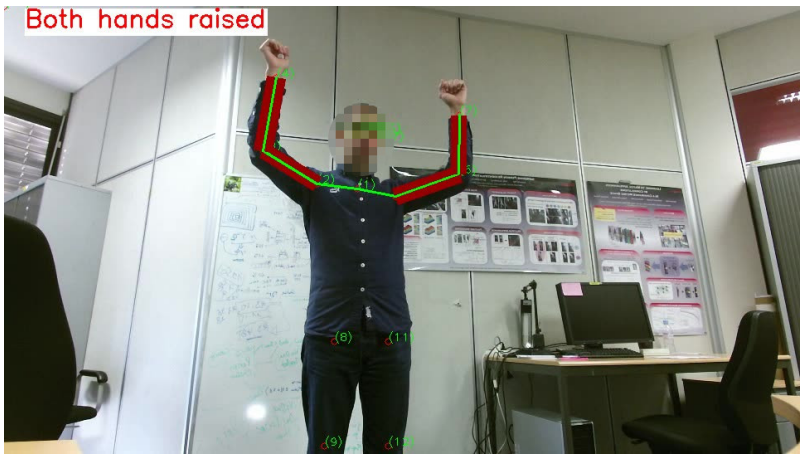


Figure 36. Examples of our application REMINARY, the actor is doing some exercise. In this case, the event that he raised both his hands is detected. The visualisation make the therapist easier to see how the system catches the event.

Figure 37. Examples of the other feature of our application REMINARY. The movement region is highlighted by a green color.

BMC is a software platform that was introduced in 2016. The goal was to create a light and easy to deploy platform in the context of gathering Business Intelligence data (BI) to help store managers monitor client trajectories and most interesting products and alleys inside a supermarket. BMC was finalized this year and had been extensively tested in real conditions with success.
The platform is divided into several modules :

- an image processing module that handles retrieving images from a camera network followed by people detection using a deep learning algorithm (SSD) and tracking. The results are stored in a MySQL database;
- a trajectory analysis module that computes statistics for BI. These statistics are computed based on the trajectories of the first module and give information regarding each alley such as the average number of customers per hour, the average time spent by the customer in the alley or the average number of stops in front of each type of product. All these data can be visualized via a web interface and are refreshed periodically;
- an automatic installation and deployment module. This module is a set of Python scripts that, given a computer with the correct hardware and OS requirements will automatically install the BMC platform and all its dependencies (OpenCV, cuda, caffe, ...). In a second time, users can enter the list of cameras they want to process with a limited set of parameters (working days and hours) and the deployment script will prepare the platform and create the CRON command line that will run BMC automatically.

An additional module was created but not included in the release : the evaluation module. It was used to find the best balance between fast processing and precise results by evaluating different parameters at the people detection level such as which model to use and which image resolution to process. After several experiments, it was decided to use a processing resolution of 480x272 to allow us to process 16 cameras simultaneously with two Nvidia GeForce GTX 980 GPU at 3-4 FPS. This framerate is sufficient for tracking in the context of a supermarket as people tends to move slowly and stops a lot. The platform was tested in a supermarket in Nice during several days and showed satisfactory results.
BMC is registered at the APP under the name BMC_1. It is intended to be used in the FUI project StoreConnect.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

- **Toyota Europ**: this project with Toyota ran from the 1st of August 2013 up to 2017 (4 years). It aims at detecting critical situations in the daily life of older adults living home alone. We believe that a system that is able to detect potentially dangerous situations will give peace of mind to frail older people as well as to their caregivers. 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 is intended to help them and their relatives to feel more comfortable because they know potentially dangerous situations will be detected and reported to caregivers if necessary. The system is intended to work with a Partner Robot (to send real-time information to the robot) to better interact with older adults.

9. Partnerships and Cooperations

9.1. Regional Initiatives

- **NeuComp** is a project of the UCA Académie d'excellence: Réseaux, Information et Société Numérique" (C@UCA). NeuComp is focusing on the model of neuron networks Leaky Integrate and Fire (LIF). The main objective of C@UCA is the brain modelling and its simulation. In this framework, the Neucomp project focuses on (1) the implementation and verification of temporal properties of neural structures; (2) the design of electronic architectures of LIF neural networks; and (3) the comparison of this electronic implementation with neuromorphic computer results. In the NeuComp project, Inria (Stars) collaborate with the LEAT (Laboratoire d'Electronique, Antennes et Télécommunications), I3S (Laboratoire d'Informatique, Signaux et Systèmes), LJAD (Laboratoire J.A. Dieudonné), Clermond Ferrand University and Arizona Unniversity.

9.2. National Initiatives

9.2.1. ANR

9.2.1.1. MOVEMENT

Program: ANR CSOSG

Project acronym: MOVEMENT

Project title: AutoMatic BiOmetric Verification and PersonnEl Tracking for SeaMless Airport ArEas Security MaNagemenT

Duration: January 2014-June 2017

Coordinator: MORPHO (FR)

Other partners: SAGEM (FR), Inria Sophia-Antipolis (FR), EGIDIUM (FR), EVITECH (FR) and CERAPS (FR)

Abstract: MOVEMENT is focusing on the management of security zones in the non public airport areas. These areas, with a restricted access, are dedicated to service activities such as maintenance, aircraft ground handling, airfreight activities, etc. In these areas, personnel movements tracking and traceability have to be improved in order to facilitate their passage through the different areas, while insuring a high level of security to prevent any unauthorized access. MOVEMENT aims at proposing a new concept for the airport's non public security zones (e.g. customs control rooms or luggage loading/unloading areas) management along with the development of an innovative supervision system prototype.

9.2.1.2. SafEE

Program: ANR TESCAN

Project acronym: SafEE

Project title: Safe & Easy Environment for Alzheimer Disease and related disorders

Duration: December 2013-May 2017

Coordinator: CHU Nice

Other partners: Nice Hospital(FR), Nice University (CobTeck FR), Inria Sophia-Antipolis (FR), Aromatherapeutics (FR), SolarGames(FR), Taichung Veterans General Hospital TVGH (TW), NCKU Hospital(TW), SMILE Lab at National Cheng Kung University NCKU (TW), BDE (TW)

Abstract: SafEE project aims at investigating technologies for stimulation and intervention for Alzheimer patients. More precisely, the main goals are: (1) to focus on specific clinical targets in three domains behavior, motricity and cognition (2) to merge assessment and non pharmacological help/intervention and (3) to propose 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.

9.2.1.3. 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 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.

9.2.2. FUI

9.2.2.1. 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 Hospitaliere Sainte-Marie.

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

9.2.2.2. StoreConnect

Program: FUI

Project acronym: StoreConect.

Project title: StoreConnect.

Duration: September 2016 - September 2018.

Coordinator: Ubudu (Paris).

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

Abstract: StoreConnect is an FUI project started in 2016 and will end in 2018. The goal 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.

9.2.2.3. ReMinAry

Program: FUI

Project acronym: ReMinAry.

Project title: ReMinAry.

Duration: September 2016 - September 2019.

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 reeducative approach, through clinical trials centered on two pathologies with immobilization: post-traumatic (surgery of the shoulder) and neurodegenerative (amyotrophic lateral sclerosis).

9.3. European Initiatives

9.3.1. FP7 & H2020 Projects

9.3.1.1. CENTAUR

Title: Crowded ENvironments moniToring for Activity Understanding and Recognition

Programm: FP7

Duration: January 2013 - December 2016

Coordinator: Honeywell

Partners:

Ecole Polytechnique Fédérale de Lausanne (Switzerland)

Honeywell, Spol. S.R.O (Czech Republic)

Neovision Sro (Czech Republic)

Queen Mary University of London (United Kingdom)

Inria contact: François Brémond

We aim to develop a network of scientific excellence addressing research topics in computer vision and advancing the state of the art in video surveillance. The cross fertilization of ideas and technology between academia, research institutions and industry will lay the foundations to new methodologies and commercial solutions for monitoring crowded scenes. Research activities will be driven by specific sets of scenarios, requirements and datasets that reflect security operators' needs for guaranteeing the safety of EU citizens. CENTAUR gives a unique opportunity to academia to be exposed to real life dataset, while enabling the validation of state-of-the-art video surveillance methodology developed at academia on data that illustrate real operational scenarios. The research agenda is motivated by ongoing advanced research activities in the participating entities. With Honeywell as a multi-industry partner, with security technologies developed and deployed in both its Automation and Control Solutions and Aerospace businesses, we have multiple global channels to exploit the developed technologies. With Neovison as a SME, we address small fast paced local markets, where the quick assimilation of new technologies is crucial. Three thrusts identified will enable the monitoring of crowded scenes, each led by an academic partner in collaboration with scientists from Honeywell: a) multi camera, multicoverage tracking of objects of interest, b) Anomaly detection and fusion of multimodal sensors, c) activity recognition and behavior analysis in crowded environments. We expect a long term impact on the field of video surveillance by: contributions to the state-of-the-art in the field, dissemination of results within the scientific and practitioners community, and establishing long term scientific exchanges between academia and industry, for a forum of scientific and industrial partners to collaborate on addressing technical challenges faced by scientists and the industry.'

9.3.2. Collaborations in European Programs, Except FP7 & H2020

Program: EIT Digital Activity

Project acronym: ELEMENT

Project title: Early detection of cognitive disorders on the basis of speech analysis

Duration: Jan 2017-Dec 2017

Coordinator: German Research Centre for Artificial Intelligence DFKI (Germany)

Other partners: Inria, Association Innovation Alzheimer (France) and University of Edinburgh (UK)

Abstract: ELEMENT is a new Innovation Activity to facilitate faster, earlier diagnosis and intervention for cognitive decline. The project aims to bring a unique new product o the European market that enables light-touch screening for cognitive decline in non-clinical settings, resulting in faster, earlier diagnosis and intervention.

9.4. International Initiatives

9.4.1. Informal International Partners

- **Collaborations with Asia:** Stars has been cooperating with the Multimedia Research Center in Hanoi MICA on semantics extraction from multimedia data. Stars also collaborates with the National Cheng Kung University in Taiwan and I2R in Singapore.
- **Collaboration with U.S.A.:** Stars collaborates with the University of Southern California.
- **Collaboration with Europe:** Stars collaborates with Multitel in Belgium, the University of Kingston upon Thames UK, and the University of Bergen in Norway.

9.4.2. Other IIL projects

9.4.2.1. The ANR SafEE (see section)

Stars collaborates with international partners such as Taichung Veterans General Hospital TVGH (TW), NCKU Hospital(TW), SMILE Lab at National Cheng Kung University NCKU (TW) and BDE (TW).

9.5. International Research Visitors

9.5.1. Visits of International Scientists

This year, Stars has been visited by the following international scientists:

- Salwa Baabou, Ecole Nationale d'Ingénieurs de Gabès, Tunisia;
- Adlen Kerboua, University of Skikda, Algeria;

9.5.1.1. Internships

Abhishek Goel

Date: Aug 2017-Dec 2017

Institution: BITS Pilani, Rajasthan, India

Supervisor: Michal Koperski

Srijan Das

Date: Jan 2017- May 2017

Institution: National Institute of Technology, Rourkela, India

Supervisor: Michal Koperski

Salwa Babou

Date: Apr 2017-Sep 2017

Institution: Laboratoire d'Electroniques et des Technologies de l'Information, à l'ENIS, SFAX, Tunisia

Supervisor: Francçois Brémond

Yu-Fen Chen

Date: Feb 2017-Aug 2017

Institution: National Tapei University of Technology, Tawain

Supervisor: Carlos Fernando Crispim Junior

Kuan-Ru Lee

Date: Aug 2017- Dec 2017

Institution: National Tapei University of Technology, Tawain

Supervisor: Carlos Fernando Crispim Junior

Chandraja Dharmana

Date: June 2017- Dec 2017

Institution: BITS Hyderabad, India

Supervisor: François Brémond

Shaira Kansal

Date: Jul 2017- Dec 2017

Institution: PEC, Chandigarh, India

Supervisor: Carlos Fernando Crispim Junior

Kartik Kartik

Date: Jul 2017- Dec 2017

Institution: PEC, Chandigarh, India

Supervisor: Carlos Fernando Crispim Junior

Rahul Pandey

Date: May 2017- Dec 2017

Institution: LMNIT, Rajasthan, India

Supervisor: Carlos Fernando Crispim Junior

Francesco Verrini

Date: Jun 2017- Dec 2017

Institution: Universita degli Studi di Genova, Italy

Supervisor: Carlos Fernando Crispim Junior, Michal Koperski

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. Member of the Organizing Committees

François Brémond was a member of the Management Committee and COST Action IC1307 in 2017.

10.1.2. Scientific Events Selection

10.1.2.1. Chair of Conference Program Committees

François Brémond was an Area and Session Chair of AVSS - 14th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Lecce, Italy, August 29th - September 1st, 2017.

François Brémond was an Area Chair of ICIAP - 19th International Conference on Image Analysis and Processing; Area Chair; in Catania, Italy, September 11-15, 2017.

Antitza Dantcheva was program and session chair at the 17th International Conference of the Biometrics Special Interest Group, BIOSIG 2017

10.1.2.2. Member of the Conference Program Committees

François Brémond was program committee member of the conferences and workshops: IWT4S 2017, COMSI17, SP-CROWD 2017, ICPRS-17, ICDP2017, AVSS17 and WACV17-18.

Antitza Dantcheva was program committee member of the conference International Conference on Biometrics (IJCB 2017).

Jean-Paul Rigault is a member of the *Association Internationale pour les Technologies à Objets* (AITO) which organizes international conferences such as ECOOP.

10.1.2.3. Reviewer

François Brémond was reviewer for the conferences : CVPR2017-18, ICCV2017, VISION4HCI17, WACV 2017, SAV17, ICVS 17, ISG 18, BER17, BMVC17, IROS17.

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- François Brémond was handling editor of the international journal "Machine Vision and Application".

10.1.3.2. Reviewer - Reviewing Activities

François Brémond was reviewer for IEEE Transactions on Circuits and Systems for Video Technology Editorial Office.

Monique Thonnat is a reviewer for the journals Artificial Intelligence in Medicine AIIM (Elsevier) and The Computer Journal (Oxford University Press).

Antitza Dantcheva reviewed for journals including IEEE Transactions on Information Forensics and Security (TIFS), The Visual Computer, Entropy, Pattern Recognition Letters.

10.1.4. Invited Talks

François Brémond was invited by Huawei Research, within Video Intelligence Event, on People Re-ID, Dublin, 21st November 2017.

Antitza Dantcheva:“Personalized facial expression and activity recognition during mnemotherapy-sessions for patients with major neurocognitive disorders”, IBME, University of Oxford, UK, November 2017

Antitza Dantcheva: “New Advances in Soft Biometrics”, Biometrics Congress, London, UK, November 2017

Antitza Dantcheva: “Can a smile reveal your gender?”, Day of Biometrics, Caen, France, July 2017

Alexandra Koenig was invited by Dr. Dimitrios Kokkinakis to give a talk on ICT for monitoring dementia patients and in particuler Automatic Speech analysis for the assessment of cognitive impairment, University of Gothenburg, Sweden, 18 September 2017.

Alexandra Koening was invited by Dr Jessica Peter to give a talk on Automatic Analysis of the Semantic Verbal Fluency performance in elderly people, University of Bern, Swiss, 18 October 2017.

10.1.5. Leadership within the Scientific Community

Monique Thonnat is President of the University Hospital Institute IHU Liryc in Bordeaux on cardiac electrical disorders since 2015.

Antitza Dantcheva serves in the Technical Activities Committee of the IEEE Biometrics Council 2017.

10.1.6. Scientific Expertise

François Brémond was expert for a research program at the Campus for Research Excellence and Technological Enterprise (CREATE) from the National University of Singapore (NUS).

François Brémond was expert for the European CHIST-ERA ERA-NET Call 2016, for the project "SUPER CARE".

François Brémond was expert for the project call "Recherche - Enseignement Supérieur - Plate-formes Mutualisées" de la Région Nouvelle-Aquitaine, March 2017.

François Brémond was expert for the "Revue Retraite et société", April 2017.

Monique Thonnat was an expert to review COFECUB research project (Scientific cooperation between Brasil and France) for the French government.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Master: Annie Ressouche, Safety in Middleware for Internet of Things, 6h, niveau (M2), Polytech Nice School of Nice University.

Master : Inès Sarray, Safety in Middleware for Internet of Things, 6h, niveau (M2), Polytech Nice School of Nice University.

10.2.2. Supervision

PhD: Minh Khue Phan Tran, Maintaining the engagement of older adults with dementia while interacting with serious game, Cote d'Azur University, 20th of April, François Brémond.

PhD: Michal Koperski, Detecting critical human activities using RGB/RGBD cameras in home environment, François Brémond.

PhD: Auriane Gros, Evaluation and Specific Management of Emotional Disturbances with Activity Recognition Systems for Alzheimer patient, 4th of December, François Brémond.

PhD in progress : Thi Lan Anh Nguyen, Complex Activity Recognition from 3D sensors, Dec 2014, François Brémond.

PhD in progress: Ines Sarray, Activity Recognition System Design, Oct 2015, Sabine Moisan.

PhD in progress: Farhood Negin, People Detection for Activity Recognition using RGB-Depth Sensors, Jan 2015, François Brémond.

PhD in progress: Ujjwal Ujjwal, Pedestrian Detection to Dynamically Populate the Map of a Crossroad, Sep 2016, François Brémond.

PhD in progress: Srijan Das, Activity Recognition in videos by combining an ontology based language with CNN networks, Dec 2017, François Brémond.

10.2.3. Juries

François Brémond was jury member of Tenure Track Selection: committee member for permanent position, Mines School - ParisTech, September 2017

François Brémond was jury member of the mid-term review for 4 PhDs - UNED Doctoral Consortium, Mr. Artaso, Mr. Timón, Mr. Pérez, and Mr. Navares, Facultad de Informática, UNED, Madrid, 12 June 2017

François Brémond was jury member of the following PhD theses:

PhD, Andrews Sobral, Université La Rochelle, 11 mai 2017.

PhD, Athira Muraleedharan Nambiar, Electrical and Computer Engineering, Instituto Superior Técnico, Institute for Systems and Robotics, Lisboa, Portugal, 1st September 2017.

PhD, Enjie Ghorbel, Université de Rouen Normandie, 12 octobre 2017.

PhD, Bassem HadjKacem, École Nationale d'Ingénieurs de Sfax, Tunisie, 5 novembre 2017.

PhD, Nabila Mansouri, Université de Sfax et l'Université de Valenciennes, 20 novembre 2017.

Monique Thonnat was member of the final selection board (jury d'admission) of junior (CR) and senior (DR) research scientists of (Cérema, ENPC, ENTPE, IFSTTAR, IGN, Météo-France).

Monique Thonnat is the research representative of Ecole Polytechnique jury for delivering the Prize of best student scientific projects since June 2011.

Monique Thonnat is member of the scientific board of ENPC, Ecole Nationale des Ponts et Chaussees since June 2008.

Sabine Moisan was reviewer in the PhD jury of Hassan Loulou, ESTACA, Paris nov. 2017

11. Bibliography

Major publications by the team in recent years

[1] A. AVANZI, F. BRÉMOND, C. TORNIERI, M. THONNAT. *Design and Assessment of an Intelligent Activity Monitoring Platform*, in "EURASIP Journal on Applied Signal Processing, Special Issue on "Advances in Intelligent Vision Systems: Methods and Applications"", August 2005, vol. 2005:14, pp. 2359-2374

[2] B. BOULAY, F. BRÉMOND, M. THONNAT. *Applying 3D Human Model in a Posture Recognition System*, in "Pattern Recognition Letter", 2006, vol. 27, n^{o} 15, pp. 1785-1796

[3] F. BRÉMOND, M. THONNAT. *Issues of Representing Context Illustrated by Video-surveillance Applications*, in "International Journal of Human-Computer Studies, Special Issue on Context", 1998, vol. 48, pp. 375-391

[4] N. CHLEQ, F. BRÉMOND, M. THONNAT. *Advanced Video-based Surveillance Systems*, Kluwer A.P. , Hangham, MA, USA, November 1998, pp. 108-118

[5] F. CUPILLARD, F. BRÉMOND, M. THONNAT. *Tracking Group of People for Video Surveillance*, Video-Based Surveillance Systems, Kluwer Academic Publishers, 2002, vol. The Kluwer International Series in Computer Vision and Distributed Processing, pp. 89-100

[6] F. FUSIER, V. VALENTIN, F. BRÉMOND, M. THONNAT, M. BORG, D. THIRDE, J. FERRYMAN. *Video Understanding for Complex Activity Recognition*, in "Machine Vision and Applications Journal", 2007, vol. 18, pp. 167-188

[7] B. GEORIS, F. BRÉMOND, M. THONNAT. *Real-Time Control of Video Surveillance Systems with Program Supervision Techniques*, in "Machine Vision and Applications Journal", 2007, vol. 18, pp. 189-205

[8] C. LIU, P. CHUNG, Y. CHUNG, M. THONNAT. *Understanding of Human Behaviors from Videos in Nursing Care Monitoring Systems*, in "Journal of High Speed Networks", 2007, vol. 16, pp. 91-103

[9] N. MAILLOT, M. THONNAT, A. BOUCHER. *Towards Ontology Based Cognitive Vision*, in "Machine Vision and Applications (MVA)", December 2004, vol. 16, n^{o} 1, pp. 33-40

[10] V. MARTIN, J.-M. TRAVERE, F. BRÉMOND, V. MONCADA, G. DUNAND. *Thermal Event Recognition Applied to Protection of Tokamak Plasma-Facing Components*, in "IEEE Transactions on Instrumentation and Measurement", Apr 2010, vol. 59, n^{o} 5, pp. 1182-1191

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

[12] S. MOISAN. *Une plate-forme pour une programmation par composants de systèmes à base de connaissances*, Université de Nice-Sophia Antipolis, April 1998, Habilitation à diriger les recherches

[13] 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

[14] A. RESSOUCHE, D. GAFFÉ, V. ROY. *Modular Compilation of a Synchronous Language*, in "Software Engineering Research, Management and Applications", R. LEE (editor), Studies in Computational Intelligence, Springer, 2008, vol. 150, pp. 157-171, selected as one of the 17 best papers of SERA'08 conference

[15] 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

[16] 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

[17] M. THONNAT. *Vers une vision cognitive: mise en oeuvre de connaissances et de raisonnements pour l'analyse et l'interprétation d'images*, Université de Nice-Sophia Antipolis, October 2003, Habilitation à diriger les recherches

[18] M. THONNAT. *Special issue on Intelligent Vision Systems*, in "Computer Vision and Image Understanding", May 2010, vol. 114, n^{o} 5, pp. 501-502

[19] V. VU, F. BRÉMOND, M. THONNAT. *Temporal Constraints for Video Interpretation*, in "Proc of the 15th European Conference on Artificial Intelligence", Lyon, France, 2002

[20] V. VU, F. BRÉMOND, 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

[21] N. ZOUBA, F. BRÉMOND, A. ANFOSSO, M. THONNAT, E. PASCUAL, O. GUERIN. *Monitoring elderly activities at home*, in "Gerontechnology", May 2010, vol. 9, n^{o} 2

Publications of the year

Doctoral Dissertations and Habilitation Theses

[22] M. KOPERSKI. *Human Action Recognition in Videos with Local Representation*, Universite Cote d'Azur, November 2017, https://hal.inria.fr/tel-01648968

[23] M. K. PHAN TRAN. *Maintaining the engagement of older adults with dementia while interacting with serious game*, Université Côte d'Azur, April 2017, https://tel.archives-ouvertes.fr/tel-01578114

Articles in International Peer-Reviewed Journals

[24] F. C. CRISPIM-JUNIOR, A. GÓMEZ URÍA, C. STRUMIA, M. KOPERSKI, A. KONIG, F. NEGIN, S. COSAR, A.-T. NGHIEM, G. CHARPIAT, F. BREMOND, D. P. CHAU. *Online recognition of daily activities by color-depth sensing and knowledge models*, in "Sensors", June 2017, vol. 17, n^{o} 7, pp. 1-15 [*DOI :* 10.3390/s17071528], https://hal.inria.fr/hal-01658438

[25] A. KONIG, L. KLAMING, M. PIJL, A. DEMEURRAUX, R. DAVID, P. ROBERT. *Objective measurement of gait parameters in healthy and cognitively impaired elderly using the dual-task paradigm*, in "Aging Clinical and Experimental Research", December 2017, vol. 29, n^{o} 6, pp. 1181 - 1189 [*DOI :* 10.1007/s40520-016-0703-6], https://hal.inria.fr/hal-01672597

International Conferences with Proceedings

[26] *Best Paper*
C. CHEN, A. DANTCHEVA, T. SWEARINGEN, A. ROSS. *Spoofing Faces Using Makeup: An Investigative Study*, in "IEEE International Conference on Identity, Security and Behavior Analysis 2017", New Delhi, India, February 2017, https://hal.archives-ouvertes.fr/hal-01430020.

[27] C. F. CRISPIM-JUNIOR, J. VLASSELAER, A. DRIES, F. BREMOND. *BEHAVE - Behavioral analysis of visual events for assisted living scenarios*, in "Assisted Computer Vision and Robotics workshop in conjunction with International Conference on Computer Vision", Venice, Italy, October 2017, https://hal.inria.fr/hal-01658665

[28] S. DAS, M. KOPERSKI, F. BREMOND, G. FRANCESCA. *Action Recognition based on a mixture of RGB and Depth based skeleton*, in "AVSS 2017 - 14-th IEEE International Conference on Advanced Video and Signal-Based Surveillance", Lecce, Italy, August 2017, https://hal.inria.fr/hal-01639504

[29] E. DE MARIA, D. GAFFÉ, A. RESSOUCHE, C. GIRARD RIBOULLEAU. *A Model-checking Approach to Reduce Spiking Neural Networks*, in "BIOINFORMATICS 2018 - 9th International Conference on Bioinformatics Models, Methods and Algorithms", Funchal Madeira, Portugal, January 2018, pp. 1-8, https://hal.archives-ouvertes.fr/hal-01638248

[30] E. DE MARIA, T. L 'YVONNET, D. GAFFÉ, A. RESSOUCHE, F. GRAMMONT. *Modelling and Formal Verification of Neuronal Archetypes Coupling* , in "CSBio 2017 - 8th International Conference on Computational Systems-Biology and Bioinformatics", Nha Trang, Vietnam, CSBio '17 Proceedings of the 8th International Conference on Computational Systems-Biology and Bioinformatics, ACM, December 2017, vol. 17, pp. 3-10 [*DOI :* 10.1145/3156346.3156348], https://hal.inria.fr/hal-01643862

[31] F. M. KHAN, F. BREMOND. *Multi-shot Person Re-Identification Using Part Appearance Mixture*, in "WACV 2017 - IEEE Winter Conference on Applications of Computer Vision", Santa Rosa, CA, United States, IEEE, March 2017, pp. 605-614 [*DOI :* 10.1109/WACV.2017.73], https://hal.inria.fr/hal-01654916

[32] N. LINZ, J. TRÖGER, J. ALEXANDERSSON, A. KONIG. *Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task*, in "IWCS 2017 - 12th International Conference on Computational Semantics", Montpellier, France, September 2017, pp. 1-7, https://hal.inria.fr/hal-01672593

[33] N. LINZ, J. TRÖGER, J. ALEXANDERSSON, A. KONIG, P. ROBERT, M. WOLTERS. *Predicting Dementia Screening and Staging Scores From Semantic Verbal Fluency Performance*, in "ICDM 2017 - IEEE International Conference on Data Mining, Workshop on Data Mining for Aging, Rehabilitation and Independent Assisted Living", New Orleans, United States, November 2017, pp. 719-728 [*DOI :* 10.1109/ICDMW.2017.100], https://hal.inria.fr/hal-01672590

[34] I. SARRAY, A. RESSOUCHE, S. MOISAN, J.-P. RIGAULT, D. GAFFÉ. *An Activity Description Language for Activity Recognition*, in "IINTEC 2017 - IEEE International Conference on Internet of Things, Embedded Systems and Communications", Gafsa, Tunisia, October 2017, https://hal.inria.fr/hal-01649674

[35] N. THI LAN ANH, F. M.KHAN, F. NEGIN, F. BREMOND. *Multi-Object tracking using Multi-Channel Part Appearance Representation*, in "AVSS 2017 : 14-th IEEE International Conference on Advanced Video and Signal-Based Surveillance", Lecce , Italy, August 2017, https://hal.inria.fr/hal-01651938

[36] S. YOON, F. M. KHAN, F. BREMOND. *Efficient Video Summarization Using Principal Person Appearance for Video-Based Person Re-Identification*, in "The British Machine Vision Conference (BMVC)", London, United Kingdom, September 2017, https://hal.inria.fr/hal-01593238

Conferences without Proceedings

[37] A. DANTCHEVA, P. BILINSKI, H. T. NGUYEN, J.-C. BROUTART, F. BREMOND. *Expression Recognition for Severely Demented Patients in Music Reminiscence-Therapy*, in "European Signal Processing Conference (EUSIPCO)", Kos island, Greece, August 2017, 5 p. , https://hal.archives-ouvertes.fr/hal-01543231

Research Reports

[38] I. SARRAY, A. RESSOUCHE, S. MOISAN, J.-P. RIGAULT, D. GAFFÉ. *Synchronous Automata For Activity Recognition*, Inria Sophia Antipolis, April 2017, n^{o} RR-9059, STARS,MCSOC, https://hal.inria.fr/hal-01505754

[39] R. TRICHET, F. BREMOND. *Dataset Optimization for Real-Time Pedestrian Detection*, Inria Sophia-Antipolis, June 2017, n^{o} RR-9084, 15 p. , https://hal.inria.fr/hal-01566517

Other Publications

[40] C. GIRARD RIBOULLEAU. *Modèles probabilistes et vérification de réseaux de neurones*, Université Nice - Sophia-Antipolis, June 2017, https://hal.inria.fr/hal-01550133

[41] A. RESSOUCHE, D. GAFFÉ, D. HAVAYARIMANA. *Études et développement de diagrammes de décision linéaires*, UNSA, September 2017, pp. 1-36, https://hal.inria.fr/hal-01665717

References in notes

[42] M. ACHER, P. COLLET, F. FLEUREY, P. LAHIRE, S. MOISAN, J.-P. RIGAULT. *Modeling Context and Dynamic Adaptations with Feature Models*, in "Models@run.time Workshop", Denver, CO, USA, October 2009, http://hal.inria.fr/hal-00419990/en

[43] M. ACHER, P. LAHIRE, S. MOISAN, J.-P. RIGAULT. *Tackling High Variability in Video Surveillance Systems through a Model Transformation Approach*, in "ICSE'2009 - MISE Workshop", Vancouver, Canada, May 2009, http://hal.inria.fr/hal-00415770/en

[44] S. H. BAE, K. J. YOON. *Robust Online Multi-object Tracking Based on Tracklet Confidence and Online Discriminative Appearance Learning*, in "2014 CVPR", June 2014, pp. 1218-1225, http://dx.doi.org/10.1109/CVPR.2014.159

[45] A. BAR-HILLEL, D. LEVI, E. KRUPKA, C. GOLDBERG. *Part-based feature synthesis for human detection*, in "ECCV", 2010

[46] R. BENENSON, M. MATHIAS, T. TUYTELAARS, L. V. GOOL. *Seeking the Strongest Rigid Detector*, in "CVPR", 2013

[47] R. BENENSON, M. OMRAN, J. HOSANG, B. SCHIELE. *Ten years of pedestrian detection, what have we learned?*, in "ECCV, CVRSUAD workshop", 2014

[48] B. BERKIN, B. K. HORN, I. MASAKI. *Fast Human Detection With Cascaded Ensembles On The GPU*, in "IEEE Intelligent Vehicles Symposium", 2010

[49] P. BILINSKI, M. KOPERSKI, S. BAK, F. BRÉMOND. *Representing Visual Appearance by Video Brownian Covariance Descriptor for Human Action Recognition*, in "AVSS - 11th IEEE International Conference on Advanced Video and Signal-Based Surveillance", Seoul, South Korea, IEEE, August 2014, https://hal.inria.fr/hal-01054943

[50] G. BRAZIL, X. YIN, X. LIU. *Illuminating Pedestrians via Simultaneous Detection & Segmentation*, in "Proceedings of the IEEE International Conference on Computer Vision", Venice, Italy, 2017

[51] Z. CAO, T. SIMON, S.-E. WEI, Y. SHEIKH. *Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields*, in "arXiv preprint arXiv:1611.08050", 2016

[52] Z. CAO, T. SIMON, S.-E. WEI, Y. SHEIKH. *Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields*, in "CVPR", 2017

[53] S. CHAN-LANG, Q. PHAM, C. ACHARD. *Bidirectional Sparse Representations for Multi-Shot Person Re-identification*, in "AVSS", 2016

[54] G. CHEN, Y. DING, J. XIAO, T. X. HAN. *Detection Evolution with Multi-Order Contextual Co-occurrence*, in "CVPR", 2013

[55] G. CHÉRON, I. LAPTEV, C. SCHMID. *P-CNN: Pose-based CNN Features for Action Recognition*, in "ICCV", 2015

[56] E. CORVEE, F. BREMOND. *Haar like and LBP based features for face, head and people detection in video sequences*, in "International Workshop on Behaviour Analysis and Video Understanding (ICVS 2011)", Sophia Antipolis, France, September 2011, 10 p. , https://hal.inria.fr/inria-00624360

[57] A. D. COSTEA, S. NEDEVSCHI. *Word Channel Based Multiscale Pedestrian Detection without Image Resizing and Using Only One Classifier*, in "CVPR", 2014

[58] 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

[59] N. DALAL, B. TRIGGS. *Histograms of oriented gradients for human detection*, in "CVPR", 2005

[60] N. DALAL, B. TRIGGS, C. SCHMID. *Human detection using oriented histograms of flow and appearance*, in "European conference on computer vision", Springer, 2006, pp. 428–441

[61] A. DANTCHEVA, F. BRÉMOND. *Gender estimation based on smile-dynamics*, in "IEEE Transactions on Information Forensics and Security", 2016, 11 p. [*DOI :* 10.1109/TIFS.2016.2632070], https://hal.archives-ouvertes.fr/hal-01412408

[62] R. DAVID, E. MULIN, P. MALLEA, P. ROBERT. *Measurement of Neuropsychiatric Symptoms in Clinical Trials Targeting Alzheimer's Disease and Related Disorders*, in "Pharmaceuticals", 2010, vol. 3, pp. 2387-2397

[63] E. DE MARIA, A. MUZY, D. GAFFÉ, A. RESSOUCHE, F. GRAMMONT. *Verification of Temporal Properties of Neuronal Archetypes Modeled as Synchronous Reactive Systems*, in "HSB 2016 - 5th International Workshop Hybrid Systems Biology", Grenoble, France, Lecture Notes in Bioinformatics series, October 2016, 15 p. [*DOI :* 10.1007/978-3-319-47151-8_7], https://hal.inria.fr/hal-01377288

[64] P. DOLLAR, S. BELONGIE, P. PERONA. *The Fastest Pedestrian Detector in the West*, in "BMVC", 2010

[65] P. DOLLAR, Z. TU, P. PERONA, S. BELONGIE. *Integral channel features*, in "BMVC", 2009

[66] P. DOLLÁR, R. APPEL, W. KIENZLE. *Crosstalk Cascades for Frame-Rate Pedestrian Detection*, in "ECCV", 2012

[67] N. DVORNIK, K. SHMELKOV, J. MAIRAL, C. SCHMID. *BlitzNet: A Real-Time Deep Network for Scene Understanding*, in "IEEE International Conference on Computer Vision (ICCV)", 2017

[68] M. FARENZENA, L. BAZZANI, A. PERINA, V. MURINO, M. CRISTANI. *Person re-identification by symmetry-driven accumulation of local features*, in "CVPR", 2010

[69] P. F. FELZENSZWALB, R. B. GIRSHICK, D. MCALLESTER. *Cascade object detection with deformable part models*, in "CVPR", 2010

[70] P. FELZENSZWALB, R. GIRSHICK, D. MCALLESTER, D. RAMANAN. *Object Detection with Discriminatively Trained Part-Based Models*, in "PAMI", 2009, vol. 32, n^{o} 9, pp. 1627–1645

[71] P. FELZENSZWALB, D. MCALLESTER, D. RAMANAN. *A discriminatively trained, multiscale, deformable part model*, in "Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on", IEEE, 2008, pp. 1–8

[72] A. GAIDON, Z. HARCHAOUI, C. SCHMID. *Temporal Localization of Actions with Actoms*, in "IEEE Trans. Pattern Anal. Mach. Intell.", 2013, vol. 35, n^{o} 11, pp. 2782–2795, https://doi.org/10.1109/TPAMI.2013.65

[73] C. GAO, J. WANG, L. LIU, J.-G. YU, N. SANG. *Temporally aligned pooling representation for video-based person re-identification*, in "ICIP", 2016, pp. 4284–4288

[74] R. GIRSHICK, J. DONAHUE, T. DARRELL, J. MALIK. *Rich feature hierarchies for accurate object detection and semantic segmentation*, in "Computer Vision and Pattern Recognition", 2014

[75] R. GIRSHICK. *Fast r-cnn*, in "Proceedings of the IEEE international conference on computer vision", 2015, pp. 1440–1448

[76] D. GRAY, H. TAO. *Viewpoint invariant pedestrian recognition with an ensemble of localized features*, in "Computer Vision–ECCV", 2008, pp. 262–275

[77] M. HIRZER, C. BELEZNAI, P. M. ROTH, H. BISCHOF. *Person Re-identification by Descriptive and Discriminative Classification*, in "Image Analysis", Springer, 2011, pp. 91–102

[78] M. HIRZER, P. ROTH, M. KÖSTINGER, H. BISCHOF. *Relaxed pairwise learned metric for person re-identification*, in "ECCV", 2012

[79] S. HOCHREITER, J. SCHMIDHUBER. *Long short-term memory*, in "Neural computation", 1997, vol. 9, n^{o} 8, pp. 1735–1780

[80] S. KARANAM, Y. LI, R. J. RADKE. *Person Re-Identification with Discriminatively Trained Viewpoint Invariant Dictionaries*, in "ICCV", 2015

[81] W. KE, Y. ZHANG, P. WEI, Q. YE, J. JIAO. *Pedestrian detection via PCA filters based convolutional channel features*, in "ICASSP", 2015

[82] M. KOPERSKI, P. BILINSKI, F. BRÉMOND. *3D Trajectories for Action Recognition*, in "ICIP - The 21st IEEE International Conference on Image Processing", Paris, France, IEEE, October 2014, https://hal.inria.fr/hal-01054949

[83] M. KOPERSKI, F. BREMOND. *Modeling Spatial Layout of Features for Real World Scenario RGB-D Action Recognition*, in "AVSS 2016", Colorado Springs, United States, August 2016, pp. 44 - 50 [*DOI :* 10.1109/AVSS.2016.7738023], https://hal.inria.fr/hal-01399037

[84] M. KOSTINGER, M. HIRZER, P. WOHLHART, P. M. ROTH, H. BISCHOF. *Large scale metric learning from equivalence constraints*, in "2012 CVPR", June 2012, pp. 2288-2295, http://dx.doi.org/10.1109/CVPR.2012.6247939

[85] C. KÄSTNER, S. APEL, S. TRUJILLO, M. KUHLEMANN, D. BATORY. *Guaranteeing Syntactic Correctness for All Product Line Variants: A Language-Independent Approach*, in "TOOLS (47)", 2009, pp. 175-194

[86] R. LAYNE, T. M. HOSPEDALES, S. GONG, Q. MARY. *Person Re-identification by Attributes*, in "Bmvc", 2012, vol. 2, n^{o} 3, 8 p.

[87] Y. LECUN, L. BOTTOU, Y. BENGIO, P. HAFFNER. *Gradient-based learning applied to document recognition*, in "Proceedings of the IEEE", 1998, vol. 86, n^{o} 11, pp. 2278–2324

[88] Y. LI, Z. WU, S. KARANAM, R. RADKE. *Multi-Shot Human Re-identification Using Adaptive Fisher Discriminant Analysis*, in "BMVC", 2015

[89] S. LIAO, Y. HU, S. Z. LI. *Joint Dimension Reduction and Metric Learning for Person Re-identification*, in "CoRR", 2014, vol. abs/1406.4216, http://arxiv.org/abs/1406.4216

[90] S. LIAO, Y. HU, X. ZHU, S. Z. LI. *Person Re-identification by Local Maximal Occurrence Representation and Metric Learning*, in "CVPR", 2015

[91] W. LIU, D. ANGUELOV, D. ERHAN, C. SZEGEDY, S. E. REED, C. FU, A. C. BERG. *SSD: Single Shot MultiBox Detector*, in "CoRR", 2015, vol. abs/1512.02325, http://arxiv.org/abs/1512.02325

[92] W. LIU, D. ANGUELOV, D. ERHAN, C. SZEGEDY, S. REED, C.-Y. FU, A. C. BERG. *SSD: Single Shot MultiBox Detector*, in "ECCV", 2016

[93] K. LIU, W. ZHANG, R. HUANG. *A Spatio-Temporal Appearance Representation for Video-Based Pedestrian Re-Identification*, in "ICCV", 2015

[94] B. MA, Y. SU, F. JURIE. *Local descriptors encoded by Fisher descriptors for person re-identification*, in "ECCV Workshops", 2012

[95] J. MARIN, D. VAZQUEZ, A. M. LOPEZ, J. AMORES, B. LEIBE. *Random Forests of Local Experts for Pedestrian Detection*, in "ICCV", 2013

[96] N. MCLAUGHLIN, J. M. DEL RINCON, P. MILLER. *Recurrent Convolutional Network for Video-based Person Re-identification*, in "CVPR", 2016

[97] P. METTES, J. C. VAN GEMERT, S. CAPPALLO, T. MENSINK, C. G. M. SNOEK. *Bag-of-Fragments: Selecting and Encoding Video Fragments for Event Detection and Recounting*, in "Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, June 23-26, 2015", A. G. HAUPTMANN, C. NGO, X. XUE, Y. JIANG, C. SNOEK, N. VASCONCELOS (editors), ACM, 2015, pp. 427–434, http://doi.acm.org/10.1145/2671188.2749404

[98] S. MOISAN, J.-P. RIGAULT, M. ACHER, P. COLLET, P. LAHIRE. *Run Time Adaptation of Video-Surveillance Systems: A software Modeling Approach*, in "ICVS, 8th International Conference on Computer Vision Systems", Sophia Antipolis, France, September 2011, http://hal.inria.fr/inria-00617279/en

[99] T. OJALA, M. PIETIKAINEN, D. HARWOOD. *A Comparative Study of Texture Measures with Classification Based on Feature Distributions*, in "Pattern Recognition", 1996, vol. 29, n^{o} 3, pp. 51–59

[100] T. OJALA, M. PIETIKAINEN, T. MAENPAA. *Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns*, in "PAMI", 2002

[101] D. ONEATA, J. VERBEEK, C. SCHMID. *The LEAR submission at Thumos 2014*, HAL CCSD, 2014

[102] W. OUYANG, X. WANG. *A discriminative deep model for pedestrian detection with occlusion handling*, in "Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on", IEEE, 2012, pp. 3258–3265

[103] W. OUYANG, X. WANG. *Joint Deep Learning for Pedestrian Detection*, in "ICCV", 2013

[104] S. PAISITKRIANGKRAI, C. SHEN, A. VAN DEN HENGEL. *Strengthening the Effectiveness of Pedestrian Detection with Spatially Pooled Features*, in "ECCV", 2014

[105] G. PAVLAKOS, X. ZHOU, K. G. DERPANIS, K. DANIILIDIS. *Coarse-to-fine volumetric prediction for single-image 3D human pose*, in "2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", IEEE, 2017, pp. 1263–1272

[106] S. PEDAGADI, J. ORWELL, S. VELASTIN, B. BOGHOSSIAN. *Local Fisher discriminant analysis for pedestrian re-identification*, in "CVPR", 2013

[107] L. PISHCHULIN, E. INSAFUTDINOV, S. TANG, B. ANDRES, M. ANDRILUKA, P. V. GEHLER, B. SCHIELE. *Deepcut: Joint subset partition and labeling for multi person pose estimation*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", 2016, pp. 4929–4937

[108] F. POIESI, R. MAZZON, A. CAVALLARO. *Multi-target tracking on confidence maps: An application to people tracking*, in "Computer Vision and Image Understanding", 2013, vol. 117, n^{o} 10, pp. 1257 - 1272 [*DOI :* 10.1016/J.CVIU.2012.08.008], http://www.sciencedirect.com/science/article/pii/S1077314212001634

[109] J. REDMON, A. FARHADI. *YOLO9000: Better, Faster, Stronger*, in "arXiv preprint arXiv:1612.08242", 2016

[110] S. REN, K. HEN, R. GIRSHICK, J. SUN. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*, in "NIPS", 2015

[111] L. M. ROCHA, S. MOISAN, J.-P. RIGAULT, S. SAGAR. *Girgit: A Dynamically Adaptive Vision System for Scene Understanding*, in "ICVS", Sophia Antipolis, France, September 2011, http://hal.inria.fr/inria-00616642/en

[112] R. ROMDHANE, E. MULIN, A. DERREUMEAUX, N. ZOUBA, J. PIANO, L. LEE, I. LEROI, P. MALLEA, R. DAVID, M. THONNAT, F. BREMOND, P. ROBERT. *Automatic Video Monitoring system for assessment of Alzheimer's Disease symptoms*, in "The Journal of Nutrition, Health and Aging Ms(JNHA)", 2011, vol. JNHA-D-11-00004R1, http://hal.inria.fr/inria-00616747/en

[113] A. ROSHAN ZAMIR, A. DEHGHAN, M. SHAH. *GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs*, in "Proceedings of the European Conference on Computer Vision (ECCV)", 2012

[114] A. SHAHROUDY, J. LIU, T.-T. NG, G. WANG. *NTU RGB+ D: A large scale dataset for 3D human activity analysis*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", 2016, pp. 1010–1019

[115] G. SHU, A. DEHGHAN, O. OREIFEJ, E. HAND, M. SHAH. *Part-based multiple-person tracking with partial occlusion handling*, in "2012 IEEE Conference on Computer Vision and Pattern Recognition", June 2012, pp. 1815-1821, http://dx.doi.org/10.1109/CVPR.2012.6247879

[116] X. SHU, F. PORIKLI, N. AHUJA. *Robust orthonormal subspace learning: Efficient recovery of corrupted low-rank matrices*, in "CVPR", 2014

[117] K. SIMONYAN, A. ZISSERMAN. *Very Deep Convolutional Networks for Large-Scale Image Recognition*, in "CoRR", 2014, vol. abs/1409.1556

[118] M. SOUDED, F. BREMOND. *Optimized Cascade of Classifiers for People Detection Using Covariance Features*, in "International Conference on Computer Vision Theory and Applications (VISAPP)", Barcelona, Spain, February 2013, https://hal.inria.fr/hal-00794369

[119] C. SU, F. YANG, S. ZHANG, Q. TIAN, L. S. DAVIS, W. GAO. *Multi-task learning with low rank attribute embedding for person re-identification*, in "ICCV", 2015, pp. 3739–3747

[120] K. D. TANG, B. YAO, F. LI, D. KOLLER. *Combining the Right Features for Complex Event Recognition*, in "IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013", IEEE Computer Society, 2013, pp. 2696–2703, https://doi.org/10.1109/ICCV.2013.335

[121] N. THI LAN ANH, F. BREMOND, J. TROJANOVA. *Multi-Object Tracking of Pedestrian Driven by Context*, in "Advance Video and Signal-based Surveillance", Colorado Springs, United States, IEEE, August 2016, https://hal.inria.fr/hal-01383186

[122] VIOLA, M. JONES. *Rapid Object Detection Using a Boosted Cascade of Simple Features*, in "CVPR", 2001

[123] S. WALK, N. MAJER, K. SCHINDLER, B. SCHIELE. *New features and insights for pedestrian detection*, in "CVPR", 2010

[124] J. WAN, Y. ZHAO, S. ZHOU, I. GUYON, S. ESCALERA, S. Z. LI. *Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops", 2016, pp. 56–64

[125] T. WANG, S. GONG, X. ZHU, S. WANG. *Person Re-identification by Video Ranking*, in "ECCV", 2014

[126] T. WANG, S. GONG, X. ZHU, S. WANG. *Person Re-Identification by Discriminative Selection in Video Ranking*, in "T-PAMI", 2016

[127] X. WANG, T. X. HAN, S. YAN. *An hog-lbp human detector with partial occlusion handling*, in "ICCV", 2009

[128] H. WANG, C. SCHMID. *Action recognition with improved trajectories*, in "Proceedings of the IEEE International Conference on Computer Vision", 2013, pp. 3551–3558

[129] C. WOJEK, B. SCHIELE. *A performance evaluation of single and multi-feature people detection*, in "DAGM Symposium Pattern Recognition", 2008

[130] T. XIAO, H. LI, W. OUYANG, X. WANG. *Learning deep feature representations with domain guided dropout for person re-identification*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", 2016, pp. 1249–1258

[131] Y. YAN, B. NI, Z. SONG, C. MA, Y. YAN, X. YANG. *Person Re-identification via Recurrent Feature Aggregation*, in "ECCV", 2016

[132] X. YANG, Y. TIAN. *Effective 3d action recognition using eigenjoints*, in "Journal of Visual Communication and Image Representation", 2014, vol. 25, n^{o} 1, pp. 2–11

[133] J. YOU, A. WU, X. LI, W. ZHENG. *Top-push Video-based Person Re-identification*, in "CVPR", 2016

[134] J. YUAN, B. NI, X. YANG, A. A. KASSIM. *Temporal Action Localization with Pyramid of Score Distribution Features*, in "2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016", IEEE Computer Society, 2016, pp. 3093–3102, https://doi.org/10.1109/CVPR.2016.337

[135] M. ZANFIR, M. LEORDEANU, C. SMINCHISESCU. *The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection*, in "Proceedings of the IEEE International Conference on Computer Vision", 2013, pp. 2752–2759

[136] M. ZENG, Z. WU, C. TIAN, L. ZHANG, L. HU. *Efficient person re-identification by hybrid spatiogram and covariance descriptor*, in "2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)", June 2015, pp. 48-56, http://dx.doi.org/10.1109/CVPRW.2015.7301296

[137] S. ZHANG, C. BAUCKHAGE, A. B. CREMERS. *Informed Haar-Like Features Improve Pedestrian Detection*, in "CVPR", 2014

[138] S. ZHANG, R. BENENSON, B. SCHIELE. *Filtered Channel Features for Pedestrian detection*, in "CVPR", 2015

[139] L. ZHANG, Y. LI, R. NEVATIA. *Global data association for multi-object tracking using network flows*, in "Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on", June 2008, pp. 1-8, http://dx.doi.org/10.1109/CVPR.2008.4587584

[140] L. ZHANG, L. LIN, X. LIANG, K. HE. *Is faster r-cnn doing well for pedestrian detection?*, in "European Conference on Computer Vision", Springer, 2016, pp. 443–457

[141] R. ZHAO, W. OUYANG, X. WANG. *Person re-identification by salience matching*, in "ICCV", 2013

[142] L. ZHENG, Z. BIE, Y. SUN, J. WANG, C. SU, S. WANG, Q. TIAN. *MARS: A Video Benchmark for Large-Scale Person Re-identification*, in "ECCV", 2016

[143] Q. ZHU, S. AVIDAN, M. YEH, K. CHENG. *Fast Human Detection using a Cascade of Histograms of Oriented Gradients*, in "CVPR", 2006