Temporally Coherent Interpretations for Long Videos Using Pattern Theory

Fillipe Souza, Sudeep Sarkar
University of South Florida
Tampa, FL USA
fillipe@mail.usf.edu,sarkar@cse.usf.edu

Anuj Srivastava
Florida State University
Tallahassee, FL USA
anuj@stat.fsu.edu

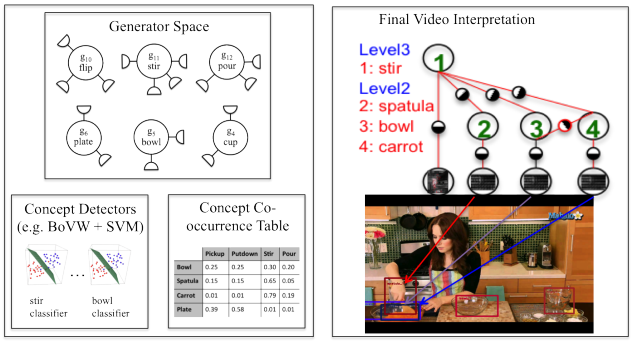
Jingyong Su
Texas Tech University
Lubbock, TX USA
jingyong.su@ttu.edu

Abstract

Graph-theoretical methods have successfully provided semantic and structural interpretations of images and videos. A recent paper introduced a pattern-theoretic approach that allows construction of flexible graphs for representing interactions of actors with objects and inference is accomplished by an efficient annealing algorithm. Actions and objects are termed generators and their interactions are termed bonds; together they form high-probability configurations, or interpretations, of observed scenes. This work and other structural methods have generally been limited to analyzing short videos involving isolated actions. Here we provide an extension that uses additional temporal bonds across individual actions to enable semantic interpretations of longer videos. Longer temporal connections improve scene interpretations as they help discard (temporally) local solutions in favor of globally superior ones. Using this extension, we demonstrate improvements in understanding longer videos, compared to individual interpretations of non-overlapping time segments. We verified the success of our approach by generating interpretations for more than 700 video segments from the YouCook data set, with intricate videos that exhibit cluttered background, scenarios of occlusion, viewpoint variations and changing conditions of illumination. Interpretations for long video segments were able to yield performance increases of about 70% and, in addition, proved to be more robust to different severe scenarios of classification errors.

1. Introduction

The problem of understanding activities in video data and providing meaningful semantic interpretations is very important. In recent years, a variety of solutions have been proposed and, among other ideas, the techniques based on encoding scene structure using graphs have shown promise in this problem area. These approaches represent items of interest – objects, actors, actions, etc. – as nodes in graphs and ascertain their interactions through graph edges. The main advantage of this framework is that one can naturally associate probability models with such graphs, thus providing statistical interpretations to solutions. Also, one can use both prior knowledge and the current data to deduce optimal interpretations in a coherent way. The main limitation in the current graph-theoretical solutions has been the rigidity of graph structures. In most cases, the graph geometries (connectivities, neighborhoods, etc) are pre-determined and only the node values are allowed to be variable. Even when the edges are allowed to change, they are usually based on a simple thresholding, or decisions that are spatio-temporally local, i.e. isolated from other nodes.



(a) (b)

Figure 1: Overview of the pattern theoretic framework proposed in [1]. (a) shows basic elements of this framework: a generator space containing basic ontological elements of representation called generators, machine learning-based concept classifiers and prior knowledge in terms of frequency tables of concept co-occurrences. (b) shows a pattern theoretic video interpretation that is a combination of generators. Connections between generators that represent ontological concepts indicate occurrence of certain interactions. Features are connected to ontological generators to support their semantic value in the interpretation.

Souza et al. [1] recently introduced a flexible graph-theoretical approach that is based on Grenander's pattern theory [2]. Here the flexibility comes from the fact that both

nodes and edges are allowed to be variables and are inferred from available knowledge. There are two dominant sources of knowledge: (1) the prior in form of frequency tables of concept co-occurrences, contextual knowledge about actions represented by the underlying ontology extracted from previous annotated videos, and (2) objects and actions detected using machine learning techniques, and their detection scores, in the current video. In [1], the authors studied short videos containing individual actions (pick up, put down, pour, stir, etc) and demonstrated the strength of this pattern theoretic approach and the flexibility of its representation. Here one does not need to explicitly model each of the variants for an action. It is capable of discovering hidden events or events not previously considered during annotation phase. An illustration of this pattern theoretic framework is shown in Figure 1.

The main limitation of [1]'s work is that it cannot perform well with large videos containing multiple actions and complex interactions between actions. If one splits large videos into smaller, disjoint time windows and performs individual inferences, then the overall interpretation can be both inconsistent and sub-optimal. In this paper we pursue a more comprehensive approach by introducing temporal bonds across sub-configurations that represent individual interactions (actions performed on/with objects). This additional structure enables us to discard (temporally) local solutions in favor of globally optimal and temporally consistent configurations, as illustrated in Figure 2.

We demonstrate these ideas on a recent challenging data set of cooking scenarios, the YouCook datasets. Its videos depict high-level activities in unconstrained scenarios, with cluttered background, clutter of objects, variable conditions of illumination, different viewpoints and camera motion.

2. Related Work

Implicit and explicit structural models have been proposed for tackling the problem of generating semantically complex video interpretations. The main difference between these methods is in the way contextual, logical and temporal dependencies are encoded. Implicit structural models [3] [4] [5] attempt to capture these information implicitly through some general form of data representation [6], such as the BoVW framework [7] and Linear Dynamical Systems (LDS) [8], or a set of coefficients learned using the max-margin framework [9]. Others compute distributions or some statistical summaries, such as the co-occurrence of concepts, which serve to indicate or corroborate with certain probability the existence of more complex ones and also used to derive semantic description through text [10, 11, 12]. These approaches do not offer flexibility in representation because if new concepts are later added to redefine the characterization of a single complex event, the models have to be reconstructed. Moreover, it is not clear how scalable these methods are for when the structure variability and semantic complexity of the target events increase.

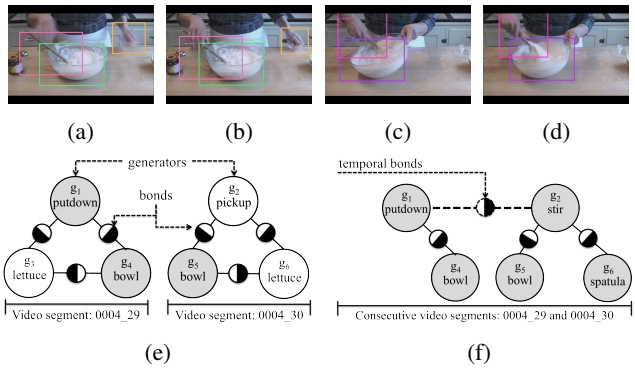


Figure 2: Illustration of advantage in using temporal bonds. Top rows shows frames from two consecutive segments of a video. The first segment depicts the interaction *put bowl down* (the small one with the left hand) and second segment depicts *stir ingredients in a bowl using spatula*. (e) shows [1]'s interpretations for both segments. (f) shows our approach's interpretation for both segments. Shaded circles denote correctly identified generators.

The traditional explicit structural models are most closely related to our approach. These models are typically hand-crafted or algorithmically learned from the training data. They can be generated in terms of dynamic Bayesian Networks (DBN) [13][14], Sum Product Networks (SPN) [15, 16], Stochastic Context-free Grammars (SCFG) [17][18][19][20], AND-OR graphs [21][22], Petri Nets [23], or general hierarchical graphical models [24] [25]. Complex events are usually sought to be composed by a set of temporally ordered sequence of sub-events [18], which can also suffer certain order variations or have optional steps [19]. These works typically consist of a low-level layer in which feature observations provide evidence for concepts from the top layers, such as sub-events or composite events. For example, Hilde *et al.* [19] use HMMs to learn models for sub-events such as *pour coffee* and *take cup* and model more semantically complex events such as *preparing coffee* using a SCFG that describes their syntax in the form of occurrence of sub-events. Other works use the SVM framework to provide confidence values as evidence for the occurrence of sub-events. Context and temporal dependence constraints are mostly supplied by coefficients learned with the max-margin framework or co-occurrence statistics of the target concepts.

Unlike the representation proposed by [1], these methods face limitations in representation caused by increase in structural complexity and non-linearity with respect to the variations in order, presence or absence of certain sub-

events, specially when structural learning is required [14]. Such models impose a typical order or a limited number of variations [19] in which a complex events can occur given a set of related sub-events. Automatic learning of structures does not seem easily scalable for when updates on the models are needed to include new knowledge and can require large amounts of data to be available for learning all possible structural variations. Additionally, it does not provide a mechanism to discover new optional sub-events. Following [1], our approach overcomes these limitations by putting forward a representation model based on rules of domain knowledge ontology that is formed mostly by simple concepts, each having a combinatorial signature. These combinatorial signatures span a space of structures that may represent either interactions between concepts or complex events of interest. By simple combinatorial rules, our method is capable of accounting for a larger space of structures, which includes identifying different structural variations and unseen structures that represent one same event.

3. Pattern Theoretic Formulation

Knowledge representation through Pattern Theory consists of using concepts and concept combination rules described by some domain-specific ontology. In Pattern Theory, these concepts become *generators* and combination rules are transformed into *bond structures* of generators. Based on the target ontology, generators are organized into levels of abstraction and at each level they are further classified into different *modality groups*, each group establishing different properties of similarity (which can be structural or semantic). A collection of generators organized in this fashion forms a domain-specific *generator space*. Such generator space forms the basis for generating complex structures by the combinations of generators. Formally,

Definition 1 Generators *are building blocks used to construct graph structures that represent patterns of interest. They will be denoted by $g_i \in \mathcal{G}$, where $\mathcal{G}$ is a chosen generator space.*

Remark 1 *Each generator g_i has a* **bond structure** $B(g_i) = (B_s(g_i), B_v(g_i))$ *that is defined by a structural arrangement $B_s(g)$ of in-bonds and out-bonds with coordinates $j = 1, 2, ..., w(g)$. Each bond has a bond values $\beta_j(g) \in B_v(g)$. $B_s(g)$ accounts for the set of bonds of a generator g and $B_v(g)$ is the set of bond values of bonds in g.*

Definition 2 *A* **modality set** $\mathcal{M}$ *is a partition of $\mathcal{G}$ such that any pair of generators in a modality $M_k \in \mathcal{M}$ holds similar properties.*

Remark 2 *A* **modality** $M_k \in \mathcal{M}$ *pertaining to $\mathcal{G}$ is induced by a similarity $s \in S$, defining in which terms the generators in that modality are similar to each other. An example is the modality utensils = { bowl, cup, pan, spoon }; the modality name is suggestive of the common property among its generators.*

The concept of modality adds flexibility to the construction of structures by allowing generators of same modality to be exchangeable and take on different roles in one same configuration. Generators combine to each other through their bond structures to form *configurations* that represent complex patterns of interest. Bond structures account for ontological constraints stemming from logical, contextual, and temporal dependence, which ensure construction of configurations with coherent pattern structures. Formally,

Definition 3 *A configuration $c = \sigma(g_1, g_2, ..., g_n)$ is a connected graph structure composed by n generators $g_i \in G$ that respect the bond relation ρ.*

Remark 3 *For each closed bond $g_{i''} \downarrow g_{i''}$ in a configuration, $\rho(\beta_{j'}(g_{i'}), \beta_{j''}(g_{i''}))$ returns TRUE to indicate that the out-bond $\beta_{j'}(g_{i'})$ and in-bond $\beta_{j''}(g_{i''})$ are compatible.*

A configuration is a graph structure whose connections are identified as closed bonds by pairs of generators. Bonds from a configuration carry energy values that collectively measure the quality of the pattern structure they compose. Such energy value is formulated as the response of an *acceptor function* $\mathcal{A}(\beta_{j'}(g_{i'}), \beta_{j''}(g_{i''}))$.

Definition 4 *The* **acceptor function** $\mathcal{A}(.)$ *measures the worth value of a closed bond $g_{i'} \downarrow g_{i''}$ in terms some ontological constraints to form a certain pattern structure. $\mathcal{A}(.)$'s computation varies with the type of compatibility between generators.*

The probability of a configuration $c = \sigma(g_1, \ldots, g_n)$ is expressed as a product of terms associated with its generators $g_i \in \mathcal{G}$ and closed bonds $g_{i''} \downarrow g_{i''}$.

$$p(\sigma(.)) = \frac{\prod_{(k,k')\in\sigma}^{n} \mathcal{A}^{1/T}(\beta_j(g_i), \beta_{j'}(g_{i'}))}{Z(T)}, \quad (1)$$

where $k = \beta_j(g_i)$ and $k' = \beta_{j'}(g_{i'})$ denote bonds of generators, $Z(T)$ is the partition function, T is set to 1, and n denotes the number of generators that form an interpretation. Thus, its energy equivalent form is $E(\sigma(.)) = -\log p(\sigma(.))Z(T)$, which results in

$$E(\sigma(.)) = -\sum_{(k,k')\in\sigma} \log \mathcal{A}(\beta_j(g_i), \beta_{j'}(g_{i'})) \quad (2)$$

Finding a certain pattern structure using some chosen generator space as basis means to look into a combinatorial search space. Since it is computationally prohibitive to enumerate all possible configurations and measure the quality of each one, it is typical to consider stochastic processes for the purpose, which consists of minimizing $E(\sigma(.))$.

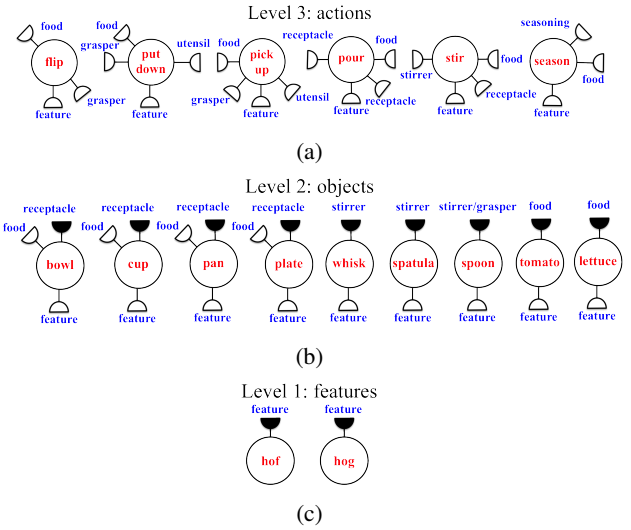


Figure 3: This illustration shows a sample of generators form the chosen generator space. They are shown with the names of ontological concepts they represent and their bond structures. (a), (b), and (c) depict generators of actions, objects and features, respectively. In-bond are shown in shaded semi-circles and out-bonds in white-filled semi- and blanket bonds are out-bonds. The bond values of these bonds are the names of modality groups.

3.1. Generator Space

The chosen generator space $\mathcal{G}$ is organized into 3 levels of abstraction. In each level resides elements with similar characteristics. Level 1 consists of generators to represent motion and shape features. Level 2 is formed by generators that represent object labels. Level 3 contains generators to represent actions. In each level generators can be further clustered into modality groups. An illustration of the generator space is shown in Figure 3 and a through description follows below.

In particular, motion features histograms of visual words (BoVW) constructed using a visual dictionary of histograms of optic flow (HOF). Shape features are given by histograms of oriented gradients. For each video segment there is a single HOF-based BoVW, formulated as a *motion feature generator*. HOGs are computed for each bounding box track available across the video segment sequence. Thus, each video segment will be associated with multiple *shape feature generators*, each corresponding to a bounding box track of some detected object.

Feature generators provide support for concepts represented by ontological generators that convey he semantics of a video interpretation.These connections are closed bonds named *support bonds*. Ontological generators of action concepts bond to ontological generators of object concepts to indicate interactions (e.g., generator *stir* bond to generator *spatula*) and ontological generators of objects connect to other generators of objects to also indicate interaction between the involved objects; these connections are called *semantic bonds*.

Generators of actions bond to other generators of actions to sustain temporal coherence between actions in interpretations of consecutive video segments.These are called *temporal bonds* and is the main contribution of this work. Both semantic and temporal bonds are considered *ontological bonds* for forming connections between ontological generators.

3.2. Bonds: Combinations of Generators

Two arbitrary generators $g_i \in G$ and $g_j \in G$ connect to each other through a single bond whose bond values conform to the bond relation $\rho(.)$. Bond values of out-bonds are names of modality groups. For example, the generator $stir$ has an out-bond with bond value $stirrer$, indicating that it can connect to any of the generators in the modality group $stirrer = \{spatula, whisk\}$. For being in the modality group $stirrer$, *spatula* has a variable number of in-bonds of bond value *stirrer*, which means that it can be connected to any other generator with an out-bond of bond value *stirrer*.

Support, semantic and temporal bonds carry bond energies that quantify how valuable for the interpretation a combination between two generators is. The bond energies are computed using the acceptor function $\mathcal{A}(\beta_{j'}(g_{i'}), \beta_{j''}(g_{i''})) = \exp(\tanh(kf(g_i, g_j)))$. For support bonds, $f(g_i, g_j)$ is the classification score output by the g_i's classifier that determine how likely is for the feature generator g_j to belong to the concept represented by g_i. Here concept classifiers are multi-class linear-SVM classification models, for both actions and objects. For semantic and temporal bonds, $f(g_i, g_j)$ come from concept co-occurrence tables.

3.3. Pattern Theoretic Video Interpretations

A long video consists of a sequence of video segments with variable lengths. Each video segment depicts a basic interaction depicting a steps of cooking a recipe. A video interpretation is a sequence of interpretations for a temporal window containing a single video segment or multiple consecutive segments. A unit of interpretation is defined by the limits of a single temporal window. Each temporal window is associated with a set of feature generators. An interpretation is generated for each temporal window by solving the optimization problem indicated in Equation 2. To this end, we implement a MCMC-based simulated annealing algorithm. This inference algorithm makes use of a local proposal function that makes small changes to vary the interpretation structure and a global proposal function

that samples interpretations from a global jump configuration built with the k best generators for each feature. Generating an interpretation is a combinatorial search problem whose elements of comibations are generators. The search space size varies with the number of features available. If a temporal window consists of one motion feature generator and two shape feature generators, then there are more than 1900 possible interpretations (6 actions $\times$ 18 objects $\times$ 18 objects); the search space size grows exponentially with the number of features.

4. Results

In this section, we analyzed the numerical performance and qualitative advantages of using the pattern theoretic approach with temporal bonds. First, we evaluated the quality of output interpretations by analyzing samples taken from the experiments. We discussed the effects of adding temporal bonds to the bond structure of generators and in which scenarios temporal bonds can lead to more interesting (desirable) interpretations. We also analyzed how critical the inclusion of temporal bonds to the model is when interpretations are based on multiple segments of videos.Then, we evaluated the performance in controlled scenarios of classification errors stemming from synthetic concept classifiers. We finalized our discussion with a comparative performance analysis on the YouCook data set when using real machine learning based concept classifiers.

For comparative analysis, we contrasted the performance profile of the proposed approach with [1]'s and a *baseline* algorithm that generates interpretations exclusively based on the best classification scores using linear-SVM classification models (*i.e.* a purely machine learning-based method). The performance metric consisted of counting the number of correct ontological generators found in the interpretation given the ground-truth's. The highest performance rate is 1 and lowest is 0. For example, the performance rate of the interpretation in Figure 6j is 0.86.

4.1. Interpretations with Temporal Bonds

Temporal bonds allow the pattern theoretic process to take into account temporal dependence information between consecutive actions, accordingly, interpretations. We found several cases in which temporal bonds helped identifying the correct actions across multiple consecutive video segment interpretations. Four of these cases of success are illustrated in Figures 4, 5 and 6.

Figures 4e and 4f show two interpretations generated by [1], each for one of two consecutive video segments. Recall that each video segment depicts some action that results in an interaction with one or multiple objects. These interpretations were generated by optimizing the energy function for each video segment, separately. [1]'s method fails to find the correct action for the first video segment, interpreting it as *pick up* instead of *put down*. Contrarily, the pattern theoretic interpretation using temporal bonds, shown in Figure 4k, successfully identifies the action *put down*, while maintaining the good semantic bonds captured by [1]. Unlike our approach, applying [1]'s method to generate an unified interpretation for the two segments at once does not produce a better interpretation than the ones generated based on separate inference of each segment. In fact, this interpretation, illustrated in Figure 4i, introduced more errors, confusing *pick up* by *put down* in both segments.

Another example of success by our approach taken from the experiments is depicted in Figure 4l, where the action *stir* was correctly inferred for describing the interaction occurring in the second segment. The approach in [1] failed to determine the action *stir*, instead inferring *pick up* in the single segment scenario and *season* in multiple segment one.

In both illustrated cases, not only was our approach able to generate improved interpretations through the addition of temporal bonds but it also preserved relevant generators of objects and bonds correctly identified in the single-segment based inference from [1]'s approach. This same effect has been observed for larger set of segments, as illustrated in Figures 5 and 6. We discussed these cases in the next section to point the benefits of generating interpretations for larger temporal windows containing multiple video segments. Since more degrees of freedom are available when considering temporal bonds, this permitted our approach to explore other possibilities of interpretations with more confidence than when they were not present.

4.2. Interpretations for Multiple Segments

We identified several cases in which interpretations generated for multiple-segment temporal windows using our approach helped determine the correct interpretations of actions for video segments that are misinterpreted by [1]'s approach. Temporal bonds allow our approach to not only search for coherent local interactions but also naturally focus on identifying the correct temporal ordering of actions in adjacent video segments.

For instance in Figures 5 and 6, our approach's interpretation (Figures 5j and 6j) was able to preserve the correctly detected objects found in interpretations generated by [1] (Figures 5e-5i and Figures 6e-6i) while fixing the action interpretations of consecutive video segments. More interestingly, the case depicted in Figure 5 shows that our approach's interpretation leveraged the confidence of the action interpretation in the third segment, *put down*, to propagate multiple corrections in the two past segments and the video segment ahead; the action interpreted sequence was *put down* $\rightarrow$ *pick up* $\rightarrow$ *put down* $\rightarrow$ *pick up* (Figure 5j). Nonetheless, the same effect was not observed when using [1]'s method, which produced the sequence *pick up* $\rightarrow$ *put down* $\rightarrow$ *put down* $\rightarrow$ *put down* (Figure 5i).

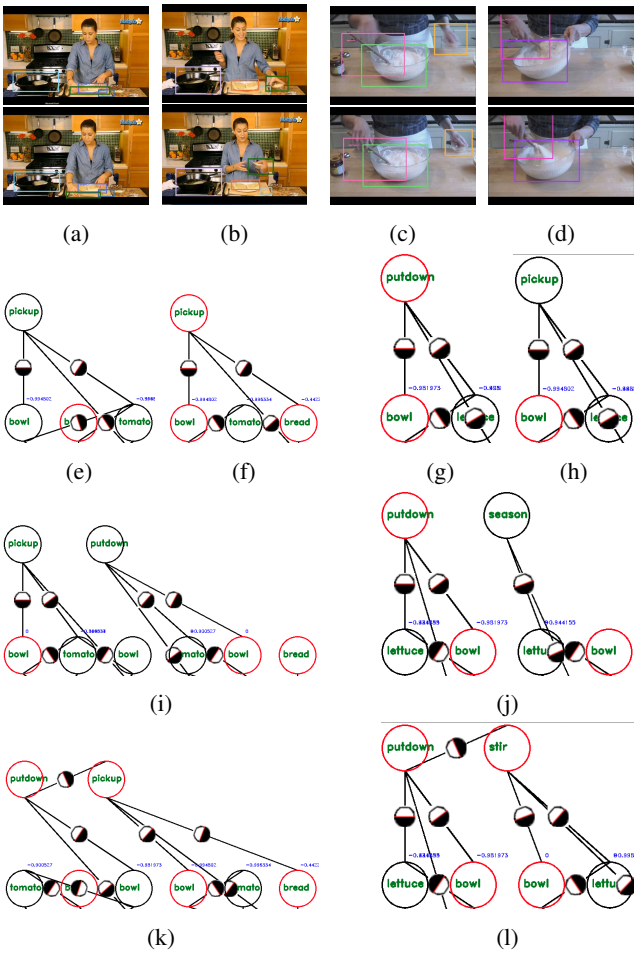


Figure 4: (a) and (b) illustrate two consecutive video segments describing steps for making french toast. (c) and (d) illustrate two consecutive video segments describing steps for making dough. The pairs (e)-(f) and (g)-(h) show the corresponding interpretations by [1] based on single-segment windows, while (i) and (j) are derived from two-segment windows. (k) and (l) present corresponding interpretations generated by our approach.

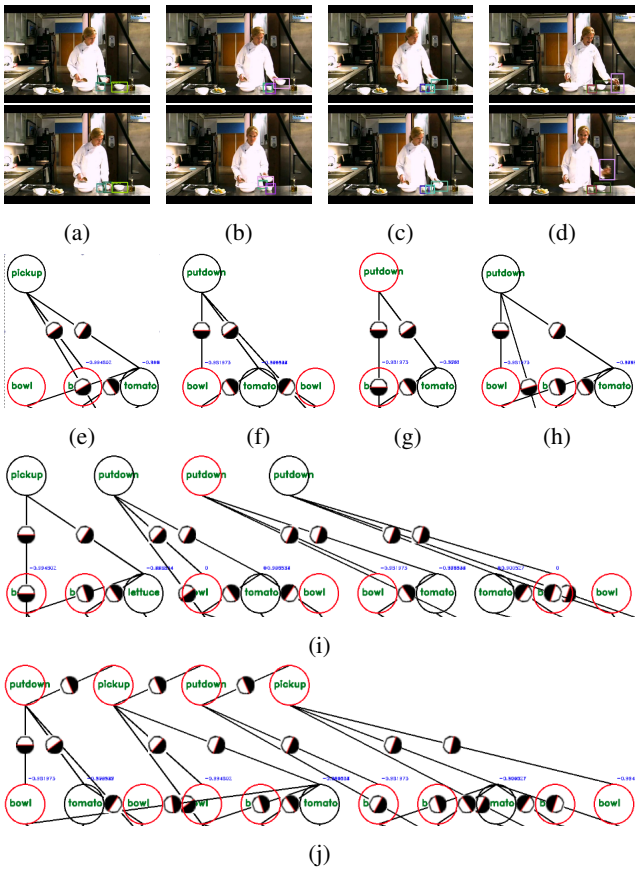


Figure 5: (a)-(d) illustrate four consecutive video segments describing steps for making salad. (e)-(h) show interpretations by [1] based on single-segment windows, and (i) for the four segments at once. (j) depicts the interpretation by our approach.

In this work, temporal bonds were only explored at the level of actions under the assumption that coherence in determining the participating objects in interactions is mostly dependent on correctly identifying the action being performed. The focus then revolves around finding temporal coherent interpretations in terms of actions. Identifying the correct sequence of actions indirectly influences on the quality of the overall sequence of interpretations.

4.3. Experiments with Synthetic Classifiers

Classification scores help measure the quality of an interpretation; therefore, they are essential for ascertaining the global optimal interpretation. We set up two controlled scenarios of degradation stemming from concept classifiers in order to evaluate our approach's tolerance to classification errors. First, we varied the classification error rates of the classifiers from 10% to 60%. In these cases, the classification score ranks of the correct labels for the affected features are the second best. Then, we fixed the classification error rate to 50% and vary the score rank of the feature's correct labels from 2 to 5.

Figure 7 shows the performance profile of the approaches for increasing rates of classification error. In Figure 7a, where interpretations are generated for each individual video segment, our approach and [1]'s were superior to the baseline's, but only for high rates of classification error (>20%). In this same case, our approach and [1]'s had comparable performance rates, since no temporal data could be explored. Our approach produced performance improvement increase of more than 7% over [1]'s for larger temporal windows, multiple video segments at once (Figure 7b).

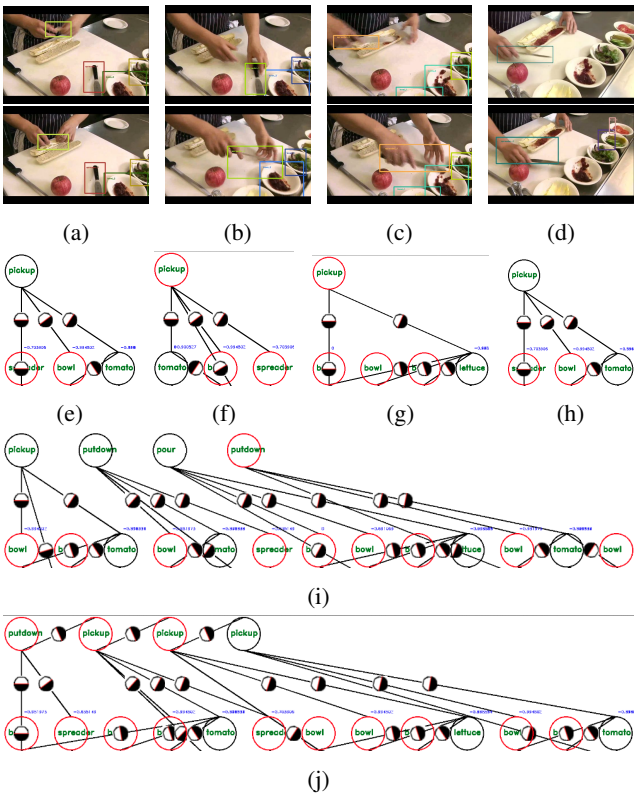


Figure 6: (a)-(d) illustrate four consecutive video segments describing steps for making sandwich. (e)-(h) show interpretations by [1] based on single-segment windows, and (i) for the four segments at once. (j) present the interpretation by our approach.

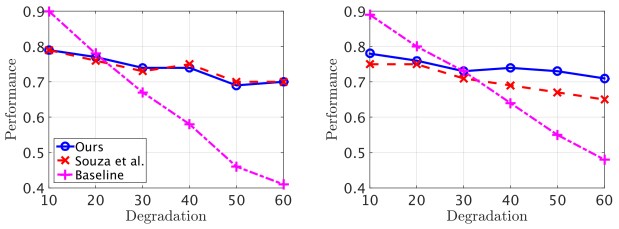
In summary, if the concept classifiers are not sufficiently good to be used alone, these results indicate that ontology-based approaches like ours and [1]'s are imperative in order to achieve reasonably sufficient performance.

Figure 8 shows the performance profiles in which approximately 50% of the features had their correct labels's classification scores as the kth best classification scores, where k varies from 2 to 5. Overall, Figures 8a-8b show that our approach was consistently capable of correcting the feature labeling, even when the feature correct label had the fifth best classification score. Figure 8b shows that for multiple-segment based inference our approach is consistently superior to [1]'s, with up to 12% increase. This suggests that under uncertain scenarios, our approach would be more advantageous because it improves performance by generating video interpretations based on multiple segments.

4.4. Experiments with Real Classifiers

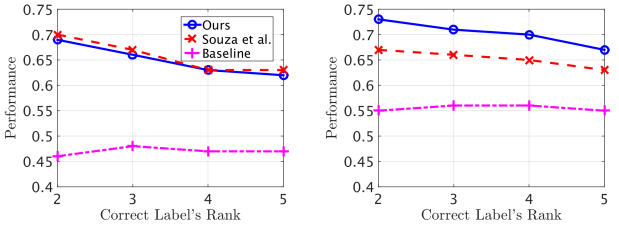
Overall our approach improved the total average interpretation performance by approximately 5 times the baseline's and by ∼30% over [1]'s when no bond types were given preference over others (Table 1). Three types of bonds contribute to measure the quality of an interpretation (see Section 3.2). Figure 9 shows how the overall average video interpretation performance varied as certain types of bonds were given more participation weight than others. In all cases, overweighting support bonds dropped the interpretation performance rate to the baseline's, which was low because of the weak concept classifiers. More weight on the participation of temporal bonds was sufficient to achieve higher performance rates for all multiple-segment cases (Figure 9b). This emphasizes our assumption that correct action interpretations should naturally lead to correct identification of the true involved objects. Overweighting semantic bonds was influential mostly in the single-segment case (Figure 9a), where temporal bonds were not relevant.

We also observed the average performance when considering the top k interpretations for describing each video. Figure 10a shows that our approach and [1]'s had comparable performance for the single-segment case when no temporal information is available. When generating multiple interpretations for temporal windows containing multiple consecutive segments, our approach provided the best overall performance. In comparison to [1]'s original idea of



(a) 1-segment temporal window (b) 4-segment temporal window

Figure 7: Interpretation performance profiles for varying scenarios of classification error rates, ranging from 10% to 60%.



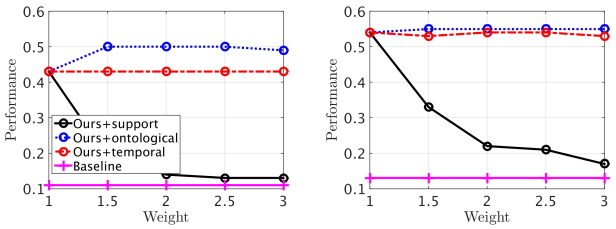
(a) 1-segment temporal window (b) 4-segment temporal window

Figure 8: Interpretation performance profiles showing the tolerance of comparative approaches for decreasing rank of classification scores of the correct label for about 50% of the video features.

Table 1: Comparison of overall average performance of video interpretations for increasing temporal window sizes (# video segments).

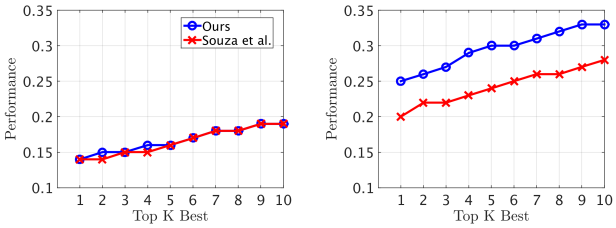
Method	1	2	3	4
Ours	**0.40**	**0.52**	**0.50**	**0.52**
Souza et al.	**0.41**	0.44	0.41	0.46
Baseline	0.11	0.12	0.13	0.13

analyzing individual segments, it nearly doubling the performance, with improvements ranging from about 67% to 73%, depending on the number k (compare Figure 10b with Figure 10b).



(a) 1-segment temporal window (b) 4-segment temporal window

Figure 9: Video interpretations performance profile of our approach when different bond types have more participation than others.



(a) 1-segment temporal window (b) 4-segment temporal window

Figure 10: Video performance interpretation when considering the top k best interpretations for describing each video.

5. Time Complexity and Running Time

Our global and local proposal functions explore the bond-constrained space of feasible solutions in an efficient manner. The time computational complexity was worked out to be $O(k * m_c * m_o + k(n_f + m_o * m_o))$, where k is the total number of sampling iterations, m_c is the number bonds from a candidate generator for replacement, m_o is the total number of open bonds in a current interpretation and n_f is the number of feature generators. For all the experiments we fixed $k = 3000$. The running time grows linearly with the number of feature generators n_f, consistent with our analysis (see Table 2).

Table 2: Average CPU+I/O time of videos per number of feature generators n_f. Machine spec: 4 16-core 2.3 GHz CPUs (AMD Opteron 6376), 16 16GB RAM units.

n_f	30	33	34	36	37	44	46	48
sec	279	342	382	399	375	477	513	550

6. Conclusion

In this paper we advanced the adaption of Grenander's pattern theory originally proposed in [1] by introducing a bond structure that captures temporal information. This allowed us to generate temporally coherent semantic interpretations of videos. Similar to [1], the basic units of interest (i.e., actions and objects) are denoted by generators that combine to each other to form graphical structures, which represent video interpretations. The quality of an interpretation is governed by the energies of its bonds. These bond energies are defined using classification scores and frequency tables of concept co-occurrences, which help define and seek optimal configurations. While previous applications have been restricted to analyzing short videos containing isolated actions, we have extended this idea to longer videos using additional bond structures, which allow interactions between actions that are adjacent in time. The aforementioned experiments, involving more than 700 video segments from the YouCook data set, demonstrated the power of adding action temporal bonds in the configurations. Not only did we improve the performance in detection of generators but we also improved the overall scene interpretations. In addition, the our approach was more robust to degradation in feature-level classification performance than its counterparts.

The use of additional bond structures clearly helped interpret more complex scenes and allowed for enhanced inferences. In view of the flexibility of this pattern theoretic framework in representing complex systems, in future work, we plan to include additional (types of) generators and bond structures. In this future extension, configurations that represent interpretations of small video segments can be turned into composite generators, naturally augmenting the representation hierarchical system, which can be used to help understand even longer videos depicting more complex activities.

Acknowledgment

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

References

[1] F. D. M. de Souza, S. Sarkar, A. Srivastava, and J. Su, "Pattern theory-based interpretation of activities," in *International Conference on Pattern Recognition (ICPR)*, 2014. 1, 2, 3, 5, 6, 7, 8

[2] U. Grenander and M. I. Miller, *Pattern theory: from representation to inference*, 2007, vol. 1. 1

[3] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2847–2854. 2

[4] A. Fathi and J. M. Rehg, "Modeling actions through state changes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2579–2586. 2

[5] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic, "Weakly supervised action labeling in videos under ordering constraints," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 628–643. 2

[6] C. C. Tan, Y.-G. Jiang, and C.-W. Ngo, "Towards textually describing complex video contents with audio-visual concept classifiers," in *ACM International Conference on Multimedia (MM)*, 2011, pp. 655–658. 2

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8. 2

[8] S. Bhattacharya, M. M. Kalayeh, R. Sukthankar, and M. Shah, "Recognition of complex events: Exploiting temporal dynamics between underlying concepts," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[9] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1250–1257. 2

[10] M. U. G. Khan, L. Zhang, and Y. Gotoh, "Towards coherent natural language description of video streams," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 664–671. 2

[11] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, U. Lowell, and S. Guadarrama, "Generating natural-language video descriptions using text-mined knowledge," *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 10, 2013. 2

[12] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2634–2641. 2

[13] B. Laxton, J. Lim, and D. Kriegman, "Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8. 2

[14] E. Swears, A. Hoogs, Q. Ji, and K. Boyer, "Complex activity recognition using granger constrained dbn (gcdbn) in sports and surveillance video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3

[15] M. R. Amer and S. Todorovic, "Sum-product networks for modeling activities with stochastic structure," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1314–1321. 2

[16] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu, "Cost-sensitive top-down/bottom-up inference for multiscale activity recognition," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 187–200. 2

[17] M. S. Ryoo and J. K. Aggarwal, "Semantic representation and recognition of continued and recursive human activities," *International Journal of Computer Vision (IJCV)*, vol. 82, no. 1, pp. 1–24, 2009. 2

[18] N. Vo and A. Bobick, "From stochastic grammar to bayes network: Probabilistic parsing of complex activity," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[19] K. Hilde, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3

[20] H. Pirsiavash and D. Ramanan, "Parsing videos of actions with segmental grammars," 2014. 2

[21] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4d human-object interactions for event and object recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 3272–3279. 2

[22] M. R. Amer, S. Todorovic, A. Fern, and S.-C. Zhu, "Monte carlo tree search for scheduling activity recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 1353–1360. 2

[23] M. Albanese, R. Chellappa, V. Moscato, A. Picariello, V. Subrahmanian, P. Turaga, and O. Udrea, "A constrained probabilistic petri net framework for human activity detection in video," *IEEE Transactions on Multimedia*, vol. 10, no. 6, pp. 982–996, 2008. 2

[24] T. Lan, L. Sigal, and G. Mori, "Social roles in hierarchical models for human activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1354–1361. 2

[25] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1226–1233. 2