2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Building Semantic Understanding beyond Deep Learning from Sound and Vision

Fillipe D M de Souza, Sudeep Sarkar
Dept. of Computer Science & Engineering
University of South Florida
Tampa, Florida 33620
Email: fillipe@mail.usf.edu, sarkar@usf.edu

Guillermo Cámara-Chávez
Dept. de Ciência da Computacão
Universidade Federal de Ouro Preto
Ouro Preto, MG Brazil
Email: guillermo@iceb.ufop.br

***Abstract*—Deep learning-based models have recently been widely successful at outperforming traditional approaches in several computer vision applications such as image classification, object recognition and action recognition. However, those models are not naturally designed to learn structural information that can be important to tasks such as human pose estimation and structured semantic interpretation of video events. In this paper, we demonstrate how to build structured semantic understanding of audio-video events by reasoning on multiple-label decisions of deep visual models and auditory models using Grenander's structures for imposing semantic consistency. The proposed structured model does not require joint training of the structural semantic dependencies and deep models. Instead they are independent components linked by Grenander's structures. Furthermore, we exploited Grenander's structures as a means to facilitate and enrich the model with fusion of multimodal sensory data; in particular, auditory features with visual features. Overall, we observed improvements in the quality of semantic interpretations using deep models and auditory features in combination with Grenander's structures, reflecting as numerical improvements of up to 11.5% and 12.3% in precision and recall, respectively.**

I. Introduction

Deep learning-based models have recently been widely successful at outperforming traditional approaches in several computer vision applications such as image classification [1], object detection [2], [3], action recognition [4], [5] and event recognition [6]; however, those deep models are learned to perform hard labeling. Another research direction seeks to combine graphical models with deep neural networks (DNN) to harness the power of deep features while tackling problems that are better modeled with knowledge of structural dependencies, such as in semantic image segmentation [7], extraction of words from noisy images [8], human pose estimation [9], and group activity analysis [10].

In this paper we demonstrate how to construct structured semantic understanding of audio-video events from auditory features and deep visual features using Grenander's structures [11]. The proposed structured model performs probabilistic reasoning on multiple decisions of low-level classification models of actions and objects based on the input audio-video features. These multiple decisions are available for each feature as a list of action and object labels, where each label is associated with one of the top k classification scores output by the classifiers. Grenander's structures use structural semantic information of the domain to weigh the feature support provided by the classifiers and therefore impose semantic consistency on their classification decisions. An illustration of our contribution is shown in Figure 1.

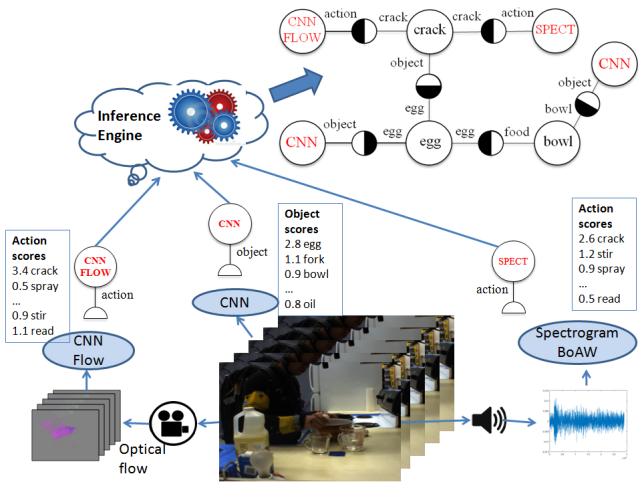

Fig. 1. Grenander's structures probabilistic reasoning on the top k labels scored on CNN and auditory features form semantically consistent interpretations.

Grenander's structures are also known as elements of pattern theory (PT) [11]. de Souza *et al.* [12] proposed a pattern theory framework for semantic understanding of video events, followed by a more detailed description in [13] and [14]. This same framework was modified in [15] to improve interpretations of long video temporal sequences by incorporating temporal dependencies in the structures. This paper extends the work in [15] to handle fusion of multimodal features and to integrate and reason on deep learning-based models for structured semantic understanding of videos.

The contribution of this paper is three fold. First, to the best of our knowledge, this is the first paper to show results on semantic interpretation of events using deep features jointly with Grenander's structures. Secondly, we exploited Grenander's structures as a means to facilitate and enrich the model with fusion of multimodal sensory data; in particular, auditory features with visual features. Thirdly, we present the first

results on structured semantic understanding of audio-video events using the CMU Kitchen dataset published in [16].

II. Feature Extraction

In this section, we describe the features used for action and object detection. The former are based on two modal features: visual and auditory, while the latter uses only appearance features. For auditory features, we utilize spectrogram features, follow by the bag-of-audio-features (BoAW) approach. Due to the success of Convolutional Neural Networks (CNN), we explore CNN features for motion description of actions and appearance description of objects. A SVM classifier categorizes action and object features.

For action recognition based on audio features, we use the BoAW, a histogram representation of frame-based audio features. These features are computed locally from short-term auditory frames. The histogram is suitable for describing the global characteristics of a sound event. The audio feature used in this work is the spectrogram due to its efficiency to identify spoken words and analyze different types of sounds [17]. The parameters used for the spectrogram generation are: a Hamming window function width of 512 samples and an overlapping of 50%. Then, these features are quantized using a codebook. We build the codebook using the K-means algorithm with $K = 400$. The spectrogram cluster histograms (*spec*) are used as input into a SVM classifier.

For action recognition based on visual information, we follow [18] and represent the frames with motion-based CNN descriptors. Feature vectors are extracted from each frame and then used an aggregated approach over time to generate a video descriptor [19]. First, we need to compute optical flow [20] for each consecutive pair of frames. Following [18], the values of the motion field v_x, v_y are mapped to color intensity interval [0, 255]. Then, the transformed flow maps are saved as color images (3 channels), where each corresponding channel contains v'_x, v'_y and the flow magnitude, respectively. The flow images are resized to 224×224 pixels to match the CNN input layer. We use the motion network provided by [20], the "VGG-f" network with 5 convolutional and 3 fully-connected layers, pretrained for action recognition task on the UCF101 dataset [21]. We utilize the output of the second fully-connected layer as our frame descriptor f_t, which consists of 4096 values. To generate the video descriptor, we follow the aggregation process [19]. For each descriptor dimension over the T frames, we compute the minimum and maximum values and concatenate them, doubling the vector size. We employ temporal differences between $f_{t+\Delta t}$ and f_t to capture the temporal evolution of frame descriptors, where $\Delta t = 4$ frames. Then, the aggregating process is also applied to frame descriptor differences. The aggregating vectors generated on feature descriptor, and feature differences are concatenated and used as the video descriptor (*CNN Flow*), this feature vector is the input into a SVM classifier.

For object detection, we also base our descriptor on CNN features. To represent the appearance, we use the second fully-connected layer of the publicly available "VGG-f" network pretrained on the ImageNet ILSVRC-2012 challenge dataset [22]. Finally, this feature vector (*CNN*) is used as input into a SVM classifier.

III. Grenander's Structures as Representation of Semantic Understanding

In the proposed framework, the building blocks of event interpretations such as actions and objects are represented as *generators*, denoted as g. We reuse the collection of generators presented in [15] and complement it with more generators representing actions and objects only available in the CMU Kitchen dataset released in [16]; thus, forming a richer *generator space* G of the cooking domain (see Figure 2). Each generator g has a structure composed of *out-bonds* $\beta'(g)$ and *in-bonds* $\beta''(g)$. Each bond carries a bond value that indicates how it is semantically related to other generators. For example, the generator *pour* has three out-bonds whose bond values are *container*, *container* and *food*. The generator *pour* can connect to the generator *oil* through the bond holding value *food* to indicate that oil is the liquid being poured in the event.

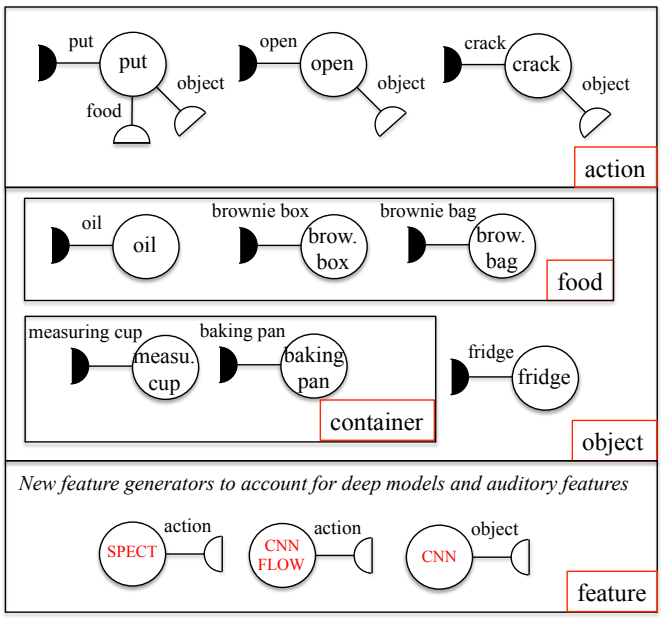

Fig. 2. New generators added to the generator space presented in [15]. The feature generators at the bottom feature the contribution of this paper, both the multimodal fusion of features and leveraging of deep models for generating structured semantic understanding of videos.

A. Coupling deep models with Grenander's structures

We introduce new types of feature generators to the generator space G to account for CNN features of objects and actions, namely, the generators *CNN* and *CNNFLOW*. The generator *CNN* has an out-bond of bond value *object* and *CNNFLOW* has an out-bond of bond value *action* (see Figure 2). Each CNN feature generator is associated with a deep classification model, either for classification of actions or for classification

of objects. These classification models are multi-class linear-SVM classifiers trained on CNN features.

B. Multimodal fusion with Grenander's structure

To enable the multimodal capability to the model, we add a new type of feature generator to account for incoming auditory features, namely, the generator *SPECT*. Its bond structure is formed by a single out-bond of bond value *action* (see Figure 2).

C. Bond Quantification

Two generators g_i and g_j connect through compatible bonds. The meaning of such connection is determined by their bond values; for example, the generator $stir$ has an out-bond of bond value *stirrer*, such that any other generator that connects to that out-bond will serve the role of a stirrer in the event. The strength of compatibility between bonds is quantified by the acceptor function

$$A(\beta'(g_i), \beta''(g_j)) = \exp(q(g_i, g_j) \tanh(f(g_i, g_j))). \quad (1)$$

where $f(.)$ is a scoring function that measures the compatibility of connecting the labels by g_i and g_j through their respective bonds β' and β''. If g_i is a feature generator and g_j is a generator representing an action or object label, then $f(.)$ responds as the classification score associated with the classifier of g_i for the label represented by g_j. If both g_i and g_j represent action/object labels, then $f(.)$ represents the entry value of the frequency table that counts the co-occurrence of labels describing events of the target domain. $q(.)$ weighs the rescaled score output by $f(.)$ depending on what type of bond is formed between g_i and g_j. If $(\beta'(g_i), \beta''(g_j))$ forms a support bond, then we let $q(.) = 1.5$, otherwise, if a semantic bond, $q(.) = 1.0$. This means we are emphasizing the support given by the classification scores more than the prior.

D. Semantic Interpretations

The collection of bonds in a generator g captures the semantic dependencies among the generators, dictating how they can combine with each other to form more complex structures referred to as *configurations*. A configuration c is a connected structure of generators used to express the semantic understanding (or semantic interpretation) of an audio-video event. Given a set of audio-video features, the goal is to discover the most probable configuration c that explains the semantics of the occurring events. To this end, we measure the quality of a configuration c using the following energy function:

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

IV. Inference of Grenander's Structures

The semantic video interpretation inference problem amounts to minimizing the energy cost function $E(c)$. Despite the similarity with the energy function for MRF models, our search space of solution is very different. In our case, both the number of generators and structures are variable, whereas MRF models have fixed number of random variables and a fixed structural connection. For this reason, this problem cannot be solved exactly, which motivates us to use a sampling strategy built on a Monte Carlo Markov Chain (MCMC) algorithm coupled with a simulated annealing scheduling. The proposal function of our MCMC algorithm works efficiently by restricting the proposal space of solution to those that can be spanned only by the top k labels pulled by the classification scores on the input features of the test video.

V. Results

A. Dataset

The Carnegie Mellon University Multimodal Activity database [16] contains multimodal measures of human activity, performing tasks that involve cooking and food preparation. The dataset contains five different recipes: brownies, pizza, sandwich, salad and scrambled eggs. The following modalities were recorded: high and low resolution videos and five microphones. We carried our experiments using brownie recipe videos only since only those videos had their fine-grained annotation of events available. Spriggs *et al.* [23] generated the ground truth for some videos and recipes. In total, there are 13 event brownie labeled videos. For training, we use videos identified by numbers 7, 8, 13, 14, 17 and 19. The test set was formed by 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 test set consisted of 233 event video segments. We evaluated and compared different combinations of features with different inference approaches (pure machine learning - ML and pattern theory - PT). For actions, we chose histograms of optical flow (HOF), convolution neural network based on motion (CNN Flow) and histograms of audio features based on spectograms (SPECT). As for object, we chose appearance convolution neural network (CNN) and histograms of oriented gradient (HOG). This way, the combination `PT cnn-cnnflow` means that the inference and representation were modeled with pattern theory using object models based on CNN features and action models based on CNN Flow features.

B. Semantic understanding based on machine learning labels only

Machine learning algorithms such as support vector machines (SVM) and neural networks (NN) are widely used to label actions and objects directly from video features. We implemented a strategy (which we called ML-based labels) based on linear-SVM classification models to generate semantic interpretations of events based on auditory and visual features. Given a set of auditory and visual features from a video, each feature is labeled according to the best classification score. The resulting set of labels represents the semantic understanding of the video. The best semantic interpretation is the one formed with all labels retrieved from best classification scores of the

features. Thus, the kth best interpretation is formed by labels retrieved from the kth best classification scores of the features. This strategy ignores the structural semantic information of the domain, relying solely on the confidence of the classification scores to build the interpretations.

C. Semantic understanding with structured output built on deep features

Grenander's structures jointly with priors leverage the evidence and confidence provided by the classification scores of deep models better than its counterpart that use no structural information and rely on support of deep models decisions alone. Figure 3 shows that the performance of interpretations by the deep models with Grenander's structures are often superior over the top 10 interpretations. Grenander's structures help the inference algorithm to exchange highly confident classifier's choices of labels by less confident ones that improve the semantic consistency of the interpretations; thus, improving the overall quality of the interpretations. Examples of these are illustrated in Figure 6, for example, the exchanges of *put* by *crack*, *bowl* by *brownie box* and *oil* by *egg*.

Although the methods employing HOF and HOG features generally show lower performance when compared to those built on CNN features, Figure 4 shows that the method with HOF and HOG features combined with Grenander's structures can achieve performance rates comparable to methods employing CNN features without Grenander's structures. This suggests that a structured model based on pattern theory can be potentially used to boost the performance of models using traditional features and have comparable performance to the state-of-the-art models using CNN features only; therefore, serving as possibly a less costly alternative if training and using deep models are computationally demanding for a specific task.

In summary, the interpretations supported by deep features and Grenander's structures had the highest performance rates, leading both recall and precision rates. Table I shows the overall performance rate of each method, considering up to the top 10 interpretations. Once more CNN features were proven to be superior to the traditional combination of feature histograms such as HOFs and HOGs.

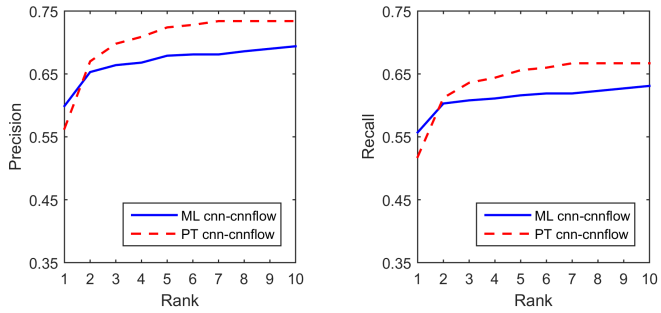

Fig. 3. Semantic understanding with deep features perform better when coupled with Grenander's structure.

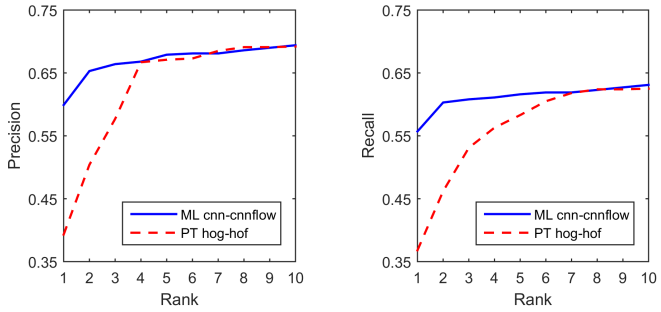

Fig. 4. Semantic understanding with traditional features (HOFs and HOGs) coupled with Grenander's structures (PT hog-hof) has achievable performance rates comparable to the semantic understanding models relying solely on classification score support of deep models (ML cnn-cnnflow).

D. Semantic understanding with sound and vision

The methods employing only auditory features for the recognition of actions were the most positively sensitive to the presence of domain knowledge imposed by Grenander's structures. For example, when ignoring motion features, there was a performance rate improvement of 11.5% and 12.3% in precision and recall (see Table I), respectively. Grenander's structures allowed the latent discriminating power of auditory features to become visible, which was reflected by having `PT cnn-spect` outperform `ML cnn-spect` by more than 10% in all performance metrics. We also observed that this method (PT cnn-spect) achieved comparable performance rates to more computationally heavier methods that depend on motion features, namely, `PT cnn-cnnflow` and `PT cnn-cnnflow-spect`. This suggests that audio features could be potential surrogates for the discriminating power offered by motion features while requiring less computational power; thus, allowing for implementation strategies of the low-level video processing layer that are computationally less expensive.

Qualitatively, this improvement was reflected mostly on selecting the right action to describe the event, correcting 47.8% of all test events. The most corrected actions were *pour* (26%), *take* (23.3%) and *stir* (21.9%). On the other hand, the sound feature support was not as positively complimentary to deep visual features (cnnflow) as we expected in building the semantic understanding of events. The method combining deep visual features with auditory features (PT cnn-cnnflow-spect) corrected just as many cases of wrongly labeled actions as did the method with deep features only (PT cnn-cnnflow) when contrasted with their counterparts supported with HOF and HOG features. Additionally, PT cnn-cnnflow-spect did not improve any interpretation case missed by PT cnn-cnnflow. In Table I, we observe that their overall performance rates were equivalent.

E. Improving semantic interpretation performance with Grenander's structure

Figure 5 shows how often labels of certain actions and objects are fixed in the interpretations due to the semantic

TABLE I
INTERPRETATION PERFORMANCE RATES FOR SEMANTIC UNDERSTANDING WITH SOUND AND VISION FEATURES.

	Recall	Precision
ML hog-hof	0.583	0.651
PT hog-hof	0.625	0.692
ML cnn-cnnflow	0.631	0.694
PT cnn-cnnflow	**0.667**	**0.734**
ML hog-spect	0.583	0.649
PT hog-spect	0.63	0.696
ML cnn-spect	0.585	0.651
PT cnn-spect	**0.657**	**0.726**
ML hog-hof-spect	0.593	0.659
PT hog-hof-spect	0.631	0.696
ML cnn-cnnflow-spect	0.626	0.689
PT cnn-cnnflow-spect	**0.667**	**0.734**

consistency imposed by Grenander's structures, shown as black bars. The gray bars indicate how often certain labels are missed by the method using Grenander's structures but correctly retrieved by the method without structure priors. The graph on the top shows that in general the most likely actions to be corrected in the semantic interpretations by Grenander's structure-based methods are *stir*, *pour*, *take* and *open*. This also dictates what object labels are most likely to be correctly selected to build the interpretations, namely, *bowl*, *fridge*, *measuring cup* and *brownie box*. Note that these objects are semantically compatible with the most likely actions to be often correctly selected by the Grenander's structure methods. For example, interpretations likely to be proposed with combination of these actions and objects include *open fridge*, *stir bowl*, *pour oil into measuring cup*, *take brownie box*, *open brownie box*, etc. The graph at the bottom, in Figure 5, shows that in fact these labels are the most likely labels to be corrected by methods using Grenander's structures. On the other hand, other object labels more likely to be corrected by methods without the structural influence, for instance, *cap* and *egg*.

In a nutshell, the graphs in Figure 5 show that the methods based on Grenander's structure are more likely to generate semantically consistent interpretations than the methods based solely on feature support. Figure 6 illustrates three interpretation cases depicted at different rows. On the first row, the correct interpretation is generated by the method based on Grenander's structures because of the structural connections (bonds) between the action and object labels, namely, $crack \rightarrow egg$ and $bowl \rightarrow egg$; thus changing the interpretation from *putting egg in a brownie box* to *cracking egg in a bowl*. Another good example of semantic consistency is illustrated in the second row of Figure 6, where the interpretation is changed from *open bowl* (which even by common sense may not be semantically coherent) to *open brownie box*.

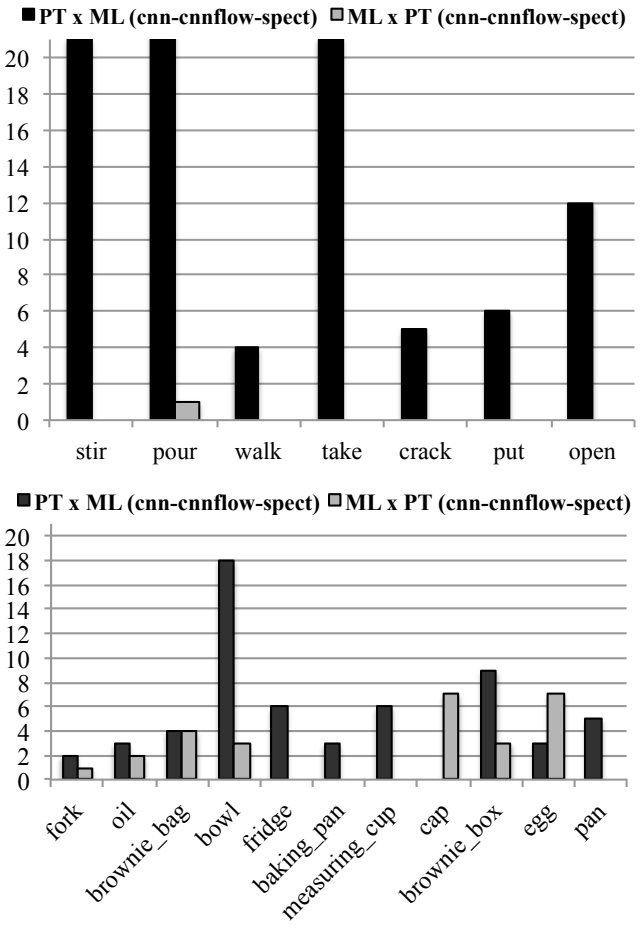

Fig. 5. Number of action (first row) and object (second row) labels that were correctly selected to build the best interpretations by the Grenander's structures-based methods and were missed by the methods (black bars). The gray bars show the opposite statistics.

VI. CONCLUSION

In this paper, we demonstrated that the predictive power of CNN features were improved by considering the structural semantic dependencies of events encoded in terms of Grenander's structures (generators, bonds and configurations). These structures carry complimentary data that encourage rectification of erroneously highly confident detections by deep classifiers of actions and objects. Auditory features were verified to be potentially a sufficient source of data for modeling actions. The semantic interpretations generated by the method built on auditory features for actions and CNN features for objects, *i.e.* PT cnn-spect, were qualitatively comparable to the ones generated by its counterparts that model actions with CNN features. This indicates that we could potentially reduce the feature pre-processing computational cost by skipping the motion analysis step. Finally, we verified that even when using features not as discriminative as CNN features, Grenander's structures can be sufficiently strong to achieve performance rates comparable to when using CNN features-based models alone.

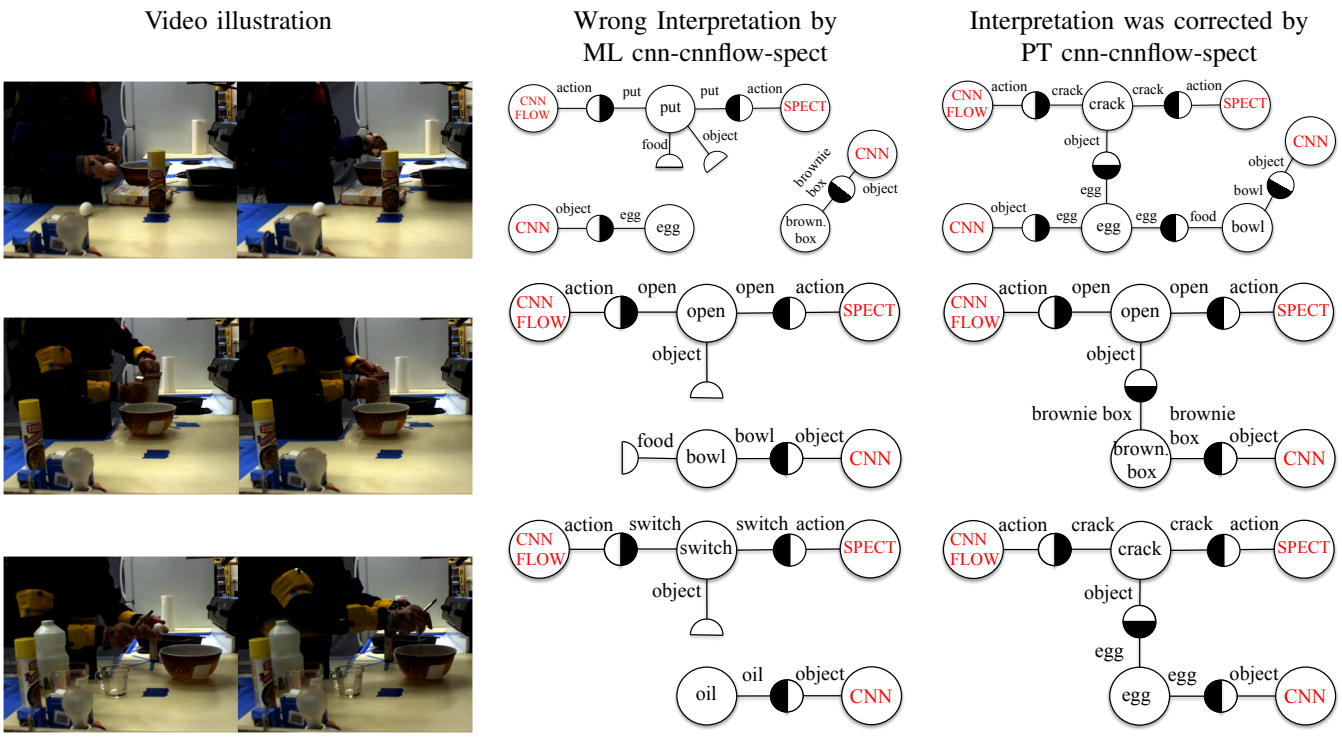

Fig. 6. Comparative illustration of video interpretations generated by the method based on deep models without structural information (second column) and the method with deep models using Grenander's structures (third column). For each case (each row), the interpretations were corrected by the method that uses Grenander's structures.

ACKNOWLEDGMENT

This research was supported in part by NSF grants 1217676. The authors would like to thank the Brazilian National Research Council - CNPq (Grant # 234272/2014-7)

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[2] 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.

[3] 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.

[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014, pp. 1725–1732.

[5] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014, pp. 568–576.

[6] X. Wang and Q. Ji, "Video event recognition with deep hierarchical context model," in *CVPR*, 2015, pp. 4418–4427.

[7] J. Xu, A. Schwing, and R. Urtasun, "Tell me what you see and i will show you where it is," in *CVPR*, 2014, pp. 3190–3197.

[8] L.-C. Chen, A. G. Schwing, A. L. Yuille, and R. Urtasun, "Learning deep structured models," *arXiv preprint:1407.2538*, 2014.

[9] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *NIPS*, 2014, pp. 1799–1807.

[10] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," *arXiv preprint:1511.04196*, 2015.

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

[12] F. D. de Souza, S. Sarkar, A. Srivastava, and J. Su, "Pattern theory-based interpretation of activities," in *ICPR*. IEEE, 2014, pp. 106–111.

[13] ——, "Pattern theory for representation and inference of semantic structures in videos," *PRL*, vol. 72, pp. 41–51, 2016.

[14] ——, "Spatially coherent interpretations of videos using pattern theory," *IJCV*, pp. 1–21, 2016.

[15] F. Souza, S. Sarkar, A. Srivastava, and J. Su, "Temporally coherent interpretations for long videos using pattern theory," in *CVPR*. IEEE, 2015, pp. 1229–1237.

[16] F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran, "Guide to the carnegie mellon university multimodal activity (cmu-mmac) database," *RI*, p. 135, 2008.

[17] W. Yang, D. Yang, X. Chen, and H. He, "Music identification based on music word model," in *ICME*, 2015, pp. 1–6.

[18] G. Gkioxari and J. Malik, "Finding action tubes," in *CVPR*, June 2015, pp. 759–768.

[19] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN Features for Action Recognition," in *ICCV*, Santiago, Chile, Dec. 2015.

[20] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *ECCV*, vol. 3024. Springer, May 2004, pp. 25–36. [Online]. Available: http://lmb.informatik.uni-freiburg.de//Publications/2004/Bro04a

[21] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

[22] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[23] 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.