Towards a Knowledge-based approach for Generating Video Descriptions

Sathyanarayanan Aakur, Fillipe DM de Souza, Sudeep Sarkar
Department of Computer Science and Engineering
University of South Florida
Tampa, Florida, USA
saakur@mail.usf.edu, fillipe@mail.usf.edu, sarkar@usf.edu

***Abstract*—Existent video description approaches advocated in the literature rely on capturing the semantic relationships among concepts and visual features from training data specific to various datasets. Naturally, their success at generalizing the video descriptions for the domain is closely dependent on the availability, representativeness, size and annotation quality of the training data. Common issues are overfitting, the amount of training data and computational time required for the model. To overcome these issues, we propose to alleviate the learning of semantic knowledge from domain-specific datasets by leveraging general human knowledge sources such as ConceptNet. We propose the use of ConceptNet as the source of knowledge for generating video descriptions using Grenander's pattern theory formalism. Instead of relying on training data to estimate semantic compatibility of two concepts, we use weights in the ConceptNet that determines the degree of validity of the assertion between two concepts based on the knowledge sources. We test and compare this idea on the task of generating semantically coherent descriptions for videos from the Breakfast Actions and Carnegie Mellon's Multimodal activities dataset. In comparison with other approaches, the proposed method achieves comparable accuracy against state-of-the-art methods based on HMMs and CFGs and generate semantically coherent descriptions even when presented with inconsistent action and object labels. We are also able to show that the proposed approach performs comparably with models trained on domain-specific data.**

***Keywords*-Activity Recognition; Reduced Training needs; Pattern Theory; ConceptNet; Semantic coherence**

I. Introduction

Enabled by the availability of large amount of annotated image and video data sets and technological advances in powerful parallel hardware, object detection and action recognition [1], [2] and image captioning [3], [4] have witnessed major progress in performance. These ideas have also been successfully extended to activity recognition in videos, in which models are trained to distinguish one category from a set of predefined human activities from another using discriminant models [5]–[8], graphical models [9]–[11] and deep learning models [12]–[18].

The dominant approach to solving the problem of human activity descriptions in videos formulates it as a video classification problem, in which discriminant models are the most commonly used methods. These approaches are simple and fast during operation, but can suffer from data annotation errors and offer no flexibility in handling

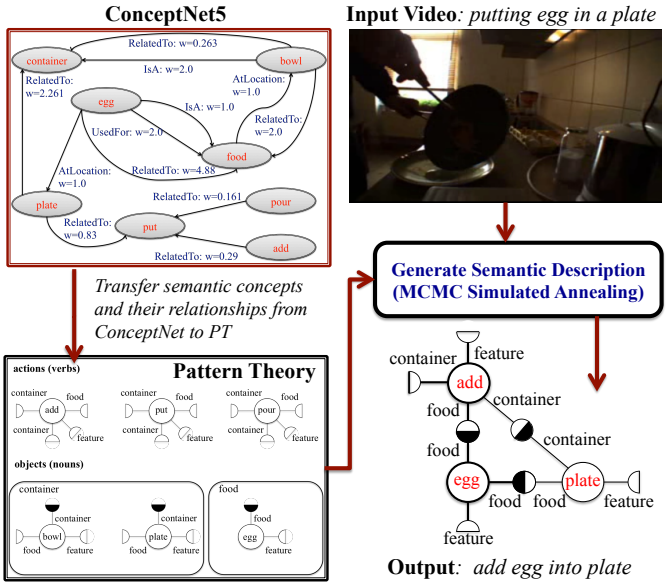


Figure 1: We use ConceptNet5 (top left) as the source of prior knowledge about the semantic relationship of concepts such as actions and objects. This alleviates the computational burden reserved for training the model.

variations that may not be captured in the training data. These methods simply enforce labels to videos based on feature observations and models learned from a specific dataset. Graphical approaches [9]–[11] attempt to explicitly model the semantic relationships that characterize human activities with context-free grammars [19], inductive logic programming [20], Markov networks [21], AndOr graphs [22] and Petri-Nets [23]. Those probabilistic models are typically prescribed as a product of conditional probability distributions. The parameter space of these models can be very large, which depends on the number of predefined random variables, their structural dependencies and the number of states of each variable. Such models are also likely to overfit and not generalize well as their performance are largely dependent on the knowledge implicitly captured in the data set available.

Deep learning models rely on multiple levels of training in order to optimize their parameter space to handle variations in scale, illumination, and pose. In addition to

training the model for identifying the basic elements of activities such as actions and objects, they often also rely on training data to capture the semantic relations between the concepts and as such constitute a significant portion of the training process. Approaches, such as those proposed in [12], [14]–[16], acquire a strong ontology specific to the dataset through the vocabulary of the training split. This allows the models to handle variations within the test and validation sets. The training sets are used at multiple levels in deep learning models such as for extracting semantic concepts from videos, learning the sentence structures and integrating both such knowledge sources to generate descriptions for videos. While the computational burden in the training step, has been alleviated with the recent surge of powerful parallel computational architectures, the ability of such systems to accurately describe the video activity is largely dependent on the availability, size and quality of training data. In order to overcome these limitations, we advocate the use of semantic knowledge available in a general human knowledge database known as ConceptNet in place of learning it from a specific data set whose knowledge is constrained to a specific domain.

In this paper, we build upon the video description framework based on pattern theory [9] to leverage semantic knowledge captured by a general human knowledge base known as ConceptNet [24]. The requirement for a large, finely annotated dataset could be significantly alleviated by leveraging the inherent knowledge from ConceptNet into the flexibility provided by Grenander's Pattern Theory [25] framework. We show that this pattern theory model when integrated with ConceptNet can produce video descriptions whose quality are comparable or better than those reported by frameworks that require domain-specific training data.

The contributions of this paper are three fold: i) We eliminate the need for domain-specific training, which as a result ii) significantly reduces the training requirements and provides flexibility in the model to adapt to variations in the expressiveness of actions and objects within the same context, while maintaining semantic coherence of the descriptions, and iii) we show that integrating ConceptNet into the pattern theory framework can produce results that are comparable with those reported by models learned from domain-specific training data.

II. Generating Video Descriptions using ConceptNet

In this section, we describe the proposed method for constructing semantic descriptions of videos using a general human knowledge base in ConceptNet5 [26]. We start with discussion about ConceptNet and then, we describe how the pattern theory constructs are derived from the ConceptNet semantic network. Following which, we explain how the semantic relationship are quantified using the ConceptNet framework and incorporated in the process of generating the video descriptions.

A. ConceptNet

ConceptNet is a database of knowledge proposed by Liu and Singh [24] to provide a common medium for storing and retrieving cross-domain semantic information from general human knowledge. This database supports both structured and commonsense knowledge as expressed by humans in natural language. Technically, it encodes and expresses knowledge using a graph, in which the nodes represent concepts and edges are assertions about the relationships involving concepts. Figure 2 illustrates these ideas; for example, the edge between nodes `egg` and `plate` represents an assertion with the relation `AtLocation` to indicate that eggs can be placed or found in plates. Notice also that each relation has a weight that determines the degree of validity of the assertion given the sources.

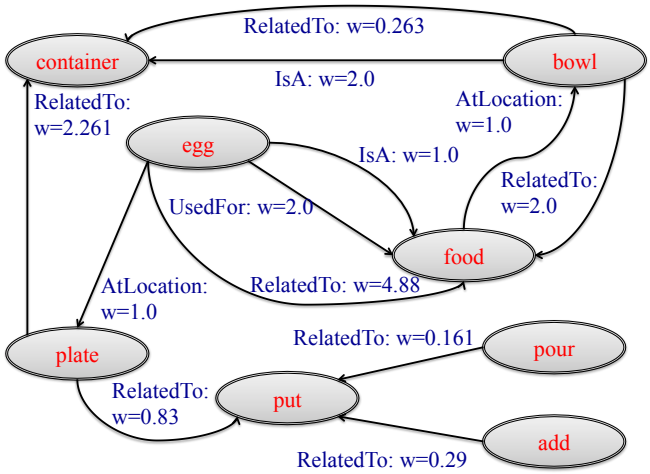


Figure 2: ConceptNet is semantic network of commonsense knowledge. Here is an example of how the semantic relationships of concepts are stored in the ConceptNet hypergraph.

As illustrated in Figure 2, concepts can be words or phrases and they are mined from multiple sources of knowledge available in the Internet, such as WordNet [27] , Wiktionary and Wikipedia to name a few. Assertions indicate semantic relations between concepts such as `UsedFor`, `RelatedTo`, `CreatedBy`, `IsA`, `PartOf`, to name a few. Assertions are validated by sources and assigned weights that measure the strength of their validity. Positive values indicate that an assertion is true, and negative determines the opposite of what is asserted.

The practicality of ConceptNet stems from providing a single, accessible framework that collates contextual information from sources with varying levels of granularity and different languages. It can be used in a variety of purposes such as natural language processing [28], common sense

reasoning [29], [30] and video retrieval [31], [32], to name a few.

In the proposed approach, we aim to exploit these semantic relationships encoded within the ConceptNet framework to guide the MCMC based inference process. The interpretations of the video activities are enforced by the inference process to provide a semantically coherent description that can sufficiently explain the presence of the different concepts (actions and objects) within the context of the video activity. This approach allows the model to generate descriptions that are able to establish semantic basis for the presence of these concepts within the video activity even when presented with inconsistent component labels from the machine learning classifiers. Additionally, the presence of the ConceptNet-based knowledge empowers the model with the ability to differentiate among the different candidate labels presented for each component (action and object) and infer the most semantically aligned interpretation of the activity. Hence, we also reduce the dependence on the accuracy of the machine learning classifiers which in turn, reduces the training requirements.

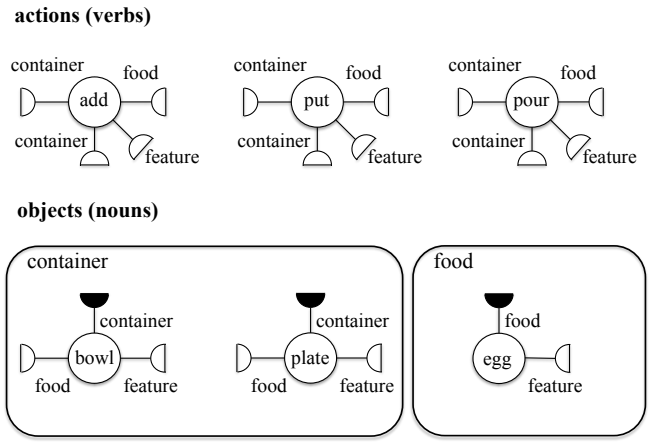


Figure 3: A sample of the generator space. The generators represents actions (on the top) and objects (at the bottom) that are commonly used to describe scenes in cooking videos. Each generators has a local bond structure with values. These values are drawn from a set of modalities: e.g., food and container.

B. Video Description Framework

In this section, we introduce a video description framework based on Grenander's Pattern Theory constructs [25] using ConceptNet as the main data source.

1) Semantic Concepts and their Relationships: Semantic concepts such as nouns (objects) and verbs (actions) are represented as *generators* g with a collection of generators forming a *generator space* G. A simplified illustration of the generator space is shown in Figure 3. Each generator has a local *bond structure* that allows them to connect with each other. The bond structure is representative of the semantic relationship between generators, with each *bond* $\beta_j(g)$ carrying some semantic relationship inherent to the noun or verb. For example, the generator that represents the action `add` has out-bonds with bond values `container` and `food`. These bond values are names of *modality groups* that partition the generator space into semantically homogeneous groups of generators; for instance, `food` is a modality grouping generators `egg`, `fruit`, `coffee` and `milk`. The modality allows the framework to connect generators that are semantically related to it by qualifying bonds with semantic relationships.

The type of relation that defines the edge (or assertion) from ConceptNet determines what the endpoint nodes are in the pattern theory algebra. For example, the node `bowl` has an incoming edge from node `food` that is marked with the relation `AtLocation`, which means that `bowl` becomes a generator g that has an out bond with bond value `food`.

2) Quantifying Semantic Relationships of Concepts: The bond between generators are quantified using ConceptNet with the strength of the semantic relationships between generators being quantified by the bond energy function:

$$a(\beta'(g_i), \beta''(g_j)) = q(g_i, g_j) \tanh(f(g_i, g_j)). \qquad (1)$$

where $f(.)$ is a weight returned by ConceptNet that determines the degree of validity of the assertion of relationship between two concepts represented by the generators g_i and g_j through their respective bonds β' and β''. The $\tanh$ function normalizes the score output by $f(.)$ to range from -1 to 1. $q(g_i, g_j)$ weights the score output by the $tanh$ function according to the bond connection type (e.g., semantic or support) β' and β'' forms.

3) Grounding Semantic Concepts with Data: Apart from connections between generators of semantic concepts, there exists a different type of generator that represents observations from video data. These are called feature generators and they represent data evidence for semantic concepts and contribute to the energy of the final configuration. When semantic concepts are proposed to be part of a video description, their participation is supported by connecting them with feature generators. These connections are called *support bonds* and are quantified using confidence scores from classification models such as SVM in this paper.

4) Generating Descriptions: Generators combine together through their local bond structures to form *configurations* c. Configurations represent video descriptions of activities and the configuration with the lowest energy $E(c)$ is chosen to be the best semantic description of a video. The energy $E(c)$ of a configuration c is the sum of the bond energies formed by the bond connections that combine the

generators in the configuration, as described in Equation 2.

$$E(c) = -\sum_{(\beta',\beta'')\in c} a(\beta'(g_i), \beta''(g_j)). \quad (2)$$

where g_i and g_j represent generators in the configuration. These generators within a configuration c represent its semantic content and hence gives rise to the description of a given video. And $a(.)$ represents the bond energies as captured by Equation 1. Note that there might be multiple bond energies between any two generators or concepts, depending on the semantic relation captured.

Searching for the best semantic description of a video involves minimizing the energy cost function $E(c)$. The space of solution spanned by the generator space is much larger than the space of solution for a typical Markov Random Field (MRF). In particular, both the number of generators and structures can be variable, contrary to MRF models, whose structure and number of random variables are fixed. A feasible optimization solution for such exponentially large space, is to use a sampling strategy. We follow the work in [9] and employ a Markov Chain Monte Carlo (MCMC) based simulated annealing process. This process relies on both a local and global proposal function. The local proposal function replaces generators in a configuration based on the contribution to the total energy of the configuration and the global proposal makes structural changes to the configuration that are reflected as jumps from a subspace to another.

III. Experimental Setup

A. Datasets

The Carnegie Mellon University Multimodal Activity dataset (CMU) contains multimodal measures of human activities such as cooking and food preparation. The dataset contains five different recipes: brownies, pizza, sandwich, salad and scrambled eggs. Spriggs *et al.* [33] generated the ground truth for some videos and recipes. The experiments were performed on the brownie recipe videos for performance comparison with [34]. In total, there are 13 labeled videos in the brownie dataset with each of the videos split into smaller units consisting of individual activities, of which we take 233 segments as the test set from the videos numbered 9, 12, 16, 20, 22 and 24. In brownie recipe dataset, we can find 12 action labels, namely, stir, crack, spray, twist, etc. and 14 object labels, including baking pan, bowl, brownie box, oil, fridge, etc. The possible combinations of actions and objects is a much larger, typically in the order of 1000's. Each action could involve more than one object.

The Breakfast Actions dataset consists of more than 1000 recipe videos, consisting of different scenarios with a combination of 10 recipes, 52 subjects and differing viewpoints captured from up to 5 cameras, which provides differing qualities of videos with varying amounts of clutter and occlusion. The units of description are temporal video segments of these videos, given by the video annotation provided along with the dataset. These units are recipe steps, for instance, pour coffee, peel fruit and fry egg; thus, we consider more than 5000 units of descriptions for evaluation purpose. The possible activity descriptions is large as they span across a combination of more than 10 action categories such as peel, crack, squeeze, fry, butter, stir and smear and more than 25 object categories such as plate, fruit, cereals, egg, orange, bun, knife, coffee and glass, respectively, to name a few.

B. Feature Extraction and Training

In order to represent individual concepts of actions and objects in the Breakfast Actions dataset, we used handcrafted features Histogram of Optical Flow (HOF) for actions and Histogram of Oriented Gradients for objects. Features for action recognition were extracted by computing dense optic flow frames from three temporally sequential segments - each representing the start, development and end of the action sequence respectively. A histogram of optic flow (weighted by magnitude) is then constructed for these temporal segments to characterize the integral stages of the action. The composite feature for action recognition is then composed by ordered concatenation of the individual HOFs. Features for object detection were generated using the well-known HOG features composed within the bounding boxes for generating object candidate labels. We also explored the impact of leveraging multimodal features on the CMU dataset along with deep features used within deep learning models such as Convolutional Neural Networks (CNN) for capturing the feature descriptions for objects (CNN) and actions (CNN-Flow) as well as auditory features such as bag-of-audio words (BoAW) and spectrogram features as auditory feature descriptors.

The only training performed in this model is the one required to train the classification models for identifying actions and objects using linear support vector machines (SVM) with the aforementioned feature descriptors. The official splits provided in the datasets were used during this training process.

C. Performance Metric

In order to measure the performance of the proposed approach, we use two metrics - namely F-measure and precision. The F-measure takes into account both precision and recall and hence provides a definitive measure of the model's ability to incorporate semantic knowledge into its descriptions. Precision is given by the ratio of the number of correctly detected generators to the total number of generators in the groundtruth interpretation. Recall is given by the ration of correctly detected generators to the total number of generators in the generated interpretation.

IV. RESULTS

In this section, we analyze the effectiveness of taking an approach of integrating ConceptNet into the pattern theory framework to generate semantic descriptions for video activity segments. First we evaluate the performance of the approach on the Breakfast Actions and CMU datasets. We then show the impact of reduced training on the original pattern theory model proposed by Souza et al [9] and contrast the performance of the ConceptNet based model against varying levels of training on the original pattern theory model to demonstrate the need for training in domain-specific approaches.

A. Semantic Description Performance

As shown in Table I, ConceptNet enables the pattern theory framework to generate semantic descriptions based on its vast knowledge base and perform comparably with other approaches that use domain-specific training. Note that precision is comparable to classification accuracy reported in [11].

Approach	Precision
DM [11]	6.10%
HMM [11]	14.90%
CFG + HMM [11]	31.8%
Souza et al (Top 10) [9]	38.60%
PT + CN (Top 10)	33.88%

Table I: Comparison of Performance with other methods. We report the precision so that we can directly compare against the performances reported by Hilde et al [11] and Souza et al [9]. PT + CN represents the ConceptNet integrated pattern theory approach proposed in this paper. Top 10 means that we consider the best 10 descriptions generated by the approach.

We can see that the ConceptNet based approach improves the performance of the HMM-based approach reported by Hilde [11] and performs comparably with the Context Free Grammar model and the original pattern theory model ([9]). It is important to note that the performance of the proposed approach is remarkable considering that the model is neither trained specifically for the kitchen domain nor on the dataset itself (other than for obtaining action/object labels) to generate the semantic descriptions. The other methods reported here use a five-fold cross-validation. Another important characteristic to note is that the Context Free Grammar method makes use of temporal information such as states and transitions between states to build the final descriptions which is not the case with our approach.

We also explored the impact of using multimodal features as the basis for generating action and object labels on the CMU Multimodal Activities dataset. The multimodal nature of the CMU dataset allows us to extract various features such as bag-of-audio words (BoAW) and spectrogram features as auditory feature descriptors (Spect) whereas the success of Convolutional Neural Networks (CNN) inspired the use of CNN features for describing the actions (CNN Flow) and objects (CNN) of concepts in the video. From Table II,

Features	PT [34]	PT + CN
Hog, Hof	0.657	0.676
CNN, CNNFlow	0.699	0.677
Hog, Spect	0.661	0.671
CNN, Spect	0.690	0.633
Hog, Hof, Spect	0.662	0.645
CNN, CNNFlow, Spect	0.699	0.646

Table II: Performance Comparison using deep features on the CMU multimodal activities dataset. We report the F-scores in order to represent the performance of each model. CNN and CNNFlow refer to the use of CNN feature descriptors for objects and actions respectively whereas Spect refers to the use of auditory features.

we can see that multimodal features have a positive impact on the performance of the ConceptNet based approach on the dataset and narrows the gap between using a domain-specific ontology and a common sense knowledge source in ConceptNet. An interesting observation is that the auditory features can be considered to be representative of actions and hence the use of auditory features alone for labeling actions have shown to achieve high levels of performance comparable to the use of visual features in action recognition as seen from rows 3 and 4 of Table II.

B. Impact of Reduced Training on Baseline

In this section, we explore the effect of reduced training on the original pattern theory model (the baseline algorithm) and compare the performance with the ConceptNet based approach, which does not involve any domain-specific training. The original approach used a domain ontology constructed from the fine annotations present within the dataset and then utilize it in multiple aspects of the algorithm to impose restrictions to arrive at a semantically correct description. This procedure constitutes a significant portion of the training process in the original pattern theory approach in addition to the training required for identifying the generators (actions and objects).

We analyzed the impact of reduced training on the model and compared the performance against our method by choosing a random sampling of varying amounts of training data. The significance of the domain ontology constructed during training can be seen in Figure 4. Random sampling was chosen when restricting the amount of training due to the fact that the various activities that involved different combinations of action and object labels, were uniformly distributed across the dataset.

C. Quality of ConceptNet-based Descriptions

An interesting observation from the results was that the descriptions obtained from the ConceptNet based approach

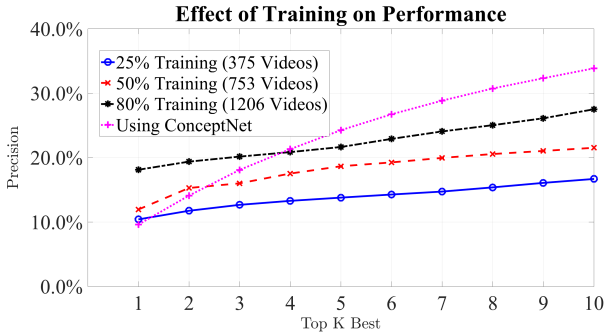


Figure 4: Comparison of performance with varying levels of training on the original pattern theory model [9].

are more semantically aligned with the ground truth and hence performs better on complex activities as seen in Figure 5. The ConceptNet based approach provides meaningful semantic generalization to the descriptions. Some descriptions are composed by labels of actions and objects that do not perfectly match the ground truth's but they are semantically closer – the error is solely from the point of view of the ground truth learned from the data set. For example, a video with ground truth description "put egg on plate" was interpreted by the ConceptNet based approach as "add egg to plate" (see Figure 5(c)). In that case, the proposed action label was "add" rather than "put", which in this context conveys the same semantics. This is a natural exchange of concepts whose semantic relationship is present in the ConceptNet semantic network, as shown in Figure 2, discussed in Section II-A. These are actual examples of how ConceptNet allows us to better generalize and enforce semantically coherent descriptions, even when not strictly matching with what would have been available in the training data.

Another interesting aspect to note is that the depth of semantic knowledge within the ConceptNet and its effect on generating descriptions for video activities allow the model to generate semantically coherent descriptions as highlighted by scenarios where the ConceptNet based approach was able to handle variation in the proposal of both action and label generators. For example, consider a video segment with the ground truth as "put fruit in a bowl" as shown in Figure 5(d). The generators proposed varied dramatically from the ground truth, but the ConceptNet based approach was able to adapt to the variation and arrive at a semantically coherent description of "put tea in a bowl" as opposed to an incoherent description of "stir tea, dough" as the ontology restriction was not able to semantically relate the presence of the generator "dough".

Approaches with domain-specific training data are restricted to the vocabulary of the annotations present in the training samples; thus, not achieving a desirable semantic generalization in its descriptions. This is strongly associated with a possible overfitting of the parameters learned from the dataset to quantify the semantic relationships of actions and objects. The use of a cross-domain knowledge base through ConceptNet underpins the ability that the model to adapt and generalize to variations in the action or object labels used to generate semantically relevant descriptions of videos. That is, although the descriptions possess variations from the ground truth, they still possess similar semantic content and hence avoid issues related to overfitting on domain-specific data.

V. Conclusion

We proposed pattern theory based approach which integrated a common sense knowledge base such as ConceptNet in order to generate semantic descriptions for video activities. We showed that the performance of the ConceptNet based approach is comparable with other approaches and possesses a significantly reduced need for training. Our approach achieved competitive results with other approaches that required high amounts of training data and improved the quality of descriptions and is superior to these approaches when the domain ontology is restricted by reducing the amount of available training data. We also show that leveraging deep features such as CNN features and auditory features such as spectrograms further narrowed the gap between domain-specific models and our approach.

References

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[2] Y. Zhang, K. Sohn, R. Villegas, G. Pan, and H. Lee, "Improving object detection with deep convolutional networks via bayesian optimization and structured prediction," in *CVPR*, 2015, pp. 249–258.

[3] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

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

[6] A. Vahdat, K. Cannons, G. Mori, S. Oh, and I. Kim, "Compositional models for video event detection: A multiple kernel learning latent variable approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1185–1192.

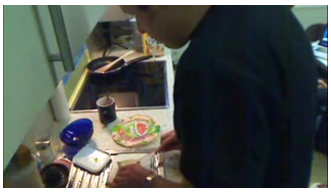
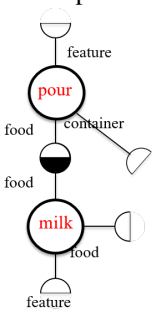
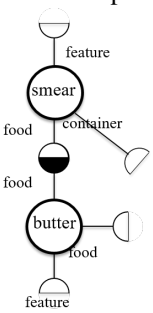
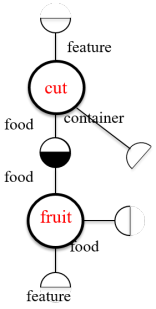
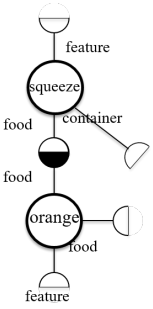
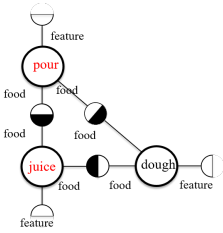
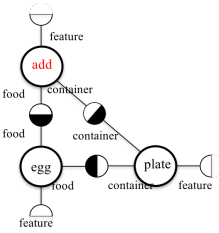
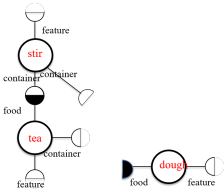
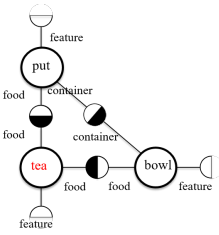
Ground truth	Description by PT + domain-specific training	Description by PT + ConceptNet
		
Smear butter	*Pour milk*	*Smear butter*
(a)		
		
Squeeze orange	*Cut fruit*	*Squeeze orange*
(b)		
		
Put egg on plate	*Pour juice to dough*	*Add egg to plate*
(c)		
		
Put fruit in a bowl	*Stir tea, dough*	*Put tea in a bowl*
(d)		

Figure 5: Comparative illustration of video descriptions generated by the original pattern theory method [9] (second column) and the proposed method using ConceptNet framework (third column) in comparison with the ground truth (first column). For each case (each row), it can be seen that the descriptions made by the ConceptNet based approach had more *semantic coherence* compared to the domain-specific Pattern Theory model [9]; thereby increasing the quality of descriptions made. Deviations from the ground-truth are highlighted in red.

[7] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2568–2577.

[8] X. Wang and Q. Ji, "Video event recognition with deep hierarchical context model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4418–4427.

[9] F. D. de Souza, S. Sarkar, A. Srivastava, and J. Su, "Spatially coherent interpretations of videos using pattern theory," *IJCV*, pp. 1–21, 2016.

[10] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2634–2641.

[11] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 780–787.

[12] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," *arXiv preprint arXiv:1412.4729*, 2014.

[15] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4507–4515.

[16] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, "Translating video content to natural language descriptions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 433–440.

[17] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis, "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos," in *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on*. IEEE, 2009, pp. 2012–2019.

[18] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4534–4542.

[19] S.-W. Joo and R. Chellappa, "Recognition of multi-object events using attribute grammars," in *2006 International Conference on Image Processing*. IEEE, 2006, pp. 2897–2900.

[20] K. S. R. Dubba, A. G. Cohn, D. C. Hogg, M. Bhatt, and F. Dylla, "Learning relational event models from video," *Journal of Artificial Intelligence Research*, vol. 53, pp. 41–90, 2015.

[21] V. I. Morariu and L. S. Davis, "Multi-agent event recognition in structured scenarios," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3289–3296.

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

[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.

[24] H. Liu and P. Singh, "Conceptneta practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.

[25] U. Grenander, *Elements of pattern theory*. JHU Press, 1996.

[26] R. Speer and C. Havasi, "Conceptnet 5: A large semantic network for relational knowledge," in *The Peoples Web Meets NLP*. Springer, 2013, pp. 161–176.

[27] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[28] V. Pinheiro, V. Furtado, T. Pequeno, and D. Nogueira, "Natural language processing based on semantic inferentialism for extracting crime information from text," in *Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on*. IEEE, 2010, pp. 19–24.

[29] H. Liu and P. Singh, "Commonsense reasoning in and over natural language," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2004, pp. 293–306.

[30] C. Havasi, R. Speer, J. Pustejovsky, and H. Lieberman, "Digital intuition: Applying common sense using dimensionality reduction," *IEEE Intelligent systems*, vol. 24, no. 4, pp. 24–35, 2009.

[31] C. G. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends in Information Retrieval*, vol. 2, no. 4, pp. 215–322, 2008.

[32] C. C. Tan and C.-W. Ngo, "The vireo team at mediaeval 2013: Violent scenes detection by mid-level concepts learnt from youtube." in *MediaEval*, 2013.

[33] E. H. Spriggs, F. D. L. Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *CVPRW*, June 2009, pp. 17–24.

[34] F. D. de Souza, S. Sarkar, and G. Camara-Chavez, "Building semantic understanding beyond deep learning from sound and vision," *ICPR 2016 Proceedings*, 2016.