Pattern Theory-based Interpretation of Activities

Fillipe D. M. de Souza, Sudeep Sarkar,
Computer Science & Engineering
University of South Florida, Tampa
fillipe@mail.usf.edu, sarkar@cse.usf.edu

Anuj Srivastava,
Department of Statistics
Florida State University, Tallahassee
anuj@stat.fsu.edu

Jingyong Su
Department of Mathematics & Statistics
Texas Tech University, Lubbock
jingyong.su@ttu.edu

***Abstract*—We present a novel framework, based on Grenander's pattern theoretic concepts, for high-level interpretation of video activities. This framework allows us to elegantly integrate ontological constraints and machine learning classifiers in one formalism to construct high-level semantic interpretations that describe video activity. The unit of analysis is a generator that could represent either an ontological label as well as a group of features from a video. These generators are linked using bonds with different constraints. An interpretation of a video is a configuration of these connected generators, which results in a graph structure that is richer than conventional graphs used in computer vision. The quality of the interpretation is quantified by an energy function that is optimized using Markov Chain Monte Carlo based simulated annealing. We demonstrate the superiority of our approach over a purely machine learning based approach (SVM) using more than 650 video shots from the YouCook dataset. This dataset is very challenging in terms of complexity of background, presence of camera motion, object occlusion, clutter, and actor variability. We find significantly improved performance in nearly all cases. Our results show that the pattern theory inference process is able to construct the correct interpretation by leveraging the ontological constraints even when the machine learning classifier is poor and the most confident labels are wrong.**

I. Introduction

Computational models for video activity recognition can be described as a hierarchy of representations, each representing larger and larger temporal scales. One such hierarchy involves transition through representations of parts and their movements (e.g. hand, head), actions (e.g. pick up, walk, run) or objects (e.g. can, bowl), interactions (e.g. person stirring butter), followed by tasks (e.g. preparing cookie dough). The input consists of low-level spatio-temporal features detected in the video. To date, few works have attempted to address all levels of this hierarchy. There are many works on action recognition and very few on the interaction or task level recognition [1]; thus, we focus on the latter levels. As has been comprehensively discussed in [1], high level event recognition varies from direct classification to fusion techniques. Most works try to bridge the hierarchical gap in one step using some form of machine learning approach that is trained with a sufficiently large dataset.

Some of the current methods in interaction and task level recognition are categorized in Table I. As we can see, most of the works start from a feature histogram representation. HOG and STIP features appear to be most common. The final outputs of these approaches are mostly sentences, arrived at using different methods that rely on a machine learning-based labeling of objects and actions. Many of the approaches try to bridge the gap between detected features and final concept-label in one step. Support vector machines (SVM) appear to be the most commonly used machine learning technique. These outputs are then used to generate sentences. A popular method for this final step involves using predefined sentence templates that are filled in using the inferred labels. Another method uses an event-interaction concept matrix [2] where an event is characterized by the occurrence of a particular subset of simpler concepts. The linear combination resulting from the multiplication of the event-interaction concept matrix and a test video concept score vector is a chain of probabilities for each possible event. The highest score in the chain is the final label of the video, which guides the generation of textual descriptions. Some approaches [3], [4] use prior knowledge derived from text mining but do not use predefined ontology.

In this paper, we present a fundamentally new approach to high level video interpretation using Grenander's pattern theory formulation [10]. The basic idea in this theory is to use subject matter knowledge, i.e. domain ontology in our case, to develop algebraic representations of the system under study, and then impose probabilistic superstructures on them to drive inferences. The inference engines are kept general to allow for simultaneous variable and model estimation. This framework is both rigorous and comprehensive. In recent years, the strengths of pattern theory have become more visible in applications involving biological growth [11] and stochastic models for human thoughts (unpublished manuscripts can be found at http://www.dam.brown.edu/ptg/publications.shtml). An integral part of our approach is the use of ontology, which refers to an organization of concepts in terms of their properties, and more importantly, relationships and dependencies with other concepts. This theory builds on well known ideas of stochastic grammars and syntactic pattern recognition.

The novelty and distinctiveness of our approach are threefold. First, our proposed method is an integrated process that evaluates ontological label confidence as needed during the search process which is in contrast to the sequential ideas presented in the literature. This helps us recover larger structures even from poor detection results. Similarly, our framework boosts weak predictions by domain knowledge likelihoods derived from dataset annotation. Second, while previous works focus on generating textual descriptions, we propose to extract graphical structures that represent the video contents. These graphical structures allow us to express video interpretations in a language independent manner and can be mapped to language constructs if so desired. Third, our

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.28

TABLE I

SUMMARY OF WORKS ON HIGH LEVEL VIDEO ANALYSIS OF INTERPRETATIONS AND/OR TASKS. LEGEND: (*) LOCATION, DISTANCE OF FACIAL FEATURES, AND TORSO AND ARM LENGTH, ANGLE, STRIDE; LC - LINEAR COMBINATION; ECM - EVENT-CONCEPT MATRIX; SDP - STANFORD DEPENDENCY PARSER; SVO - SUBJECT-VERB-OBJECT; DPM - DEFORMABLE PART MODELS; DT - DECISION TREE

Hierarchy	Component	[5]	[6]	[2]	[7]	[8]	[9]	[3]	[4]
	Prior Knowledge	-	-	-	-	Text Mining	-	SVO mining (SDP[1])	Video Textual Description
Level 1	*Spatial Features*	edges, color	Haar, Color, Geometry*, EOH	SIFT	HOG, body post.	BoVW + HoG	OpponentSIFT	HOG	HOG3D, HOG, color
	Motion Features	-	-	STIP (HOG,HOF)	OF	BoVW + STIP	STIP (HOG,HOF)	STIP (HOG,HOF)	-
Level 2	*Object Recognition*	Matching	Cascade	SVM	DPM	DPM	SVM	DPM, Verb Expansion	MMLDA, DPM
Level 3	*Action Recognition*	Matching	HMM	SVM	HMM	DT	SVM	SVM, Verb Expansion	MMLDA
	Activity Recognition	-	-	LC[2] w/ ECM[3]	HMM	Chain Rule	-	Content Planning	Tripartite Template Graph
	Task Recognition	-	-	LC w/ ECM	-	-	SVM + ECM	-	-
	Sentence Generation	concept hierarchy of actions	Probabilistic Template Filling	Template Filling	Template Filling	-	-	BerkeleyLM	MMLDA + Ranking + NN

proposed algorithm is also easily extendable to recognition of semantically higher-level events by simply adding on ontological layers.

II. PATTERN THEORY BASED COMPUTATIONAL STRUCTURE

The basic units of representation in pattern theory are generators that represent ontological labels as well as the detected features in the video. Each ontological generator represents either an action, an interaction, or a task level concept and are associated with possible bonding links that can connect to other generators. Each feature generator represents a group of video features detected in the image, e.g. a bag of features describing a detected object or a tracked action in the video. These generators are linked using bonds with different constraints. The bonds between the ontological generators are constrained by the domain ontology whereas the bonds between the ontological and feature generators are determined by the strength of classification and represent signal level support for concepts. An *interpretation* of a video is a configuration of these connected generators that results in a richer graph structure with more expressive power than those commonly used in computer vision. These interpretative structures link object and action labels that are tied to observed features. The scene inference is guided by an energy function that is optimized using Markov Chain Monte Carlo (MCMC) based simulated annealing. This framework allows us to elegantly capture ontological constraints and bottom-up machine learned label outputs as a means to construct high-level semantic interpretations that describe video activities. While integrating several stages, this inference procedure simultaneously performs high level event recognition. In the following subsections, we outline this theory customized to our context using more rigorous language.

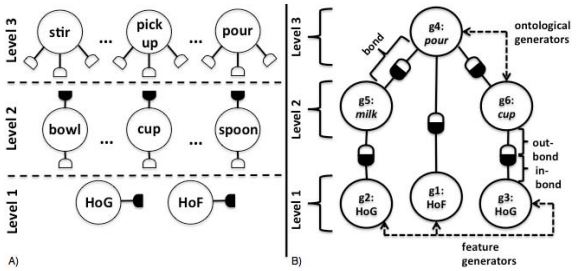


Fig. 1. A) Hierarchy of generators divided by level. Out-bonds are shown in white semicircles and in-bonds are shown in dark semicircles. B) An example of pattern theoretic interpretation of a video. Generators represent detected features, objects, and actions. Bonds between ontological and feature generators are called grounding links. Bonds between ontological generators are referred to as ontological links.

A. Generators - Morphology and Equivalence Classes

A generator $g \in G$ is the basic unit used to construct regular structures that represent information about patterns, where G accounts for the whole space of generators. Some example generators are shown via a schematic in Figure 1A). Any generator g has a bond structure $B(g) = (B_s(g), B_v(g))$ that is defined by an arrangement $B_s(g)$ of (in- and out-) bonds with coordinates in the range $j = 1, 2, ..., w(g)$ (such that $w(g) = w_{in}(g) + w_{out}(g)$) and a set $B_v(g)$ of bond values $\beta_j(g)$. Any generator has a variable number of *in-bonds* $w_{in}(g)$ (or *in-bond arity*) and a constant number of *out-bonds* $w_{out}(g)$ (or *out-bond arity*). The set of out-bonds can be thought of as the signature of a generator.

The domain ontology establishes the morphology of the generators, their roles, and how the connections should be made to express contextually relevant structures. These structures shall represent human-perceivable patterns in the real

world (e.g., pour milk in bowl, slice carrots, etc.).

Each level of the hierarchy represents a categorical concept, as show in Figure 1A). Thus, the set of generators can be seen as a composition of generator subspaces, $G = \cup_\alpha G^\alpha$, where $\alpha \in \{1, ..., M\}$ correspond to the levels. The lowest level in the hierarchy is composed of generators that represent feature groups detected in the video, such as bags of similar histogram of oriented gradients (HOG) features or histogram of optic flow (HOF) features. These generators do not have out-bonds, but only in-bonds, connected to higher level concepts. The intermediate level of the hierarchy represents inanimate object concepts, such as bowl, cup, etc. The top level representing actions (that interact with objects), such as stir, pickup, etc. This hierarchy can be extended upward with new generators representing task level concepts. In this paper, we restrict ourselves to three levels.

Disjoint subsets of the generators, each corresponding to a hierarchy level, can also contain equivalence classes that will allow us to switch among generators transformed by a valid group. Let S be a similarity group that induces an equivalence relation on G^α, and contains similarities $s \in S$. This means that G^α is partitioned into equivalence classes G^α_γ such that any pair of generators in a class holds similar properties. Therefore, an equivalence class is a subset $G^\alpha_\gamma \subset G^\alpha$ of generators in which any two generators $g_i, g_j \in G^\alpha_\gamma$ possess similar properties and hold the same bond structure. For instance, the group $\{water, juice, oil, milk\}$ is an equivalence class, representing *liquids*, which can be interchanged.

B. Interpretations

Generators combine and relate to each other by connecting their out-bonds to in-bonds of others, forming structures referred to as *interpretations*. Formally, an *interpretation* $\sigma(g_1, g_2, ..., g_n)$ is a connected structure σ that represents a composition of n generators $g_i \in G$ while respecting the bond relation ρ. The bond relation ρ is a truth function that determines whether two bonds β' and β'' are compatible. This is determined from the ontological constraints for ontological bonds.

Figure 1B) depicts a possible interpretation of a video where three groups of features were detected: an HOF feature generator g^1_f (accounting for the motion found in the scene) and two HOG feature generators (g^2_f and g^3_f). At the ontological level, the generator $pour$ has two out-bonds (with values $milk$ and cup) and no in-bonds, while the generators $milk$ and cup have both one out-bond to explain features and one in-bond to sustain the significance or existence of other generators.

We can define the concept of a regular or irregular interpretation based on the bonds that are connected. Note that an interpretation need not have all the bonds connected. An interpretation $c = \sigma(g_1, g_2, ..., g_n)$ with all out-bonds connected is called a *regular interpretation* (or completely regular interpretation). Interpretations with only some of their out-bonds connected have incomplete meanings and are called irregular interpretations. For instance, if the out-bond of the *cup* generator was not connected to the HOG generator in Figure 1B), we would have an irregular interpretation (i.e. an interpretation with gaps).

C. Probability of Interpretation

The probability of an interpretation is expressed as a product of terms associated (i) with each individual generator and (ii) each bond. To each generator, we associate a positive-valued quality function $Q(g)$ that captures its importance or preponderance in the domain ontology. In this paper, we have chosen these values to be the same constant; however, they can be relaxed in the future.

A bond between two generators is formalized as the response of an acceptance function, $A(g_i, g_j)$, that determines the degree of compatibility in the connection $g_i \downarrow g_j$. We have two kinds of bonds: bonds between two ontological concepts ($g^3_i \downarrow g^2_i$) and bonds between an ontological concept and detected feature generator, e.g $g^3_i \downarrow g^1_i$ and $g^2_i \downarrow g^1_i$ is depicted in Figure 1. The intra-ontology links are derived from the object-action co-occurrence tables and ontology domain constraints. We compute the fraction of times that a particular pair of ontology labels (i.e. action-object, object-object, and action-action category pairs) exists in the training data. We also make use of domain knowledge in terms of links that we know should and should not exist between ontologies. This simple estimation procedure can perhaps be made more rigorous using a Bayesian framework; however, it suffices for our current illustration of the pattern theory framework. We quantify the second type of links (i.e. between feature generators and ontological labels) using confidence values returned by machine learning classifiers. These grounding links indicate that a certain ontological concept denoted by generator g^2_i has its occurrence supported by a generator g^1_j that holds data of a specific kind of feature (e.g., shape features, appearance features). We use the soft outputs of inference algorithms (e.g., SVM) that are run to detect the concepts represented by the generators. For example, in the connection $g^2_i = bowl \downarrow g^i_j = HOG$, we could run SVM predictor that knows a classification model of $bowl$ learned with HOG features; the prediction confidence value would be used as the weight value supporting the connection $g^2_i \downarrow g^1_j$.

The probability of interpretation $\sigma(g_1, \cdots, g_n)$ is given by

$$p(\sigma) = \frac{k_n}{n! Z(T)} \prod_{i=1}^{n} Q(g_i) \prod_{(k,k') \in \sigma}^{n} (A(\beta_j(g_i), \beta_{j'}(g_{i'})))^{\frac{1}{T}}, \tag{1}$$

where $Z(T)$ is the partition function and the ratio $\frac{k_n}{n!}$ captures the complexity of the interpretation. The energy of an interpretation is then written as $E(\sigma) = -T \log p(\sigma) Z(T)$. The number of generators in the interpretation is denoted by n, and k_n is a constant. This multiplicative ratio factor helps the probability function have a desirable property - probability of an interpretation that is the simple union of two smaller interpretations is lower than the product of the probabilities of the individual interpretation. In other words, for two independent groups $\sigma = (\sigma_1, \sigma_2)$ we can show that $E(\sigma) = E(\sigma_1) + E(\sigma_2) + B(s_1 + s_2, s_1)$ where $B(.,.)$ is

the binomial coefficient which is always ≥ 1. Thus, there is an in-built bias against simple addition of two interpretations. We can also show that the partition function, which actually represents a *grand* Gibbs ensemble, is bounded if ρ is set to less than a value that is dependent on the maximum link values.

D. MCMC-based Simulated Annealing Inference

Computationally, the crucial task is to be able to compute the probabilities of possible interpretations without computing the notorious partition function, $Z(T)$. The inference engines will be kept general to allow for simultaneous variable and model estimation; a particular tool in this setup is the random sampling of posterior. We can synthesize random possible interpretations by an iterative procedure of simple and/or composite moves through the interpretation space. We employ a MCMC based simulated annealing process to maximize the probability function or equivalently minimize the corresponding energy function $E(\sigma)$. The MCMC process uses a global proposal function that suggests significant changes in the interpretation structure and a local proposal function that suggests simple moves. We generate the following global proposal. For feature generator g_i^1, select an ontological generator $g_j^{2,3}$ that can be possibly linked, based on top k classification results. Then we add the ontological connections between the selected explanations to create the interpretation. For local proposal moves, we randomly select a generator from the interpretation and replace it with the proposed surrogate that yields the lowest local energy. At each iteration of the simulated annealing, we randomly choose between these moves based on a prechosen preference for making global moves. The simulated annealing process is run on a linearly decreasing temperature schedule.

III. Experimental Setup

We discuss the dataset, its challenges, and then outline details about the key pattern theoretic concepts. This is followed by an outline of the baseline algorithm considered for comparative analysis and, subsequently, a description of the evaluation framework.

A. Data

We validated the proposed framework on a challenging, recent YouCook dataset [4], consisting of instructional videos of different cooking styles, such as assembling, baking, grilling, etc. From a computer vision point of view, this dataset poses enormous challenges for vision algorithms due to presence of camera motion, diverse background scenes and contexts, clutter, and differences in subjects. In this study, we focus solely on the evaluation of labeling and interpretation problem by using a subset of the dataset whose objects and actions locations have been annotated (44 of 88 videos). Out of this subset, we selected 22 videos to form a new training set and the remaining 22 for the test set such that they both had roughly the same object and action categories. There are 6 action categories ($stir$, $pickup$, $putdown$, $season$, $flip$, and $pour$) and 18 object categories ($bowl$, cup, $spatula$, $knife$, pan, $tongs$, $plate$, oil, $pepper$, $tomato$, $butter$, $spreader$, $bread$, $spoon$, $lemon$, $carrot$, $meat$, and egg).

Each video consists of several shots that depict the different steps of a cooking recipe. These shots form our units for interpretation. There are 309 training shots and 359 testing shots. Each shot exhibits one of the studied actions and displays some objects that might be participating in the action or just appearing in the scene. For example, a shot could depict a cook picking up a whisk to stir ingredients in a bowl while a slice of meat and a knife are on the table.

B. Object and Action Feature Generators

As you might recall, in our pattern theoretic framework, for each detected object and action, we instantiate a feature generator, characterized by its feature representation. An action is represented by a sequence of three stages of motion pattern captured by histograms of optic flow (HOF). Dense optic flow frames are computed for the pairs of consecutive frames in a shot. Each shot is further divided in three segments. An HOF, weighed by the magnitude, is then assembled for each segment to characterize the motion patterns of the action start, its development, and its ending. The action is then represented by the ordered concatenation of the three HOFs extracted from the shot segments. As for the objects generators, we use the histograms of oriented gradients (HOG) representation. Other more sophisticated features are possible; however, these suffice for now to demonstrate the power of using ontology.

C. Ontology to Feature Bond Quantification

We have two kinds of bonds: i) bonds between two ontological concepts and ii) bonds between an ontological concept and detected feature generator. As we described earlier, for the latter, we use multiclass classification models for the action and for objects built using linear support vector machines with LibSVM [12]. Because the number of training instances across categories is uneven, we generated synthetic samples using SMOTE [13] for the minority categories (e.g., categories $season$ and $flip$ had only 4 and 7 training instances). We noted that the confusion matrices of the classifier's diagonals are only partially dominant. This means that the learned models are weak if recognition simply consists of labeling based upon the model's best prediction scores. There is a noticeable confusion between overlapping categories. For example, there is a great deal of the picking up action involved in flipping action ($pickup$ and $flip$). Additionally, training instances of objects typically used together add confusion to the classification models. For instance, spatulas are commonly used to stir ingredients in a bowl ($bowl$ and $spatula$). This is the kind of confusion that we expect to be alleviated by the inclusion of prior knowledge ontology. For instance, the action $pickup$ could be ruled out in favor of $pour$ to label the action happening in a scene after inferring the presence of cup and oil. Additionally, ontology constraints and prior knowledge derived from the co-occurrence of objects and actions can also be used to boost the inference process.

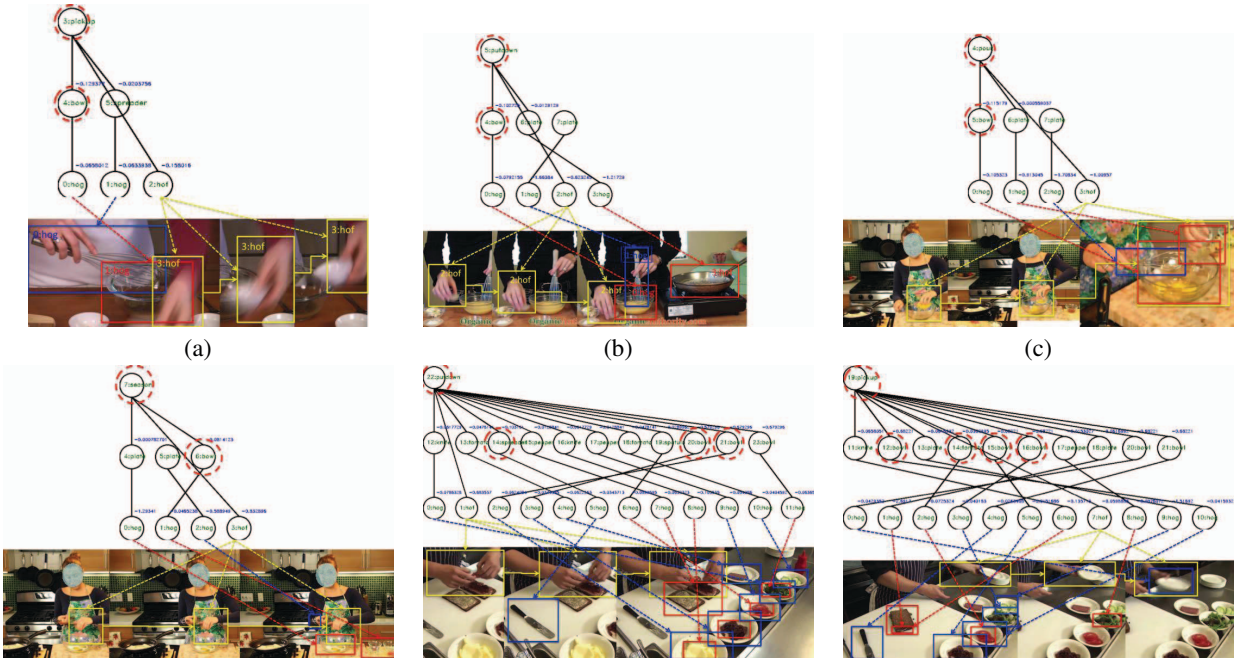
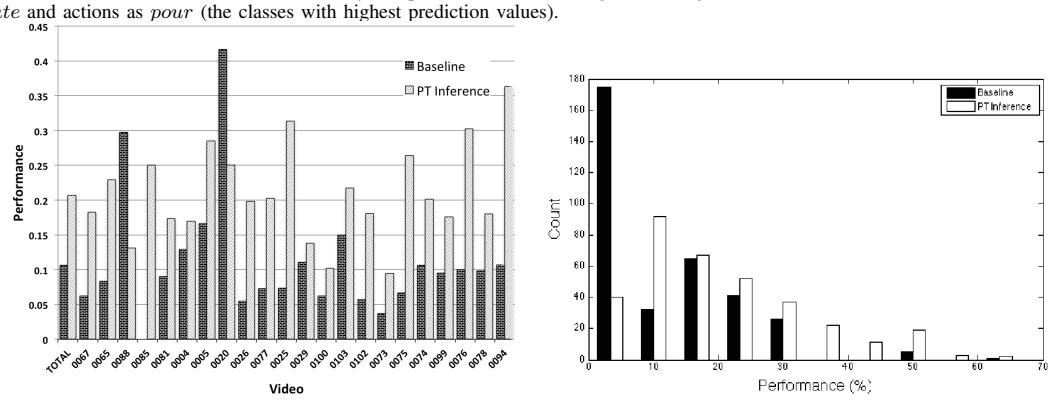


Fig. 2. Interpretations generated by the PTI algorithm for different video shots (a-f) are depicted in terms of circles and links forming graph structures above their corresponding sample frames. Shot (a-f) increase in visual complexity from left to right and top to down. The feature generators are grounded to their corresponding regions in the images. The correct ontological labels are outlined with dashed, red circles. Shots (c-d) and (e-f) are pairs of consecutive shots from the same video. Shots (e-f) exhibit the most visually complex scenario containing several objects. For these cases, the baseline erroneously labeled all objects as *plate* and actions as *pour* (the classes with highest prediction values).



(a) (b)

Fig. 3. (a) Percentage of correct labels by the pattern theoretic inference process and by the baseline algorithm, computed over videos depicting full recipes. (b) The number of shots with a particular performance for the baseline and PT Inference. The horizontal axis represents the performance scores with good values to the right.

D. A Baseline Interpretation Algorithm

We are not aware of any approach that we can use as a basis for comparison. Our output consists of a graphical representation involving ontological concepts, grounded by observed features. Most competing approaches output sentence level interpretation. So, we constructed a baseline approach that has a machine learning flavor, which is the dominant paradigm in computer vision. The baseline algorithm simply returns the object action nodes connected to the observed features. There are no connections between the object and the action labels; thus, no judgments are made about what action is performed on which object.

E. Evaluation Metric

An interpretation consists of ontological and feature generators connected by bonds. The performance is measured by calculating the fraction of the number of correct ontological generator labels associated with each feature generator to the total number of feature generators. Using the examples depicted in Figure 2, this would be the number of correct label

associations (highlighted with red dashed circles) divided by the total number of feature generators.

IV. Results

In Figure 2 we present a few examples of output by the pattern theoretic inference (PT Inference) method. The shots display visually complex scenarios containing objects as well as their interactions with other existing objects (e.g., egg yolk in a bowl) and their participation in the ongoing action (e.g., person grab plates). In case of shots (a) and (b), the best machine learned output labels are all wrong (i.e., the baseline algorithm score is 0). Thus, relying on purely bottom up, machine learned, labels will result in very poor performance owing to the notoriously weak classifiers. Ontological constraints help to pull out labels that are not necessarily the ones declared with the highest confidence by the SVM classifier. The PT Inference process attempts to make sense of explanations given by weak prediction values, instead of discarding them, by considering other co-relevant predictions. These improvements come from co-support provided by the labels predicted in the scene.

In Figure 3 (a) we compare the average performance of interpretations by the purely bottom-up baseline and PT Inference methods for each test video (a complete recording of one recipe). The interpretation performances are computed based upon the percentage of correct labels predicted on each shot. With the exception of 2 test videos, the baseline approach is outperformed by the PT Inference approach. On average PT Inference improved the recognition performance by 10% . Figure 3 (b) further demonstrates the superiority of PT Inference. It shows the histogram of scores for the two algorithms. We observe that the baseline algorithm results in mostly zero scores, whereas the PT Inference algorithm results in higher scores.

On first consideration there are similarities between the Pattern Theory (PT) framework and Probabilistic Graphical Models (PGMs). Both employ the use of graphs structures, such as nodes and links, and reasoning with both of them can be cast in energy minimization terms. However, there are fundamental differences in the representation, which imparts certain advantages to PT over PGM. The nodes in PGM represent random variables, but nodes in PT (i.e. generators) can be seen to represent specific outcomes. The links in PGM capture dependencies and are quantified by probability values, whereas the links in PT (i.e. bonds) capture compatibility between outcomes in terms of energy values. Modeling in terms of random variables makes the PGM specific for each condition. For instance, the PGM structure to model the interaction of a person with an object will be different from one that is used to model interaction of two persons with an object. We will have to pre-specify (or learn from prior labeled data) these structures. However, for PT, we just need to specify the energy model between people and object generators and it will be able to handle interaction among any number of people and any number of objects. In other words, PT is easily extensible to handling new situations.

V. Conclusion

We find that Grenander's pattern theoretic concepts offer an elegant framework for building high-level interpretation of video activities in terms of probabilistic algebraic structures. Ontological constraints and machine learning labeling approach, which is the current computer vision paradigm, are naturally integrated into this framework. The resulting interpretative structures link object and action labels, grounded to observed features in the sequences. The MCMC based simulated annealing performs well in optimizing this energy function. We demonstrated the superiority of our approach over a purely machine learning based approach (SVM) using more than 650 video shots from the YouCook dataset. Our results show that the PT Inference algorithm constructs nearly correct interpretations by leveraging the ontological constraints even when the machine learning classifier is poor and the most confident labels are often incorrect.

Acknowledgment

This research was supported in part by NSF grants 1217515 and 1217676.

References

[1] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *International Journal of Multimedia Information Retrieval*, pp. 1–29, 2012.

[2] C. C. Tan, Y.-G. Jiang, and C.-W. Ngo, "Towards textually describing complex video contents with audio-visual concept classifiers," in *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011, pp. 655–658.

[3] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, U. Lowell, and S. Guadarrama, "Generating natural-language video descriptions using text-mined knowledge," *NAACL HLT 2013*, p. 10, 2013.

[4] P. Das, C. Xu, R. F. Doell, and J. J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2634–2641.

[5] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *International Journal of Computer Vision*, vol. 50, no. 2, pp. 171–184, 2002.

[6] M. U. G. Khan, L. Zhang, and Y. Gotoh, "Towards coherent natural language description of video streams," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 664–671.

[7] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi *et al.*, "Video in sentences out," *arXiv preprint arXiv:1204.2742*, 2012.

[8] T. S. Motwani and R. J. Mooney, "Improving video activity recognition using object recognition and text mining." in *ECAI*, 2012, pp. 600–605.

[9] J. Guo, D. Scott, F. Hopfgartner, and C. Gurrin, "Detecting complex events in user-generated video using concept classifiers," in *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*. IEEE, 2012, pp. 1–6.

[10] U. Grenander, *General Pattern Theory*. Oxford University Press, 1993.

[11] U. Grenander, A. Srivastava, and S. Saini, "A pattern-theoretic characterization of biological growth," *IEEE Transactions on Medical Imaging*, vol. 26, no. 5, pp. 648–659, 2007.

[12] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.