Co-segmentation for Space-Time Co-located Collections – Supplementary Material –

For interactive results and comparisons on our space-time co-located image collections, please open `index.html`.

Contents

1. Extending convex belief propagation

The variational interpretation for approximating Gibbs marginal probabilities follows the optimization program

$$\begin{aligned} &\arg\max_{b_i, b_{i,j}} \sum_{i,j,y_i,y_j} b_{i,j}(y_i,y_j)\theta_{i,j}(y_i,y_j) + \sum_{i,y_i} b_i(y_i)\theta_i(y_i) + \sum_{i\in V} c_i H(b_i) + \sum_{i,j\in E} c_{i,j} H(b_{i,j}) \qquad (1)\\ &\text{s.t.} \quad b_i(y_i), b_{i,j}(y_i,y_j) \geq 0, \quad \sum_{y_i,y_j} b_{i,j}(y_i,y_j) = 1, \sum_{y_i} b_i(y_i) = 1, \sum_{y_j} b_{i,j}(y_i,y_j) = b_i(y_i). \end{aligned}$$

We denote by $N(i)$ the set of edges that connect to node i and set $H(b_i) = -\sum_{y_i} b_i(y_i) \log b_i(y_i)$. Whenever the edges E compose a graph without cycles and $c_{i,j} = 1, c_i = 1 - |N(i)|$ are the Bethe coefficients, the above variational program results in exact inference: the optimal beliefs $b_i(y_i)$ are the Gibbs marginal probabilities. For graphs with cycles this program approximates the inference, as its optimal beliefs approximate the Gibbs marginal probabilities. Importantly, when one uses positive entropy coefficients $c_i, c_{i,j} > 0$ the program is everywhere concave and have a unique global optimum. This global optimum can be attained efficiently, e.g., by message-passing algorithms [7, 3]. In our work we consider an extension of this program, to count for interactions across images:

Claim 1 *Consider a program that augments Equation (1) with the non-linearities across images* $\sum_{i,j\in E_b} \sum_{y_i} b_i(y_i) b_j(y_j)$.

$$\begin{aligned} &\arg\max_{b_i, b_{i,j}} \sum_{(i,j)\in E, y_i, y_j} b_{i,j}(y_i,y_j)\theta_{i,j}(y_i,y_j) + \sum_{i\in V, y_i} b_i(y_i)\theta_i(y_i) + \qquad (2)\\ &\qquad \sum_{i,j\in E_b} \sum_{y_i} b_i(y_i) b_j(y_j) + \sum_{i\in V} c_i H(b_i) + \sum_{i,j\in E} c_{i,j} H(b_{i,j})\\ &\textit{s.t.} \quad b_i(y_i), b_{i,j}(y_i,y_j) \geq 0, \quad \sum_{y_i,y_j} b_{i,j}(y_i,y_j) = 1, \sum_{y_i} b_i(y_i) = 1, \sum_{y_j} b_{i,j}(y_i,y_j) = b_i(y_i). \end{aligned}$$

Then this program is strictly concave if $c_{i,j} > 0$ *and for any* i *there holds* $c_i > \lambda_{max}(E_b)$ *where* E_b *is the adjacency matrix between images and* $\lambda_{max}(N_b)$ *is its maximal eigenvalue. Moreover, a message-passing algorithm (performing block coordinate ascent over the beliefs of this program) is guaranteed to converge to the program's optimum.*

Proof: The concavity of the program is determined by the eigenvalues of its Hessian. A function is strictly concave if the eigenvalues of its Hessian are negative. The Hessian of the linear terms vanishes and we do not consider it. The Hessian of

Figure 1: Comparison to the correspondences provided by NRDC. For each template-target image pair, the reliable correspondences are colored in either yellow (in foreground regions) or blue (in background regions). The silhouettes of our multi-target inference results are provided in red.

	BRIDE		SINGER		TODDLER	
	P	J	P	J	P	J
(i)	66.6	16.2	78.2	09.1	92.0	37.6
(i)+	68.5	19.8	77.8	11.5	93.6	47.1
(ii)	89.4	78.9	97.2	88.6	96.3	75.0
(iii)	**92.2**	**83.0**	**98.4**	**93.8**	**96.9**	**79.9**

Table 1: **Evaluation of our inference approach**. We break-down the performance of the different stages (numbered according to Section 2) of a single inference iteration.

the entropy function is a diagonal matrix whose entries are the minus of the inverse beliefs. Therefore the eigenvalues of the entropy's Hessian are at most -1. The Hessian of the mixing terms $\sum_{i,j\in E_b}\sum_{y_i} b_i(y_i)b_j(y_j)$ is dominated by the Hessian of E_b. This is a convex-concave function whose convex element is determined by $\lambda_{max}(E_b)$. To overcome this convexity it is sufficient to set $c_i > \lambda_{max}(E_b)$ for every i. The convergence to the global optimum follows the strict concavity and Proposition 2.7.1 in [1].

The maximal eigenvalue of an adjacency matrix is at most its maximal degree, hence $\lambda_{max}(E_b) \leq \max_j |N_b(j)|$. This gives a simpler condition to guarantee convergence which does not require the maximal eigenvalue of E_b, namely for any i with $c_i > \max_j |N_b(j)|$.

2. Evaluating our inference technique

To evaluate our inference technique, we provide both qualitative and quantitative results on the following intermediate stages: (i) single inference from correspondences only; (ii) a complete direct template-target inference application; and (iii) joint multi-target inference. In Table 1(a), we report the average P and J scores per dataset. These scores average over all the images which are adjacent to a template seed, and further average over three random template seeds per dataset. To evaluate how well the correspondences provided by NRDC capture the object in the target images, we report two different scores. The first (indicated by (i) in Table 1(a)) considers foreground regions directly according to the foreground correspondences. The yellow regions in Figure 1 illustrate these regions. The second (indicated by (i)+ in Table 1(a)) considers foreground regions that are obtained after performing graph-cuts, with these regions provided for initialization.

As can be observed both in Table 1(a) and Figure 1, our technique achieves a significant improvement over the initial set of reliable correspondences. Note that our technique is robust a to few outliers (false positives), e.g., in the right-most image of Figure 1, and recovers the object thanks to the joint multi-target inference stage. For a more extensive assessment of our inference technique, please refer to the interactive html files.

3. Results and comparisons on sampled video collections

In Figure 2 we provide additional qualitative results and comparisons on the first five frames from the sparse sequences sampled from the Davis datasets [5]. See Table 1 in the paper for a quantitative evaluation on these datasets.

References

[1] D. Bertsekas. Nonlinear programming. 1999. 2

[2] A. Faktor and M. Irani. Co-segmentation by composition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1297–1304, 2013. 3

[3] T. Heskes. Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. *Journal of Artificial Intelligence Research*, 26(1):153–190, 2006. 1

[4] G. Kim and E. P. Xing. On multiple foreground cosegmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 837–844. IEEE, 2012. 3

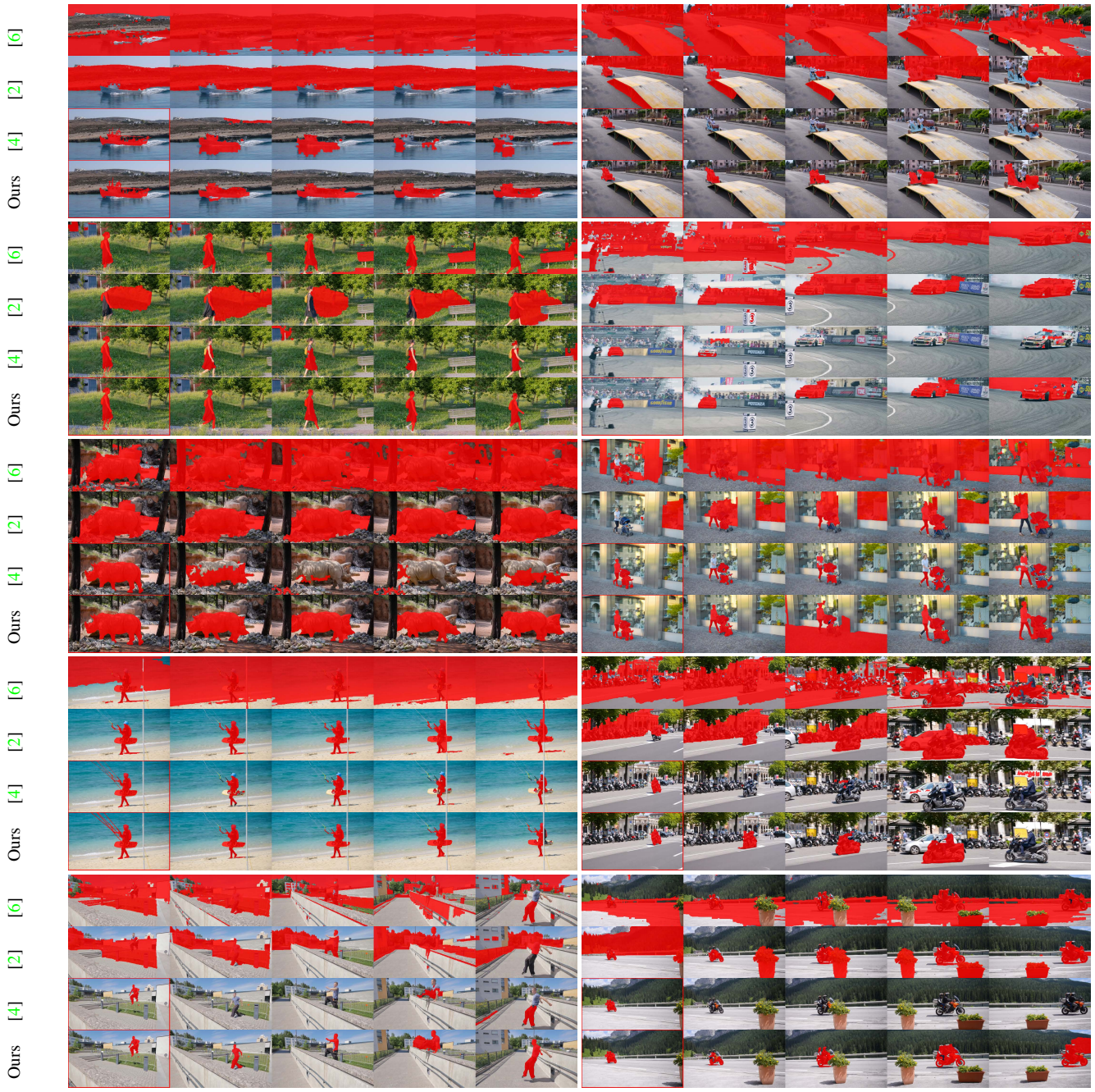

Figure 2: Qualitative comparison to state-of-the-art co-segmentation techniques on the first five frames from the sparse sequences sampled from the following Davis datasets [5]: boat, soapbox, lucia, drift-turn, rhino, stroller, kite-walk, scooter-black, parkour, motorbike. The first frame is provided as template for the semi-supervised techniques.

[5] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 2, 3

[6] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1939–1946. IEEE, 2013. 3

[7] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. A new class of upper bounds on the log partition function. *Trans. on Information Theory*, 51(7):2313–2335, 2005. 1