Supplementary Material For rCRF: Recursive Belief Estimation over CRFs in RGB-D Activity Videos

Ozan Sener
School of Electrical & Computer Eng.
Cornell University

Ashutosh Saxena
Department of Computer Science
Cornell University

I. Introduction

In this supplementary material, we give the detailed derivation of the posterior belief we state in equation (8) of the main paper as well as the detailed algorithm to solve the optimization problem in equation (10) of the main paper. We also present the experimental results we did not include in the main paper due to the space limit. Furthermore, we explain the details of the computational-efficiency experiment we conducted in order to obtain the results in Table 3 of the main paper.

II. Supplementary Derivations

A. Posterior Belief

Here, we present the derivation of the posterior belief (main paper Eq 8).

We start with the definition of a belief as $bel^t(\mathbf{y}) \propto \alpha^t(\mathbf{y})\beta^t(\mathbf{y})$ and substitute the definition of the messages from (main paper Eq 3). The resulting belief in log likelihood form is,

$$\log\ bel^t(\mathbf{y}) = \log p(\mathbf{x}^t|\mathbf{y}^t) + \log \sum_{\mathbf{y}^{t-1}} \alpha^{t-1}(\mathbf{y}^{t-1}) p(\mathbf{y}^t|\mathbf{y}^{t-1}) + \log \sum_{\mathbf{y}^{t+1}} p(\mathbf{x}^{t+1}|\mathbf{y}^{\mathbf{t+1}})\beta^{\mathbf{t+1}}(\mathbf{y}^{\mathbf{t+1}})\mathbf{p}(\mathbf{y}^{\mathbf{t+1}}|\mathbf{y}^{\mathbf{t}}) \tag{1}$$

Here, we observe that forward message $\alpha^{t-1}(\mathbf{y}^{t-1}) = p(\mathbf{y}^{t-1}|\mathbf{x}^1, \dots, \mathbf{x}^{t-1})$ is a probability distribution, and backward message $\frac{1}{\gamma}\beta^{t+1}(\mathbf{y}^{t+1})p(\mathbf{x}^{t+1}|\mathbf{y}^{\mathbf{t+1}})$ can be considered as unnormalized density with a normalization term γ. Similar to the measurement equation, we also approximate the belief with its lower bound via Jensen inequality and compute the belief as;

$$\log\ bel^t(\mathbf{y}) \approx \log p(\mathbf{x}^t|\mathbf{y}^t) + \sum_{\mathbf{y}^{t-1}} \alpha^{t-1}(\mathbf{y}^{t-1}) \log p(\mathbf{y}^t|\mathbf{y}^{t-1}) + \frac{1}{\gamma} \sum_{\mathbf{y}^{t+1}} \beta^{t+1}(\mathbf{y}^{t+1}) p(\mathbf{x}^{t+1}|\mathbf{y}^{t+1}) \log p(\mathbf{y}^{t+1}|\mathbf{y}^t) + \log \gamma \tag{2}$$

Here, we also substitute $p(\mathbf{x}^t|\mathbf{y}^t)$ with (main paper Eq 7) and obtain the belief up to a scale as;

$$bel(\mathbf{y}^{\mathbf{t}}) \propto \exp\Bigg[\sum_{i,j \in E^t} \Big(\theta_{x_i^t, x_j^t}(y_i^t, y_j^t) - \tilde{\theta}(y_i^t, y_j^t) \Big) \sum_{i \in V^t} \Bigg(\theta_{x_i^t}(y_i^t) - \tilde{\theta}(y_i^t) + \sum_{\mathbf{y}^{t-1}} \alpha^{t-1}(\mathbf{y}^{t-1}) \log p(y_i^t|y_i^{t-1}) + \frac{1}{\gamma} \sum_{\mathbf{y}^{t+1}} \beta^{t+1}(\mathbf{y}^{t+1}) p(\mathbf{x}^{t+1}|\mathbf{y}^{t+1}) \log p(y_i^{t+1}|y_i^t) \Bigg) \Bigg] \tag{3}$$

where $\gamma = \sum_{\mathbf{y}^{t+1}} \beta^{t+1}(\mathbf{y}^{t+1}) p(\mathbf{x}^{t+1}|\mathbf{y}^{t+1})$

B. Solving Div-M-Best for rCRF

In this section, we explain how to solve the optimization problem in (main paper Eq 10) by using the Lagrange relaxation proposed by Batra et al.[1].

We are interested in the following optimization problem;

$$\begin{aligned} \mathbf{y}^{t,i} &= \arg\max_{\mathbf{y}} bel^t(\mathbf{y}) \\ &s.t.\ \Delta(\mathbf{y}, \mathbf{y}^{t,j}) \geq \delta \quad \forall\, j < i \end{aligned} \tag{4}$$

We first take the logarithm of the objective function since $\log$ is a monotonic function. We then follow the Div-M-Best procedure [1]. Div-M-Best uses Lagrange relaxation after dualizing the $\Delta(\mathbf{y}, \mathbf{y}^{\mathbf{t,j}}) \geq \delta$ constraints. Hence, the relaxed unconstrained optimization problem is,

$$\mathbf{y}^{t,i} = \arg\max_{\mathbf{y}} \log bel^t(\mathbf{y}) + \sum_{m=0}^{i-1} \lambda_m (\Delta(\mathbf{y}, \mathbf{y}^{t,m}) - \delta) \tag{5}$$

Please note that we use the hamming distance for $\Delta(\cdot,\cdot)$ within all of our experiments. Hence, we substitute the Hamming distance in the optimization objective with $\Delta(\cdot,\cdot)$. We

further substitute the (3) as,

$$\begin{aligned}\mathbf{y}^{t,i} = \arg\max_{\mathbf{y}} & \sum_{i,j \in E^t} \left(\theta_{x_i^t, x_j^t}(y_i^t, y_j^t) - \tilde{\theta}(y_i^t, y_j^t)\right) \\ & \sum_{i \in V^t} \Bigg(\theta_{x_i^t}(y_i^t) - \tilde{\theta}(y_i^t) + \sum_{m=0}^{i-1} \lambda_m \mathbb{1}_{y_i^t \neq y_i^{t,m}} \\ & + \frac{1}{\gamma} \sum_{\mathbf{y}^{t+1}} \beta^{t+1}(\mathbf{y}^{\mathbf{t+1}}) p(\mathbf{x}^{t+1} | \mathbf{y}^{t+1}) \log p(y_i^{t+1} | y_i^t) \\ & + \sum_{\mathbf{y}^{t-1}} \alpha^{t-1}(\mathbf{y}^{t-1}) \log p(y_i^t | y_i^{t-1}) \Bigg)\end{aligned} \tag{6}$$

Where $\mathbb{1}_A$ is an indicator function, and it is 1 when A is true and 0 otherwise. Thus, the final optimization problem in (6) is equivalent to finding the MAP solution of a CRF with modified energy function. Moreover, we solve it by using the original inference method (Mixed Integer Programming) following [3].

III. Supplementary Results

A. Computational-Efficiency of the Inference

We evaluated the computational-efficiency of our algorithm experimentally by recording its run-time. While recording the run-time, we did not include the feature extraction and pre-processing times since feature extraction and pre-processing steps are identical for all the competing algorithms. In other words, the recorded numbers are the inference time of the algorithms. We perform our experiments on an Intel i7 3.0 GHz laptop with 6Gb RAM running Ubuntu operating system using Python programming language. While implementing the algorithms, we only used a single core. Therefore, optimizing and parallelizing the code will give future gains.

Inference time during detection: We recorded the runtime of the inference algorithm for each temporal segment while estimating the belief for observed frames in the detection setting. We present the results in Table I.

Since the optimization algorithm, we define in (main paper 8) uses the inference procedure repeatedly for each sample, we expect to have a constant *number of diverse samples* multiplicative factor in our computation time. As shown in Table I, resulting inference time is approximately ten times the MAP solution although the belief is over $M = 15$ diverse samples. We believe this is due to the effective initialization of the mixed integer program with the previous results while iteratively computing the samples. Moreover, temporal segments are longer than 1 second long hence the resulting algorithm is still real-time.

TABLE I: Computation time for computing a belief over $M = 15$ samples per temporal segment excluding pre-processing.

MAP Sol. [3]	$30ms$	Full Belief	$367ms$

Inference time during anticipation: We experiment over the 3 seconds into the future anticipation setting and summarize the results in the Table II. We observed that our method had average computation time of $1.41s$ and [2] had average computation time of $34.1s$. This behavior is the result of efficient and accurate sampling of the belief space. Since our samples are more accurate, we need fewer samples than ATCRF [2]. Hence, our computation time is significantly better. Indeed, our inference algorithm for anticipation is operating at about 2X real-time.

TABLE II: Computation time for anticipating 3 seconds in the future excluding pre-processing.

ATCRF [2]	34.1s	rCRF	1.41s

B. More Visual Results

We also include additional visual results for activity anticipation in Figure 1 covering various environments. Similar to the visual results in the main paper, we anticipate four temporal segments into the future. Each segment is between 1 to 3 seconds long and the activity at the middle of the segment is anticipated via rCRF using the rest of the video. We present the belief of each anticipated segment as well as the middle frame of the segment. Please note that the visual frames are not visible to the algorithm and only included for evaluation purposes.

As shown in Figure 1, anticipated belief is highly accurate. In the first row, the subject is moving the apple, and our algorithm anticipates the next activity accurately as eating. For its consecutive segment, our algorithm relies on the periodic activity of moving and eating/drinking. Hence, it anticipates that the activity following moving should be eating or drinking. In the third row, the initial belief is flatter. Hence, consecutive beliefs are flatter and less informative yet accurate. Moreover, in the fifth row, our algorithm accurately anticipates that box need to be placed after being moved.

References

[1] Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich. Diverse m-best solutions in markov random fields. In *Proc. ECCV*, pages 1–16. Springer, 2012.

[2] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *RSS*, 2013.

[3] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 32(8):951–970, 2013.

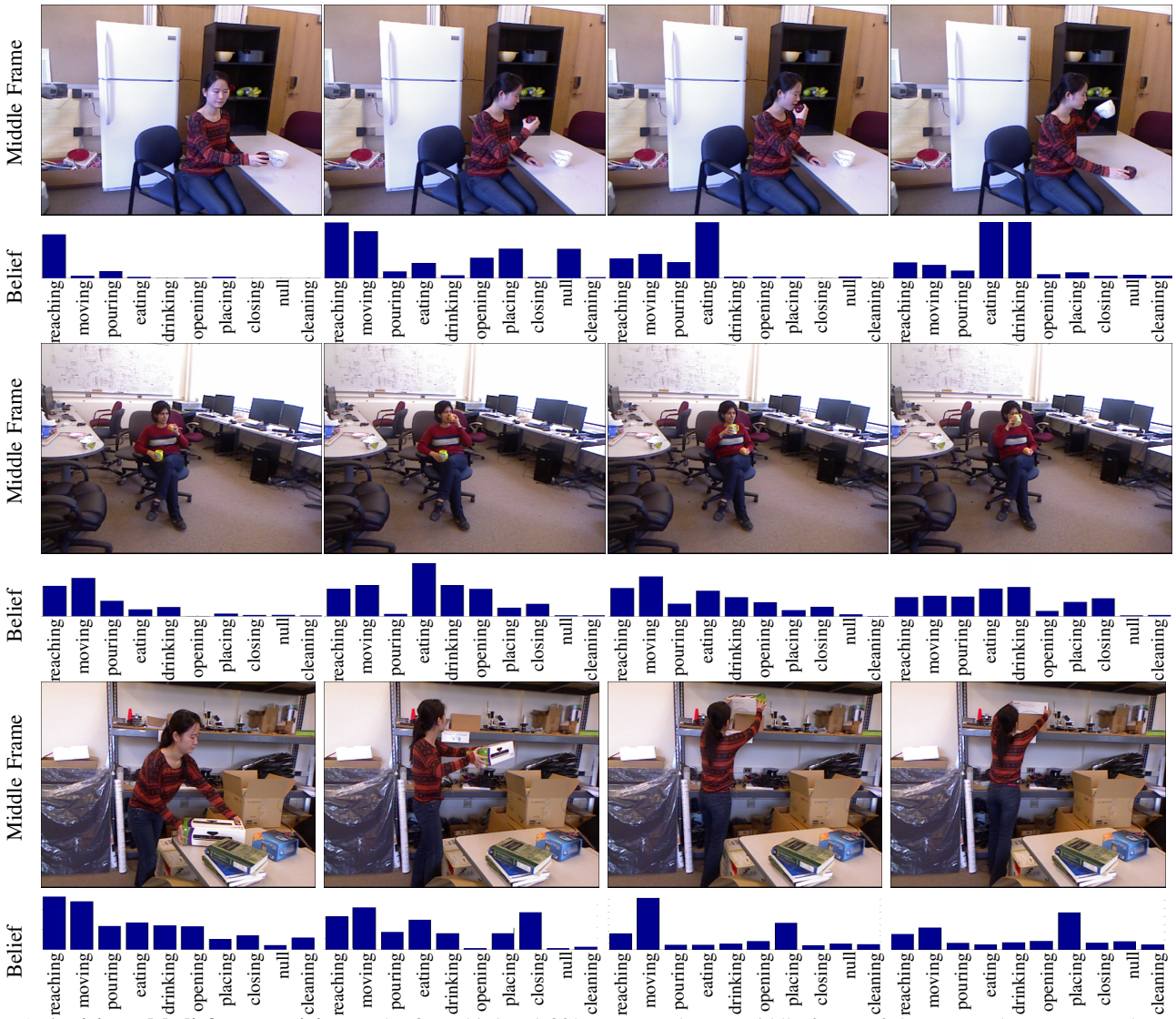

Fig. 1: **Anticipated belief over activity.** In the first, third and fifth row, we show a middle frame of the temporal segment. In the second, fourth and sixth row, we show the anticipated belief. Note that frames are not visible to the algorithm and only included for evaluation.