Autonomous object recognition in videos using Siamese Neural Networks

Nawel Medjkoune, Frédéric Armetta, Mathieu Lefort, Stefan Duffner

Autonomous object recognition in videos using Siamese Neural Networks

Nawel Medjkoune[1,2], Frédéric Armetta[1,2], Mathieu Lefort[1,2], and Stefan Duffner[1,3]

[1]Université de Lyon, CNRS
[2]Université Lyon 1, LIRIS, UMR5205, F-69622, France.
[3]INSA-Lyon, LIRIS, UMR5205, F-69621, France.

***Abstract*— For a robot to be deployed in unconstrained real world environments, it needs to be autonomous. In this preliminary work, we focus on the capacity of an autonomous robot to discover and recognize objects in its visual field. Current existing solutions mainly employ complex deep neural architectures that need to be pre-trained using large datasets in order to be effective. We propose a new model for autonomous and unsupervised object learning in videos that does not require supervised pre-training and uses relatively simple visual filtering. The main idea relies on the saliency-based detection and learning of objects considered similar (thanks to a spatio-temporal continuity). For this purpose the learning of objects is based on a Siamese Neural Network (SNN). We demonstrate the capacity of the SNN to learn a good feature representation despite the deliberately simple and noisy process used to extract candidate objects.**

I. Introduction

Nowadays, service and social robots are still limited in the quality and quantity of tasks they can accomplish. These limitations are mainly due to the poorly structured or unconstrained open environments they are operating in. In order to be efficient in these real-world settings, they would need to become more autonomous. Ideally, a robot progressively acquires informations from the outside world during its lifetime, similarly to a baby that is learning. This gain in terms of autonomy raises, amongst others, the problems of unsupervised, on-line and life-long learning in non-stationary environments [1]. In this work, we focus on the capacity of an autonomous robot to discover and recognize objects localized in its visual field. Two main issues arise in this case: how to detect and localize new objects in this open environment and how to learn meaningful visual representations incrementally in an unsupervised way.

Multiple robust solutions exist for detecting, tracking and recognizing objects or persons in images and video streams, *e.g.* [2], [3], [4], [5], [6]. However, these methods are designed and pre-trained for specific object categories, like faces, persons or cars, and in a dedicated environment, and they are not able to accommodate new object categories during operation. To provide the visual perception algorithms with more autonomy, other works proposed unsupervised learning methods, based on neural networks, that are able to learn discriminative internal representations from unlabelled data and to adapt to a given context using transfer learning. For example, Wang *et al.* [7] use a large unlabelled dataset of videos and specific low-level object patch extraction and tracking techniques to learn visual representations and then exploit this model to learn specific objects with labelled data. Another approach proposed by Liang *et al.* [8] consists in continuously updating the learned visual representation using recognized objects from videos. However, their method incorporates a considerable amount of prior knowledge by using a Convolutional Neural Network (CNN), pre-trained on a large labelled dataset.

These existing methods are similar to our proposed approach in that they use unsupervised or weakly supervised learning techniques for visual representation learning exploiting the temporal coherence in videos. Although showing a high accuracy of object detection and recognition, they have some inherent limitations for an autonomous agent, *e.g.* they require a supervised pre-training on a given large image dataset in order to be effective. Moreover, the pre-training as well as their operation are computationally expensive due to the used complex Deep Neural Networks architectures. In our work, we address these issues by proposing a weakly supervised approach for object recognition using a convolutional SNN not requiring any labelled data or supervised pre-training. Our method uses no prior knowledge about the environment and only employs relatively simple visual filtering techniques inspired by human cognition (like saliency) to extract candidate objects.

II. Model

We propose a new model for autonomous and unsupervised object learning and recognition in videos using neural networks, illustrated in Fig. 1.

In our approach, we consider an autonomous agent equipped with a visual sensor (the camera) and an internal structure (the cognitive system). The agent is exposed to a visual flow coming from the environment and processes it frame by frame using low-level filters.

The main idea relies on the learning of objects considered similar (thanks to the saliency-based detection [9] and spatio-temporal continuity [10]) as detailed below. For this purpose, we form pairs of images representing similar and dissimilar objects and learn this similarity metric using a convolutional SNN [11], [12]. In a bootstrapping process, this learned similarity space can then be used to recognize previously seen objects. However, here, we are particularly interested in the

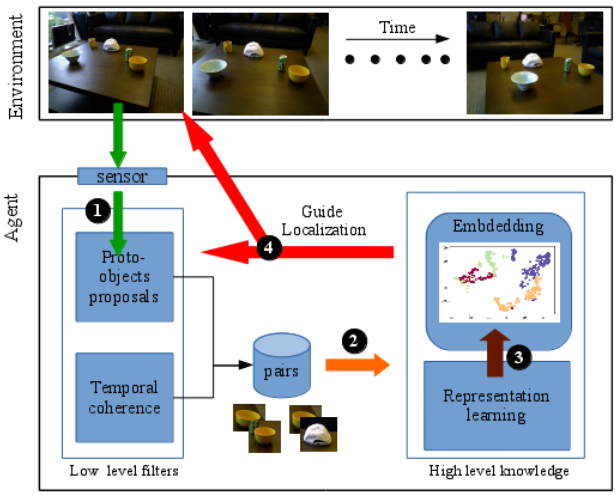

Fig. 1. Overview of the proposed approach. The agent captures the visual field frame by frame and processes it to discover proto-objects (arrow 1). Using spatio-temporal coherence, pairs of similar and dissimilar objects are created and given to the neural network (arrow 2) that learns an internal visual representation of the objects (arrow 3).

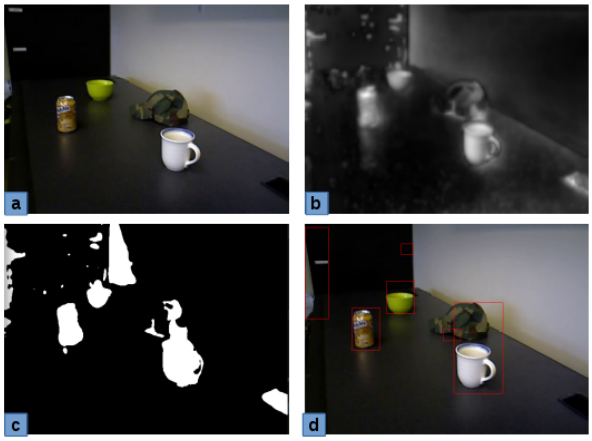

Fig. 2. Steps of object discovery using saliency attention model. a: original image, b: saliency map, c: binary map after thresholding the saliency map, d: bouding boxes around proto-objects

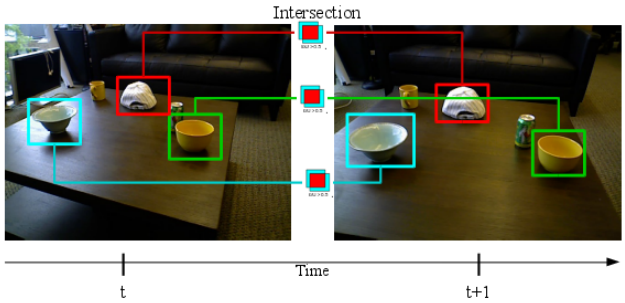

Fig. 3. Pair association using spatio-temporal coherence. Detections that are spatially close in two consecutive frames are considered the same object.

first placement used for data. Indeed, without any knowledge of the objects and due to the relatively simple temporal filtering to detect candidate objects, the first extraction is coarse and highly sensitive to environment noise.

First, proto-object proposals are extracted. As we want the system to be functional without prior knowledge, we suggest a simple method for object proposals inspired by the human visual system. In a common visual scene, objects differ from their surroundings in color, intensity, texture, orientation, depth and other features. The human visual system is able detect regions that differ from their surroundings using bottom-up saliency [13], [14]. Based on this observation, we propose to compute saliency as a cue to quickly form candidates of possible objects that we want to recognize, as illustrated in Fig. 2. A saliency map is computed from the original image and a thresholding is applied resulting in a binary map, where potential object candidates are highlighted with white blobs. Bounding boxes are then drawn around these blobs, representing object candidates.

Given that the input images are not labeled, we use the principle of spatio-temporal coherence of the objects present in the video sequences in order to provide a weakly supervised signal. The idea is that detections that are close in space and time are likely to correspond to the same object. Therefore, we associate detections in consecutive frames using spatial locality (see Fig. 3) and create "tubelets" of patches each one representing one object. Then, positive and negative pairs of patches (*i.e.* supposedly coming from the same and from different objects) are created based on these tubelets by associating respectively two patches from the same tubelet, and two patches from different tubelets. The learning is performed off-line by extracting all pairs from input videos.

The extracted pairs form the input of a SNN [12], which transforms the data into a representation that brings closer similar data points and pulls apart dissimilar data points in the feature space by minimizing a contrastive loss function:

$$\mathcal{L}(W) = \sum_{i=1}^{M} L(W, Y^i, X_1^i, X_2^i) = (Y^i)\frac{1}{2} D_2(X_1^i, X_2^i)^2 + (1 - Y^i)\frac{1}{2}(\max(0, m - D_2(X_1^i, X_2^i)))^2 \,, \quad (1)$$

where (X_1^i, X_2^i) are pair elements and Y^i their label (positive or negative), M is the minibatch size, W the shared parameters, m an arbitrary margin and $D_2(X_1^i, X_2^i)$ is the Euclidean distance between feature vectors X_1^i and X_2^i. In our experiments, we used a very light architecture for the SNN consisting of a 3 blocks of 10, 20 and 30 channel 3×3 convolution layer with input size 32×32 and ReLU activation function followed by a 2×2 max-pooling, and a fully-connected layer of size 20.

III. Results

We evaluate the proposed approach on the RGB-D Scenes Dataset Version 2 [15]. The evaluation is done on both manually labeled classes of objects and automatically extracted training pairs which we augment randomly using data generation. The transformations made on the augmented pairs are: rotation, horizontal and vertical translation, shearing,

zoom and horizontal flip. We train the SNN using several thousands of pairs and report the results after convergence.

First, we use the representations learned by the SNN in order to get the feature representation of images in four manually labeled classes. A TSNE dimensionality reduction is then applied to project the transformed data points, illustrated in Fig. 4. In this example, training is done on data pairs extracted from video 12 (a representative example) and the projection on manually labeled data from video 12 and 14 respectively (both videos present the same objects in different backgrounds). The plots show that there is a certain structure when projecting manually labeled data images, with four separated clusters with a small confusion when generalizing.

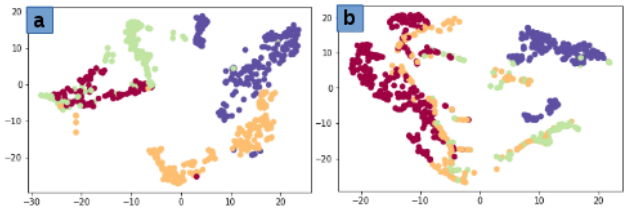

Fig. 4. TSNE dimensionality reduction of feature representations learned by the Siamese Neural Network. Each color represents one object: Cup, Bowl, hat and soda can. **a**: manually labaled dataset from the video 12. **b:** manually labeled dataset from video 14 (same objects, different background).

Secondly, we compare the classification accuracy of the manually labeled dataset using two different feature representations: one learned from a supervised CNN trained on the same dataset, and the other learned using our model. The two models have similar architectures and are trained with the same parameters. Table I shows the results of three experiments where we chose 2, 4 or 6 different objects as follow: first we take two very distinct classes (cup and hat) and then add gradually different classes of objects that can look alike (*e.g.* cups with different colors, cup and bowl with the same color).

	2 classes	4 classes	6 classes
Supervised CNN	99.8%	99.8%	98.73%
Our approach	98.05%	92.15%	74.15%

TABLE I

CLASSIFICATION RESULTS ON RGB-D MANUALLY LABELED DATASET COMPARED TO BASELINE APPROACHES. RESULTS SHOW ACCURACY OF CLASSIFICATION IN THREE DIFFERENT EXPERIMENTS.

The SNN gives comparable results to the supervised CNN with few classes, and falls to 74.15% accuracy in the presence of similar objects by color or shape.

These experiments show two results: when performing weakly supervised training of an SNN, with a relatively simple architecture and no prior pre-training, good visual representation can be learned despite the noise in the unlabeled input data, with a fair generalization to objects present in different backgrounds. Moreover, these results also show that this weakly supervised approach, only based on the feedback from spatio-temporal coherence, can be used to bootstrap the learning process.

This work is a preliminary step in order to design an autonomous system for object recognition using neural networks. We showed with these first experiments the capacity of a simple neural network in learning visual feature representations of objects in simple visual scenes. In the next step, we aim to validate these results in more complex real-world scenes before directing it towards on-line incremental learning with a feedback loop (Fig. 1, arrow 4) to guide the localization of objects using the trained neural network.

IV. ACKNOWLEDGEMENT

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

REFERENCES

[1] Sébastien Mazac, Frédéric Armetta, and Salima Hassas. On bootstrapping sensori-motor patterns for a constructivist learning system in continuous environments. In *Alife 14: Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, 2014.

[2] B. Leibe, K. Schindler, N. Cornelis, and L. V. Gool. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[3] A. Ess, B. Leibe, K. Schindler, and L. V. Gool. A mobile vision system for robust multi-person tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, 2008.

[4] O. H. Jafari, D. Mitzel, and B. Leibe. Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[5] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese CNN for robust target association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 33–40, 2016.

[6] Ali Borji, Saeed Izadi, and Laurent Itti. iLab-20M: A large-scale controlled object dataset to investigate deep learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2221–2230, 2016.

[7] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

[8] Xiaodan Liang, Si Liu, Yunchao Wei, Luoqi Liu, Liang Lin, and Shuicheng Yan. Towards computational baby learning: A weakly-supervised approach for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 999–1007, 2015.

[9] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[10] Jeff Hawkins and Sandra Blakeslee. *On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines*. Macmillan, 2007.

[11] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, pages 737–744, 1994.

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 1735–1742. IEEE, 2006.

[13] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[14] Hans-Christoph Nothdurft. Salience of feature contrast. In Laurent Itti, Geraint Rees, and John K. Tsotsos, editors, *Neurobiology of Attention*, pages 233 – 239. Academic Press, Burlington, 2005.

[15] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1817–1824. IEEE, 2011.