Temporally Coherent Interpretations for Long Videos Using Pattern Theory

Fillipe Souza[1], Sudeep Sarkar[1], Anuj Srivastava[2], Jingyong Su[3]
[1]Computer Science & Engineering, University of South Florida. [2]Statistics, Florida State University. [3]Mathematics & Statistics, Texas Tech University.

Graph-theoretical methods have successfully provided semantic and structural interpretations of images and videos. A recent paper [1] 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.

In this paper, 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.

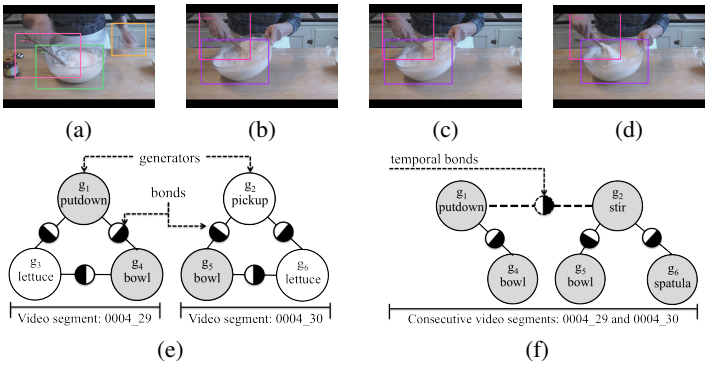

Figure 1: Illustration of advantage in using temporal bonds. Top rows shows frames from two consecutive segments of a video. The first segment depicts the interaction *put bowl down* (the small one with the left hand) and second segment depicts *stir ingredients in a bowl using spatula*. (e) shows [1]'s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

In pattern theory [2], the basic units of representations are generators g. We define the space of generators $G = \{g_1, ..., g_n\}$ hierarchically - from (image) features at the bottom level to (human) actions at the highest, such that $G = G_{features} \cup G_{objects} \cup G_{actions}$. Such generators g_i combine with each other through bonds $\beta_j(g_i)$ that satisfy predefined constraints (Figure 2). The resulting configurations of connected generators $\sigma(g_1, \cdots, g_n)$ represent a semantic understanding of the video content, an interpretation. The inference goal is to generate high-probability interpretations given a set of features. Probabilistic structures are imposed using energies that have contributions from both data (classification scores) and prior information (ontological constraints). The probability of an interpretation $\sigma(g_1, \cdots, g_n)$ is expressed as a product of terms associated with generators and bond interactions, written as

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

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

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

The search for optimal interpretations is accomplished by minimizing the energy function $E(\sigma)$ using an MCMC-based simulated annealing algorithm that uses simple moves to propose interpretation changes and to accept or reject them according to the posterior energy. A summary of the performance analysis in shown in Figure 3. Using temporal bonds and analyzing longer videos allowed us to achieve up to 83% performance improvement.

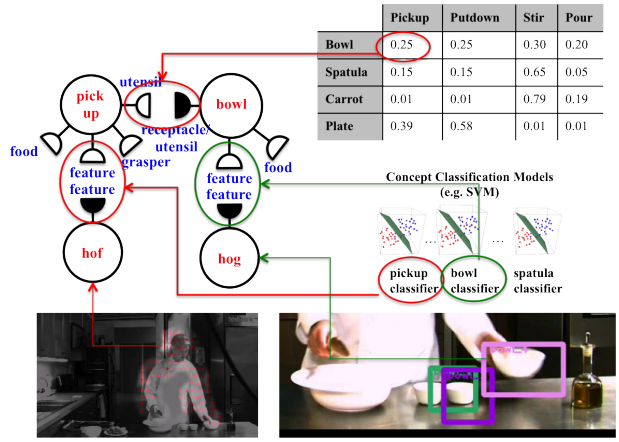

Figure 2: An illustration showing how generators are combined with each other using bonds and have their bond interaction energies computed.

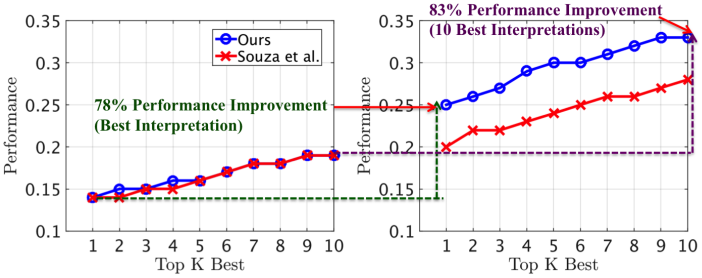

Figure 3: Left:performance rate comparison for 1-segment video sequences. Right: performance rate comparison for 4-segment video sequences. Note that temporal bonds are crucial to improve interpretation performance.

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

[1] Fillipe D M de Souza, Sudeep Sarkar, Anuj Srivastava, and Jingyong Su. Pattern theory-based interpretation of activities. In *ICPR*, 2014.

[2] Ulf Grenander. *General pattern theory-A mathematical study of regular structures*. Clarendon Press, 1993.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.