ERC Starting Grant 2020

Research proposal [Part B1]

Expectational Visual Artificial Intelligence

EVA

|  |  |
| --- | --- |
| Personal details |  |
| Principal Investigator (PI) | Efstratios Gavves |
| Host Institution for the project | University of Amsterdam |
| Proposal full title | Expectational Visual Artificial Intelligence |
| Proposal short name | EVA |
| Proposal duration in months | 60 |

**Summary**

Visual artificial intelligence automatically interprets what happens in visual data like videos. Today’s research strives with queries like: *“Is this person playing basketball?”*; *“Find the location of the brain stroke”*; or *“Track the glacier fractures in satellite footage”*. All these queries are about visual observations already taken place. Today’s algorithms focus on explaining past visual observations. Naturally, not all queries are about the past: *“Will this person draw something in or out of their pocket?”; “Where will the tumour be in 5 seconds given breathing patterns and moving organs?”*; or, *“How will the glacier fracture given the current motion and melting patterns?”.* Forthese queries and all others,the next generation of visual algorithms must expect what happens next given past visual observations. Visual artificial intelligence must also be able to prevent before the fact, rather than explain only after it. I propose an ambitious 5-year project to design algorithms that learn to expect the possible futures from visual sequences.

The main challenge for expecting possible futures is having visual algorithms that learn temporality in visual sequences. Today’s algorithms cannot do this convincingly. First, they are time-deterministic and ignore uncertainty, part of any expected future. I propose time-stochastic visual algorithms. Second, today’s algorithms are time-extrinsic and treat time as an external input or output variable. I propose time-intrinsic visual algorithms that integrate time within their latent representations. Third, visual algorithms must account for all innumerable spatiotemporal dynamics, despite their finite nature. I propose time-geometric visual algorithms that constrain temporal latent spaces to known geometries.

EVA addresses fundamental research issues in the automatic interpretation of future visual sequences. Its results will serve as a basis for ground-breaking technological advances in practical vision applications.

**Section a: Extended Synopsis of the scientific proposal (5/5 pages)**

Visual artificial intelligence automatically interprets what happens in visual data like videos. Today’s research strives to tackle queries like: *“Is this person playing basketball?”* (action recognition) [G1, G4, 21, 22, 32]; *“Find the location of the brain stroke”* (video object segmentation) [48]; or *“Track the glacier fractures in satellite footage”* (visual object tracking) [G10, G15, 42-44]. All these queries are about predictions on visual observations that have already taken place. In other words, **today’s algorithms** focus on **explaining past** visual observations. Naturally, not all queries are about the past: *“Will this person draw something in or out of their pocket?”; “Where will the tumour be in the next 1,3, or 5 seconds given breathing patterns and moving organs?”*; or *“How will the glacier fracture given the current motion and melting patterns?”.* Forthese queries and all others,the **next generation of visual algorithms** must be able to **expect** what will happen next in the **future** given past visual observations. Visual artificial intelligence must also be able to prevent before the fact, rather than explain only after the fact. I propose an ambitious 5-year project to design **algorithms that learn to expect the possible futures** from visual sequences.

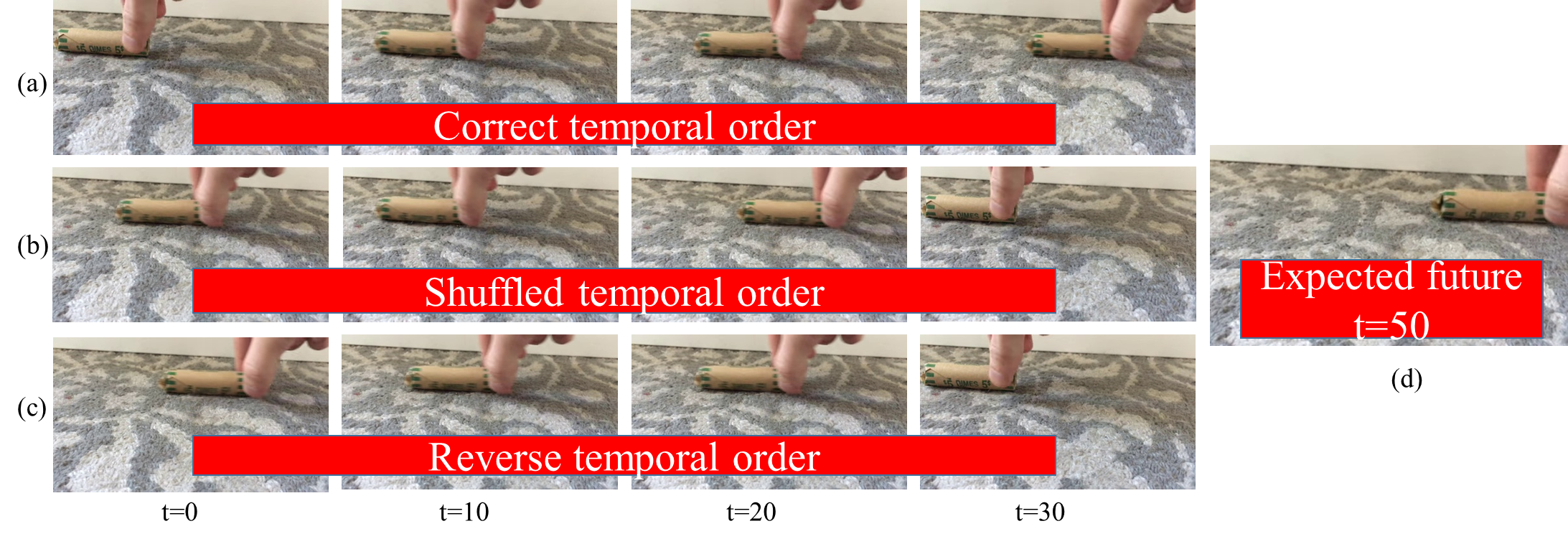
****Learning to expect the possible futures in visual sequences requires understanding of the sequence of events, *i.e.*, learning the **temporality** underpinning visual sequences. Consider the toy sequence in Figure 1 from the popular Something-Something [28] dataset, showing a paper cylinder moving. In Figure 1a the video frames are in the correct temporal order, in Figure 1b the frames are shuffled. From the shuffled video frames, it is impossible to predict the expected future location of the cylinder two seconds ahead of time. From the correct temporal order, however, it becomes immediately clear that in two seconds the cylinder is expected in the right side of the image (Figure 1d). While in this toy example the importance of temporality is perhaps superficial[[1]](#footnote-1), temporality will be critical in tomorrow’s tasks. It will be critical for determining whether a suspect is expected to draw something in or out of their pocket in surveillance videos. It will be critical for determining where the tumour is expected to move in the next few seconds in MRI sequences. And, it will be critical to determine the rate at which a glacier is expected to melt and fracture in the near future. To be able to expect the possible futures, it is imperative to have visual algorithms that model time and learn the temporality in visual sequences.

Figure 1. Understanding time is critical for algorithms to expect the future given past observations. For instance, when moving a paper cylinder, only with the correct temporal order it is possible to predict the expected future in (d), while (b)shuffled or (c) reverse create confusion.

**Scientific challenges.** Most of today’s algorithms that interpret visual sequences are **time-deterministic**. However, time and the expected future is anything but. In learning the future, multiple sources of uncertainty and stochasticity exist. A first source of stochasticity is that there exist many plausible futures given a visual sequence; eventually only one of them is observed. *Given a person going to the kitchen, will she make coffee, or tea or prepare dinner?* A second source of stochasticity is due to the uncertainty in the spatiotemporal appearances of the observed visual sequences. *In the given visual sequence, are these two persons fighting or they are having a conversation?* A third source of stochasticity relates to the so called “jumpy imagination” principle [49], namely when the various futures are expected to happen. When making coffee or tea, for instance, the immediate future steps (boiling water, stirring, and so on) will happen within seconds. When preparing dinner, however, the immediate future steps (open the fridge, wash vegetables, and so on) take place at different and non-uniformly spread timestamps. Today’s visual recognition algorithms do not incorporate stochasticity and uncertainty. They operate on restricted setups, where visual sequences are available in their entirety in advance. Each sequence example is always annotated with respect to a predefined set of classes. And, their predictions are in fixed and pre-determined time spans. Indeed, if visual artificial intelligence is to expect the future, **the first fundamental challenge** is to design **stochastic spatiotemporal** **visual algorithms that** **can handle uncertainty in** visual sequences. As a promising direction, I propose time-stochastic visual algorithms with temporally invertible generative models.

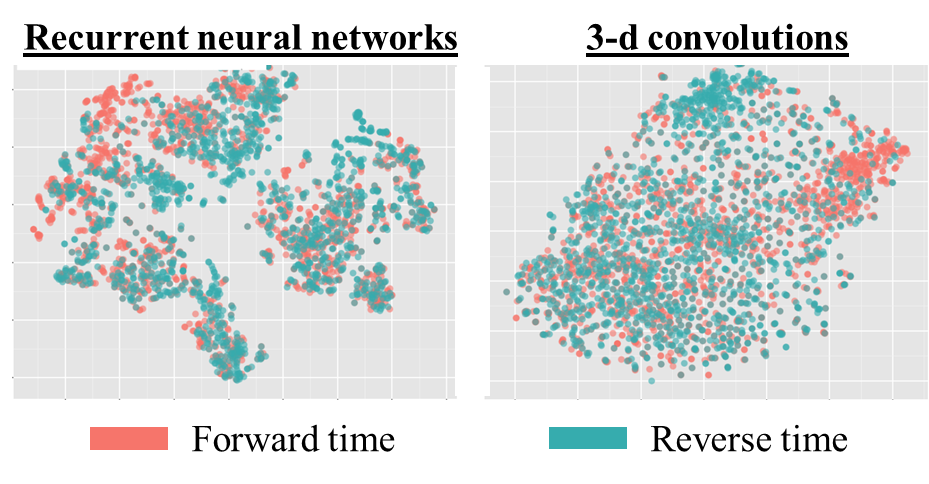
In the recent years, the video understanding literature [G1, G4, G14, 21, 31-33, 35-39] has been exploring ways of breaking with the manual supervision paradigm in visual sequences. Instead, the capabilities of temporal self-supervision for training visual algorithms have been tested. Despite the larger training sizes available with temporal self-supervision, to date self-supervised algorithms have not been able to surpass manual annotation. Recently, [36] hypothesize and show that temporal self-supervision is more fruitful when having algorithms with an extended temporal footprint. This makes sense, as the goal of video understanding is different (quantify appearance and temporal nuances) than the goal of purely image understanding (focus only on appearance, ignore all nuances including temporal that are irrelevant to object semantics). Unfortunately, most of today’s visual algorithms do not model time convincingly. The fundamental problem is that they view time as an extrinsic variable, to be added either as an extra input or output dimension in their spatiotemporal neural networks. That is**, today’s video algorithms** are **time-extrinsic** by design. They rely on two components: *(i)* convolutions (2-d [20, 34], or 3-d [21, 33]), followed by *(ii)* a temporal aggregation mechanism [G4, G14, 21, 33]. Both components **sideline time** either on purpose or by consequence. 3-d convolutions can at best only encode short-range motions [21] that discriminate between coarse movements (*e.g.*, *“eating an apple”* vs. *“throwing an apple”*)but struggle with fine-grained movements (*e.g.*, *“apple falling down from the tree to the ground”* vs. *“apple moving from the ground up to the tree”*). Furthermore, the most popular temporal aggregation is taking a statistical average or max over time, both **ignoring temporality** by design. A close second for a temporal aggregation is recurrent nets [G13, G16, 22]. Many have shown [G13, 22] that recurrent networks ignore temporality in practice, as a consequence of chaotic optimization [50]. Be it 3-d convolutions or recurrent networks, today’s algorithms have trouble even telling apart whether a video is played in forward (like Figure 1a) or in the reverse (like in Figure 1c) as shown in [G13], see Figure 2. Indeed, to escape from manual supervision and move towards temporal self-supervision, **the second fundamental challenge** is to design visual recognition algorithms **that encode time-intrinsically** in their latent representations**.** As a promising direction,I propose time-intrinsic visual algorithmsby aligning their neural network structure with time**.**

Figure 2. Today’s visual recognition algorithms have trouble modelling time in visual sequences; note the overlap even in forward vs reverse time representations in videos.

A core reason why encoding time in visual sequences is hard is the conflict between what needs to be learnt and what can be learnt by algorithms. On one hand, **any visual recognition algorithm** –be it a recurrent neural network or 3-d convolutions– **is by nature finite**; finite number of parameters, finite number of convolutions, finite number of frames and finite number of labels. On the other hand, even in simple videos there exist several hundreds or thousands of frames with hundreds thousands of pixels per frame. To encode time, an algorithm must account for the practically innumerable dynamics between pixels within a frame and between frames. Today’s video algorithms learn their temporal representations directly on open and unconstrained latent spaces. This is problematic, as the algorithms learn to overfit to the innumerable spatiotemporal dynamics [G13, 33, 50] in the training data, rather than generalize to the general spatiotemporal dynamics patterns. Thus, **the third fundamental challenge** **is to design visual recognition algorithms that** despite their finite nature **can account for the** **innumerable spatiotemporal dynamics.** As a promising direction,I propose time geometric visual algorithms, by fitting known geometries to the spatiotemporal latent space.

**Aim and objectives.** **The overall aim** of this project is to design **visual recognition algorithms** that **given past** video observations **predict the expected future**, be it future object locations, future people’s actions, even future’s cause and effect patterns**.** I want to learn what people are likely to do next given their past actions and behavior. I want to learn where objects are expected given their past and co-located objects. Ideally, I want to even learn cause and effect mechanisms in visual interactions and reuse them in novel visual scenes. Notably, these are few examples only. **I propose to learn the expected futures given past observations by designing visual algorithms that are time-stochastic, time-intrinsic and time-geometric.** Starting from the scientific challenges, the proposal has the following objectives.

|  |
| --- |
| O1. Study, develop and evaluate time-stochastic visual algorithms (first challenge)  O2. Study, develop and evaluate time-intrinsic visual algorithms (second challenge)  O3. Study, develop and evaluate time-geometric visual algorithms (third challenge)  O4. ShowTime: evaluation, dissemination & benchmarking |

The approach is timely and builds upon three key recent technological advances: *(i)* the rapid developments in video understanding with an emphasis on action recognition [G1, G4, G14, 21, 31-33, 35-39] and long-term tracking [G10, G15, 42-44], *(ii)* the immense progress in deep generative [51-54], geometric [30] and dynamical [29] learning, and *(iii)* the emergence of a plethora of rich video datasets [21, 28, 44, 45, 55-57] ranging from short [21] to very long [55], including specialized datasets that focus on temporality [28] or temporal relational reasoning [56].

**State-of-the-art.** From the pioneering spatiotemporal local descriptors [31] to spatiotemporal trajectories for actions [32], the interest in videos has a long history in visual recognition. With the arrival of deep learning the focus moved towards deep neural networks. Early attempts with recurrent neural networks [57] showed that pre-deep learning methods are surprisingly competitive and large datasets are key to better video understanding. Today, 3-d convolutions [21, 23] including spatiotemporal decompositions in my previous work [G14], and algorithms that consider spatiotemporal context [27] are the frontrunners in several video benchmarks. Generative models of visual sequences [54] with video pixel networks have also been explored. They model the entire visual sequence by a joint probability distribution, a design choice that is hard to satisfy for longer sequences and larger number of pixels. The above methods focus on making predictions in videos in past video observations.

Expecting the future by visual algorithms has received interest very recently. [59-63] explore the role of motion, locality and pose for future actions, pose, trajectories and segmentations, also from an egocentric [58] point of view. [64] proposed early action recognition where the beginning and end of the action is known in advance, whereas in previous work [G19] we explored online action prediction, a task that proved demonstrably harder. Predicting future trajectories [70, 71] has been also popular, assuming that perfect past trajectories are available a priori. Several works [65-67] proposed future frame prediction as a proxy task for unsupervised learning of representations, usually focusing on aesthetically pleasing visual reconstructions for the immediate next frames. Most of these works explore the possible applications with predicting the future and propose application-specific algorithms using domain knowledge.

This proposal is interested in a general framework for learning to expect the future, with learning temporality as the central focus. Learning temporality has attracted sudden interest from a vision [35, 36, G13] and learning [49, 75, 76] perspective. Vision algorithms that attempt to learn temporality show promising results when quantifying some properties of temporality, like arrow-of-time [36, G13] or temporal causality [G13]. However, they cannot compete with strong supervision, they typically assume predetermined environments, and they cannot generalize to visual sequences longer than a few seconds. From a learning point of view [49, 75, 76], learning temporal abstractions has also provided promising evidence. These models, however, assume artificial setups of synthetic environments that can be manipulated by reinforcement learning. Further future research is required to generalize to realistic data and without assuming interactive environments.

Today’s visual algorithms are time-deterministic, time-extrinsic and with unconstrained temporal latent spaces. This paradigm suffices when making predictions about past visual observations. However, it is unsuited when the goal is predictions about the expected future. The methodological originality of EVA comes from breaking with this paradigm. The proposition is to explore time-stochastic, time-intrinsic and time geometric visual algorithms. By the end of the project, EVA will enable predicting the expected futures in long and complexvisual sequences with minimal manual supervision.

**Methodology.** EVA proposes four workpackages, one per objective.

**WP1. Time-stochastic visual algorithms (PhD1+PostDoc+PI).** We explore with PhD1 time-stochastic visual algorithms accounting for various sources of uncertainty. For addressing the first source of stochasticity, namely uncertainty in the observed visual sequences, we recast popular deterministic visual algorithms for sequences like I3D [21] or our Timeception [G14] within the variational inference framework [51]. Variational autoencoders effectively rely on deterministic encoders and decoders, interleaved with stochastic layers. As shown [51] the deterministic encoders and decoders can be implemented by any feature extractor networks. We investigate I3D and Timeception as feature extractors (task 1.1). Next, we explore deep generative models for temporal sequences. Specifically, deterministic time-aligned DenseNets from previous work [G13] have a temporally residual structure. We build upon recent advances on deep invertible residual models and normalizing flows [52] to study, develop and evaluate temporally-invertible stochastic residual networks (task 1.2). Third, we study, develop and evaluate “jumpy futures” in visual sequences by integrating Poisson generative processes [29] to the temporally-invertible residual networks. Poisson distributions are typically used for modelling series of discrete events where time differences are unknown and must be inferred. They were shown to work well with simple time series, we focus on high-dimensional, noisy visual sequences. Last, we focus on integration and efficiency (task 1.4). The algorithms are developed and evaluated using standard video datasets. The final outcome is visual algorithms incorporating stochasticity in their learning to generate expected futures.

**WP2. Time-intrinsic visual algorithms (PhD2+PostDoc+PI).** With the ultimate goal of temporal self-supervision in mind, with PhD2 we break with the tradition of time-extrinsic visual algorithms. The focus is time intrinsic visual algorithms that incorporate time in their latent structure. We start from my previous work [G13], where we show that one way to integrate time within the neural network structure is to align neural network layers with time steps. This yields time-aligned neural networks trained with strong manual supervision. We modify time-aligned networks to be amenable to temporal self-supervision. Specifically, we explore the addition of temporal constraints within the layers (viewed as time steps) of the spatiotemporal neural network (task 2.1). Next, we note the similarity of time-aligned models to the last year’s best NeurIPS paper, NeuralODEs [29]. Time-aligned networks view neural network weights as a discrete function of time. NeuralODEs view neural network activations as a continuous function of time and express them with ordinary differential functions. Inspired by the similarity, we explore modelling visual algorithms whose weights are continuous functions of time, allowing for greater modelling flexibility and predicting expected futures at arbitrary temporal scales (task 2.2). Continuing on this direction, we further explore the use of said models to specific applications of predicting expected futures in various spatiotemporal tasks. We focus on areas where my group has strong presence, specifically action recognition and object tracking. We modify the researched algorithms to predict the expected future locations and actions during video object tracking and action recognition (task 2.3). Last, we focus on integration and efficiency (task 2.4). The algorithms are developed and evaluated using video datasets with spatial and temporal localization annotations. The final outcome is time-intrinsic visual algorithms that incorporate time in their latent structure.

**WP3. Time-geometric visual algorithms (PhD3+PostDoc+PI).** With PhD3 we investigate visual algorithms that constrain the latent temporal spaces to known geometries. The idea is that the algorithm learns to project the innumerable spatiotemporal dynamics on specific geometries. Thus, any new spatiotemporal dynamics will have to obey the same geometry, avoiding overfitting and allowing for better generalization. In order to introduce geometry in temporal latent spaces, the foundational work on geometric deep learning [30] sheds light. They argue that geometry can be expressed in the form of graphs or geometric manifolds. We first explore visual algorithms that learn undirected graph temporal representations (task 3.1). The graph nodes represent salient temporal events, that are expected to be common for both past and future visual sequences across different videos. The graph edges capture the temporal connectivity between key temporal events. The expected future can then be predicted using graph linkage completion [40]. While undirected graphs capture the global temporal context, they lack local temporal context that discriminates between fine-grained temporal differences. Thus, we continue with directed temporal graphs for temporal representations to enable local temporal context (task 3.2). Parting from graphs as a definition of geometry, next we explore geometric manifolds as representations of temporality learned from visual sequences. A trajectory of coordinates on the manifold represents a visual sequence. The expected future is then recovered by continuing the trajectories on the geometric manifold (task 3.3). Last, we focus on integration and efficiency (task 3.4). For this workpackage we rely on video datasets where one or few objects are recorded, like object tracking or video object segmentation datasets. The final outcome is time-geometric visual algorithms, whose temporal latent spaces conform to known geometries.

**WP4. ShowTime: evaluation & benchmarking (PostDoc+PI).** Proper evaluation and benchmarking is key for learning successful and robust visual algorithms, as history shows [45]. Evaluating the expected future given past observations in a visual sequence has advantages and disadvantages. An advantage is that given any video dataset, one can always split videos into “past subsequences” and “future subsequences” for free (single-mode future evaluation). A disadvantage is that single-mode future evaluation only represents one of the possible futures; there are several other plausible futures that an algorithm could predict. Solutions have been proposed, *e.g.*, best-of-many samples evaluation [70], all with limitations. Pragmatically speaking, this disadvantage incapacitates direct evaluation without resorting to synthetic visual environments, which in turn are hard to make realistic enough. Instead, with the postdoctoral fellow we explore an indirect evaluation of expecting visual futures, specifically by evaluating temporality in algorithms. As temporality is key to predicting the expected future, temporality can be used as an indirect evaluation platform. This indirect evaluation complements the direct single-mode future evaluation. The plan is to devise and organize a temporal visual decathlon benchmark. This decathlon evaluates the capacity of visual algorithms in satisfying key properties of temporality, like spatiotemporal continuity, spatiotemporal directionality, spatiotemporal localization and others. The plan is not to make one more dataset; rather reuse existing excellent resources and collaborate with other leading researchers to share the effort and maximize impact (task 4.1). Further, with the PostDoc and the PhD students we evaluate existing visual algorithms on the decathlon, open-sourcing all relevant results, code and data (task 4.2).

**Datasets.** Recently, there is an explosion of video datasets that address several aspects and challenges relevant to the proposal, from length [55] and complexity [56] to temporality [28]. We start with the Charades-Ego [55] and the Something-Something [28] datasets. Charades-Ego contains continuous video recordings of people performing sequential actions from an egocentric perspective. Something-Something contains classes only distinguishable by visual algorithms that encode temporality.

**Research context.** As part of this proposal, I have agreed with Prof. A. Zisserman from the Visual Geometry Group in the University of Oxford for a research visit and collaboration. This will complement local ongoing collaborations with Prof. M. Welling and Prof. C. Snoek.

**Feasibility.** Theproposed novel direction starts from several published works and experience I have collected over the years. Furthermore, the proposed research will be conducted within the context of world-class computer vision and machine learning specialists, including a research visit to one of the top research groups in computer vision. I am confident that the proposed research and direction is feasible.

**Expected results, impact and knowledge utilization.** The overall aim of this project **ambitiously** seeks to contribute to shifting the field of visual recognition from explaining what has already been observed to predicting what is to be expected in the future. That is, answer questions *where* an object *will be* and *how will* it affect other objects, *what will* the person perform next or *what effect will* a particular visual intervention have. The high gain of this shift is evident, as it will allow for artificial intelligence that **prevents before the fact** and **plans ahead**, rather than explain after the fact.

My research in the past 5 years was on the frontier of another shift: taking the field of video understanding from handcrafted manual features to deep representation learning. I am confident that I can do a similar shift towards with the help of my new independent research team and close collaborators.

**Impact and knowledge utilization.** EVA addresses fundamental research challenges in visual recognition. Its results will serve as a basis for ground-breaking technological advances in practical applications. With the advent of deep learning, nowadays major technology and service vendors generate massive amounts of video sequence data. The demand (and hope) for improving or inventing technologies and services by leveraging this data is high. Learning to expect futures, as explored by EVA, will unlock high-impact research and applications. A few examples:

I1. Today, video understanding faces the major problem of over-reliance on manual supervision. Also, algorithms have trouble with longer and complex videos. Advances from the proposed research will benefit greatly the state-of-the-art in video understanding.

I2. Monitor social media for security purposes by recognizing eminent malicious activities or threats in videos

I3. Monitor environmental changes, by analysing water levels and glacier melting rates from satellite footage and predicting their expected behaviour.

I4. Anticipate risky behaviour either from pedestrians, motorcyclists or other cars, thus improving the safety of autonomous driving platforms.

I5. Help with cancer analysis by recognizing biomarkers in medical video imaging that specifies the tumour.

Regarding practical applications, I have been taking concrete steps with collaborators for I3 and I4.

**Cancer analysis with spin off Ellogon.AI[[2]](#footnote-2).** The visual algorithms in the proposal will be used for two technologies for the spin off I co-founded, Ellogon.AI. The first technology is cross-referencing biomarker signatures in radiology and genome sequences. The second technology is tracking tumours live in MRI sequences and guiding adaptive radiotherapy devices[[3]](#footnote-3) for optimal radiation delivery. The goal is that early technology prototypes will be available within two years for valorisation.

**Safe autonomous driving with BMW.** I am working with BMW for forecasting the future trajectories of road users, including pedestrians, cyclists and other cars. The researched visual algorithms will be used for modelling pedestrian trajectories from regular RGB videos. The algorithms will be tested with the closed-loop BMW simulators. Algorithms will further be tested with actual autonomous BMW vehicles, with the help of the BMW partner, senior research engineer N.S. Nagaraja (Naveen-Shankar.Nagaraja@bmw.de).

# References by PI without supervisors

|  |
| --- |
| [G1] B. Fernando, E. Gavves, J.M. Oramas, A. Ghodrati and T. Tuytelaars, “Modeling video evolution for action recognition,” in *CVPR*, 2015. |
| [G2] B. Fernando, H. Bilen, E. Gavves and S. Gould, “Self-Supervised Video Representation Learning with Odd-One-Out Networks,” in *CVPR*, 2017. |
| [G3] X. Jia, E. Gavves, B. Fernando and E. Gavves, “Guiding the long-short term memory model for image caption generation,” in *CVPR*, 2017. |
| [G4] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi and S. Gould, “Dynamic image networks for action recognition,” in *CVPR*, 2017. |
| [G5] B. Fernando, E. Gavves, J.M. Oramas, A. Ghodrati and T. Tuytelaars, “Rank pooling for action recognition,” in *CVPR*, 2015. |
| [G6] C. Oh, E. Gavves and M. Welling, “[BOCK: Bayesian Optimization with Cylindrical Kernels](javascript:void(0)),” in ICML, 2018. |
| [G7] C. Oh, J. Tomczak, E. Gavves and M. Welling, “Combinatorial Bayesian Optimization using Graph Representations,” in *NeurIPS*, 2019. |

# References by PI as main supervisor

|  |
| --- |
| [G8] T. Davidson, J. Tomczak and E. Gavves, “Increasing Expressivity of a Hyperspherical VAE,” in NeurIPS, Workshop on Bayesian Deep Learning , 2019. |
| [G9] A. Pervez, T. Cohen and E. Gavves, “Low Bias Gradient Estimates for Very Deep Boolean Stochastic Networks,” ICLR (openreview preprint), 2019. |
| [G10] J. Valmadre, L. Bertinetto, J. Henriques, R. Tao, A. Vedaldi, A. Smeulders, P. Torr and E. Gavves, “Long-Term Tracking in the Wild: A Benchmark,” in 2018, ECCV |
| [G11] Y. Chen, T. Mensink and E. Gavves, “3D Neighborhood Convolution: Learning Depth-Aware Features for RGB-D and RGB Semantic Segmentation,” 3DV, 2019. |
| [G12] S. Shkodrani, M. Hoffman and E. Gavves, “Dynamic Adaptation on Non-Stationary Visual Domains,” ECCVW, 2018. |

# References by PI with supervisors

|  |
| --- |
| [G13] A. Ghodrati, E. Gavves and C. Snoek, “Video Time: Properties, Encodings, Evaluation,” in BMVC, 2018. |
| [G14] N. Hussein, E. Gavves and A. Smeulders, “Timeception for Complex Action Recognition,” in CVPR, 2019 |
| [G15] R. Tao, E. Gavves and A. Smeulders, “Siamese Instance Search for Tracking,” in CVPR, 2016 |
| [G16] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain and C. Snoek, “VideoLSTM convolves, attends and flows for action recognition,” CVIU, 2018 |
| [G17] E. Gavves, T. Mensink, T. Tommasi, C. Snoek and T. Tuytelaars, “Active transfer learning with zero-shot priors: Reusing past datasets for future tasks,” in ICCV, 2015. |
| [G18] E. Gavves, B. Fernando, C. Snoek, A. Smeulders and T. Tuytelaars, “[Fine-grained categorization by alignments](javascript:void(0)),” in ICCV, 2013. |
| [G19] R. de Geest, E. Gavves, A. Ghodrati, Z. Li, C. Snoek, T. Tuytelaars, “[Online action detection](javascript:void(0)),” in ECCV, 2016. |

# Other references

|  |
| --- |
| [20] Y. LeCun, Y. Bengio and G. Hinton, “Deep learning,” Nature, 2015 |
| [21] J. Carreira and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” in CVPR, 2017. |
| [22] B. Zhou, A. Andonian, A. Oliva and A. Torralba, “Temporal Relational Reasoning in Videos,” in ECCV, 2018 |
| [23] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks,” in ICCV, 2015 |
| [24] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, 1997 |
| [25] J. Chung, Ç. Gülçehre, K. Cho and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” CoRR abs/1412.3555, 2014. |
| [26] G. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev and A. Gupta, “Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding,” in ECCV, 2016. |
| [27] X. Wang, R. Girshick, A. Gupta and K. He, “Non-Local Neural Networks,” in CVPR, 2018 |
| [28] R. Goyal and e. al., “The "Something Something" Video Database for Learning and Evaluating Visual Common Sense,” in ICCV, 2017 |
| [29] T. Q. Chen et al., “Neural Ordinary Differential Equations,” in NeurIPS, 2018 |
| [30] M. Bronstein, J. Bruna, Y. LeCun, A. Szlam and A. Vandergheynst, “Geometric Deep Learning: Going beyond Euclidean data,” IEEE Signal Processing, 2017 |
| [31] I. Laptev, “On Space-Time Interest Points,” IJCV, 2005. |
| [32] H. Wang and C. Schmid, “Action Recognition with Improved Trajectories,” in ICCV, 2013 |
| [33] A. we, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, “Large-Scale Video Classification with Convolutional Neural Networks,” in CVPR, 2014 |
| [34] C. Szegedy and e. al., “Going deeper with convolutions,” in CVPR, 2015. |
| [35] L. Pickup, Z. Pan, D. Wei, Y.-C. Shih, C. Zhang, A. Zisserman, B. Schölkopf and W. Freeman, “Seeing the Arrow of Time,” in CVPR, 2014. |
| [36] D. Wei, J. Lim, A. Zisserman and W. Freeman, “Learning and Using the Arrow of Time,” in CVPR, 2018. |
| [37] I. Misra, C. L. Zitnick and M. Hebert, “Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification,” in ECCV, 2016. |
| [38] G. Varol, I. Laptev and C. Schmid, “Long-Term Temporal Convolutions for Action Recognition,” PAMI, 2018. |
| [39] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic and B. Russell, “ActionVLAD: Learning Spatio-Temporal Aggregation for Action Classification,” in CVPR, 2017. |
| [40] T. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks,” in ICLR, 2017. |
| [41] M. Aubry and B. Russell, “Understanding Deep Features with Computer-Generated Imagery,” in ICCV, 2015. |
| [42] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik and P. Torr, “Fully-Convolutional Siamese Networks for Object Tracking.,” in ECCC Workshop, 2016. |
| [43] D. Held, S. Thrun and S. Savarese, “Learning to Track at 100 FPS with Deep Regression Networks,” in ECCV, 2016. |
| [44] M. Kristan and e. al., “The Sixth Visual Object Tracking VOT2018 Challenge Results,” in ECCV Workshop on Visual Object Tracking, 2018. |
| [45] A. Smeulders, D. Manh Chu, R. Cucchiara, S. Calderara, A. Dehghan and M. Shah, “Visual Tracking: An Experimental Survey,” PAMI, 2014. |
| [46] M. Hossain, A. K. Gostar, A. Bab-Hadiashar and R. Hoseinnezhad, “Visual Mitosis Detection and Cell Tracking Using Labeled Multi-Bernoulli Filter,” IEEE Fusion, 2018. |
| [47] L. Cerviño, D. J. and S. Jiang, “MRI-guided tumor tracking in lung cancer radiotherapy,” Physics in Medicine and Biology, 2011. |
| [48] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, “A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation,” in CVPR, 2016. |
| [49] T. Kim, S. Ahn and Y. Bengio, “Variational Temporal Abstraction,” arXiv preprint arXiv:1910.00775 [cs.LG], 2019. |
| [50] R. Pascanu, T. Mikolov and Y. Bengio, “On the difficulty of training Recurrent Neural Networks,” in ICML, 2013. |
| [51] D. Kingma, M. Welling, “Autoencoding Variational Bayes,” in ICLR, 2014. |
| [52] D. Rezende, S. Mohamed, “Variational Inference with Normalizing Flows,” in ICML, 2015. |
| [53] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, K. Kavukcuoglu, “Conditional Image Generation with PixelCNN Decoders,” in ICML, 2013. |
| [54] N. Kalchbrenner, A. van den Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, K. Kavukcuoglu, “Video Pixel Networks,” in ICML, 2017. |
| [55] G. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, K. Alahari, “Charades-Ego: A Large-Scale Dataset of Paired Third and First Person Videos,” in CVPR, 2018. |
| [56] R. Girdhar, D. Ramanan, “CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning,” arXiv preprint arXiv:1910.04744, 2019. |
| [57] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, “Large-scale Video Classification with Convolutional Neural Networks,” in CVPR, 2014. |
| [58] D. Jayaraman, “Learning image representations tied to ego-motion,” in ICCV, 2015. |
| [59] X. Liang, L. Lee, W. Dai and E. Xing, “Dual Motion GAN for Future-Flow Embedded Video Prediction,” in ICCV, 2017. |
| [60] Z. Liu, S. Wu, S. Jin, Q. Liu, S. Lu, R. Zimmermann and L. Cheng, “Towards Natural and Accurate Future Motion Prediction of Humans and Animals,” in CVPR, 2019. |
| [61] X. Qi, Z. Liu, Q. Chen and J. Jia, “3D Motion Decomposition for RGBD Future Dynamic Scene Synthesis”, in CVPR, 2019. |
| [62] J. Liang, L. Jiang, J. Niebles, A. Hauptmann and L. Fei-Fei, “Peeking Into the Future: Predicting Future Person Activities and Locations in Videos”, in CVPR, 2019. |
| [63] J. Walker, K. Marino, A. Gupta and M. Hebert, “The Pose Knows: Video Forecasting by Generating Pose Futures.”, in ICCV, 2017. |
| [64] M. Hoai and F. de la Torre, “Max-Margin Early Event Detectors”, in CVPR, 2012. |
| [65] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, “Video (language) modeling: a baseline for generative models of natural videos”, arxiv preprint arXiv:1412.6604, 2014. |
| [66] D. Jayaraman, F. Ebert, A. Efros, S. Levine, “Time-Agnostic Prediction: Predicting Predictable Video Frames”, in ICLR, 2019. |
| [67] Y. Kwon, M. Park, “Predicting Future Frames Using Retrospective Cycle GAN”, in CVPR, 2019. |
| [68] R. Krishnan, U. Shalit, David Sontag, “Deep Kalman Filters”, arxiv preprint arXiv:1511.05121, 2015. |
| [69] M. Gemici et al., “Generative Temporal Models with Memory”, arxiv preprint arXiv:1702.04649, 2017. |
| [70] N. Lee, W. Choi, P. Vernaza, C. Choy, P. Torr, M. Chandraker, “DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents”, in CVPR, 2017. |
| [71] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, S. Savarese, “Social LSTM: Human Trajectory Prediction in Crowded Spaces”, in CVPR, 2017. |
| [72] I. Gulrajani, k. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, A. Courville, “PixelVAE: A Latent Variable Model for Natural Images”, in ICLR, 2017. |
| [73] Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, Jörn-Henrik Jacobsen, “Invertible Residual Networks”, arxiv preprint arXiv:1811.00995, 2019. |
| [74] D. Ha, A. Dai, Q. Le, “Hypernetworks”, in ICLR, 2017. |
| [75] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, Y. Bengio, “A Recurrent Latent Variable Model for Sequential Data”, in NeurIPS, 2015. |
| [76] K. Gregor, G. Papamakarios, F. Besse, L. Buesing, T. Weber, “Temporal Difference Variational Autoencoder”, in ICLR, 2019. |
| [77] L. Wiskott, T. Sejnowski, “Slow Feature Analysis: Unsupervised Learning of Invariances”, in Neural Computation, 2002. |

**Section B: Curriculum Vitae (2/2 pages)**

|  |  |  |  |
| --- | --- | --- | --- |
| Personal Information | | | |
| Family name, First name: Gavves, Efstratios | | Nationality: Greek | Date of birth: 14-01-1985 |
| Research ID: [Google Scholar](https://scholar.google.nl/citations?user=QqfCvsgAAAAJ&hl=en) | Website: [www.egavves.com](http://www.egavves.com) | | |

|  |  |
| --- | --- |
| Education | |
| Sep 2008-Sep 2014 | PhD (Thesis: *Nuances in Visual Recognition*) Institute of Informatics, University of Amsterdam, The Netherlands Supervisor: Cees G.M. Snoek, Promotor: Arnold W.M. Smeulders |
| Sep 2002-Sep 2008 | Diploma degree (BSc & MSc). (Thesis: *Gesture Recognition from Depth Images*) Electrical & Computer Engineering, Aristotle University of Thessaloniki, Greece Supervisor: Sotiris Malassiotis |

|  |  |  |
| --- | --- | --- |
| Current Position | | FTE |
| Sep 2015- | Assistant Professor (Tenure Track) - University of Amsterdam, The Netherlands (Institute of Informatics) - **Supervising** 9 PhD students | 1.0 |
| Sep 2015- | Scientific Manager of QUVA Lab - Joint lab between University of Amsterdam and Qualcomm - Directed by Professors M. Welling, A. Smeulders and C. Snoek. - **Leading team of 12** PhD and Postdocs, coordinating with 3 full professors |  |
| May 2019- | **Co-founder**/Chief Science Officer of Ellogon.AI - Spin-off of University of Amsterdam and Netherlands Cancer Institute - Mission: use visual recognition on histopathology & MRI sequences for biomarker recognition (biomedical imaging patterns) for cancer | 0.1 |

|  |  |  |
| --- | --- | --- |
| Previous Positions | | FTE |
| Feb 2014-Aug 2015 | Post-Doctoral Researcher with Tinne Tuytelaars (ERC StG laureate) - KU Leuven (ESAT-PSI). **Guiding 3** PhD students. | 1.0 |
| Jan 2019-Aug 2019 | Scientific Consultant  - NICO.Lab. Medical AI startup. **Guiding team** of 5 research engineers |  |

|  |  |
| --- | --- |
| Fellowships and Awards | |
| 2019-23 | NWO LIFT. Grant for 1 PhD for Safe Self Driving, co-funded by BMW |
| 2018 | Best paper award. MSc thesis with S. Shkodrani published in ECCV Workshop on Transferring & Adapting Source Knowledge in Computer Vision & VISDA |
| 2015- | Invited to lead QUVA Lab. Invited to lead 12 PhD students, coordinate 3 professors |
| 2015- | AAA Scholarship. 6-year funding with tenure track position at University of Amsterdam |
| 2015 | FWO Postdoc Researcher Award. Declined in favour of tenure track assistant professorship |
| 2014 | CVPR Doctoral Consortium Award. To most promising recent PhDs in Computer Vision |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Supervision of Graduate Students and Postdoctoral Fellows | | | | | |
| Type |  |  | Name |  | Since |
| 100% supervision PhD | | | 2x: A. Pervez (Deep temporal learning), Y. Chen (3D vision) | | 2019- |
| 50% cosupervision PhD | | | 7x: N. Hussein (Action recognition), C. Oh (Bayesian optimization), R. Zoetmulder (Medical transfer learning), S. Liao (Fine-grained recognition), M. Kilickaya (Interaction recognition), M. Reisser (Distributed machine learning), P. O’Connor (Spiking neural nets, completed) | | 2015- |
| 50% cosupervision Postdoc | | | 4x: Ongoing: D. Gupta, Completed: A. Ghodrati, R. Tao, Z. Li | | 2017-19 |

\* All above graduate students and postdoctoral fellows were with the University of Amsterdam

|  |  |
| --- | --- |
| Education Activities | |
| MSc in AI | Deep Learning. Attendance from 15 to 200+ students. Course grade: 8.6/10 Invited to give lectures in Computer Vision course | 2016- 2016-18 |
| PhD level | ASCI Computer Vision by Learning. Course grade: 8.3/10 | 2017-19 |
| Summer Schools | Vision, Understanding and Machine Intelligence, Portugal (declined) TomTom Summer School on Autonomous Driving Russian Summer School in Information Retrieval | 2019 2019 2017 |

|  |  |
| --- | --- |
| Organization of Scientific Meetings | |
| Tutorials | Action Classification and Video Modelling with IEEE CVPR (main organizer)  Zero-shot learning with IEEE CVPR and ECCV (co-organizer) | 2019 2016-18 |
| Workshops | Brave New Ideas for Video Understanding & Motion Representations with IEEE CVPR and ECCV (main organizer) | 2016-18 |
| Seminar lead | Initiated first Deep Learning meetup in the Netherlands, close to 2,000 members. Brought speakers like: I. Laptev (INRIA Paris), O. Vinyals (Google DeepMind), M. Bethge (MPI Tuebingen), C. Maddison (University of Oxford) | 2016-19 |

|  |  |
| --- | --- |
| Commissions of Trust | |
| PhD Committee | K. Tran, D. Kingma, Z. Li, R. Tao, S. Puttemans, S. Kordumova | 2016-18 |
| Exam board | BSc & MSc in AI. Evaluating & guaranteeing course and exam quality | 2017- |
| Admissions Board | MSc in AI. Evaluating student applications for admissions | 2017-18 |
| Examiner | Examining quality of MSc theses and projects | 2016- |

|  |  |
| --- | --- |
| Reviewing Activities | |
| Guest editor | Invited for the special issue on Multi-Modal AI for the Machine Vision and Applications Journal | 2019 |
| Senior reviewer | Conference of Association for Advancement of Artificial Intelligence, British Machine Vision Conference, International Conference on Image Analysis and Processing, ACM Multimedia, European Conference on Computer Vision | 2016-19 |
| Reviewer | 30+ times invited reviewer in all major computer vision conferences/journals like: IEEE Conference in Computer Vision and Pattern Recognition, IEEE International Conference in Computer Vision, IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision | 2014-19 |

|  |  |
| --- | --- |
| Membership of Scientific Societies |  |
| European Laboratory for Learning and Intelligent Systems (personal invitation) | 2018- |
| Computer Vision Foundation member | 2015- |

|  |  |
| --- | --- |
| Major Collaborations Abroad |  |
| Prof. A. Zisserman (agreed for future research visit and collaboration as part of this proposal) |  |
| Prof. A. Vedaldi (University of Oxford, ERC StG Laureate) | 2016-18 |
| Prof. P. Torr (University of Oxford) | 2018 |
| Research Scientist B. Fernando (A\* Institute, Singapore) | 2013-17 |
| Assist. Prof. H. Bilen (University of Edinburgh) | 2016-17 |
| Assoc. Prof. S. Gould (Australian National University) | 2016-17 |
| Prof. L. van Gool (KU Leuven, ETH) | 2017 |
| Prof. T. Tuytelaars (University of Oxford, ERC StG Laureate) | 2013-16 |

**Appendix: Current research grants and any on-going applications related to the proposal of the PI (Funding ID)**

Mandatory information (does not count towards page limits)

**Current grants (Please indicate "No funding" when applicable):**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Project Title* | *Funding source* | *Amount*  *(Euros)* | *Period* | *Role of the PI* | *Relation to current*  *ERC proposal[[4]](#footnote-4)* |
| FLORA: Future prediction of obstacle locations in traffic scenes for collision avoidance | NWO, BMW | 239,000 | 2019-23 | PI | Expecting future trajectories of pedestrians |
| QUVA Lab | Qualcomm,  UvA | 12 fte | 2015-20 | Invited to lead as Scientific Manager | PhD topics related to WPs (action recognition, generative models) |
| MEDIFOR | SRI, DARPA | 1fte | 2016-20 | Did not win myself, stepped in because of colleague who left the University | Focus on 3D vision |

**On-going and submitted grant applications (Please indicate "None" when applicable):**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Project Title* | *Funding source* | *Amount*  *(Euros)* | *Period* | *Role of the PI* | *Relation to current*  *ERC proposal2* |
| TIMING: Learning Time in Videos | NWO | 800,000 | 2021-26 | PI | Sharing some similar ideas with WP3. |
| DeepBrain: Deep Temporal Decomposition Learning in CT Scans for Early Brain Stroke Detection | NWO | 239,000 | 2020-24 | PI | Focus on medical domain and temporal sequences |

**Section C: Early achievements track-record (Max. 2 pages)**

**Summary:**

* **3,086 citations total**, **h-index: 23**, **i10-index: 32.**
* Total of **44 international refereed publications**

38 international conferences

|  |  |
| --- | --- |
| Publishing in: |  |
| IEEE CVPR | **12x** |
| IEEE ICCV | **5x** |
| ICLR | **4x** |
| IEEE TPAMI | **3x** |
| ECCV | **3x** |
| CVIU | **2x** |
| NeurIPS | **1x** |
| ICML | **1x** |
| IJCV | **1x** |

6 international journals

Conferences are the most important output indicator[[5]](#footnote-5) in computer vision

* 12 international patents
* Publishing with **85% frequency in A+** conferences and journals

(CORE Australian Association report).

* Roughly **50% of my publications without my PhD supervisors**
* **7 publications** where **I** **initiated** and **guided international** collaboration
* **13 publications** as a **co-supervisor** (2nd, 3rd author)in QUVA Lab

including the “Siamese Instance Search for Tracking”

published in CVPR 2016 (389 citations), paradigm adopted by

majority of modern object trackers (VOT competition)

|  |  |
| --- | --- |
| Publication statistics without supervisors | |
| Publications with my PhD supervisors | **23x** |
| Publications without my PhD supervisors | **21x** |
| Publications with me as main supervisor | **4x** |

* **Main supervisor in 6 publications**

(I am leading supervisor of PhDs since 2018)

(3 of them with MSc students)

* 24 invitedtalks
* 1 best paper awardin workshop

**Brief summary of my research in the past 5 years:** My scientific curiosity evolves around the role of time in visual recognition by learning; my research goal is to translate this to algorithms. Five years ago, I was a PostDoc in KU Leuven with Prof. T. Tuytelaars. During my PostDoc I had six publications in IEEE CVPR, IEEE TPAMI, ICCV, including a personal invitation to IEEE TPAMI. I was invited to join the newly founded Facebook AI research group in Paris; I was also invited as a tenure-track Assistant Professor and Scientific Manager of QUVA Lab, with the responsibility to lead twelve PhD and PostDocs and coordinate with three full professors. I chose the latter. QUVA was the first academic-industry lab on artificial intelligence in the Netherlands. Its success inspired the start of the national Innovation Centre for Artificial Intelligence.

In the last five years I co-authored 23 publications, including 9x CVPR, 3x ICLR, 2x ICCV, 2x TPAMI. 13 of them have me as a co-supervisor with QUVA students, 6 of them have me as a main supervisor, the rest is with international collaborations. My most visible work in these 5 years was the invention of the siamese tracking paradigm, independently proposed by University of Oxford and Stanford University. It is the foundation for most trackers nowadays in the leading long-term tracking VOT benchmark. Another relevant work was time-aligned DenseNets, which learn representations with strong temporal capacity. Last, another research line that inspires me is Spiking Neural Networks, which can potentially be as energy-efficient as the human brain; I co-authored three ICLR papers on the subject. I further initiated a research collaboration with BMW and secured NWO funding on safe autonomous driving.

Selected publications without PhD supervisors

|  |
| --- |
| * + - 1. B. Fernando, E. Gavves, J.M. Oramas, A. Ghodrati, T. Tuytelaars, Modeling video evolution for action recognition, IEEE Conference in Computer Vision and Pattern Recognition (CVPR), 2015,   Extended to IEEE Journal on Pattern Analysis and Machine Intelligence after personal invitation from the editor, 2016, [Oral Presentation Top 3%] - [421 citations] [relates to WP2]. |
| * + - 1. H. Bilen\*, B. Fernando\*, E. Gavves\*, A. Vedaldi, S. Gould, Dynamic image networks for action recognition (\*Shared workload). IEEE Conference in Computer Vision and Pattern Recognition (CVPR), 2016. Extended to IEEE Journal on Pattern Analysis and Machine Intelligence, 2017   [Oral Presentation Top 3%] - [344 citations] [relates to WP2, WP3]. |
| C. Oh, E. Gavves, M. Welling, *BOCK: Bayesian Optimization with Cylindrical Kernels*  International Conference in Machine Learning, 2018 [Oral Presentation top 8%] [relates to WP1]. |

Selected publication with PI as first author

|  |
| --- |
| * + - 1. E. Gavves, B. Fernando, C.G.M. Snoek, A.W.M. Smeulders, T. Tuytelaars, Fine-grained categorization by alignments, IEEE International Conference on Computer Vision, 2013 (ICCV). Extended to International Journal of Computer Vision, 2015. [243 citations] [relates to WP1]. |

Selected publication with PI as leading supervisor

|  |
| --- |
| Jack Valmadre, Luca Bertinetto, Joao Henriques, Ran Tao, Andrea Vedaldi, Arnold WM Smeulders, Philip HS Torr, Efstratios Gavves, [*Long-term tracking in the wild: A benchmark*](javascript:void(0))*.* European Conference on Computer Vision, 2018. [21 citations] [relates to WP3, WP4]. |

Selected invitations, honors & awards

|  |
| --- |
| Invited by the Editor to submit to IEEE Journal on Pattern Analysis and Machine Intelligence, 2016. |
| Invited speaker by British Computer Society (British Machine Vision chapter) on Video Understanding, together with leading researchers including A. Zisserman, C. Schmid, I. Laptev, M. Shah, R. Shuthankar, C. Snoek. The purpose was to discuss the present and future of the visual understanding field. |
| My deep learning course was shortlisted internationally among top ones, next to courses by G. Hinton’s and A. Ng’s on neural networks and deep learning. |
| Best paper award awarded to publication based on MSc thesis by S. Shkodrani that I supervised. Published in ECCV Workshop in Transferring and Adapting Source Knowledge in Computer Vision and VISDA Challenge, 2018. |
| Invited to *European Laboratory for Learning and Intelligent Systems* (ELLIS), chaired by B. Caputo, N. Oliver, B. Scholkopf, M. Welling. ELLIS is an initiative for “involving the very best European academics while working together closely with basic researchers from industry.” |

Selected invited presentations

|  |  |  |
| --- | --- | --- |
| To public | Interview for NWO I/O Magazine Presentation to BMW on spatiotemporal deep learning Presentation to NICO.Labs about computer vision on medical imaging Presentation to Amsterdam Arena about computer vision on surveillance  Presentation to ING on Generative Adversarial Networks | 2019  2018  2018  2017  2017 |
| To researchers | Presentation to British Computer Society on Video Understanding  Presentation to Qualcomm  Presentation to INRIA Paris Presentation to Xerox Grenoble Presentation to Scyfer on Object Tracking Presentation in Columbia University, New York | 2019  2018  2017  2017  2017 2016 |

Selected international patents

|  |
| --- |
| Tao, E. Gavves, A.W.M. Smeulders, *Adapting to appearance variations when tracking a target object in video sequence*, US Patent App. 10/019,631, 2018  After publication: R. Tao, E. Gavves, A.W.M. Smeulders, *Siamese Instance Search for Tracking,* CVPR, 2016[389 citations] |
| M. Jain, Z. Li, E. Gavves, C.G.M. Snoek, *Action localization in sequential data with attention proposals from a recurrent network,* US Patent App. 15/250,755, 2017. |
| R. Tao, E. Gavves, A.W.M. Smeulders, *Generic mapping for tracking target object in video sequence*, US Patent App. 15/192,935, 2017 |
| Z. Li, M. Jain, E. Gavves, C.G.M. Snoek, *Video analysis with convolutional attention recurrent neural networks*, US Patent 9,830,709, 2017.  After publication: *Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, C.G.M. Snoek*, VideoLSTM convolves, attends and flows for action recognition*,* CVIU, 2018[162 citations] |
| Z. Li, M. Jain, E. Gavves, C.G.M. Snoek, *Recurrent networks with motion-based attention for video understanding*, IEEE Conference on Computer Vision, 2017. |

1. *One can hardcode handcrafted yet brittle rules to make predictions in simple videos like this with static background.* [↑](#footnote-ref-1)
2. I co-founded the University spin off, Ellogon.AI. The mission I sto use visual recognition algorithms on histopathology, MRI and CT sequences for biomarker recognition. Other co-founders: Dr. E. Kanoulas (associate professor, University of Amsterdam), Dr. J. Teuwen (research scientist, Netherlands Cancer Institute), Dr. H. Horlings MD (head computational pathologist, Netherlands Cancer Institute), Dr. S. Dhawan (fund manager at Oncode Fund). [↑](#footnote-ref-2)
3. Adaptive radiotherapy devices are manufactured with Netherlands Cancer Institute and associated companies like Elekta. [↑](#footnote-ref-3)
4. Describe clearly any scientific overlap between your ERC application and the current research grant or on-going grant application. [↑](#footnote-ref-4)
5. Publications in prestigious international conferences is the most important output indicator in computer vision. Specifically, among conferences and journals in all sciences, ranks according to [google scholar](https://scholar.google.nl/citations?view_op=top_venues&hl=en): IEEE CVPR (rank 10), ECCV (rank 56), ICCV (rank 71), IEEE TPAMI (rank 76). Context: Nature ranks 1st, the journal of High Energy Physics ranks 30th, the Cancer Research Journal 87th. [↑](#footnote-ref-5)