Computer Vision and Image Understanding 146 (2016) 9–26

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Semantic video labeling by developmental visual agents[☆]

Marco Gori[a], Marco Lippi[b,*], Marco Maggini[a], Stefano Melacci[a]

[a] Department of Information Engineering and Mathematics, University of Siena, Via Roma 56, Siena 53100, Italy
[b] Department of Computer Science and Engineering, University of Bologna, Viale Risorgimento 2, Bologna 40136, Italy

ARTICLE INFO

Article history:
Received 8 August 2015
Accepted 16 February 2016
Available online 26 February 2016

Keywords:
Learning from constraints
Life-long learning
Scene understanding
Motion estimation
Deep learning

ABSTRACT

In the recent years, computer vision has been undergoing a period of great development, testified by the many successful applications that are currently available in a variety of industrial products. Yet, when we come to the most challenging and foundational problem of building autonomous agents capable of performing scene understanding in unrestricted videos, there is still a lot to be done. In this paper we focus on semantic labeling of video streams, in which a set of semantic classes must be predicted for each pixel of the video. We propose to attack the problem from bottom to top, by introducing Developmental Visual Agents (DVAs) as general purpose visual systems that can progressively acquire visual skills from video data and experience, by continuously interacting with the environment and following lifelong learning principles. DVAs gradually develop a hierarchy of architectural stages, from unsupervised feature extraction to the symbolic level, where supervisions are provided by external users, pixel-wise. Differently from classic machine learning algorithms applied to computer vision, which typically employ huge datasets of fully labeled images to perform recognition tasks, DVAs can exploit even a few supervisions per semantic category, by enforcing coherence constraints based on motion estimation. Experiments on different vision tasks, performed on a variety of heterogeneous visual worlds, confirm the great potential of the proposed approach.

1. Introduction

Computer vision systems have nowadays shown outstanding results in several specific tasks that range from object recognition, detection, localization, to segmentation and tracking. The whole research field of computer vision is facing a great development, and related technologies can be found today in a variety of low-cost commercial devices such as cameras, tablets, and smartphones, as well as in highly advanced systems, as in the case of autonomous vehicles, augmented reality environments, medical diagnosis assistants, video surveillance controllers.

Despite this notable achievements, the general problem of constructing an automatic agent capable of performing visual scene understanding in unrestricted domains is far from being solved. As a matter of fact, the basic task of video semantic labeling (or scene parsing), which consists in assigning a semantic label to each pixel of a given video stream, has mostly been carried out only at the frame level, as the outcome of well-established pattern recognition methods working on images [1–5]. Conversely, we maintain that there are strong arguments to start exploring the more challenging problem of semantic labeling in unrestricted video streams, by developing automatic visual systems which can continuously interact with the environment and improve their skills, in a lifelong process which in principle never ends. Rather than exploiting huge amounts of labeled examples at once [6], which could be extremely costly to obtain in case of videos, we argue that these agents should be capable of using even only a few pixel-wise supervisions per semantic category, but exploiting the intrinsic information coming from motion in order to virtually extend supervisions, as well as to enforce coherence in predictions. Roughly speaking, once a pixel has been labeled, the constraint of motion coherent labeling virtually offers tons of other supervisions, that are essentially ignored in most machine learning approaches working on big databases of labeled images. This process resembles the visual interaction experienced in their own life by humans, who progressively acquire knowledge and competence, and can perform scene understanding after receiving just a few supervisions.

Following this idea, in this paper we introduce Developmental Visual Agents. DVAs continuously develop their visual skills by processing videos coming from any kind of source, and by interacting with users, from which they can ask for and receive supervisions. These agents implement a lifelong learning mechanism which

[☆] This paper has been recommended for acceptance by Gang Hua.

[*] Corresponding author.

E-mail addresses: marco@diism.unisi.it (M. Gori), marco.lippi3@unibo.it, marcolippi83@gmail.com (M. Lippi), maggini@diism.unisi.it (M. Maggini), mela@diism.unisi.it (S. Melacci).

http://dx.doi.org/10.1016/j.cviu.2016.02.011
1077-3142/© 2016 Elsevier Inc. All rights reserved.

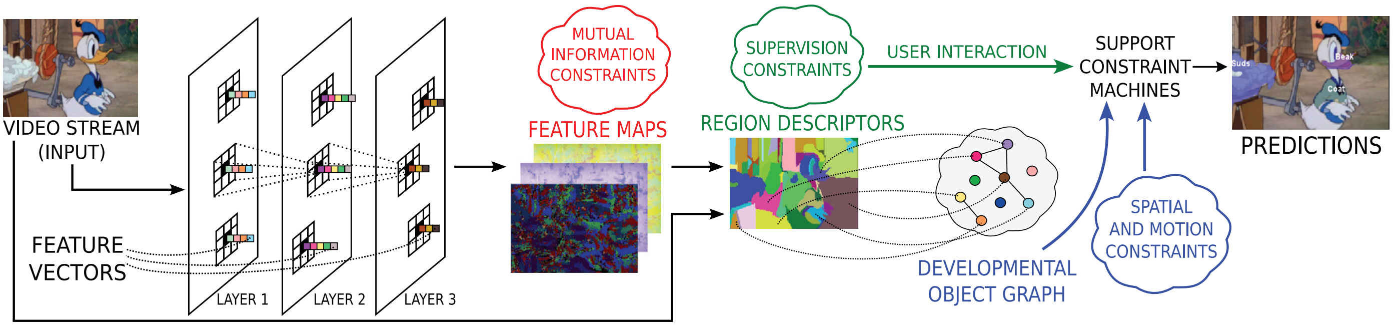

Fig. 1. An overview of the DVA architecture, which implements a complete visual scene labeling system, from feature extraction up to the symbolic levels where users interact, and where the agent reports its predictions (pixel-wise). Section 3 details the left-side of this picture up to the feature maps, while Section 4 focuses on the remaining rightmost blocks.

proceeds asynchronously with respect to the processing of the input stream and the acquisition of external information. We aim at devising these systems *from bottom to top*, starting from the low-level feature extraction process, up to the symbolic layer where interaction with users occurs, and scene parsing predictions are shown. On top of the representation gained by motion coherence, the mapping to linguistic descriptions is dramatically simplified.

In this scenario, the learning framework indeed plays a crucial role. The work described in this paper is rooted on the theory of *learning from constraints* [7] that allows us to model the interaction of intelligent agents with the environment by means of constraints on the tasks to be learned. The notion of constraint is very well-suited to express both visual and linguistic granules of knowledge. In the simplest case, a visual constraint is just a way of expressing the supervision on a labeled pixel, but the same formalism is used to express motion coherence, as well as complex dependencies on real-valued functions. In principle one could also include abstract logic formalisms, such as First-Order-Logic (FOL) formulae [8,9].

The main contributions of the paper can be summarized as follows:

- DVAs are introduced as general-purpose visual agents for semantic video labeling in unrestricted domains, in a complete bottom-to-top pipeline from feature extraction up to the symbolic layers;
- A lifelong learning paradigm for on-line video processing is defined, which exploits motion and temporal coherence throughout the life of the agent;
- An incremental development of the system is proposed, so that DVAs continuously interact with external users, who provide new supervisions as the learning process asynchronously proceeds;
- In order to perform real-time responses, motion coherence is widely exploited, and time budgets are introduced for handling and processing the input video stream.

A few ideas at the basis of the DVA architecture and the acronym “DVA” have been introduced in our previous works [10–12], yet they have been limited to a preliminary feature extraction process only, and have never been applied to video semantic labeling. The focus of this paper is on semantic predictions based on visual patterns and motion dynamics. In order to move closer to a real understanding of the scene, higher level reasoning mechanisms should be added to relate the predictions on the frame pixels. While these mechanisms are out of the scope of this work, the selected grounding theory of *learning from constraints* [7] is generic enough to naturally embed new types of knowledge into the DVA architecture.

The next Section of the paper will shortly describe the whole architecture of a DVA, while the subsequent Sections 3 to 4 will describe in detail the computational blocks regarding feature extraction and the symbolic levels, respectively. Section 5 will relate the DVA paradigm to existing works in the literature, while in Section 6 several experiments will show the performance of the proposed approach.

The software for running experiments with DVAs can be downloaded at the website of the project:

http://dva.diism.unisi.it

together with several videos and other supplementary material which illustrate the behavior of DVAs in different scenarios.

2. Developmental visual agents

A DVA is a system designed to perform semantic labeling in unrestricted domains, by living in its own environment and by continuously processing videos, following a lifelong learning paradigm. The agent is devised so as to implement all the levels of a truly on-line vision system, starting from feature extraction up to the symbolic layers where interaction with users occurs, and predictions on semantic categories are attached to visual patterns. The system architecture is depicted in Fig. 1.

Given a video stream $\mathcal{V}$, we indicate with $\mathcal{V}_t$ the video frame at time t. The first element of the DVA pipeline consists of computational blocks hierarchically organized into multiple layers, that will be described in Section 3. In each layer, a set of features are progressively learned and extracted from each pixel x of $\mathcal{V}_t$. The features of the $\ell+1$th layer are built upon the ones extracted at layer ℓ, and they pretty much resemble the responses to convolutional filters, with a local support that is limited to a small area around x, also referred to as *receptive field* [13], indicated with a small grid in Fig. 1. Filter responses on x are encoded and collected into a *feature vector* (a group of colored boxes in Fig. 1), and the response to the same feature for all pixels is referred to as *feature map* (Fig. 1).

DVAs parametrize receptive fields by considering also their transformed instances under the class of affine transformations, paired by a criterion that allows us to get an affine invariant representation of the data covered by the field. This choice allows DVAs to compactly represent such data by a fixed-length vector called *receptive input* (Section 3.1). As the on-line video processing advances, *motion estimation* is yielded by matching receptive inputs among consecutive frames (Section 3.2). Given the receptive inputs of $\mathcal{V}$ observed up to time t, a set of features are learned in an unsupervised setting, following information theoretical principles of Minimal Entropy Encoding [14] (Section 3.3). Features inherit motion coherence by the aforementioned matching scheme, and, within a given layer, they can be grouped to encode different properties of the input. Before being fed as input to the next layer, features are projected onto a space of lower dimensionality, by estimating the principal components over a time window with the