Inria Project-Team Proposal
STARS
Spatio-Temporal Activity Recognition of Social interactions
Reconnaissance d'Activités Spatio-Temporelles pour les interactions Sociales

Francois Bremond

March 2024

Theme: Vision, Perception and Multimedia Understanding

Domain: Perception, Cognition, Interaction.

Type: EPI
Inria Sophia Antipolis - Méditerranée
Team leader: Francois Bremond,
francois.bremond@inria.fr, www-sop.inria.fr/members/Francois.Bremond

Abstract

The new STARS research project-team will focus on the design of computer vision methods for real-time understanding of social interactions observed by sensors. Our objective is to propose new algorithms to analyze the behavior of people suffering from behavioral disorders, in order to improve their quality of life. We will study long-term spatio-temporal interactions performed by humans in their natural environment. We will address this challenge by proposing novel deep-learning architectures to model behavioral traits such as facial expression, gaze, gestures, body behavior, and body language. To cope with the limited amount of available data and the privacy issues of medical data, we propose data generation for data augmentation and anonymization. Another important challenge is to make the link between collected data, medical diagnosis, and ultimately treatments. To validate our research we will work closely with our clinical partners, in particular those of the Nice Hospital.

Keywords

Computer vision, Deep learning, Video understanding, Human behavior and cognition, Social interaction

Contents

1 The Team

Permanent scientific members

- **Team Leader**: François Brémond, DR1, HDR
- **Research scientists INRIA**:

  Michal Balazia, ISFP [from 01/10/2023][1]
  Antitza Dantcheva, CRCN, HDR
  Monique Thonnat, DR0-2, HDR (Emeritus from 01/06/2024)
- **Team Assistant**: Sandrine Boute

Non-permanent team members

- **Postdoc**:

  Baptiste Chopin (Inria Action Exploratoire XGAN) [01/07/2023 – 30/08/2025]
  Olivier Huynh (Inria Heroes project) [01/01/2023 – 30/11/2024]
- **PhD Students**:

  Tanay Agrawal (Inria, GAIN project) [01/10/2023 – 30/09/2026]
  Abid Ali (UCA) [01/05/2021 – 30/10/2024]
  Yuan Gao (Inrae, PEPR) [01/09/2023 – 31/08/2026]
  Mohammed Guermal (Inria, AirBus project) [01/12/2020 – 29/05/2024]
  Snehashis Mahji (Inria, Toyota project) [01/01/2023 – 31/12/2025]
  Thomasz Stanczyk (3IA) [01/02/2023 – 31/01/2026]
  Valeriya Strizhkova (3IA) [01/10/2021 – 30/09/2024]
  Di Yang (Inria, Toyota project) [01/11/2020 – 01/02/2024]
- **Research engineers INRIA**:

  Mahmoud Ali (Inria, ACTIVIS project) [01/05/2023 – 30/06/2024]

2 Context & Motivations

The new research project-team will focus on the long-term spatio-temporal interactions performed by humans in their natural environment. Our objective is to propose new algorithms to analyze the *behavior* of people suffering from *behavioral disorder*, in order to improve their quality of life. Deep learning techniques are highly successful for *simple* action recognition, nevertheless several important challenges remain in activity recognition in general and specifically for our target medical application domain.

To validate our research, we will work closely with our clinical partners. We have a **strategic partnership** named CoBTek[2] with the clinicians of Nice Hospital (CHU Nice) to study the impact of video understanding approaches for cognitive disorders. This partnership started in January 2012 and has evolved to a University Côte d'Azur team and

[1] Michal Balazia is at STARS since 01/09/2019 first as postdoc then SRP4ERC before getting a permanent position.

[2] See more information on this strategic partnership in section 4.2 Medical Collaborations.

joint work with monthly regular meetings between STARS and the clinicians of Institut Claude Pompidou (ICP), Lenval, and Pasteur hospitals. The two directors of CoBTek are François Brémond and Florence Askenazy (PU-PH) at Lenval. Our objective to deepen research in social interaction is motivated by the needs of our clinician partners. A typical use-case of social interactions observed by sensors appears in the clinical assessments of psychiatric patients, such as people suffering from conditions like major depression, bipolar disorder, or schizophrenia [1]. In these clinical assessments, interactions between the patient and the clinician are recorded with multi-modalities, i.e., with video, audio, and physiological sensors. The goal is to extract digital markers (defined by formal interaction models), which are indicators that can characterize a digital phenotype. The digital markers are automatically extracted from the recorded data and the digital phenotypes could lead to a treatment improving the patient's behavioral disorders.

Social interaction as a new study target. An abundance of valuable diagnostic relevant information is extracted from the interaction between clinician and patient. This clinical interaction (e.g., conversation between patient and clinician including verbal and nonverbal communication) is traditionally a clinician's most important source of information about patients' social skills, mood, and motivation levels. However, a comprehensive clinical interview requires sufficient consultation time as well as strong clinical competencies and expertise to be able to detect early subtle signs of changes in communication. Moreover, for detecting these changes during a clinical conversation, no standardized objective measures exist, leaving a lot of room for speculation and subjective biases. Introducing methodologies to assess in a quantitative manner behavioral dynamics during real-life social interaction could help indicate for instance level of reciprocity and therapeutic alliance, which until now is merely left to clinical intuition as we pointed out recently [1].

Need for precise and sensitive digital markers. To develop and test new measures of mental illness, a movement from traditional markers and phenotyping to digital markers and digital phenotyping is needed. 'Digital phenotyping' refers to the moment-to-moment quantification of human behavior in everyday life using data from digital devices. Digital phenotyping suggests collecting patient data allowing for non-intrusive and continuous monitoring of behavioural and mental states, ultimately revealing clinically relevant information. Similarly, 'digital markers' (e.g., frequency of eye contact) are digitally-obtained disease indicators that can be used to define a digital phenotype (e.g., eye gaze). Interaction-based phenotyping could provide various additional data to generate an observer-independent assessment of behavior during a social interaction which reflects as a mirror the current symptomology of a patient. Additionally, interaction-based measures such as social synchrony may have predictive value for treatment outcomes. Recent progress in computer vision, speech processing, and machine learning has enabled detailed and objective characterization of human interaction behavior [2]. Applying these advanced methods of artificial intelligence provides new opportunities to identify digital markers of patient behavior. Such markers have the potential to provide objective and continuous assessments of symptomatology in the context of patients' daily lives [3, 4], thereby allowing to precisely tailor treatment to the concrete patient trajectory. So far, many developed techniques are based solely on verbal information during interviews; however interpersonal communication often occurs non-verbally. Thus, merging computer vision-based measurement in a multi-modal approach would enhance the quality of analysis by allowing the detection of changes in the quality of communication as alterations in the dyadic interaction patterns.

Digital markers and methods. In recent years, behavior recognition methods based on artificial intelligence (i.e., machine or deep learning) have become increasingly effective in a variety of tasks, including action classification [5], body language and gestures [6], gaze estimation [7], eye contact detection, facial action units, facial expression [8], as well as affect extracted from single or multiple modalities [9]. A growing number of approaches make use of this progress in human behavior sensing to analyze clinical interaction data (e.g., therapy sessions), linguistic and paralinguistic characteristics from speech. As psychiatric disorders (depression, bipolar, schizophrenia) impact the quality of social interactions, there is an emphasis on studying these quantifiable behavioral dynamics in real-life social interaction at the dyadic level rather than solely individual behavior [1]. While these initial results are promising, this research needs to be accelerated by further development of digital phenotyping technology focusing on scalability and equity, by establishing shared longitudinal data repositories and by fostering multidisciplinary collaborations between clinical stakeholders, including patients, computer scientists, and medical researchers.

Sensors for analyzing human interactions. We are planning to keep using mainly RGB monocular cameras for video analysis. These off-the-shelf sensors are affordable, and very precise with a large dynamic range and high resolution. They are easily deployable in elderly homes and in hospitals. However, we will also investigate new types of sensors (e.g. RGB-D and infrared cameras, physiological sensors, and microphones) to capture complementary information and depending on the use-cases. These new sensors can open up new avenues of research. As we do not want to disturb the everyday activities of the end-users, we will first train our models with a large variety of sensors in dedicated locations, such as laboratories. Second, we will distill the learned weights into lighter models trained only with RGB video streams. These lighter RGB models are more convenient and less intrusive, as they can be processed only using standard RGB cameras. Third, we will use only these lighter RGB models at run-time in embedded devices directly at the end-users' locations. Therefore, we will only use the sensors and cameras pertinent to the end-users.

From STARS to the new team. M. Balazia, F. Bremond, A. Dantcheva and M. Thonnat are computer vision scientists and members of the STARS project-team (for more information see section 9 Short CVs of Researchers). STARS was created in January 2013 and will reach soon the 12 years duration limit. Two former members of the STARS team, S. Moisan and A. Ressouche, who specialized in software engineering, are now retired. The deep learning techniques we develop for activity recognition are based on standardized toolboxes and libraries available for the research community. Consequently, the STARS research direction on software engineering for activity recognition will not be pursued.

As detailed in section 3 *"Research Axes and Applications"*, the new team will have a new research axis on *Data Generation for Augmentation and Anonymisation.* In *activity recognition*, we have kept *Body Analysis*, expanded by *Face Analysis* and *Emotion Recognition.* We have also added a new subpart on *Multimodal Recognition.* In addition we now focus not just on the behavior of one individual but on the interactions between several persons.

3 Research Axes and Applications

Our research objective is related to the recognition of human actions, facial expressions, and body language in social interactions. Therefore we plan to work on two main research axes:

Axis 1: Human Interaction Recognition based on body and face analysis,

Axis 2: Data Generation for Augmentation and Anonymization for solving data limitation and privacy issues.

3.1 Axis 1: Human Interaction Recognition

Participants: F. Bremond, M. Balazia, A. Dantcheva, M. Thonnat

3.1.1 Body Language Analysis

Participants: F. Bremond, M. Balazia, M. Thonnat

Body language has been actively researched by psychologists for decades. Early work by Mehrabian [10] found that, among other signals, backward leaning of the torso is indicative of liking. A study by [11] indicated that people believe power is expressed with nonverbal cues like open posture (i.e., no arms crossed or legs crossed), more gesturing, and less self-touching (both hands and face). Displacement behaviors such as grooming, face touching or fumbling are related to anxiety and stress regulation [12]. As a consequence of these manifold connections of body language with important personal and social attributes, body language analysis has been a focus of automatic approaches attempting to infer high-level attributes such as emotion [13], leadership role, or personality type [14]. In contrast to the human science studies discussed above, these automatic approaches commonly lack an explicit intermediate representation of functional bodily behavior categories. Instead, they rely on a generic feature representation, encoding body postures and movements or on deep learning approaches [13] without clear interpretable internal structure. While such representations can be effective in prediction scenarios, they often lack interpretability and may miss subtle but meaningful differences, e.g., between fumbling and scratching.

Recognition of Actions and Body Language. RGB-based human action recognition has often been addressed by three main approaches. Two-stream 2D Convolutional Neural Networks [15, 16, 17] generally contain two 2D CNN branches taking different input features extracted from the RGB videos for action recognition. Recurrent Neural Networks (RNN) [18, 19, 20] usually employ 2D CNNs as feature extractors for an LSTM model. 3D CNN based methods [21, 22, 23] extend 2D CNNs to 3D structures, to simultaneously model the spatial and temporal context information in videos that is crucial for action recognition. For instance, a two-stream 2D CNN architecture, proposed by Wang et al. [24] divides each video into three segments and processes each segment with a two-stream network, fusing the individual classification scores by an average pooling method to produce the video-level prediction of the action class. Also, the two-stream Inflated 3D CNN (I3D) has been introduced by Carreira and Zisserman [25] to inflate the convolutional and pooling kernels of a 2D CNN with an additional temporal dimension to process at once a 3D block of pixels. The transformer method proposed by Liu et al. [26] that was designed for natural language processing has been recently extended to computer vision tasks [27, 28] to recognize human activities. In contrast to action recognition, which typically considers freely moving people [29, 30, 31], limited work on body language recognition addressed more constrained social interaction scenarios. For example, Yang et al. [32] generated sequences of body language predictions from estimated human poses and fed them to an RNN for emotion interpretation and psychiatric symptom prediction. Kratimenos et al. [33] extracted a holistic 3D body shape, including hands and face, from a single image and fed them also to an RNN for sign language recognition. Singh et al. [34] used handcrafted features to analyze body language for estimating person's emotions and state of mind. Santhoshkumar et al. [35] use Feedforward Deep CNNs for detecting emotions from full body motions. We observe that the common denominator of body language analysis methods are the employment of a general action recognition method without handling the specificity of body language such as subtle motions or micro facial expressions.

To summarize, these body language analysis methods enable us to measure objectively the behavior of humans by recognizing their Activities of Daily Living (ADL), their emotions, eating habits, and lifestyle. Human behavior can be modeled by learning from a large number of data, collected from a variety of sensors, to improve and optimize for instance, the quality of life of people suffering from behavior disorders, such as anxiety or apathy. In previous work, STARS successfully detected the everyday life activities performed by an individual living alone at home and we were able for instance, to detect breakfast activities, such as "preparing coffee", and "cutting bread", with sufficient accuracy [5, 36, 37].

Short-term research. We plan to draw upon our collaboration with clinicians within various projects and also upon the ethological rating scheme, such as the functional body language categories described in [38], to derive a set of *bodily behaviors* that are intuitively interpretable and allow to train models for fine-grained behavior distinctions and finally to link them for instance to *behavior disorders*. We already have started to *collect several datasets* depicting typical behavioral traits characterizing human interactions. These datasets will be used to evaluate the abilities of state-of-the-art methods to extract meaningful behavioral traits and to design new ones.

Mid-term research. We will study *long-term social interactions* performed by humans *in* their *natural environment*, by modeling their behavioral traits such as fine gestures and body language [6]. In particular, we will focus on spatio-temporal interactions between two individuals, to measure how well their behaviors are aligned with each other, and how well they are synchronized. This is critical for instance, to understand the *therapeutic alliance between patient and clinician* and to predict the success of a therapy [39].

To tackle these issues, we would like to go beyond Deep Learning by incorporating some *semantic modeling* within the Deep Learning pipeline, which today consists of a combination of CNN and transformers [40] to be able to model the complex action patterns in untrimmed videos. These complex action patterns include composite actions and concurrent actions occurring in long untrimmed videos. Existing methods [40, 41, 37] have mostly focused on modeling the variation of visual cues across time locally or globally within a video. However, these methods consider the temporal information without any further semantics. Videos may contain rich semantic information such as objects, actions, and scenes. Real-world videos contain also many complex actions with inherent relationships between action classes at the same time steps or across distant time steps. Modeling such class-temporal relationships can be extremely useful for locating actions in those videos. Therefore, semantic relational reasoning can help determine the action instance occurrences and locate the actions in the video, especially for complex actions in the video. To handle these challenges, we proposed a Class-Temporal Relational Network (CTRN) [42] that explores both, the class and temporal relations of detected actions.

To go beyond classical Deep CNN, a first attempt will consist of (1) effectively extracting action-relevant semantics from real-world untrimmed videos and (2) modeling the cross-semantic relations to enhance the action detection performance. To extract the relevant semantics, large Language-Vision models could be used. There are many foundation models for Large Language Models (LLM) and Large Language and Vision Models

(VLM) available for training our models on our target datasets. The main issue is how to adapt these models to the specificities of our datasets. We have already successfully used some of these foundation models (e.g. CLIP) to add semantics to the extracted visual features for activity detection [43]. In particular, we are using prompted engineering (e.g. learnable prompts) to guide the semantics extraction along the specificities of our application domains (e.g. homecare monitoring).

Long-term research. Using supervised techniques to learn *interactive behavior models* is often not possible in the medical domain, as data are scarce (in particular due to privacy issues) and subjectively annotated. We will address these problems by proposing novel architectures and taking advantage of other *learning paradigms*, such as weakly-supervised, unsupervised, self-learning [44, 45], and *lifelong learning*[3] to get more generic models able to adapt to multiple datasets. Other learning mechanisms will be studied: fusion, multi-tasks, guided-Attention, Self-Attention [40], Knowledge Distillation, and contrastive learning [36] to better benefit of multi-modalities [14], and to train with less data and with less annotation.

3.1.2 Face Analysis and Emotion Recognition

Participants: F. Bremond, M. Balazia, A. Dantcheva

An emotion is a mental state that arises spontaneously and is often accompanied by cognitive, physical, and physiological changes. Due to the complexity of human reactions, recognizing emotions is still limited and remains the target of many relevant scientific researches. In fact, Emotion Recognition is a highly multidisciplinary field where psychology meets deep learning. Emotions are typically divided in basic categories, as theorized by Ekman who identified basic discrete emotions [46]. Such categorization has been extended considering the interconnection between emotions and multiple intensities [47], [48].

Predicting emotions has been attempted via facial expression analysis in videos [49], which has been widely adopted both, in research and in industry owing to its ease of use with just a camera. However, the accuracy of computer vision algorithms, as in the case of CNN, is typically limited in identifying *real* emotions. Facial micro-expression recognition recently reported state-of-the-art performances when implemented with a transformer-based architecture [50]. While the FaceReader system, launched in late 2005 [50], is used worldwide in institutes and companies, there are still some limitations as image quality and facial angulation. Other main open challenges in the field are small available datasets and subjective annotations. Typical datasets range between some hundreds of videos to a few thousands and the annotations are often noisy due to the human complexity. A person may be happy even if he/she is not smiling and people differ widely in how expressive they are in showing their inner emotions. So, emotion annotations are very subjective and need to be adequately addressed. Moreover, emotions have multiple nuances, with different intensities.

Regarding emotional models, various architectures have been used as RNNs, LSTMs, CNNs [51, 52, 53, 54], with the aim of capturing the spatio-temporal information. In order to improve the recognition accuracy, multimodal transformers have been introduced [55], exploiting self- and cross-attention. Knowledge distillation from multimodal to unimodal (video) transformers has been reported, to reduce the acquisition complexity at inference time [56]. The state-of-the-art is achieved today with multimodal transformers, using

[3] Lifelong learning is also known as Incremental Learning or Continual Learning. It aims to adapt a model to new data without catastrophic forgetting on old knowledge and has become a key point for training a generalizable and robust model.

video, audio, and language cues [57, 58]. Here, the video and the audio are processed by small transformer encoders receiving as input features pre-trained on other datasets. The model extracting features is frozen and therefore it cannot be adapted to a new targeted dataset. For the video transformer, the inputs are fixed representations, such as DLN features in [59], IResNet and DenseNet features in [60], Facet/Openface features in [58], R(2+1)D-152 features in [57] and landmarks and action units in [61]. Such feature extractors and shallower encoders are typically used when small datasets are targeted. The main limitations of this approach are two. First, frozen representations are less appropriate for raw data than end-to-end trainable models. Second, smaller models are less accurate for recognizing specific expressions. In order to use raw data and bigger encoders, proper pre-training is needed to limit overfitting. While self-supervised techniques, such as VideoMAE [62] can be used for that purpose, they may miss the little details necessary to recognize facial micro-expressions. They are therefore not well adapted for the emotion recognition task.

Short-term research. One way to address these issues is to propose a novel and efficient self-supervised pre-task to enable end-to-end video emotion classification. For instance, we will propose a novel pre-training scheme that will allow for the use of small datasets, while preserving end-to-end classification with a given transformer-based backbone such as ViT-B [63]. In an early approach, we expand VideoMAE [62] by reconstructing different views of the masked input video. Based on one view (e.g., the frontal one) the autoencoder reconstructs other views (*e.g.,* top, down, laterals), generating augmented information in the latent space due to a more detailed representation, addressing this way low intensity and subtle emotions. Once the model is pre-trained on a large dataset, it can be fine-tuned to another small dataset and this will improve the performance of the emotion recognition task.

Video-based emotion recognition entails challenges related to small-size datasets, subjective annotation, as well as continuous head-pose variations and occlusions. We envision working with larger public datasets, as well as generated data (Axis 2) of diverse nature that can train models well and contribute to higher accuracy. On a similar note, models trained with a large variety of datasets, involving a balanced set of ethnicities, age groups, and gender are instrumental in developing robust automated emotion recognition models. The latter can be accustomed to data generation, see Axis 2.

Most current algorithms have been tested on datasets developed in a controlled environment. In contrast, *real-time and real-life emotion recognition*, which we will focus on, is exponentially more challenging, as it is prone to noise and may degrade the performance drastically.

Long-term research. Limitations related to face analysis and face recognition have to do with controlled conditions, compound emotions, and micro-expressions.

Current emotion recognition focuses on either classifying emotions into seven distinct categories, namely ‘happy’, ‘sad’, ‘surprise’, ‘angry’, ‘disgust’, ‘neutral’, and ‘fear’ and/or characterizing them by ‘valence’ and ‘arousal’. We will consider additional and potentially *composite categories of emotions.* A related issue here has to do again with more balanced dataset for each class, which can be availed by data generation, see Axis 2.

Generally speaking, models for emotion recognition employ the same datasets for training and testing. Cross-dataset evaluation is fundamental in increasing the robustness of models. However, this again brings us to the challenges that new learning techniques need to be designed (e.g. lifelong learning) or large amounts of training data are needed, which can be mitigated by data generation (see Axis 2). Traditional learning methods only consider the performance of a single fixed target domain. In the single target domain scenario, people usually assume that all training data are available before training and deploying a trained model. However, a real-world video monitoring system can record new data every day and from new locations when new cameras are added to an existing system. How to adapt a model to new settings without catastrophic forgetting on old knowledge has become a key point for training a generalizable and robust model. This research field, called lifelong learning (also known as Incremental Learning or Continual Learning) could be useful to adapt a previously trained model to a new dataset or to add new classes of emotion. We have already investigated this new research area on the Re-ID topic [64] and we are planning to extend it to emotion recognition.

3.1.3 Multimodal Recognition of Human Interactions

Participants: F. Bremond, M. Balazia, M. Thonnat

Behavior traits can be detected in self-presentation videos based on the acoustic and visual, non-verbal features such as pitch, intensity, movement, head orientation, posture, fidgeting, and eye-gaze. According to [65, 66], modalities such as audiovisual, text, and demographic features are important for personality prediction. Emotion recognition has generated specific approaches for multimodal data processing [67]. Deep bimodal models give state-of-the-art results on Multimodal Language Analysis in the Wild [68]. [69, 70] have shown that body gestures, head movements, expressions, and speech lead to an effective diagnosis of apathy. Few models have dealt with trimodal fusion of features [71, 72]. Although multimodal approaches are commonly used to recognize personality traits, there does not exist a comprehensive method to optimize and combine the considerable amount of informative features. All modality features may be concatenated together for behavior prediction; this approach is referred to early fusion. However, most of the multimodal approaches perform late fusion on heterogeneous data, as it outperforms other techniques. Present research in the field aims to find efficient ways for feature extraction and combination. We aim to design new approaches able to utilize all possible information available in an optimal manner [14]. The objective is to develop and test Human Behavior Coding algorithms using RGB video cameras at test time [36, 65], but using multi-modalities at training time with multiple datasets with various modalities to better characterize human behavior during interactions. As it is challenging to be an expert in all modalities, we will rely on open-source code (when available) or on our partners (when needed) to obtain the most effective backbone models for extracting multi-modal features. For instance, we are collaborating with DFKI [70] to extract audio and text features for measuring neuropsychiatric symptoms in patients with early cognitive decline. For electrophysiological signals, we are working with the Biorobotic Institute - Scuola Superiore Sant'Anna (Pontedera, Italy) [73] to compute more objective measurements of emotion.

Short-term research. Recent multimodal sentiment analysis approaches focus on deep neural networks and propose multi-sensor data fusion methods. As emotions are a complex set of reactions with multiple components [46], the idea is to compare/infuse/combine salient information from different modalities, coming from video cameras, audio and from biosensors. To lift the ambiguity, bio-signals (or Galvanic Skin Conductance (GSC) or electrodermal activity (EDA), electrocardiography (ECG), electroencephalography (EEG), respiration and heart rate, etc.) will be used as guides for emotion recognition. The objective here is to develop and test Human Behaviour Coding algorithms using multi-modalities at training time with multiple datasets with various modalities to better characterize human behavior during interactions, using the minimum or available modalities (e.g., only RGB video cameras) at test time. We will implement Transformer-based methods, comparing various strategies such as multi-task learning, Knowledge Elicitation (infusion) using the Student-Teacher paradigm, contrastive learning, and co-training techniques. Several levels of ground truth supervision will be used to train the model. We will also explore multi-source transfer learning, by using multiple sets of data in the training stage. We will further evaluate adversarial learning and data augmentation techniques in order to increase the generalizability of the model (see Axis 2).

To conduct such research, we plan to acquire and make publicly available a consistent multimodal dataset to stimulate different levels of emotion to complement existing datasets (e.g., DEAP, AMIGOS) that do not have complete or accurate annotations. This dataset will be composed of video and multiple electrophysiological signals. We intend to design a model, able to merge the salient information from heterogeneous data, especially combining video and electrophysiological signals to get more robust models and more objective measurements. The use of this new dataset will permit it to work with subtle emotions thus creating a latent space, where diverse kinds of information (subtle and strong emotions) are represented. The models will be generalizable, in order to work on diverse datasets, ensuring high accuracy even on small datasets. Having a well-trained model that can be tuned on various conditions will permit to translate emotion recognition in real-life settings.

Long-term research. Although multimodal approaches are commonly used to recognize social interactions, there does not exist a comprehensive method utilizing efficiently the considerable amount of information. Most of the multimodal approaches perform late fusion. Thus, we aim at designing new approaches able to utilize all possible information available in a more optimal manner [14]. Most state-of-the-art work is specific to a given task, to corresponding datasets, or to a subset of these modalities. So, it is tedious to modify a given work for a new task, such as a medical diagnosis. There is also a lot of existing overlapping work that could be combined to improve initial results. Thus, there is a need for a general framework that can handle this increase in data with high variability in structure and which learns robust relations that are shareable among tasks and datasets. Transformer-based backbones are state-of-the-art general feature extractors when they are carefully trained on large multimodal datasets. But, finetuning these models is resource-intensive and does not converge for small datasets. Finetuning any general feature extractor involves learning the environment in which the new data are recorded and the intricacies of the new task, while the basic spatial understanding of the video remains the same. Thus, we propose to use this basic understanding on these pre-trained large models and employ only a few additional parameters to learn the new information provided for finetuning. We will take a frozen pre-trained backbone (e.g., Video Swin Transformer) and add existing and newly developed plugins (called adapters) in parallel without changing the backbone and train only these plugins to transfer these pre-trained large models for downstream tasks (e.g., medical diagnosis) and small datasets (e.g., medical ones) with only sparse annotation. Moreover, these models will provide mid-level features for the users to understand the data processing and the taken decisions. For instance, during the consultation, the patient may have specific body gestures (e.g., scratching), head movements (e.g., looking down), facial expressions (e.g., sad), and sounds (e.g., sigh) showing effective personality and emotional cues (e.g., discomfort and stress) at this particular time of the conversation. Specific Graphical User Interfaces GUIs will be designed to show the doctors these mid-level features jointly with their impact along the conversation in an interactive manner.

3.2 Axis 2: Data Generation for Augmentation and Anonymization

Participants: A. Dantcheva, F Bremond

3.2.1 Data Generation

Participants: A. Dantcheva, F Bremond

In the past decade, computer vision has witnessed remarkable progress fuelled by the triptych of (a) algorithms for training computer vision models (e.g., backpropagation), (b) increased computational power (think of powerful graphical processing units (GPUs)), but very importantly by (c) *increased volumes of training data*. For example, MegaFace [74], with millions of facial images, has rapidly driven progress in face recognition, showcasing that better models are empowered by bigger data. Even in the occasional abundance of raw data, there is a plethora of remaining challenges in designing data-driven intelligence approaches such as deep neural networks (DNNs). These challenges stem from the fact that data must be processed; for example, data must be annotated (e.g., annotation of facial expressions in facial videos), in order to optimize the millions of network-parameters. To make things worse, the curation of large datasets is tedious, costly, time-consuming and is fundamentally bounded by the population sizes of such data, as well as by the ever-increasing privacy and usage considerations that have been recently highlighted by the General Data Protection Regulation (GDPR). The resulting real data and associated real-life datasets are scarce, private, and they inherit human biases. As such, these limitations threaten to bring any advances in computer vision to a dramatic halt. Therefore, we are now at a point, where the availability of annotated data is the main bottleneck in the development of data-hungry DNN models [75]; a bottleneck that far exceeds any algorithmic or computational bottleneck. Based on the premise that computer vision data-driven intelligence is heavily influenced by the underlying data, we here seek to understand how one can actually create data that will augment the learning space and the learning capabilities of computer vision models. Generated data or synthetic data provides a promising solution to the above challenges, as it is easier to obtain, it is inexhaustible, pre-annotated, and less expensive. In addition, synthetic data has the potential to avoid ethical and privacy concerns, as well as practical issues related to security. Further, synthetic data brings to the fore unique opportunities, allowing for the surgical injection of training data in scenarios where collecting real data may be impractical or impossible (e.g., talking dogs, faces that do not exist, etc.). Indeed synthetic data allows for new training paradigms in computer vision models. We will design methods that allow synthetic data to be dynamically generated, directly as a function of the needs of learning algorithms.

Past attempts for synthetic images and videos. Computer vision-generative models of images have received unprecedented attention, owing to recent breakthroughs in the underlying modeling methodology. The most powerful models today are built on generative adversarial networks (GAN) [76, 77], autoregressive transformers [78], and most recently diffusion models [79, 80, 81]. Diffusion models (DM) constitute neural networks, which were trained to denoise images successively blurred with Gaussian noise by learning to reverse such diffusion process. After training, such a model can generate data by simply passing randomly sampled noise through the learned de-noising process. This synthesis procedure can be interpreted as an optimization algorithm that follows the gradient of the data density to produce likely samples. In its denoising process, conditional features like class labels of data can be applied to the network for specializing its sampling process.

Such DMs outperform previous generative methods [82], as they offer robust, stable and scalable training procedures. DMs are largely unaffected by training limitations such as overfitting, as it is the case in GANs (mode collapse). In addition, DMs generally involve fewer parameters than transformer-based counterparts that typical require massive amounts of data and thus experience a performance plateau. As diverse synthetic data is a primary need for computer vision, DMs have been rapidly adopted in several settings such as image and video generation, image deblurring, high-resolution image generation, and image editing.

Challenges in video generation. However, while the image domain has seen great progress, video has proven to be more challenging due to (i) significant computational costs associated with training on video data, as well as due to (ii) the lack of large-scale, general, and publicly available video datasets. In regards to the computational challenge in (i), it is indeed the case that training current state-of-the-art image generation models is already extremely expensive computationally, making it exceedingly hard to generate videos, particularly videos of variable length. Similarly, w.r.t. the second challenge in (ii), it is the case that while in image generation there are datasets with billions of images (think of LAION-5B [83]) – in video, datasets are much smaller (think of the VoxCeleb2 [84] dataset of about 1M videos) and thus cannot support the higher complexity of open domain videos.

Limited settings of generated videos. Very recently, video generation methods such as DM-based Imagen Video [85] and Make-a-Video [86], showcased the stunning potential of generative AI. However, to date, the generated videos remain heavily constrained in quality, resolution, as well as length, mainly due to having video encoders that only encode fixed size videos or encode frames independently. Such video generation methods are further limited as they currently produce results only depicting single persons, performing simple motions in highly constrained settings with mostly a neutral background. Crucial in our effort will be our goal of generating videos that encompass complex settings of multiple subjects, able to interact in front of a non-uniform background.

Control. While we are already beginning to know a few things regarding DMs - like for example that in terms of reconstruction and encoding, DMs are superior to GANs - it is indeed the case that understanding the limits of control of such models, still lies at its infancy. In an effort to control generated images, recent works explored the discovery of semantically meaningful directions in the latent space of pre-trained GANs, where linear navigation corresponds to the desired manipulation of images. In this context and in terms of control, supervised [5], as well as unsupervised approaches [87] were proposed to edit semantics such as facial attributes, colors and basic visual transformations (e.g., rotation and zooming) in generated or inverted real images [88]. The latest addition of Latent Diffusion Models (LDMs [79]) are a positive development in this direction, as such LDMs are able to reduce the heavy computational burden when training on high-resolution images. In addition, our own work revealed - in the context of autoencoder generation models - how to disentangle motion and appearance in videos, as well as how to manipulate decomposed semantically meaningful motion-directions [89]. However, in the context of LDMs, disentanglement and manipulation of semantic attributes remains a key open research challenge of substantial potential impact and these are indeed challenges that we will explore.

Short term research. We will design algorithms that learn controllable and generalizable video representations that allow for generating an abundance of diverse videos that anticipate complex and interactive visual events. In this context, diverse relates to creating multiple futures of the same situation / same input, which then can be transferred onto another setting, the latter we refer to as generalizable.

Limiting computation costs and video data requirements To alleviate the massive computational requirements of training powerful DMs that guarantee high resolution and high frame quality, we intend to explore pre-trained, fixed image generation DMs and to endow them with temporal awareness by introducing additional temporal neural network layers, interleaved with the existing spatial layers. This has the advantage that large image datasets can be utilized in training of the spatial layers, while limited video data can be utilized for training temporal layers. This approach that we propose has the potential to allow such models to scale gracefully to the setting of video generation.

Control. Drawing from the world of GANs, latent spaces were found to contain rich semantically meaningful directions such as "zoom in", we intend to discover sets of directions in the latent space of LDMs that encode different semantics and facilitate related high-level manipulations. Such manipulations will involve editing foreground, background, identity, pose, illumination, expression, motion including blinking and gaze, as well as global motion, to mention a few. We will aim to manipulate such semantic attributes without changing the semantic content, while guaranteeing that modification parameters generalize to different videos.

Long term research. Long term goals include the design of models able to generate complex datasets, depicting social *interaction* between (generated) humans. Currently, this is a largely open research direction of high impact. We will explore the merge of individual networks that have learned different subjects. A merging network will bring together different subjects that we will coordinate by mechanisms such as multi-agent reinforcement learning.

An additional long term goal has to do with tackling the question of "How can synthetic data generation be integrated into computer vision training to provide content that is finely tailored to the learner's needs?".

We envision synthetic data being a key ingredient in reaching the holy-grail in designing computer vision models, which is to be able to replicate the learning process of humans in order to produce rich, robust and generalizable knowledge about the world. Synthetic data introduces a unique opportunity to enable a new class of visual learning techniques that are not limited by the quality and volume of the available data; in the words of [90], such techniques would only be limited by imagination and beyond.

Enhancing views. This is a challenging objective and indeed we have recently demonstrated - for a specific exploratory setting - that GAN-generated images can be used for data augmentation for novel 'views' in unsupervised person ReID [91]. Building on such progress, we aim to explore video-related tasks such as expression and activity recognition. One candidate approach will be to first interpret the latent space in order to discover the representations related to a general view concept (including factors like lighting, shifting, viewpoint), and to then augment the training data by manipulating these same factors. Our objective is more general, as we intend to learn discriminative spatio-temporal features from synthetic data in an entirely self-supervised learning manner, without the need for leveraging structural information.

Interleaving of on-the-fly generation and learning. One of our main goals is motivated by the fact that in real world settings, computer vision requires lifelong, structured and continual learning. What makes this problem even more enticing is the fact that, while the volume of data (such as the number and duration of videos) is indeed a crucial factor, an additional aspect is that not all training samples are equally important. How to properly address this variability in the information that each sample carries, remains a wide-open problem at the core of computer vision. We believe that synthetic data can play a key role in resolving this long-lasting open problem. We intend to introduce training data sampling strategies that will maximize the training progress by surgically identifying pockets of importance. In essence, we intend to carefully insert a proper number of synthetic data, in order to alter and augment the informational structure of learned classes which would have otherwise been under-represented. An on-the-fly and recurring generation (of such carefully designed samples) will be then interleaved with the learning of spatial and temporal regions. This is indeed a very novel approach at its infancy, which allows for simplified learning, as well as for data to be digested in an efficient manner.

New learning paradigms. We then intend to explore new learning paradigms that employ on-the-fly generation. These will be explored in the context of continual learning, where infinite series of tasks are being learned (e.g., detection of a growing number of categories), and for which case we propose the use of synthetic data for alleviating the challenges of catastrophic forgetting and overfitting [92]. Similarly, in the context of hierarchical learning, we will interleave large visual representation learner networks with generation networks, in order to improve down-stream tasks such as recognizing facial dynamics, and more evolved mental states by exploiting different levels of fine-grained synthetically generated representations.

3.2.2 Data Augmentation and Anonymization

Participants: A. Dantcheva, F Bremond

We aim to apply data generation models proposed in the previous section in two domains of application, namely data augmentation and data anonymization, which are catering the needs of Axis 1 (Human Interaction Recognition).

Data augmentation. The general focus of data-driven computer vision algorithms has to do with the automated extraction of patterns by finding complex data representations from large volumes of input data without human interference, utilizing the patterns to detect or classify unseen data. The powerful twist that we are envisioning is that data generation places full control over the distribution of the generated data, thus endowing us with the ability to ensure quality and diversity, while saving cost, and mitigating bias. As a consequence, we foresee that such synthetic data will allow for nothing less than a paradigm shift in training. For example, as inspired by human systems, synthetic data will bring continual, multimodal, interactive, embodied learning to the next level, providing richer and more sophisticated representations. This applies directly toward the grand goal of allowing computer vision to approach human-level intelligence; a long-term goal that will require the grasping of key concepts related to the physical world and its composition, as well as will have to entail a non-diluted ability to learn continually, interactively and multimodally [90]. We aim to identify entirely new perception models and related learning paradigms, which will exploit synthetic data in an entirely new, efficient and dynamic manner. We consider such models for a variety of recognition settings that can target a broad spectrum of facial behaviors including expressions and micro-expressions. By exploring the fundamental properties of learning with synthetic data, we anticipate computer vision models that generalize onto a large class of human actions.

Data anonymization. Privacy-preserving data-processing has obtained increased attention in the past years, with challenges having to do with data anonymization, while maintaining the image quality. The General Data Protection Regulation (GDPR) came to effect as of 25th of May, 2018, affecting all processing of personal data across Europe. GDPR requires regular consent from the individual for any use of their personal data. However, if the data does not allow to identify an individual, companies are free to use the data without consent. To effectively anonymize images, we require a robust model to replace the original face, without destroying the existing data distribution; that is: the output should be a realistic face fitting the given situation.

Anonymizing images, while retaining the original distribution is challenging, as it entails the removal of all privacy-sensitive information, generation of a highly realistic face, while providing a seamless transition between original and anonymized parts. This requires a model that can perform complex semantic reasoning to generate a new anonymized face. For practical use, we desire the model to be able to manage a broad diversity of images, poses, backgrounds, and different persons. Our proposed solution can successfully anonymize images in a large variety of cases, and create realistic faces to the given conditional information.

Short term objectives We have started the activity of *data augmentation*. Specifically, we have focused on data augmentation for person re-identification and have some early results [93, 94] which showcase the significant potential of such approach. We plan to extend this results onto other vision problems such as emotion recognition and face analysis, as well as designing new augmentation methods, which support such down-stream tasks.

Long term objectives. We have not started the activity associated to anonymization and we consider it in the state of planning and discussing with possible partners, e.g., Eurecom and BITS Pilani Hyderabad.

3.3 Applications

Video understanding consists of a complex pipeline made of various tasks, such as object detection, people tracking, pose estimation, and event detection. So, many tasks are generic, and can be shared between different application domains. The behavior analysis techniques we develop for other applications (for instance for sport or security domains) can be applied to medical applications and vice-versa.

3.3.1 Medical Applications

Our main motivation as explained before is to help clinicians to diagnose, monitor and provide pertinent treatment to patients with behavior disorders. The applications we target are not general medical diseases but the ones related to the brain and more precisely to psychiatric disorders. These disorders can appear very early in the life of the patient (for instance autism spectrum disorder [4]), they can concern adults (depression, bipolar, schizophrenia [39]) or the elderly (for instance Alzheimer disease). We have been working for the elderly patients since the creation of the CoBTek joint team in January 2012. More recently, we have extended our study to the two other categories of age. Now we have some clinical trials within these three categories of patients.

3.3.2 Other Applications

Sport applications: Sport is an interesting application domain for human activity understanding for three reasons. First, data are often publicly available, so with less ethical concerns than medical ones. Moreover, many data have been recorded and annotated to be part of international challenges[4]. Second, human activities are complex at the level

[4] https://www.soccer-net.org/challenges

of individuals, of a team and along time. Third, many companies are interested to fund research to advance the field of human activity understanding for sport. For instance, we have a collaboration with a local company, *Fairvision*[5] on football games.

Security applications: The interest and investment in vision based security systems is large and rapidly growing and is fueled by applications ranging from autonomous vehicles to personalization of customer service. Accordingly, numerous companies, military and public organizations are interested in research in this context.

4 Collaborations

We have different kinds of collaboration, we plan to pursue within the new team which are pertinent to our new objectives.

4.1 Academic

Concerning academic partnerships, we focus on partners with excellent methodological expertise. We collaborate with IDIAP in Switzerland (Jean-Marc Odobez) for gaze estimation in the wild. We are also working with DFKI COS in Germany (Jan Alexandersson) for multimedia analysis with particular expertise in audio and natural language processing. Another essential collaboration is the one with the Chinese Academy of Science (CAS) in the context of face analysis (Prof. Shiguang Shan, Prof. Xilin Chen, Prof. Hu Han). In the USA we have a long-lasting and fruitful collaboration with the Michigan State University (Prof. Arun Ross) related to biometrics. We also collaborate with BITS Pilani in India (Prof. Abhijit Das) on the topic of face analysis.

STARS could collaborate in the coming period with new teams. In particular, we could collaborate with Inria teams **(for instance Wimmics, Acentauri, RobotLearn or Chroma)** and with some teams listed in section 5 Positioning. For instance, we have already started collaborating with Michael Ryoo[6] (Associate Professor at Stony Brook University and affiliated with Google Brain) on action detection.

Moreover, we are also open to collaborations with experts in other topics (e.g. semantics).

4.2 Medical

STARS aims to develop ambitious research activities for healthcare monitoring. More precisely, we are mostly working with psychiatric physicians specialized in cognitive disorders including psychiatric experts for children, adults, and older people.

We have a strategic partnership named CoBTek with the Nice Hospital (CHU Nice) to study the impact of video understanding approaches for cognitive disorders. CoBTeK (Cognition Behavior Technology) is a team of Université Côte d'Azur. It was created in partnership with the Inria STARS team in January 2012. Since its creation, CoBTeK's mission has been to develop clinical research on assessment and care based on new technologies. CoBTeK's central theme has been to develop research on novel computer science technologies for the prevention, diagnosis, and treatment of neuropsychiatric pathologies. CoBTeK brings together all of Nice's university psychiatry departments but also supports several university courses, such as the Masters 2 in Speech Therapy and Psychology - Psycho-trauma. CoBTeK is a multi-disciplinary unit, with computer scientists working

[5] `https://www.fairvision.fr/`

[6] `http://michaelryoo.com/`

alongside clinicians, thanks to its strong links with Inria. CoBTeK is structured as a single University team (Unité Propre de Recherche), with teaching clinician-researchers focusing their research on children, adolescents, adults, and seniors. Since its creation, the team has been directed by Prof. Philippe Robert (PU-PH) and co-directed by François Brémond (DR1 Inria). The team has been headed since March 2022 by Prof. Florence Askenazy (PU-PH).

The CoBTeK team operates on multiple sites in Nice (Institut Claude Pompidou, Hôpital Lenval, Hôpital Pasteur, Inria Sophia Antipolis, JL Noisier daycare center). All hospital sites are authorized as clinical research sites. CoBTeK is directly integrated as one of the structuring units of the EUR Healthy and the Faculty of Medicine. CoBTeK was a pioneer in setting up a technical platform combining clinical and technological expertise on the same site as a care activity. The equipment on the first technical platform at the Institut Claude Pompidou is regularly made available to other University units. Its usage by industrial partners has been limited to those with whom the team has a joint research project (MindMaze, Ki Element). Thanks to the participation of Inria, several researchers contributed directly to the design of technical equipment. Soon, the implementation of the platform on the Lenval site (child and adolescent psychiatry) and on the Pasteur site (adult psychiatry) will be finalized.

The CoBTeK's scientific output follows the principles of clinical research; drafting of a protocol shared by permanent team members and Ph.D. students, submissions to the People Protection Committee (CPP), the CNIL, the French National Agency for the Safety of Medicines (depending on the type of protocol), recording of data on a secure health server, monitoring and control of inclusions, data quality and compliance with ethical rules by a clinical research assistant (ARC), retention of files for a period of 15 years. Respect for the human being is the basic principle of the clinicians involved in research, who work directly with patients and families on a daily basis. The team's added value lies in having applied this principle to the field of new technologies too.

4.3 Industrial

Our strategy is to collaborate with industrials, in order to obtain funding, to capture data and to transfer our research software into products. These collaborations are critical in our medical domain because the hospitals cannot have the engineers to maintain experimental software.

- Toyota: we have had a long-term partnership with Toyota Research Europe (Brussels) since 2013, in order to study the monitoring of older people at home. They finance regularly Ph.D. fellowships (4 fellowships) and engineers (3 engineers); we have jointly created a new public benchmark dataset for assisted living in a realistic flat. Finally, they integrate the software developed by our Ph.D. students into their HCR partner robots.

- Local companies: we will continue to work jointly with local companies (SME or start-ups) to increase their technological expertise in particular by transferring our software and by hiring our former students and engineers. For instance, for medical applications, we plan to work with Nively (Nice), Fantastic Sourcing (Cagnes-sur-Mer), Ekinnox (Sophia Antipolis) and Klava (Paris and Nice).

4.4 Collaborative Projects

STARS is used to get collaborative projects at the national, European, or international levels. Currently, we have three European projects (Mephesto, Gain, and Heroes), three

ANR projects (Respect, Activis, and FSHD), and one associated team with India. We will continue to propose new projects for funding and international visibility.

5 Positioning

As described in the related work section, social interactions have been overlooked by the computer vision community. Only a few teams focus on this domain. In video understanding the frontier between academic and industrial research is narrow and there are many top research scientists working both at universities and large private companies. As referenced in section 3, most research studies are performed on generic action recognition from web videos, such as YouTube, where only generic classes of action are targeted. More precisely, most of the researchers in action detection topics are working on general video datasets, very few on Activity of Daily Living (ADL), and even less on cognitive disorders and real-world homecare. The main reason is due to the unavailability of benchmarks made of cognitive disorder video datasets. Therefore, very few researchers are working on the exact same problems related to the objectives of STARS. Only Prof. Éric Granger[7] in Canada is working on the exact same problems (see next section).

Moreover, video understanding consists of a complex pipeline made of various tasks, such as object detection, people tracking, pose estimation, and event detection. So, some researchers are at the top of the state-of-the-art on some tasks, but not on the complete pipeline. For instance, Prof. Tang Xiaoou from the Chinese University of Hong Kong, is an expert in ADL action detection from pose sequences (see next section). These researchers could be both collaborators or competitors depending on the opportunities. For instance, we could combine the pose-based action detection from MMlab with a novel processing of RGB sequences to improve the global detection performance.

5.1 International

In the USA, one of the main researchers in action detection is Michael Ryoo[8], Associate Professor at Stony Brook University and affiliated with Google Brain. His focus is on action detection from web videos and he has also some strong results on home-care monitoring videos. In Qualcomm, Prof. Fatih Porikli[9] is an expert in Computer Vision and Machine Learning and leads a group on action recognition. This team has a large background in embedded systems on smartphones. The Department of Computer Science and Engineering of the Michigan State University (MSU)[10] has a longstanding focus and leadership in computer vision and facial analysis. In this department there are three teams led by Prof. Anil Jain, Prof. Arun Ross, and Prof. Xiaoming Liu specialized in biometrics.

In Canada, Prof. Éric Granger[11] is working on action detection for monitoring patients with cognitive decline. He has a research chair in artificial intelligence and digital health for health behavior change at Universite du Quebec. His team objectives are very close to our research interests.

In China, a large group MMlab[12], led by Prof. Tang Xiaoou (founder of the visual-surveillance Sensetime company) at the Chinese University of Hong Kong, is the pioneer in many approaches in video understanding and biometrics. In this lab, Professor Yu Qiao is working on action recognition and video generation.

[7] https://www.etsmtl.ca/recherche/professeurs-chercheurs/egranger

[8] http://michaelryoo.com/

[9] https://www.porikli.com/

[10] https://www.cse.msu.edu/

[11] https://www.etsmtl.ca/recherche/professeurs-chercheurs/egranger

[12] https://mmlab.ie.cuhk.edu.hk/

The Institute for Computing Technology at the Chinese Academy of Sciences[13] and specifically Prof. Hu Han, Prof. Shiguang Shan (who is also CTO at SeetaTech), and Prof. Xilin Chen are at the forefront of computer vision and face analysis.

5.2 Europe

In the UK, in Bristol, the team led by Prof. Dima Damen[14] is specialized in action detection and recognition. This team is active in monitoring people with health issues and is well-known for creating a benchmark dataset of egocentric videos in kitchens. The team led by Prof. Andrea Cavallaro[15] at the Queen Mary University of London works on Computer Vision and Machine Learning. In particular, this team is working on action detection in videos recorded by several cameras from multiple viewpoints. This is interesting for making the video understanding process independent from the camera's viewpoint.

In Germany, Prof. Dr. Jürgen Gall[16] is leading a group on action detection. He benchmarks his algorithms on our Toyota Smart-Home dataset. In particular, we share the same concerns about real-time software embedded in robot to collaborate with humans.

In Italy, at the University of Trento Prof. Nicu Sebe[17] and Prof. Elisa Ricci are involved in high-level research related to computer vision, action recognition, facial analysis, and generation.

5.3 France

In Paris, the MLIA team at the ISIR lab of Sorbonne University led by Prof. Matthieu Cord[18] (who is a part-time principal scientist at valeo.ai) is dedicated to Computer Vision and Machine Learning in general. They have experience in the area of action recognition.

At IMT Lille Nord Europe, the Image group in CRIStAL Laboratory (UMR CNRS 9189) led by Prof. Mohamed Daoudi[19] is working on Computer Vision and specializes in 3D facial analysis and skeleton-based video generation.

Within Inria in Domain 4, the teams with the closest topics are Thot in Grenoble and Willow in Paris. Thot and Willow are teams related to computer vision and deep learning, mainly focusing on static images and video retrieval for web applications. In contrast, STARS aims at fine-grained human interaction for medical applications (cognitive disorder). Concerning Inria teams in Domain 5 the closest teams are Epione and Aramis. These teams are specialized in studying computer vision and machine learning techniques for medical images.

6 Technology Transfer

Our strategy for technology transfer is motivated by having a real impact on medical applications. Therefore, we will follow a three-step strategy. First, we will advance fundamental research issues to propose to our medical partners new techniques in human interaction for behavior disorders taking into account their requirements. Second, we will conduct clinical trials with our medical partners to collect data on patients, and assess the quality of the results of our new methods. Third, we will collaborate with industrial partners to design

[13] http://english.ict.cas.cn/

[14] http://people.cs.bris.ac.uk/~damen/

[15] http://www.eecs.qmul.ac.uk/~andrea/

[16] https://pages.iai.uni-bonn.de/gall_juergen/

[17] https://disi.unitn.it/~sebe/

[18] https://webia.lip6.fr/~cord/

[19] https://sites.google.com/view/mohameddaoudi

and deploy new products to effectively improve the quality of life of patients with cognitive disorders. The results of the clinical trials and the feedback from the industrial partners will raise new research questions and enrich our scientific agenda acting as a virtuous circle.

Our software strategy is to provide open-source code to the research community to disseminate our results widely. For specific application domains (for instance assisted living robots) we will fill patents together with our industrial partners. We will continue to transfer our software to Nice hospital premises involved in CoBTek (Lenval Hospital, Pasteur Hospital, and Institut Claude Pompidou). We will also evaluate our software at a larger scale in private homes at Cannes Mandelieu through Nodeus Solutions[20]. We already have built strong partnerships with local and international industrial partners such as our previous start-up Ekinnox (in Sophia Antipolis) or Toyota (Europe and Japan).

STARS is also interested in the creation of startups. Antitza Dantcheva is working on the creation of the startup MovU. MovU is focused on generating videos of talking heads for corporations and is hence related to Axis 2. The goal is to make a product based on an algorithm [95] developed in the STARS team. Two former interns of the STARS team, namely Tashvik Dhamija and Pranav Balaji, have started working in MovU fulltime (since October 2023). Antitza Dantcheva's involvement will entail half a day per week for the creation period of the startup.

7 Education

The STARS research scientists teach regularly courses in computer vision and machine learning at the University Côte d'Azur and at thematic summer schools at the international level. In particular, François Brémond teaches at the Master of Science in Data Science and Artificial Intelligence at Université Côte d'Azur, the Course on Computer Vision and Deep Learning since 2019. Antitza Dantcheva teaches classes at Polytech Nice Sophia - Université Côte d'Azur in applied artificial intelligence, Master 2. Michal Balazia is planning to teach computer vision at UCA as ISFP research scientist. We plan to continue these important activities for training future Ph.D. students with pertinent skills for the team.

8 Ethical Questions

For the healthcare domain, a goal is to better diagnose people with psychiatric disorders. Comprehensive analyses of clinical interviews are deemed necessary so that the clinicians can benefit from a detailed view of the patient's behavior, body language and emotional capabilities. The development and ultimate use of novel symptom detection technologies of a vulnerable user group such as individuals with depression or schizophrenia, and the assessment methodologies are not free of ethical or legal concerns, despite the usefulness of the information technologies and their acceptance by people with or without impairments. Thus, one goal of the team is to design the right technologies that can provide the appropriate information to the patients and the medical carers, while preserving people's privacy and dignity. So, we 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. The project raises several ethics issues:

[20] `https://www.nodeus.solutions`

Human Participation: The inclusion of patients are performed by our medical partners after having obtained the authorization from the patients and the ethical authorities.

Personal Data Protection: We will record patient data, but we will take care of safety concerns. The recording computer will always be offline. To move or copy the data to a secure online data processing platform, we will first put the data on a portable hard drive which we will later connect to a computer with Internet connection. Upload will be made directly from the drive without copying to the other computer. The portable hard drive will be immediately deleted after the upload. As per secure online platforms, the INRIA Sophia center has NEF[21]. This platform does not require additional layers of encryption for data storage.

Use of Artificial Intelligence: As presented above, we aim at designing multimodal cognitive systems with perceptual capabilities to understand people's facial and body language. As a matter of fact, vision sensors can be seen as intrusive. New communication paradigms and other sensors (e.g., physiological and other new sensors to come in the future) are therefore additionally envisaged to provide the most appropriate services to the observed people while preserving their privacy. To better understand ethical issues, we have already contacted several ethical organizations: "Commission Ethique et Droit" (ODEGAM), a local association in France for ethical issues related to older people, and the French scientific counsel for the national seminar on "La maladie d'Alzheimer et les nouvelles technologies – Enjeux éthiques et questions de société". This counsel has in particular proposed a chart and guidelines for conducting researches with dementia patients. For addressing the acceptability issues, focus groups and human-computer interaction experts will be consulted on the most adequate range of mechanisms to interact and display information to patients. Ethics approval will be sought at each clinical trial for human participation in our research activities, and for data protection.

References

[1] A. König, P. Müller, J. Tröger, H. Lindsay, J. Alexandersson, J. Hinze, M. Riemenschneider, D. Postin, E. Ettore, A. Lecomte, M. Musiol, M. Amblard, R. Hurlemann, F. Bremond, and M. Balazia, "Multimodal phenotyping of psychiatric disorders from social interaction: Protocol of a clinical multicenter prospective study," *Personalized Medicine in Psychiatry*, vol. 33-34, p. 100094, Jul. 2022. [Online]. Available: https://hal.inria.fr/hal-03724844

[2] S. Chen, Y. Cho, K. Yu, L. Ferrari, and F. Bremond, "Editorial: Recognizing the state of emotion, cognition and action from physiological and behavioral signals," *Frontiers in Computer Science*, vol. 4, Aug. 2022. [Online]. Available: https://hal.science/hal-03906373

[3] R. Zeghari, R. Guerchouche, M. Tran-Duc, F. Bremond, K. Langel, I. Ramakers, N. Amiel, M. P. Lemoine, V. Bultingaire, V. Manera, P. Robert, and A. König, "Feasibility Study of an Internet-Based Platform for Tele-Neuropsychological Assessment of Elderly in Remote Areas," *Diagnostics*, vol. 12, Apr. 2022. [Online]. Available: https://hal.science/hal-03968301

[4] A. Ali, F. F. Negin, F. Bremond, and S. Thümmler, "Video-based Behavior Understanding of Children for Objective Diagnosis of Autism," in *VISAPP 2022 - 17th International Conference on Computer Vision Theory and Applications*, Online, France, Feb. 2022. [Online]. Available: https://hal.inria.fr/hal-03447060

[21] `https://wiki.inria.fr/ClustersSophia/Clusters\_Home`

[5] S. Das, R. Dai, D. Yang, and F. Bremond, "VPN++: Rethinking Video-Pose embeddings for understanding Activities of Daily Living," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Dec. 2021. [Online]. Available: https://hal.science/hal-03485766

[6] M. Balazia, P. Muller, A. L. Tanczos, A. V. Liechtenstein, and F. Bremond, "Bodily Behaviors in Social Interaction: Novel Annotations and State-of-the-Art Evaluation," in *MM '22: The 30th ACM International Conference on Multimedia*, ACM. Lisbon, Portugal: ACM, Oct. 2022, pp. 70–79. [Online]. Available: https://hal.science/hal-03936267

[7] N. Sinha, M. Balazia, and F. Bremond, "Flame: Facial landmark heatmap activated multimodal gaze estimation," in *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).* IEEE, 2021, pp. 1–8.

[8] V. Strizhkova, Y. Wang, D. Anghelone, D. Yang, A. Dantcheva, and F. Brémond, "Emotion Editing in Head Reenactment Videos using Latent Space Manipulation," in *FG 2021 - IEEE International Conference on Automatic Face and Gesture Recognition*, Jodhpur, India, Dec. 2021. [Online]. Available: https://hal.science/hal-03530150

[9] T. Agrawal, D. Agarwal, M. Balazia, N. Sinha, and F. Bremond, "Multimodal personality recognition using cross-attention transformer and behaviour encoding," in *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, INSTICC. SciTePress, 2022, pp. 501–508.

[10] A. Mehrabian, "Relationship of attitude to seated posture, orientation, and distance." *Journal of personality and social psychology*, vol. 10, no. 1, p. 26, 1968.

[11] D. R. Carney, Dana, "Beliefs about the nonverbal expression of social power." *Journal of nonverbal behavior.*, vol. 29, no. 2, 2005-06-01.

[12] M. Bardi, T. Koone, S. Mewaldt, and K. O'Connor, "Behavioral and physiological correlates of stress related to examination performance in college chemistry students," *Stress*, vol. 14, no. 5, pp. 557–566, 2011.

[13] M. Romeo, D. Hernández García, T. Han, A. Cangelosi, and K. Jokinen, "Predicting apparent personality from body language: benchmarking deep learning architectures for adaptive social human–robot interaction," *Advanced Robotics*, vol. 35, no. 19, pp. 1167–1179, 2021.

[14] T. Agrawal, M. Balazia, P. Müller, and F. Bremond, "Multimodal Vision Transformers with Forced Attention for Behavior Analysis," in *WACV '23: IEEE International Winter Conference on Applications in Computer Vision.* Waikoloa, United States: IEEE, Jan. 2023. [Online]. Available: https://hal.science/hal-03936484

[15] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 568–576.

[16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[17] M. Zong, R. Wang, X. Chen, Z. Chen, and Y. Gong, "Motion saliency based multi-stream multiplier resnets for action recognition," *Image and Vision Computing*, vol. 107, p. 104108, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885621000135

[18] Z. Liu, Z. Li, R. Wang, M. Zong, and W. Ji, "Spatiotemporal saliency-based multi-stream networks with attention-aware lstm for action recognition," *Neural Comput. Appl.*, vol. 32, no. 18, p. 14593–14602, sep 2020. [Online]. Available: https://doi.org/10.1007/s00521-020-05144-7

[19] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[20] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017.

[21] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

[22] J. C. Stroud, D. A. Ross, C. Sun, J. Deng, and R. Sukthankar, "D3d: Distilled 3d networks for video action recognition," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 614–623.

[23] M. Fayyaz, E. Bahrami, A. Diba, M. Noroozi, E. Adeli, L. V. Gool, and J. Gall, "3d cnns with adaptive temporal feature resolutions," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2021, pp. 4729–4738. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00470

[24] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Val Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*, 2016.

[25] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[28] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, dec 2021. [Online]. Available: https://doi.org/10.1145/3505244

[29] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017. [Online]. Available: https://arxiv.org/abs/1705.06950

[30] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca, "Toyota smarthome: Real-world activities of daily living," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019, pp. 833–842.

[31] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. K. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," *ArXiv*, vol. abs/1604.01753, 2016.

[32] Z. Yang, A. Kay, Y. Li, W. Cross, and J. Luo, "Pose-based body language recognition for emotion and psychiatric symptom interpretation," *CoRR*, vol. abs/2011.00043, 2020. [Online]. Available: https://arxiv.org/abs/2011.00043

[33] A. Kratimenos, G. Pavlakos, and P. Maragos, "Independent sign language recognition with 3d body, hands, and face reconstruction," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4270–4274.

[34] S. Singh, V. Sharma, K. Jain, and R. Bhall, "Edbl - algorithm for detection and analysis of emotion using body language," in *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, 2015, pp. 820–823.

[35] R. Santhoshkumar and M. K. Geetha, "Deep learning approach for emotion recognition from human body movements with feedforward deep convolution neural networks," *Procedia Computer Science*, vol. 152, pp. 158–165, 2019, international Conference on Pervasive Computing Advances and Applications-PerCAA 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050919306908

[36] R. Dai, S. Das, and F. Bremond, "Learning an Augmented RGB Representation with Cross-Modal Knowledge Distillation for Action Detection," in *ICCV 2021 - IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, Oct. 2021. [Online]. Available: https://hal.science/hal-03314575

[37] R. Dai, S. Das, L. Minciullo, L. Garattoni, G. Francesca, and F. Bremond, "PDAN: Pyramid Dilated Attention Network for Action Detection," in *WACV 2021 - Winter Conference on Applications of Computer Vision 2021*, Waikoloa / Virtual, United States, Jan. 2021. [Online]. Available: https://hal.inria.fr/hal-03026308

[38] A. Troisi, "Ethological research in clinical psychiatry: the study of nonverbal behavior during interviews," *Neuroscience & Biobehavioral Reviews*, vol. 23, no. 7, pp. 905–913, 1999.

[39] A. König, P. Müller, J. Tröger, H. Lindsay, J. Alexandersson, J. Hinze, M. Riemenschneider, D. Postin, E. Ettore, A. Lecomte, M. Musiol, M. Amblard, F. Bremond, M. Balazia, and R. Hurlemann, "Multimodal phenotyping of psychiatric

disorders from social interaction: Protocol of a clinical multicenter prospective study," *Personalized Medicine in Psychiatry*, vol. 33-34, p. 100094, 2022. [Online]. Available: https://hal.science/hal-03968278

[40] R. Dai, S. Das, K. Kahatapitiya, M. Ryoo, and F. F. Bremond, "Ms-tct: Multi-scale temporal convtransformer for action detection," in *CVPR 2022 - International Conference on Computer Vision and Pattern Recognition*, 2022. [Online]. Available: https://hal.inria.fr/hal-03682969

[41] M. Guermal, R. Dai, and F. F. Bremond, "Thorn: Temporal human-object relation network for action recognition," in *ICPR 2022 - International Conference on Pattern Recognition*, 2022. [Online]. Available: https://hal.science/hal-03698623

[42] R. Dai, S. Das, and F. Bremond, "CTRN: Class Temporal Relational Network For Action Detection," in *BMVC 2021 - The British Machine Vision Conference*, Virtual, United Kingdom, Nov. 2021, oral Presentation. [Online]. Available: https://hal.inria.fr/hal-03383140

[43] R. Dai, S. Das, M. Ryoo, and F. Bremond, "AAN : Attributes-Aware Network for Temporal Action Detection," in *BMVC 2023 - The 34th British Machine Vision Conference*, Aberdeen, United Kingdom, Nov. 2023. [Online]. Available: https://hal.science/hal-04241623

[44] J.-C. Hou, A. Mcgonigal, F. Bartolomei, and M. Thonnat, "A self-supervised pre-training framework for vision-based seizure classification," in *2022 IEEE International Conference on Acoustics, Speech, and Signal Processing proceedings*, 2022. [Online]. Available: https://hal.science/hal-03817281

[45] D. Yang, Y. Wang, Q. Kong, A. Dantcheva, L. Garattoni, G. Francesca, and F. Bremond, "Self-supervised video representation learning via latent time navigation," in *AAAI*, 2023.

[46] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, 1992.

[47] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, 2017.

[48] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2021.

[49] J. A. Russell, J.-A. Bachorowski, and J.-M. Fernández-Dols, "Facial and vocal expressions of emotion," *Annual Review of Psychology*, 2003.

[50] J. Hong, C. Lee, and H. Jung, "Late fusion-based video transformer for facial micro-expression recognition," *Applied Sciences*, 2022.

[51] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," *Proceedings of the ACM on international conference on multimodal interaction (ICMI)*, 2015.

[52] P. Khorrami, T. L. Paine, K. Brady, C. Dagli, and T. S. Huang, "How deep neural networks can improve emotion recognition on video data," *IEEE International Conference on Image Processing (ICIP)*, 2016.

[53] B. Mohan and M. Popa, "Temporal based emotion recognition inspired by activity recognition models," *International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2021.

[54] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[55] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[56] D. Agarwal, T. Agrawal, L. M. Ferrari, and F. Bremond, "From multimodal to unimodal attention in transformers using knowledge distillation," *IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS)*, 2021.

[57] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[58] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L. philippe Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," *ACM International Conference on Multimodal Interaction (ICMI)*, 2021.

[59] W. Zhang, F. Qiu, S. Wang, H. Zeng, Z. Zhang, R. An, B. Ma, and Y. Ding, "Transformer-based multimodal information fusion for facial expression analysis," *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2022.

[60] T. Zhang, C. Liu, X. Liu, Y. Liu, L. Meng, L. Sun, W. Jiang, F. Zhang, J. Zhao, and Q. Jin, "Multi-task learning framework for emotion recognition in-the-wild," *Eur. Conf. Comput. Vis. Worksh.*, 2022.

[61] H. Mao, Z. Yuan, H. Xu, W. Yu, Y. Liu, and K. Gao, "M-sena: An integrated platform for multimodal sentiment analysis," *Annual Meeting of the Association for Computational Linguistics System Demonstration Track (ACL)*, 2022.

[62] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *NeurIPS*, 2022.

[63] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6824–6835.

[64] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, "Learning invariance from generated variance for unsupervised person re-identification," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 06, pp. 7494–7508, jun 2023.

[65] D. Agarwal, T. Agrawal, L. Ferrari, and F. Bremond, "From Multimodal to Unimodal Attention in Transformers using Knowledge Distillation," in *AVSS 2021 - 17th IEEE International Conference on Advanced Video and Signal-based Surveillance*, Virtual, United States, Nov. 2021. [Online]. Available: https://hal.science/hal-03389126

[66] T. Agrawal, D. Agarwal, M. Balazia, N. Sinha, and F. Bremond, "Multimodal Personality Recognition using Cross-Attention Transformer and Behaviour Encoding," in *VISAPP '22: International Conference on Computer Vision Theory and Applications*, IAPR. virtual, United States: IEEE, Feb. 2022, pp. 501–508. [Online]. Available: https://hal.science/hal-03519184

[67] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[68] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[69] A. Das, X. Niu, A. Dantcheva, S. L. Happy, H. Han, R. Zeghari, P. Robert, S. Shan, F. Bremond, and X. Chen, "A spatio-temporal approach for apathy classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2561–2573, 2022.

[70] A. König, E. Mallick, J. Tröger, N. Linz, R. Zeghari, V. Manera, and P. Robert, "Measuring neuropsychiatric symptoms in patients with early cognitive decline using speech analysis," *European Psychiatry*, vol. 64, no. 1, p. e64, 2021. [Online]. Available: https://hal.science/hal-03477227

[71] S. Aslan and U. Güdükbay, "Multimodal video-based apparent personality recognition using long short-term memory and convolutional neural networks," *CoRR*, vol. abs/1911.00381, 2019. [Online]. Available: http://arxiv.org/abs/1911.00381

[72] D. Curto, A. Clapés, J. Selva, S. Smeureanu, J. C. S. J. Junior, D. Gallardo-Pujol, G. Guilera, D. Leiva, T. B. Moeslund, S. Escalera, and C. Palmero, "Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 2177–2188.

[73] L. M. Ferrari, G. Abi Hanna, P. Volpe, E. Ismailova, F. Bremond, and M. A. Zuluaga, "One-class autoencoder approach for optimal electrode set-up identification in wearable EEG event monitoring," in *EMBC 2021 - 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Virtuel, France, Oct. 2021. [Online]. Available: https://hal.science/hal-03367919

[74] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7044–7053.

[75] C. M. de Melo, A. Torralba, L. Guibas, J. DiCarlo, R. Chellappa, and J. Hodgins, "Next-generation deep learning based on simulators and synthetic data," *Trends in cognitive sciences*, 2021.

[76] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.

[77] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *CVPR*, 2020.

[78] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in *ICML*, 2021.

[79] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*, 2022.

[80] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, “Diffusion autoencoders: Toward a meaningful and decodable representation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 619–10 629.

[81] G. Kim, H. Shim, H. Kim, Y. Choi, J. Kim, and E. Yang, “Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding,” *arXiv preprint arXiv:2212.02802*, 2022.

[82] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[83] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, “Laion-5b: An open large-scale dataset for training next generation image-text models,” *arXiv preprint arXiv:2210.08402*, 2022.

[84] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in *INTERSPEECH*, 2018.

[85] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, “Imagen video: High definition video generation with diffusion models,” *arXiv preprint arXiv:2210.02303*, 2022.

[86] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, “Make-a-video: Text-to-video generation without text-video data,” in *ICLR*, 2023.

[87] A. Voynov and A. Babenko, “Unsupervised discovery of interpretable directions in the gan latent space,” *arXiv preprint arXiv:2002.03754*, 2020.

[88] Y. Shen, J. Gu, X. Tang, and B. Zhou, “Interpreting the latent space of gans for semantic face editing,” in *CVPR*, 2020.

[89] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, “Latent image animator: Learning to animate images via latent space navigation,” in *ICLR*, 2022.

[90] I. Joshi, M. Grimmer, C. Rathgeb, C. Busch, F. Bremond, and A. Dantcheva, “Synthetic data in human analysis: A survey,” *arXiv preprint arXiv:2208.09191*, 2022.

[91] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, “Learning invariance from generated variance for unsupervised person re-identification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[92] R. Kemker, M. McClure, A. Abitino, T. Hayes, and C. Kanan, “Measuring catastrophic forgetting in neural networks,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[93] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, “Learning Invariance from Generated Variance for Unsupervised Person Re-identification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, Dec. 2022. [Online]. Available: https://hal.science/hal-03931340

[94] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, “Joint Generative and Contrastive Learning for Unsupervised Person Re-identification,” in *CVPR 2021 - IEEE Conference on Computer Vision and Pattern Recognition*, Virtual, United States, Jun. 2021. [Online]. Available: https://hal.science/hal-03349257

[95] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, “Latent image animator: Learning to animate images via latent space navigation,” in *ICLR 2022 - The International Conference on Learning Representations*, 2022. [Online]. Available: https://hal.inria.fr/hal-03714584

9 Short CVs of Researchers

Francois BREMOND
DR1 Inria
French citizen, 24/04/1968
email:Francois.Bremond@inria.fr
http://www-sop.inria.fr/members/Francois.Bremond/

Professional career

2013 - now	Creation and direction of Inria STARS team
2012 - 2013	Inria Senior scientist (DR) in Pulsar team in Sophia-Antipolis
2007 - 2011	Inria research scientist (CR) in Pulsar team in Sophia-Antipolis
2000 - 2007	Inria research scientist (CR) in Orion team in Sophia-Antipolis
1997 - 1999	Post-doctoral position at University of Southern California (USA)

Education

2007:	**Habilitation à diriger des recherches**, University of Nice Sophia-Antipolis
1997:	**PhD in Computer Science** from University of Nice Sophia-Antipolis
1992:	**Master** in Computer Science from École Normale Supérieure Lyon and Université Claude Bernard (Lyon I).
1989:	**Graduated** in Computer Science from the École Normale Supérieure.

Scientific research

- **Computer vision:** object detection and tracking, motion analysis, pattern recognition.
- **Deep learning:** deep learning architectures, datasets for deep learning, Self-Attention, Knowledge Distillation, contrastive learning, self-learning, lifelong learning.
- **Cognitive vision:** video understanding, scene understanding, event recognition, behavior analysis, activity monitoring, multi-sensor fusion.
- **Artificial intelligence:** knowledge-based systems, spatio-temporal reasoning, machine learning, scenario modeling, context representation, uncertainty handling, knowledge acquisition, ontology.
- **PhD training:** Supervision of **26 PhDs** between 2007 and 2023.
- **Start-up:** Founding member in 2005 of the Keeneo start-up in visual-surveillance, Keeneo has been part of Digital Barriers since 2011 (www.digitalbarriers.com/).

Publications

- Author of **287 international publications**
  www-sop.inria.fr/members/Francois.Bremond/topicsText/myPublications.html

Awards

2019:	3AI Côte d'Azur, Chair holder AI for integrative computer medecine

Monique THONNAT
DR0-2 Inria
French citizen, 31/05/1957
email:Monique.Thonnat@inria.fr
http://www-sop.inria.fr/members/Monique.Thonnat/

Professional career

2017 - now	Research scientist at Inria STARS team in Sophia Antipolis
2013 - 2017	Director of Inria Bordeaux South-West
2009 - 2013	Inria Deputy Scientific Director, member of Pulsar Inria team.
2008 - 2009	Creation and direction of Inria Pulsar team in Sophia Antipolis
1995 - 2007	Creation and direction of Inria Orion team in Sophia Antipolis
1983 - 1995	Inria Research scientist at Pastis team in Sophia Antipolis

Education

2003:	**Habilitation à diriger des recherches** University of Nice Sophia Antipolis
1982:	**PhD in Optics and Signal Processing** from University Aix-Marseille III
1980:	**Engineer** ENSPM and DEA Signal and Spatio-temporal Systems from University of Marseille, France.

Scientific research

- **From 1980 to 1982: Image processing** for astronomical data reduction : background estimation and radial velocity computation.
- **From 1983 to 2007: Pattern recognition** and **Artificial intelligence** for complex object recognition (galaxy, zooplankton, fish). **Computer vision:** stereovision and interpretation of indoor scenes and urban scenes for obstacle avoidance. **Cognitive vision:** visual ontology, machine learning, classification and intelligent control. Real-time video analysis and **video understanding** (4D analysis, event recognition, activity recognition) and applications in visual surveillance and security.
- **From 2008: Human activity recognition** and their applications in healthcare.
- **PhD training:** Supervision of **29 PhDs** between 1989 and 2021. All my former students have got either an academic position (in France, USA, Singapore, Vietnam, Chili) or in the industry
- **Start-up:** Founding member in 2005 of the Keeneo start-up in visualsurveillance, Keeneo is part of Digital Barriers since 2011 (www.digitalbarriers.com/).

Publications and Patents

Publications: Author of **210 international publications**
(www-sop.inria.fr/members/Monique.Thonnat/index.html)
2 Patents:
Method and Apparatus for the Automatic Detection and Recognition of Pollen in 2000.
Method for building a geological model by seismic interpretation with cognitive vision techniques in 2008.

Awards

2008:	Appointed Knight of the National Order of the Legion of Honour by the Ministry of Research

Antitza DANTCHEVA
CRCN Inria
Austrian citizen, 11/06/1979
email:Antitza.Dantcheva@inria.fr
http://antitza.com

Professional career

2014 - now	Research scientist at Inria STARS team in Sophia Antipolis
2013 - 2014	Postdoctoral Fellow at Michigan State University, USA
2012	Postdoctoral Fellow at West Virginia University, USA
2006 - 2008	Researcher at Telecommunications Research Center Vienna (FTW), Austria

Education

2021	**Habilitation à diriger des recherches** Côte d'Azur University
2011	**Ph.D.** Signal and Image Processing from Eurecom / Télécom ParisTech, France.
2007	**M.S.** Electrical Engineering – Telecommunication Engineering from Vienna University of Technology, Austria.

Scientific research

- **From 2006 to 2008: Human Computer Interaction** for telecommunications.
- **From 2008 to 2013: Pattern recognition** and **Image Analysis** for face analysis in biometrics.
- **From 2014: Computer vision** facial video analysis for security and healthcare. Image and video generation.
- **PhD training:** Supervision of **3 PhDs** between 2018 and 2023.

Publications and Patents

Publications: Author of **70 international publications**
(https://scholar.google.com/citations?user=ZMggPHMAAAAJ&hl=en)
3 Patents: System and Method of Unveiling High-Resolution visible face images from Low-Resolution face images in 2022.
Thermal Face and Landmark detection method in 2022.
Cross-spectral face recognition training and cross-spectral face recognition method in 2022.

Honors and Awards

2017	Prestigious **ANR Jeunes chercheuses / Jeunes chercheurs** (JCJC) personal grant, Nov 2017 - Dec 2021.
2018	**Winner of ECCV'18 challenge on bias estimation in face analysis (BEFA)** at European Conference on Computer Vision (ECCV 2018). The team was composed of A. Das (Inria), A. Dantcheva (Inria), F. Brémond (INRIA).
2017	**Best Paper Award (Runner Up)** for *Spoofing Faces Using Makeup: An Investigative Study* at the IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2017).
2013	**Best Spoofing Attack Award** awarded by the European Project "Tabula Rasa".
2011	**Best Presentation Award** for *Female facial aesthetics based on soft biometrics and photo-quality* at the IEEE International Conference on Multimedia and Expo (ICME) 2011.

Michal BALAZIA
ISFP Inria
Slovakian citizen, 09/08/1988
email:Michal.Balazia@inria.fr
researchgate.net/profile/Michal-Balazia

Professional Career

2023 - now	Permanent researcher, Inria STARS team in Sophia Antipolis
2019 - 2023	Postdoctoral fellow, Inria STARS team in Sophia Antipolis
2018 - 2019	Postdoctoral fellow, University of South Florida, USA

Education

2018	**Ph.D.** in Computer Science, Masaryk University, Czech Republic
2016	**RNDr.** in Information Systems, Masaryk University, Czech Republic
2013	**Mgr.** in Information Technology Security, Masaryk University, Czech Republic

Scientific Research

- **From 2012 to 2018: Gait Recognition** from motion capture data in biometrics, signal processing, similarity search, linear machine learning models
- **From 2018 to 2019: Surveillance Event Detection** in extended video for video surveillance, object detection and tracking, activity recognition, deep learning
- **From 2019 to 2021: Face analysis** in biometrics, face uniqueness, deep neural networks
- **From 2019: Face and body analysis** for detection of psychiatric disorders, body language understanding, emotion recognition, behavior detection, transformers

Publications

- Author of **20 international publications**
  https://scholar.google.com/citations?user=idIT1iYAAAAJ

Grants and Awards

2021	**IDEX$^{\textbf{JEDI}}$ Fellowship for Young Researchers**, *Automated Face and Gesture Analysis for Digital Health Monitoring*, UCA, Sep 2021 - Jun 2023
2019	**IDEX$^{\textbf{JEDI}}$ Thematic Postdoctoral Grant**, *Deep Neural Networks: Assisted Face Analysis for Health Monitoring*, UCA, Oct 2019 - Feb 2021
2019	**Rector's Award for an Outstanding Doctoral Thesis**, *Gait Recognition from Motion Capture Data*, Masaryk University
2018	**Joseph Fourier Prize**, 1st place, *Gait Recognition from Motion Capture Data*, Institut Francais
2018	**IET Biometrics Premium Award (Best Paper)**, *Human Gait Recognition from Motion Capture Data in Signature Poses*, IET Biometrics Journal
2013	**Swedish Innovation Prize**, 1st place in Civil Security, *Gait Recognition for Biometric Surveillance*, Embassy of Sweden in Prague

10 Selection of Recent Team Publications on the Topic

Selected STARS Publications 2021

[A1] R. Dai, S. Das, L. Minciullo, L. Garattoni, G. Francesca, and F. F. Bremond, “PDAN: Pyramid Dilated Attention Network for Action Detection,” in *WACV 2021 - Winter Conference on Applications of Computer Vision 2021*, Waikoloa / Virtual, United States, Jan. 2021. [Online]. Available: https://inria.hal.science/hal-03026308

[A2] P. Robert, C. Albrengues, R. Fabre, A. Derreumaux, M. P. Pancrazi, I. Luporsi, B. Dubois, S. Epelbaum, G. Mercier, P. Foulon, V. Manera, and F. F. Bremond, “Efficacy of serious exergames in improving neuropsychiatric symptoms in neurocognitive disorders: Results of the X-TORP cluster randomized trial,” *Alzheimer’s & Dementia: Translational Research & Clinical Interventions*, vol. 7, no. 1, p. e12149, 2021. [Online]. Available: https://hal.umontpellier.fr/hal-03665103

[A3] I. Joshi, A. Utkarsh, R. Kothari, V. K. Kurmi, A. Dantcheva, S. Dutta, and P. K. Kalra, “Data Uncertainty Guided Noise-aware Preprocessing Of Fingerprints,” in *IJCNN 2021 - International Joint Conference on Neural Networks*, ser. 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen (online), China, Jul. 2021. [Online]. Available: https://hal.science/hal-03524646

[A4] S. Das, R. Dai, D. Yang, and F. F. Bremond, “VPN++: Rethinking Video-Pose embeddings for understanding Activities of Daily Living,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Dec. 2021. [Online]. Available: https://hal.science/hal-03485766

[A5] D. Yang, R. Dai, Y. Wang, R. Mallick, L. Minciullo, G. Francesca, and F. F. Bremond, “Selective Spatio-Temporal Aggregation Based Pose Refinement System: Towards Understanding Human Activities in Real-World Videos,” in *WACV 2021 – IEEE Winter Conference on Applications of Computer Vision*, Virtual, United States, Jan. 2021, wACV 2021. [Online]. Available: https://hal.science/hal-03121883

[A6] J.-C. Hou, “Quantified analysis for video recordings of seizure,” Theses, Université Côte d’Azur, Dec. 2021. [Online]. Available: https://theses.hal.science/tel-03565677

[A7] S. Majhi, S. Das, F. Brémond, R. Dash, and P. Kumar, “Weakly-supervised Joint Anomaly Detection and Classification,” in *FG 2021 - IEEE International Conference on Automatic Face and Gesture Recognition*, Jodhpur, India, Dec. 2021. [Online]. Available: https://hal.science/hal-03523563

[A8] R. Zeghari, R. Guerchouche, M. Tran Duc, F. F. Bremond, M. P. Lemoine, V. Bultingaire, K. Langel, Z. de Groote, F. Kuhn, E. Martin, P. Robert, and A. König, “Pilot Study to Assess the Feasibility of a Mobile Unit for Remote Cognitive Screening of Isolated Elderly in Rural Areas,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, p. 6108, Jun. 2021. [Online]. Available: https://hal.science/hal-03477302

[A9] N. Sinha, M. Balazia, and F. F. Bremond, “FLAME: Facial Landmark Heatmap Activated Multimodal Gaze Estimation,” in *AVSS 2021 - 17th IEEE International Conference on Advanced Video and Signal-based Surveillance*, Virtual, United States, Nov. 2021. [Online]. Available: https://inria.hal.science/hal-03386581

[A10] L. M. Ferrari, G. Abi Hanna, P. Volpe, E. Ismailova, F. F. Bremond, and M. A. Zuluaga, “One-class autoencoder approach for optimal electrode set-up identification in wearable EEG event monitoring,” in *EMBC 2021 - 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Virtuel, France, Oct. 2021. [Online]. Available: https://hal.science/hal-03367919

[A11] S. Majhi, S. Das, and F. Brémond, “DAM : Dissimilarity Attention Module for Weakly-supervised Video Anomaly Detection,” in *AVSS 2021 - 17th IEEE International Conference on Advanced Video and Signal-based Surveillance*, online, United States, Nov. 2021. [Online]. Available: https://hal.science/hal-03523616

[A12] J.-C. Hou, A. Mcgonigal, F. Bartolomei, and M. Thonnat, “A Multi-Stream Approach for Seizure Classification with Knowledge Distillation,” in *AVSS 2021 - 17th IEEE International Conference on Advanced Video and Signal-based Surveillance*, Virtual, United States, Nov. 2021. [Online]. Available: https://hal.science/hal-03433317

[A13] Y. Wang, “Learning to Generate Human Videos,” Theses, Inria - Sophia Antipolis ; Université Cote d’Azur, Sep. 2021. [Online]. Available: https://theses.hal.science/tel-03551913

[A14] ——, “Learning to generate human videos,” Theses, Université Côte d’Azur, Sep. 2021. [Online]. Available: https://theses.hal.science/tel-03662376

[A15] A. Das, S. Das, and A. Dantcheva, “Demystifying Attention Mechanisms for Deepfake Detection,” in *FG 2021 - IEEE International Conference on Automatic Face and Gesture Recognition*, virtual, India, Dec. 2021. [Online]. Available: https://hal.science/hal-03536498

[A16] A. Das, H. Lu, H. Han, A. Dantcheva, S. Shan, and X. Chen, “BVPNet: Video-to-BVP Signal Prediction for Remote Heart Rate Estimation,” in *FG 2021 - IEEE International Conference on Automatic Face and Gesture Recognition*, Jodhpur (virtual), India, Dec. 2021. [Online]. Available: https://hal.science/hal-03536497

[A17] L. M. Ferrari, U. Ismailov, F. Greco, and E. Ismailova, “Capacitive Coupling of Conducting Polymer Tattoo Electrodes with the Skin,” *Advanced Materials Interfaces*, vol. 8, no. 15, p. 2100352, Jul. 2021. [Online]. Available: https://hal.science/hal-03503743

[A18] M. Tabejamaat, F. Negin, and F. F. Bremond, “Guided Flow Field Estimation by Generating Independent Patches,” in *BMVC 2021 - 32nd British Machine Vision Conference*, Virtual, United Kingdom, Nov. 2021. [Online]. Available: https://hal.science/hal-03522489

[A19] H. Lindsay, J. Tröger, and A. König, “Language Impairment in Alzheimer’s Disease-Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through Multilingual Machine Learning,” *Frontiers in Aging Neuroscience*, vol. 13, no. 11, p. 6108, May 2021. [Online]. Available: https://hal.science/hal-03477304

[A20] A. König, E. Mallick, J. Tröger, N. Linz, R. Zeghari, V. Manera, and P. Robert, “Measuring neuropsychiatric symptoms in patients with early cognitive decline using speech analysis,” *European Psychiatry*, vol. 64, no. 1, p. e64, 2021. [Online]. Available: https://hal.science/hal-03477227

[A21] R. Dai, S. Das, and F. F. Bremond, “Learning an Augmented RGB Representation with Cross-Modal Knowledge Distillation for Action Detection,” in *ICCV 2021 - IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, Oct. 2021. [Online]. Available: https://hal.science/hal-03314575

[A22] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. F. Bremond, “Self-Supervised Video Pose Representation Learning for Occlusion-Robust Action Recognition,” in *FG 2021 - IEEE International Conference on Automatic Face and Gesture Recognition*, Jodhpur (Virtual), India, Dec. 2021. [Online]. Available: https://hal.science/hal-03476564

[A23] D. Agarwal, T. Agrawal, L. Ferrari, and F. F. Bremond, “From Multimodal to Unimodal Attention in Transformers using Knowledge Distillation,” in *AVSS 2021 - 17th IEEE International Conference on Advanced Video and Signal-based Surveillance*, Virtual, United States, Nov. 2021. [Online]. Available: https://hal.science/hal-03389126

[A24] R. Dai, S. Das, and F. F. Bremond, “CTRN: Class Temporal Relational Network For Action Detection,” in *BMVC 2021 - The British Machine Vision Conference*, Virtual, United Kingdom, Nov. 2021, oral Presentation. [Online]. Available: https://inria.hal.science/hal-03383140

[A25] T. L’Yvonnet, E. de Maria, S. Moisan, and J.-P. Rigault, “Probabilistic Model Checking for Activity Recognition in Medical Serious Games,” in *SEH 2021 - 3rd ICSE Workshop on Software Engineering for Healthcare*, Madrid, Spain, Jun. 2021. [Online]. Available: https://inria.hal.science/hal-03180187

[A26] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. F. Bremond, "Joint Generative and Contrastive Learning for Unsupervised Person Re-identification," in *CVPR 2021 - IEEE Conference on Computer Vision and Pattern Recognition*, Virtual, United States, Jun. 2021. [Online]. Available: https://hal.science/hal-03349257

[A27] V. Strizhkova, Y. Wang, D. Anghelone, D. Yang, A. Dantcheva, and F. Brémond, "Emotion Editing in Head Reenactment Videos using Latent Space Manipulation," in *FG 2021 - IEEE International Conference on Automatic Face and Gesture Recognition*, Jodhpur, India, Dec. 2021. [Online]. Available: https://hal.science/hal-03530150

[A28] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. F. Bremond, "UNIK: A Unified Framework for Real-world Skeleton-based Action Recognition," in *BMVC 2021 - The British Machine Vision Conference*, Virtual, United Kingdom, Nov. 2021, code is available at: https://github.com/YangDi666/UNIK. [Online]. Available: https://hal.science/hal-03476581

[A29] T. L'Yvonnet, E. de Maria, S. Moisan, and J.-P. Rigault, "Probabilistic Model Checking for Human Activity Recognition in Medical Serious Games," *Science of Computer Programming*, vol. 206, p. 102629, Jun. 2021. [Online]. Available: https://inria.hal.science/hal-03182420

[A30] E. de Maria, A. Bahrami, T. L'Yvonnet, A. Felty, D. Gaffé, A. Ressouche, and F. Grammont, "On the Use of Formal Methods to Model and Verify Neuronal Archetypes," *Frontiers of Computer Science*, vol. 16, p. 28, Oct. 2021. [Online]. Available: https://hal.science/hal-03053930

Selected STARS Publications 2022

[B1] I. Joshi, T. Dhamija, R. Kumar, A. Dantcheva, S. D. Roy, and P. K. Kalra, "Cross-Domain Consistent Fingerprint Denoising," *IEEE Sensors Letters*, vol. 6, no. 8, Jul. 2022. [Online]. Available: https://hal.science/hal-03966789

[B2] A. König, P. Müller, J. Tröger, H. Lindsay, J. Alexandersson, J. Hinze, M. Riemenschneider, D. Postin, E. Ettore, A. Lecomte, M. Musiol, M. Amblard, F. Bremond, M. Balazia, and R. Hurlemann, "Multimodal phenotyping of psychiatric disorders from social interaction: Protocol of a clinical multicenter prospective study," *Personalized Medicine in Psychiatry*, vol. 33-34, p. 100094, May 2022. [Online]. Available: https://hal.science/hal-03968278

[B3] C. Chen, D. Anghelone, P. Faure, and A. Dantcheva, "Attention-Guided Generative Adversarial Network for Explainable Thermal to Visible Face Recognition," in *IEEE International joint conference on biometrics*, Abu Dhabi, United Arab Emirates, Oct. 2022. [Online]. Available: https://hal.science/hal-03936358

[B4] S. Gregory, N. Linz, A. König, K. Langel, H. Pullen, S. Luz, J. Harrison, and C. W. Ritchie, "Remote data collection speech analysis and prediction of the identification of Alzheimer's disease biomarkers in people at risk for Alzheimer's disease dementia: the Speech on the Phone Assessment (SPeAk) prospective observational study protocol," *BMJ Open*, vol. 12, Mar. 2022. [Online]. Available: https://hal.science/hal-03967842

[B5] A. König, J. Tröger, E. Mallick, M. Mina, N. Linz, C. Wagnon, J. Karbach, C. Kuhn, and J. Peter, "Detecting subtle signs of depression with automated speech analysis in a non-clinical sample," *BMC Psychiatry*, vol. 22, Dec. 2022. [Online]. Available: https://hal.science/hal-03968260

[B6] R. Zeghari, R. Guerchouche, M. Tran-Duc, F. Bremond, K. Langel, I. Ramakers, N. Amiel, M. P. Lemoine, V. Bultingaire, V. Manera, P. Robert, and A. König, "Feasibility Study of an Internet-Based Platform for Tele-Neuropsychological Assessment of Elderly in Remote Areas," *Diagnostics*, vol. 12, Apr. 2022. [Online]. Available: https://hal.science/hal-03968301

[B7] M. Galliani, L. Ferrari, and E. Ismailova, "Interdigitated Organic Sensor in Multimodal Facemask's Barrier Integrity and Wearer's Respiration Monitoring," *Biosensors*, vol. 12, no. 5, p. 305, May 2022. [Online]. Available: https://hal.science/hal-03906369

[B8] R. Dai, S. Das, S. Sharma, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca, “Toyota Smarthome Untrimmed: Real-World Untrimmed Videos for Activity Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. [Online]. Available: https://hal.science/hal-03698616

[B9] R. Roy, I. Joshi, A. Das, and A. Dantcheva, “3D CNN Architectures and Attention Mechanisms for Deepfake Detection,” in *Handbook of Digital Face Manipulation and Detection : From DeepFakes to Morphing Attacks*, ser. Advances in Computer Vision and Pattern Recognition. ACVPR, S. I. Publishing, Ed., 2022. [Online]. Available: https://hal.science/hal-03524639

[B10] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, “Learning Invariance from Generated Variance for Unsupervised Person Re-identification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, Dec. 2022. [Online]. Available: https://hal.science/hal-03931340

[B11] M. Balazia, P. Müller, Á. L. Tánczos, A. V. Liechtenstein, and F. Brémond, “Bodily Behaviors in Social Interaction: Novel Annotations and State-of-the-Art Evaluation,” in *MM’22: The 30th ACM International Conference on Multimedia.* Lisbon, Portugal: ACM, Oct. 2022, pp. 70–79. [Online]. Available: https://hal.science/hal-03936267

[B12] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, “Latent Image Animator: Learning to Animate Images via Latent Space Navigation,” in *ICLR 2022 - The International Conference on Learning Representations*, Virtual, France, Apr. 2022. [Online]. Available: https://inria.hal.science/hal-03714584

[B13] I. Joshi, T. Prakash, B. Jaiswal, R. Kumar, A. Dantcheva, S. D. Roy, and P. K. Kalra, “Context-Aware Restoration of Noisy Fingerprints,” *IEEE Sensors Letters*, vol. 6, no. 10, p. 6003704, Oct. 2022. [Online]. Available: https://hal.science/hal-03966800

[B14] J.-C. Hou, M. Thonnat, F. Bartolomei, and A. Mcgonigal, “Automated video analysis of emotion and dystonia in epileptic seizures,” *Epilepsy Research*, vol. 184, Aug. 2022. [Online]. Available: https://hal.science/hal-03817305

[B15] F. Negin, M. Tabejamaat, F. F. Bremond, and R. Fraisse, “Transforming Temporal Embeddings to Keypoint Heatmaps for Detection of Tiny Vehicles in Wide Area Motion Imagery (WAMI) Sequences,” in *CVPR 2022 - IEEE /CVF - Computer Vision and Pattern Recognition Conference*, New Orleans, Louisiana, United States, Jun. 2022. [Online]. Available: https://inria.hal.science/hal-03936192

[B16] T. Agrawal, D. Agarwal, M. Balazia, N. Sinha, and F. F. Bremond, “Multimodal Personality Recognition using Cross-Attention Transformer and Behaviour Encoding,” in *VISAPP ’22: International Conference on Computer Vision Theory and Applications*, IAPR. virtual, United States: IEEE, Feb. 2022, pp. 501–508. [Online]. Available: https://hal.science/hal-03519184

[B17] S. Chen, Y. Cho, K. Yu, L. Ferrari, and F. F. Bremond, “Editorial: Recognizing the state of emotion, cognition and action from physiological and behavioral signals,” *Frontiers in Computer Science*, vol. 4, Aug. 2022. [Online]. Available: https://hal.science/hal-03906373

[B18] I. Joshi, A. Utkarsh, P. Singh, A. Dantcheva, S. D. Roy, and P. K. Kalra, “On Restoration of Degraded Fingerprints,” *Multimedia Tools and Applications*, vol. 81, no. 24, pp. 35 349–35 377, Oct. 2022. [Online]. Available: https://hal.science/hal-03966796

[B19] E. Ettore, P. Mueller, J. Hinze, M. Benoit, B. Giordana, D. Postin, A. Lecomte, H. Lindsay, P. Robert, and A. König, “Digital phenotyping for differential diagnosis of Major Depressive Episode: A literature review (Preprint),” *JMIR Mental Health*, Feb. 2022. [Online]. Available: https://hal.science/hal-03968289

[B20] H. Chen, “Towards unsupervised person re-identification,” Theses, Université Côte d’Azur, May 2022. [Online]. Available: https://theses.hal.science/tel-03783651

[B21] J. D. Gonzales Zuniga, U. Ujjwal, and F. F. Bremond, “DeTracker: A Joint Detection and Tracking Framework,” in *VISAPP 2022 - 17th International Conference on Computer Vision Theory and Applications*, online, France, Feb. 2022. [Online]. Available: https://hal.science/hal-03541517

[B22] T. L’Yvonnet, “Relationships between human activity models and brain models : application to clinical serious games,” Theses, Université Côte d’Azur, Mar. 2022. [Online]. Available: https://theses.hal.science/tel-03685758

[B23] J.-C. Hou, A. Mcgonigal, F. Bartolomei, and M. Thonnat, “A Self-Supervised Pre-Training Framework for Vision-Based Seizure Classification,” in *IEEE ICASSP 2022 : IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, Singapore, May 2022. [Online]. Available: https://hal.science/hal-03817281

[B24] M. Guermal, R. Dai, and F. F. Bremond, “THORN: Temporal Human-Object Relation Network for Action Recognition,” in *ICPR 2022 - International Conference on Pattern Recognition*, Montreal, Canada, Aug. 2022. [Online]. Available: https://hal.science/hal-03698623

[B25] A. Ali, F. F. Negin, F. F. Bremond, and S. Thümmler, “Video-based Behavior Understanding of Children for Objective Diagnosis of Autism,” in *VISAPP 2022 - 17th International Conference on Computer Vision Theory and Applications*, Online, France, Feb. 2022. [Online]. Available: https://inria.hal.science/hal-03447060

[B26] R. Dai, “Action detection for untrimmed videos based on deep neural networks,” Theses, Université Côte d’Azur, Sep. 2022. [Online]. Available: https://theses.hal.science/tel-03827178

[B27] R. Dai, S. Das, K. Kahatapitiya, M. Ryoo, and F. F. Bremond, “MS-TCT: Multi-Scale Temporal ConvTransformer for Action Detection,” in *CVPR - Conference on Computer Vision and Pattern Recognition*, New Orleans, United States, Jun. 2022. [Online]. Available: https://inria.hal.science/hal-03682969

[B28] D. Anghelone, S. Lannes, V. Strizhkova, P. Faure, C. Chen, and A. Dantcheva, “TFLD: Thermal Face and Landmark Detection for Unconstrained Cross-spectral Face Recognition,” in *IJCB 2022 - IEEE International joint conference on biometrics*, Abu Dhabi, United Arab Emirates, Oct. 2022. [Online]. Available: https://hal.science/hal-03936331

[B29] L. Domain, M. Guillery, N. Linz, A. König, J.-M. Batail, R. David, I. Corouge, E. Bannier, J.-C. Ferré, T. Dondaine, D. Drapier, and G. Robert, “Multimodal MRI cerebral correlates of verbal fluency switching and its impairment in women with depression,” *Neuroimage-Clinical*, vol. 33, p. 102910, 2022. [Online]. Available: https://hal.science/hal-03477309

Selected STARS Publications 2023

[C1] P. Müller, M. Balazia, T. Baur, M. Dietz, A. Heimerl, D. Schiller, M. Guermal, D. Thomas, F. F. Bremond, J. Alexandersson, E. André, and A. Bulling, “MultiMediate ’23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions,” in *MM 2023 - The 31st ACM International Conference on Multimedia*, ACM. Ottawa, Canada: ACM, Oct. 2023, pp. 9640–9645. [Online]. Available: https://hal.science/hal-04330332

[C2] R. Dai, S. Das, M. Ryoo, and F. Bremond, “AAN : Attributes-Aware Network for Temporal Action Detection,” in *BMVC 2023 - The 34th British Machine Vision Conference*, Aberdeen, United Kingdom, Nov. 2023. [Online]. Available: https://hal.science/hal-04241623

[C3] D. Anghelone, S. Lannes, and A. Dantcheva, “ANYRES: Generating High-Resolution visible-face images from Low-Resolution thermal-face images,” in *IEEE ICME 2023 - IEEE International Conference on Multimedia and Expo*, ser. 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane (AU), Australia, Jul. 2023. [Online]. Available: https://hal.science/hal-04391831

[C4] I. Joshi, A. Utkarsh, R. Kothari, V. K. Kurmi, A. Dantcheva, S. D. Roy, and P. K. Kalra, "On Estimating Uncertainty of Fingerprint Enhancement Models," in *Digital Image Enhancement and Reconstruction*, ser. Hybrid Computational Intelligence for Pattern Analysis series. Elsevier, Jan. 2023, no. Chapter 2, pp. 29–70. [Online]. Available: https://hal.science/hal-04391813

[C5] R. D. Labati, A. Ross, and A. Dantcheva, "Soft Biometrics," 2023. [Online]. Available: https://hal.science/hal-04391864

[C6] V. Thamizharasan, A. Das, D. Battaglino, F. F. Bremond, and A. Dantcheva, "Face Attribute Analysis from Structured Light: An End-to-End Approach," *Multimedia Tools and Applications*, vol. 82, no. 7, pp. 10 471–10 490, Mar. 2023. [Online]. Available: https://hal.science/hal-04391848

[C7] B. Chopin, M. Daoudi, and A. Bartolo, "Avatar Reaction to Multimodal Human Behavior," in *ICIAP 2023- 22nd International Conference on Image Analysis and Processing*, vol. 14365, Udine, Italy, Sep. 2023, pp. 494–505, workshops session. [Online]. Available: https://hal.science/hal-04474370

[C8] T. Agrawal, M. Balazia, P. Müller, and F. F. Bremond, "Multimodal Vision Transformers with Forced Attention for Behavior Analysis," in *WACV '23: IEEE International Winter Conference on Applications in Computer Vision*. Waikoloa, United States: IEEE, Jan. 2023. [Online]. Available: https://hal.science/hal-03936484

[C9] S. Majhi, R. Dai, Q. Kong, L. Garattoni, G. Francesca, and F. F. Bremond, "Human-Scene Network: A Novel Baseline with Self-rectifying Loss for Weakly supervised Video Anomaly Detection," Jan. 2023, working paper or preprint. [Online]. Available: https://inria.hal.science/hal-03946181

[C10] D. Yang, Y. Wang, A. Dantcheva, Q. Kong, L. Garattoni, G. Francesca, and F. Bremond, "LAC - Latent Action Composition for Skeleton-based Action Segmentation," in *ICCV 2023 - IEEE/CVF International Conference on Computer Vision*, Paris, France, Oct. 2023. [Online]. Available: https://hal.science/hal-04236097

[C11] T. Stanczyk and F. F. Bremond, "Current Challenges with Modern Multi-Object Trackers," in *ACVR 2023 - Eleventh International Workshop on Assistive Computer Vision and Robotics*, Paris, France, Oct. 2023, aCVR 2023 in conjonction with ICCV 2023. [Online]. Available: https://hal.science/hal-04323242

[C12] P. Balaji, A. Das, S. Das, and A. Dantcheva, "Attending Generalizability in Course of Deep Fake Detection by Exploring Multi-task Learning," in *2023 IEEE/CVF 6 International Conference on Computer Vision Workshops (ICCVW)*, ser. 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, Oct. 2023. [Online]. Available: https://hal.science/hal-04397222

[C13] B. Chopin, M. Daoudi, and A. Bartolo, "Avatar Reaction to Multimodal Human Behavior," in *GHB 2023 - Generation of Human Face and Body Behavior Workshop in conjonction with ICIAP*, ser. Image Analysis and Processing – ICIAP 2023 22nd International Conference, ICIAP 2023, Udine, Italy, September 11–15, 2023, Proceedings, Part I, no. LNCS-14233, Udine, Italy, Sep. 2023. [Online]. Available: https://hal.science/hal-04305332

[C14] E. de Maria, B. Lapijover, T. l'Yvonnet, S. Moisan, and J.-P. Rigault, "A Formal Probabilistic Model of the Inhibitory Control Circuit in the Brain," in *BIOINFORMATICS 2023 - 14th International Conference on Bioinformatics Models, Methods and Algorithms*, Lisbonne, Portugal, Feb. 2023. [Online]. Available: https://inria.hal.science/hal-03999574

[C15] D. Anghelone, "Computer vision and deep learning applied to face recognition in the invisible spectrum," Theses, Université Côte d'Azur, Jun. 2023. [Online]. Available: https://theses.hal.science/tel-04224480

[C16] H. Chaptoukaev, V. Strizhkova, M. Panariello, B. D'alpaos, A. Reka, V. Manera, S. Thümmler, E. Ismailova, N. Evans, F. F. Bremond, M. Todisco, M. A. Zuluaga, and L. M. Ferrari, "StressID: a Multimodal Dataset for Stress Identification," in *NeurIPS 2023*

- *37th Conference on Neural Information Processing Systems.* New Orleans, United States: NIST, Dec. 2023. [Online]. Available: https://hal.science/hal-04245507

[C17] D. Yang, Y. Wang, Q. Kong, A. Dantcheva, L. Garattoni, G. Francesca, and F. F. Bremond, "Self-Supervised Video Representation Learning via Latent Time Navigation," in *AAAI 2023 - AAAI Conference on Artificial Intelligence*, ser. Proceedings of the 37th AAAI Conference on Artificial Intelligence, vol. 37, no. 3, Washigton, D.C., United States, Feb. 2023, aAAI 2023. [Online]. Available: https://hal.science/hal-04236128

[C18] E. Ettore, P. Müller, J. Hinze, M. Riemenschneider, M. Benoit, B. Giordana, D. Postin, R. Hurlemann, A. Lecomte, M. Musiol, H. Lindsay, P. Robert, and A. König, "Digital Phenotyping for Differential Diagnosis of Major Depressive Episode: Narrative Review," *JMIR Mental Health*, vol. 10, p. e37225, Jan. 2023. [Online]. Available: https://hal.science/hal-04402917

Selected STARS Publications 2024

[D1] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. Brémond, "View-invariant Skeleton Action Representation Learning via Motion Retargeting," *International Journal of Computer Vision*, Jan. 2024, project website: https://walker-a11y.github.io/ViA-project. [Online]. Available: https://hal.science/hal-03906649

[D2] S. Majhi, R. Dai, Q. Kong, L. Garattoni, G. Francesca, and F. Bremond, "Human-scene network: A novel baseline with self-rectifying loss for weakly supervised video anomaly detection," *Computer Vision and Image Understanding*, vol. 241, p. 103955, 2024.

[D3] B. Chopin, H. Tang, and M. Daoudi, "Bipartite Graph Diffusion Model for Human Interaction Generation," in *WACV 2024 - IEEE/CVF Winter Conference on Applications of Computer Vision*, WAIKOLOA (Hawaii), United States, Jan. 2024. [Online]. Available: https://hal.science/hal-04274209

[D4] A. Ali, M. Ashish, and F. F. Bremond, "P-Age: Pexels Dataset for Robust Spatio-Temporal Apparent Age Classification," in *WACV 2024 - Winter Conference on Applications of Computer Vision*, HAWAII, United States, Jan. 2024. [Online]. Available: https://inria.hal.science/hal-04356537

[D5] S. Majhi, R. Dai, Q. Kong, L. Garattoni, G. Francesca, and F. Brémond, "Oe-ctst: Outlier-embedded cross temporal scale transformer for weakly-supervised video anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 8574–8583.

[D6] M. Guermal, A. Ali, R. Dai, and F. Brémond, "Joadaa: Joint online action detection and action anticipation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 6889–6898.