Pattern Recognition Letters 72 (2016) 41–51

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Pattern theory for representation and inference of semantic structures in videos☆

Fillipe D.M. de Souza[a,*], Sudeep Sarkar[a], Anuj Srivastava[b], Jingyong Su[c]

[a] Computer Science & Engineering, University of South Florida, 4202 E. Fowler Avenue, ENB 118, Tampa, FL 33620, USA
[b] Department of Statistics, Florida State University, 117 N. Woodward Ave., Tallahassee, FL 32306, USA
[c] Department of Mathematics & Statistics, Texas Tech University, 2500 Broadway, Lubbock, TX 79409, USA

ARTICLE INFO

Article history:
Available online 26 February 2016

Keywords:
Pattern theory
Graphical methods
Compositional approach
Video interpretation
Activity recognition

ABSTRACT

We develop a combinatorial approach to represent and infer semantic interpretations of video contents using tools from Grenander's pattern theory. Semantic structures for video interpretation are formed using generators and bonds, the fundamental units of representation in pattern theory. Generators represent features and ontological items, such as actions and objects, whereas bonds are threads used to connect generators while respecting appropriate constraints. The resulting configurations of partially-connected generators are termed scene interpretations. Our goal is to parse a given video data set into high-probability configurations. The probabilistic models are imposed using energies that have contributions from both data (classification scores) and prior information (ontological constraints, co-occurrence frequencies, etc). The search for optimal configurations is based on an MCMC, simulated-annealing algorithm that uses simple moves to propose configuration changes and to accept/reject them according to the posterior energy. In contrast to current graphical methods, this framework does not preselect a neighborhood structure but tries to infer it from the data. The proposed framework is able to obtain 20% higher classification rates, compared to a purely machine learning-based baseline, despite artificial insertion of low-level processing errors. In an uncontrolled scenario, video interpretation performance rates are found to be double that of the baseline.

1. Introduction

An automated system for providing semantic interpretations of videos will be greatly useful in many vision-based applications, including cyber-security and video content-based search applications. However, the current solutions in the literature (see [9] for a comprehensive survey) lack domain generalization and the ability to address multiple challenges at the same time. The difficulties in developing a fully-automated systems are multiple. Some relate to low-level detection and categorization, while others relate to high-level processing and interpretation. The challenges originate from: (1) handling errors stemming from the low-level processing layer, (2) lack of a flexible representation scheme to accurately model semantic structures of interest, and their variability, and (3) need for a principled mechanism to integrate multiple data sources and efficient algorithms to estimate interpretations in new data.

Table 1 summarizes the most recent approaches on the problem of video understanding. On one side, the simplest approaches concentrate on choosing the right combination of features and machine learning algorithms to infer the semantic contents as elements belonging to a pre-defined collection of semantic labels. In these approaches, feature detectors are assumed to be noise-free and learned classification models to be sufficiently discriminative with the chosen types of features. They rely on the robustness of learning algorithms to handle low-level feature errors. Additionally, no useful information about the structures of interest are explicitly encoded in the models – it is assumed that the chosen feature representation implicitly captures useful structural information descriptive of the target collection of semantic labels. Note that these classification models do not account well for structural variabilities inherent to real semantic structures. On the other hand, explicit models (or structured models) partially solves the later issues by providing complex structured models such as Bayesian networks to represent the semantic structures of interest and small variabilities. These proposed fixed structures are not

☆ This paper has been recommended for acceptance by Anders Heyden.

* Corresponding author. Tel.: +1 813 974 3652; fax: +1 813 974 5456.
E-mail address: fillipe@mail.usf.edu, fdms18@gmail.com (F.D.M. de Souza).

http://dx.doi.org/10.1016/j.patrec.2016.01.028
0167-8655/

Table 1
Summary of most recent related works on video understanding.

		Work	Features	Representation	Learning	Inference
Implicit models	Labeling	[15]	STIP, HOG	BoVW	Obj-Act Co-occurrence prob. distr.	DPM, bagged REP decision, Bayes' rules
		[7]	OpponentSIFT, STIP	BoVW-SVM	Concept score combination + SVM	SVM
		[10]	MPEG-Flow	BoVW-SVM	SVM	SVM
	Text description	[11]	Haar, color, EOH, COH	High-level features	Statistical language modeling	Probabilistic parser
		[21]	SIFT, STIP, MFCC	BoVW-SVM, event-concept relevancy matrix R	Manual definition of R	Linear combination of classification scores with R
		[12]	HOG, STIP	DPM, SVM	DPM, SVM, SVO LM	Linear interpolation of scores
		[4]	HOG3D, HOG, color	MMLDA, DPM	MMLDA, DPM, tripartite template graph	POS w/NLP tools + MMLDA + ranking
Explicit models	Labeling	[14]	Not mentioned	ADBN	Action detectors	Approx. viterbi
		[22]	STIP-BoVW	Variable duration HMM	LS-SVM	MAP w/dynamic programming
		[1]	STIP-BoVW	SPN	EM	MPE w/graph parsing
		[13]	HOG	Hierarchical PGM	S-SVM	Coordinate ascent + loopy BF
		[20]	Not mentioned	GC-DBN	Adaboost + EM	Maximum likelihood
		[16]	Not mentioned	SRG (CFG)	LS-SVM	FSM (Viterbi)
		[24]	Not mentioned	SCFG+BN	Manual + CFG compilation	Message passing
		[8]	STIP, trajectons-BoVW (actions as HMMs)	SCFG	Manual	Graph parsing w/HTK
	Structured output	[17]	Not mentioned	AOG	IP[a], MDLP[b]	Earley's-like parser
		[23]	Not mentioned	AOG	Manual	hierarchical cluster sampling + Earley's parser
		[25]	RGBD HOG, Kinects 3D joint motion vectors	4DHOI (Hierarchical graph)	Manual	DN Beam Search
		[19]	HOG, HOF	Pattern theory	Concept co-occurrence tables	MCMC-SA

[a] Information projection
[b] Minimum description length principle

obviously scalable for larger variations of target semantic structures and may require an enormous amount of training data to estimate good parameter values that characterize those variations well.

In this paper, we present a novel way of approaching the problem of video semantic interpretation. Here, the elements of Grenander's pattern theory [5] are used to model structures of semantic interpretations. This framework has been used in the past for applications involving computational anatomy, parsing language structures, and modeling biological growth [6]. In pattern theory (PT), generators are the most fundamental units of representation. In the context of video interpretation, they represent items pertaining to some domain-specific knowledge ontology, which are called ontological generators, and features extracted from videos, known as feature generators. Generators have bond structures that allow them to combine with each other, while satisfying ontological rules and forming connected structures that represent higher-level inferences. Thus, in our application, generators represent observable actions, imaged objects and video features. They can also potentially represent more complex concepts. The connecting bonds are termed in-bonds or out-bonds, determining the types of interactions between generators. Comparing to the terminology used in graphs, generators are nodes and bonds are edges. However, the generator structure is richer as it also allows for explicit representation of unconnected or dangling bonds, unlike in graphs where there is no concept of a dangling edge. The PT approach provides us with a framework for flexible representation of video semantic structures and a principled mechanism to infer these structures in new data. Such framework incorporates both prior knowledge and data in the form of machine learning-based classification models.

The contribution of this framework in advancing the state-of-the-art tools for video understanding is four-fold. First, it provides a principled and mathematically-grounded mechanism to model and infer semantic interpretations for video content descriptions. Second, it provides a very flexible and comprehensive structural representation scheme for describing complex semantic structures. Third, it is capable of overcoming errors generated by low-level classifiers, with the help of ontological constraints encoded in the representation; we shall demonstrate this capability using experiments. Fourth, the space of feasible interpretations does not grow exponentially in the number of features/concepts in our representation, yet it is rich and comprehensive. Consequently, inferences in our representation space are generated using polynomial time algorithms, built on a combination of MCMC (Markov Chain Monte Carlo) sampling and simulated annealing.

A short version of this work was presented in [19]; thus, this paper is an extended account presenting a more elaborate description of the theory, about the mapping of its conceptual elements to the practical scenario of modeling video semantic interpretations (so that it easily accessible for implementation) as well as more experimental studies for different scenarios of degradation (in terms of error rates from the low-level processing layer). We also published an extension of this work in [18] that studies how to handle inference for long sequences of videos by considering temporal information.