Supplementary Material for
The Shape-Time Random Field for Semantic Video Labeling

Andrew Kae, Benjamin Marlin, Erik Learned-Miller
School of Computer Science
University of Massachusetts, Amherst MA, USA
{akae,marlin,elm}@cs.umass.edu

This document contains supplementary material for the main paper [1]. We first describe the inference and learning procedures for the temporal SCRF and STRF models in more detail in Sections 1, 2 and then show additional qualitative results comparing STRF to baselines in Section 3.

1. Temporal SCRF

1.1. Inference

For the first frame (time $t = 1$), the SCRF is used for inference, since it does not depend on previous frames. Afterward, inference in the temporal SCRF is computed using a mean-field approximation, as shown in Algorithm 1. In step 1, the variational parameters $\boldsymbol{\mu}^{(0)}$ are initialized to the logistic regression guess, which depends only on node potentials. In step 2, the temporal potentials are computed using label guesses from the previous frame (denoted as $\boldsymbol{\alpha}$ in Algorithm 1). Recall that $Int(\mathcal{V}^{(t-1)}, s)$ refers to superpixels in the previous frame that intersect with superpixel s in the current frame. In step 4, the node, edge, and temporal potentials are used together to update the parameters $\boldsymbol{\mu}^{(i)}$. The algorithm then iterates until the parameters $\boldsymbol{\mu}^{(i)}$ either no longer change or a maximum number of iterations ($MaxIter$) is reached.

There is not much additional cost for inference (compared to inference in the SCRF) because the labels from the previous frame $t-1$ are assumed fixed, and thus the temporal potentials only need to be computed once. In step 4, the node and temporal potentials are included in the update for $\boldsymbol{\mu}^{(i)}$ but only the edge potentials change during iteration. Average inference time per frame for the temporal SCRF is about 0.78 (sec) compared to about 0.74 (sec) for the SCRF, on an Intel i7.

Algorithm 1 Mean-Field inference for the temporal SCRF

1: Initialize $\boldsymbol{\mu}^{(0)}$ as follows:

$$\mu_{sl}^{(0)} = \frac{\exp\left(f_{sl}^{\text{node}}\right)}{\sum_{l'} \exp\left(f_{sl'}^{\text{node}}\right)}$$

where

$$f_{sl}^{\text{node}}\left(\mathcal{X}_{\mathcal{V}}, \{p_{sn}\}, \Gamma\right) = \sum_{n,d} p_{sn} x_{sd} \Gamma_{ndl}$$

2: Let $\boldsymbol{\alpha}$ be the mean-field estimates from the previous frame where

$$f_{sl}^{\text{tpot1}}\left(\boldsymbol{\alpha}; \mathcal{X}_{\mathcal{T}}^{(t,t-1)}, \Phi\right) = \sum_{b \in Int(V^{(t-1)},s),} \sum_{l,l'=1}^{L} \sum_{e=1}^{D_{temp}} \alpha_{bl'} \Phi_{ll'e} x_{sbe}$$

$$f_{sl}^{\text{tpot2}}\left(\boldsymbol{\alpha}; \Pi\right) = \sum_{b \in \mathcal{V}^{(t-1)}} \sum_{l,l'=1}^{L} \alpha_{bl'} \Pi_{ll'} \left[s = b\right]$$

3: **for** i=0:*MaxIter* (or until convergence) **do**

4: update $\boldsymbol{\mu}^{(i+1)}$ as follows: $\mu_{sl}^{(i+1)} =$

$$\frac{\exp\left(f_{sl}^{\text{node}} + f_{sl}^{\text{edge}}\left(\boldsymbol{\mu}^{(i)}\right) + \kappa_1 f_{sl}^{\text{tpot1}} + \kappa_2 f_{sl}^{\text{tpot2}}\right)}{\sum_{l'} \exp\left(f_{sl'}^{\text{node}} + f_{sl'}^{\text{edge}}\left(\boldsymbol{\mu}^{(i)}\right) + \kappa_1 f_{sl'}^{\text{tpot1}} + \kappa_2 f_{sl'}^{\text{tpot2}}\right)}$$

where

$$f_{sl}^{\text{edge}}\left(\boldsymbol{\mu}; \mathcal{X}_{\mathcal{E}}, \mathcal{E}, \Psi\right) = \sum_{j:(s,j)\in\mathcal{E}} \sum_{l',e} \mu_{jl'} \Psi_{ll'e} x_{sje}$$

5: **end for**

1.2. Learning

In step 4 in Algorithm 1, the scalar parameters κ_1, κ_2 are used to weight the contribution of the temporal potentials relative to the node and edge potentials. In our experiments, we tried a variety of values between $\{0..1\}$ and chose κ_1, κ_2 based on which values performed best on the validation set. These κ_1, κ_2 parameters are then combined with a pre-trained SCRF. The temporal weights Φ, Π are initialized from the edge weights Ψ.

2. STRF

2.1. Inference

We use a feed-forward inference procedure which depends only on a window of W previous frames at time t. This approach is computationally efficient since the history of W previous frames is assumed *fixed* at time t, so the only latent variables at time t are the hidden units of the CRBM and the label variables. During inference, the first W frames are computed using the GLOC [2] model, which does not depend on previous frames. Afterward, we use a mean-field approximate inference approach for each frame as described in Algorithm 2. Inference proceeds in a sliding window fashion until all frames in the video are labeled. In step 1, the variational parameters $\boldsymbol{\mu}^{(0)}$ are initialized to the logistic regression guess (which depends only on node potentials) and $\boldsymbol{\gamma}^{(0)}$ are initialized using $\boldsymbol{\mu}^{(0)}$. Step 2 computes the temporal and CRBM potentials from previous frames in the history. Note that $p^{(t-w)}$ denotes the corresponding projection matrices from previous frames in the history. In step 4, the node, edge, temporal, and CRBM potentials are used together to update the parameters $\boldsymbol{\mu}^{(i)}$. The algorithm then iterates between updating the parameters $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\gamma}^{(i)}$, until either a maximum number of iterations is reached ($MaxIter$) or the parameters no longer change after updates.

In addition, we may use a parameter S to determine how many frames to "skip" in the history. For example, if $W = 3$ and $S = 2$, then from the previous 6 frames, every other frame is used in the history. Skipping frames may still allow us to model the temporal dependencies properly since there may not be a large change between consecutive frames $t-1$ and $t-2$. In addition, skipping frames in the history allows us to use a larger window of previous frames while still keeping the number of parameters tractable.

2.2. Learning

The STRF model is learned using a piecewise learning scheme. That is, the temporal SCRF and CRBM components are learned separately and then a scalar parameter λ is used to weight the contribution between them. In our experiments, we tried a variety of λ values between $\{0..1\}$ and chose λ based on which value performed best on the validation set. Note that in step 4 of Algorithm 2, the same κ_1, κ_2 parameters are used from Section 1.2. It is possible that jointly training all the model parameters may perform better than a piecewise model.

Algorithm 2 Mean-Field approximate inference for the STRF model at time t

1: Initialize $\boldsymbol{\mu}^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$ as follows:

$$\mu_{sl}^{(0)} = \frac{\exp\left(f_{sl}^{\text{node}}\right)}{\sum_{l'} \exp\left(f_{sl'}^{\text{node}}\right)}$$

$$\gamma_k^{(0)} = sigmoid\left(\sum_{r,l}\left(\sum_s p_{rs}\mu_{sl}^{(0)}\right) W_{rlk} + b_k\right)$$

where

$$f_{sl}^{\text{node}}\left(\mathcal{X}_{\mathcal{V}}, \{p_{sn}\}, \Gamma\right) = \sum_{n,d} p_{sn} x_{sd} \Gamma_{ndl}$$

2: Let $\boldsymbol{\alpha}$ be the mean-field estimates from the history

$$f_{sl}^{\text{tpot1}}\left(\boldsymbol{\alpha}; \mathcal{X}_{\mathcal{T}}^{(t,t-1)}, \Phi\right) = \sum_{b \in Int(V^{(t-1)}, s),} \sum_{l,l'=1}^{L} \sum_{e=1}^{D_{temp}} \alpha_{bl'}^{(t-1)} \Phi_{ll'e} x_{sbe}$$

$$f_{sl}^{\text{tpot2}}\left(\boldsymbol{\alpha}; \Pi\right) = \sum_{b \in \mathcal{V}^{(t-1)}} \sum_{l,l'=1}^{L} \alpha_{bl'}^{(t-1)} \Pi_{ll'} \left[s = b\right]$$

$$f_{sl}^{\text{crbm}}\left(\boldsymbol{\gamma}; \{p_{rs}\}, \mathbf{W}, \mathbf{C}, \mathbf{A}, \boldsymbol{\alpha}\right) = \sum_r p_{rs}\left(c_{rl} + \sum_k W_{rlk}\gamma_k + \sum_{q,w} A_{qwrl}\left(\sum_{s'} p_{rs'}^{(t-w)} \alpha_{s'l}^{(t-w)}\right)\right)$$

3: **for** i=0:*MaxIter* (or until convergence) **do**

4: update $\boldsymbol{\mu}^{(i+1)}$ as follows: $\mu_{sl}^{(i+1)} =$

$$\frac{\exp\left(f_{sl}^{\text{node}} + f_{sl}^{\text{edge}}\left(\boldsymbol{\mu}^{(i)}\right) + \kappa_1 f_{sl}^{\text{tpot1}} + \kappa_2 f_{sl}^{\text{tpot2}} + \lambda f_{sl}^{\text{crbm}}\left(\boldsymbol{\gamma}^{(i)}\right)\right)}{\sum_{l'} \exp\left(f_{sl'}^{\text{node}} + f_{sl'}^{\text{edge}}\left(\boldsymbol{\mu}^{(i)}\right) + \kappa_1 f_{sl'}^{\text{tpot1}} + \kappa_2 f_{sl'}^{\text{tpot2}} + \lambda f_{sl'}^{\text{crbm}}\left(\boldsymbol{\gamma}^{(i)}\right)\right)}$$

where

$$f_{sl}^{\text{edge}}\left(\boldsymbol{\mu}; \mathcal{X}_{\mathcal{E}}, \mathcal{E}, \Psi\right) = \sum_{j:(s,j)\in\mathcal{E}} \sum_{l',e} \mu_{jl'} \Psi_{ll'e} x_{sje}$$

5: update $\boldsymbol{\gamma}^{(i+1)}$ as follows:

$$\gamma_k^{(i+1)} = sigmoid\left(b_k + \sum_{r,l}\left(\sum_s p_{rs}\mu_{sl}^{(i+1)}\right) W_{rlk} + \sum_{w,r,l}\left(\sum_{s'} p_{rs'}^{(t-w)} \alpha_{s'l}^{(t-w)}\right) B_{wrlk}\right)$$

6: **end for**

3. Qualitative Results

We present an expanded version of the qualitative results shown in the main paper [1]. As described in the paper, there are 50 videos and for each video, we have 11 labeled, consecutive frames. For all six models in the paper, we show the results for every other frame (from time t through $t+10$) for five cases, in Tables 1-5.

In Table 1, models with hidden units (bottom four rows) tend to result in a cleaner label shape than models without hidden units, possibly due to the global shape prior. In addition to a shape prior, the STRF incorporates temporal

dependencies and results in the best overall label shape and consistency. In Table 2, STRF results in a significantly better overall labeling compared to other models as both the hair and skin shapes are more "filled out" and realistic. Table 3 shows a more subtle improvement made by the STRF model compared to other approaches.

Tables 4-5 show cases in which STRF may be propagating errors from previous frames. For example, STRF makes a mistake in Table 4 where the necktie region is consistently labeled as skin. In Table 5, the temporal potentials may be propagating an incorrect hair shape from previous frames. It is possible that information from future frames may be helpful in mitigating the effects of this error propagation. In the case of the example from Table 4, there may be a confident labeling in the future which may discourage the skin labeling around the necktie region and we would like to propagate this confident labeling to previous frames. The inference procedure can be revised to incorporate both forward and backward passes through the frames, which may lead to better labeling performance, but at the cost of complicating the inference and higher computation time.

References

[1] A. Kae, B. Marlin, and E. Learned-Miller. The shape-time random field for semantic video labeling. In *CVPR*, 2014.

[2] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller. Augmenting CRFs with Boltzmann machine shape priors for image labeling. In *CVPR*, 2013.

Model	t	t+2	t+4	t+6	t+8	t+10
Original						
Ground Truth						
SCRF						
SCRF + Temporal						
SCRF + RBM						
SCRF + RBM + Temporal						
SCRF + CRBM						
STRF						

Table 1. **Successful Case**. Models with hidden units (bottom four rows) tend to result in a cleaner label shape than models without hidden units. In particular, models with the CRBM (bottom two rows) have a more well-rounded, complete hair shape, while the STRF (bottom-most row), which incorporates both the CRBM and temporal potentials, has the best overall label shape that is temporally consistent.

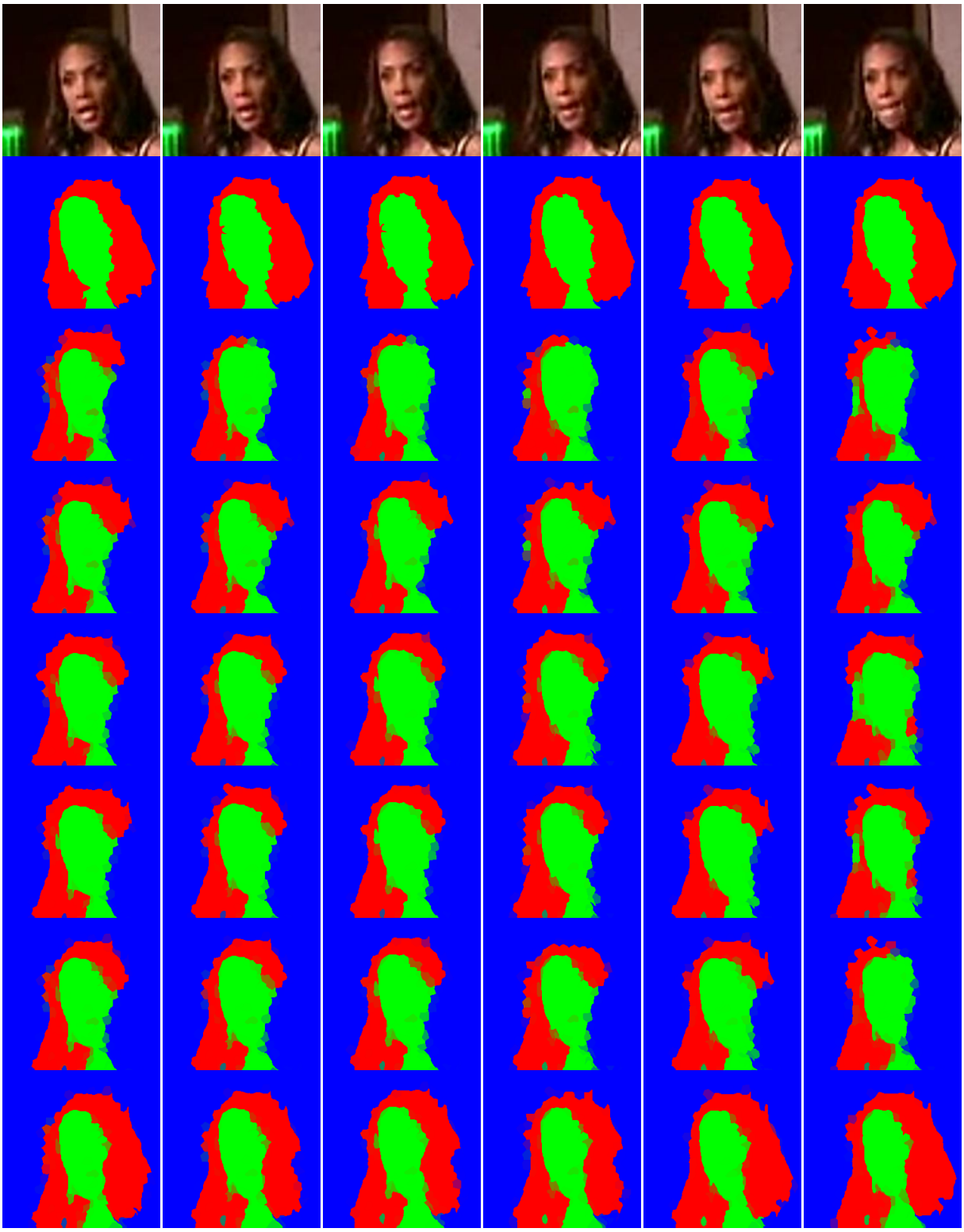
Model	t	t+2	t+4	t+6	t+8	t+10
Original						
Ground Truth						
SCRF						
SCRF + Temporal						
SCRF + RBM						
SCRF + RBM + Temporal						
SCRF + CRBM						
STRF						

Table 2. **Successful Case**. In this case, the STRF model results in a significantly better overall labeling as both the hair and skin shapes are more "filled out" and realistic, compared to the labelings by other models.

Model	t	t+2	t+4	t+6	t+8	t+10
Original						
Ground Truth						
SCRF						
SCRF + Temporal						
SCRF + RBM						
SCRF + RBM + Temporal						
SCRF + CRBM						
STRF						

Table 3. **Successful Case**. In this case, the STRF model makes a more subtle improvement compared to other models. While the other models might have a good labeling for one or two frames, STRF is more consistent overall. The small improvement in the hair labeling from STRF looks noticeably better than other labelings but this improvement may not result in a large improvement in accuracy.

Model	t	t+2	t+4	t+6	t+8	t+10
Original						
Ground Truth						
SCRF						
SCRF + Temporal						
SCRF + RBM						
SCRF + RBM + Temporal						
SCRF + CRBM						
STRF						

Table 4. **Failure Case**. In this case, the temporal consistency seems to result in worse performance for models like STRF and SCRF+Temporal, because an incorrect labeling may be propagated through time. In both of these models, the necktie region is consistently confused for the skin class.

Model	t	t+2	t+4	t+6	t+8	t+10
Original						
Ground Truth						
SCRF						
SCRF + Temporal						
SCRF + RBM						
SCRF + RBM + Temporal						
SCRF + CRBM						
STRF						

Table 5. **Failure Case**. In this case, both STRF and SCRF+Temporal consistently make an error in the hair shape. It is possible that this error in hair shape may be propagated through time by the temporal potentials.