3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Ph.D. research topic

- Title of the proposed topic (en anglais): **Robust Detection of DeepFakes**
- PhD or post-doc: **PhD and a Post-doc**
- Research axis of the 3IA: **l'IA et les territoires intelligents et sécurisés**
- Supervisor (name, affiliation, email): **Antitza Dantcheva, INRIA Sophia Antipolis, antitza.dantcheva@inria.fr**
- Potential co-supervisor (name, affiliation): **Francois Bremond**
- The laboratory and/or research group: **STARS Team of Inria**

- The description of the topic:

Manipulated images and videos, i.e., *deepfakes* have become increasingly realistic due to the tremendous progress of deep convolutional neural networks (CNNs). While technically intriguing, such progress raises a number of social concerns. In particular, such manipulations can fabricate animations of subjects involved in actions that have not taken place and such manipulated data can be circumvented nowadays rapidly via social media. Hence, we cannot trust anymore, what we see or hear on video, as *deepfakes* betray sight and sound, the two predominantly trusted human innate senses, posing a threat of distorting what is perceived as reality. To further fuel concern, deepfake techniques have become open to the public via phone applications such as FaceApp and ZAO.

We differentiate two cases of concern: the first one has to do with *deepfakes being perceived as real*, and the second relates to *real videos being misdetected for fake*, the latter referred to as "liar's dividend".

Such concerns necessitate the introduction of robust and reliable methods for fake image and video detection.

Motivated by the above, we propose research which studies detection of manipulated videos.

We note that the detection of deepfakes is challenging for several reasons: (a) it evolves a ``cat-and-mouse-game'' between the adversary and the system designer, (b) deep models are highly domain-specific and likely yield big performance degradation in cross-domain deployments, especially with large train-test domain gap.

Considering this, we intend to investigate three strategies of detection by designing algorithms that can successfully generalize onto unknown manipulations.

(a) - **3D CNNs**. Our intuition is that current state of the art forgery detection techniques omit a pertinent clue, namely *temporal information* by investigating only spatial information. In our extensive work on video generation, we have found out that generative models have exhibited difficulties in *preserving appearance* throughout generated videos, as well as *motion consistency* and we intend to exploit these factors in deepfake detection.

(b) - **Few-shot learning**. We will focus on learning *talking-patterns* pertaining to a set of enrolled subjects. We intend to classify the integrity of videos by analyzing the likelihood that a portrayed talking-behavior to belong to an enrolled person. By few-shot learning we intend to generalize this method onto unseen subjects.

(c) - **Generative noise**. Images and videos acquired by sensors incorporate a noise-pattern specific to the sensor (caused by the physical properties of the sensor). Hence, such noise-pattern is absent in generated data. In this context, we intend to study (a) whether such noise-pattern is instrumental in classifying deepfakes, (b) whether such noise pattern can be added onto generated data and hence renders videos not detectable, see (a), as well as (c) whether generative adversarial networks incorporate a generative pixel-level noise into generated data.

Please find a extended version under http://antitza.com/DFD.pdf