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

Abstract

We propose a novel discriminative model for semantic labeling in videos by incorporating a prior to model both the shape and temporal dependencies of an object in video. A typical approach for this task is the conditional random field (CRF), which can model local interactions among adjacent regions in a video frame. Recent work [16, 14] has shown how to incorporate a shape prior into a CRF for improving labeling performance, but it may be difficult to model temporal dependencies present in video by using this prior. The conditional restricted Boltzmann machine (CRBM) can model both shape and temporal dependencies, and has been used to learn walking styles from motion-capture data. In this work, we incorporate a CRBM prior into a CRF framework and present a new state-of-the-art model for the task of semantic labeling in videos. In particular, we explore the task of labeling parts of complex face scenes from videos in the YouTube Faces Database (YFDB). Our combined model outperforms competitive baselines both qualitatively and quantitatively.

1. Introduction

The task of semantic labeling is an important problem in computer vision. Labeling semantic regions in an image or video allows us to better understand the scene itself as well as properties of the objects in the scene, such as their parts, location, and context. This knowledge may then be useful for tasks such as object detection or scene recognition.

Semantic labeling in video is particularly interesting to study because there is typically more information available in a video of an object than a static image of an object. For example, we can track the motion of an object in video and learn properties such as the way the object moves and interacts with its environment, which is more difficult to infer from a static image. In addition, there are many videos available online on sites such as YouTube, which make this analysis increasingly useful.

In this work, we focus on the semantic labeling of face videos into hair, skin, and background regions, as an intermediate step to modeling face structure. We build on recent work by [16, 14] that incorporated a label prior into a conditional random field (CRF) [15] model and showed improvement in labeling accuracy over a baseline CRF. In particular, they used a restricted Boltzmann machine (RBM) [23] to model label shape and combined this with a CRF to model local regions. This model accounts for both local and global dependencies within an image, but it may be difficult to account for temporal dependencies present in a video.

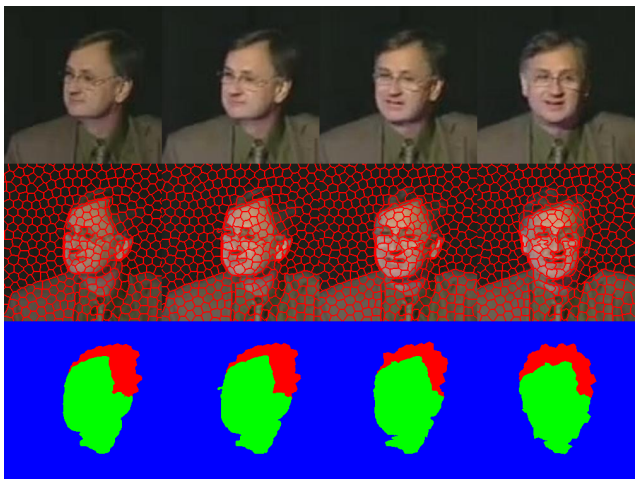


Figure 1. **YFDB Clip.** Rows correspond to 1) video frames, 2) superpixel segmentations, and 3) ground truth. Red represents hair, green represents skin, and blue represents background.

In order to model both shape and temporal dependencies, we use the conditional restricted Boltzmann machine (CRBM) [24], which is an extension of the RBM. The CRBM has been used to learn walking styles from motion-capture data and was able to generate novel, realistic motions. In our model, we incorporate the CRBM as a temporal shape prior into a CRF framework which already provides local modeling. We refer to this combined model as the Shape-Time Random Field (STRF).

Our main contributions are summarized as follows:

- The STRF model, a strong model for face labeling in

videos. STRF combines CRF and CRBM components to model local, shape, and temporal dependencies.

- Efficient inference and training algorithms for STRF.
- STRF outperforms competitive baselines, both qualitatively and quantitatively.
- The code and labeled data will be publicly available.

2. Related Work

The conditional random field (CRF) [15] has been widely used for tasks such as image labeling [8, 7, 22, 2] where nodes correspond to image regions (such as pixels or superpixels), and edges are added between adjacent regions. One straightforward way to extend the CRF to labeling in videos is to define temporal potentials between frames such as in [28], which is an approach adopted in our work.

There are several related works on using a restricted Boltzmann machine (RBM) [23] (or their deeper extensions) for shape modeling. He et al. [7] proposed multi-scale CRFs to model both local and global label features using RBMs. Specifically, they used multiple RBMs at different scales to model the regional or global label fields separately, and combined those conditional distributions multiplicatively. Salakhutdinov et al. [19] trained a DBM to learn and generate novel digits and images of small toys. Recently, Eslami et al. [4] introduced the Shape Boltzmann Machine (SBM) as a strong model of object shape, in the form of a modified DBM. The SBM was shown to have good generative performance in modeling simple, binary object shapes. They later extended the SBM to perform image labeling within a generative model [5]. There has also been work in combining hidden unit models such as the RBM within a discriminative model such as a CRF [16, 14] for the labeling task.

Because we are interested in modeling object shape over time, we use the conditional restricted Boltzmann machine (CRBM) by Taylor et al. [24]. The CRBM is an extension of the RBM with additional connections to a history of previous frames. They demonstrated that the CRBM can learn different motion styles from motion-captured data, and successfully generated novel motions. In this work, we incorporate the CRBM into a discriminative framework for semantic labeling in face videos.

Regarding the specific problem of hair, skin, background labeling, there have been several related works [21, 27, 26, 12, 14] in the literature. Scheffler et al. [21] learn separate color models for each of the hair, skin, background classes within a Bayesian framework. Wang et al. [27, 26] focus on hair labeling within a parts-based framework while Huang et al. [12] learn a CRF using color, texture and location features. This CRF model is used as a baseline in our work. While all of these approaches present new ways to label face images, none of them incorporate global shape modeling except for the GLOC model [14]. In this work we extend the GLOC model for semantic labeling in videos.

3. Models

In the following sections we introduce the Shape-Time Random Field (STRF) and its components. Note that the notation closely follows the notation used in [14].

Notation. We use the following definitions:

- A video v consists of $F^{(v)}$ frames, where $F^{(v)}$ can vary over different videos. Let each frame in video v be denoted as $v^{(t)}$ where $t \in \{1 \cdots F^{(v)}\}$.
- A video frame $v^{(t)}$ is pre-segmented into $S^{(v,t)}$ superpixels, where $S^{(v,t)}$ can vary over different frames. The superpixels represent the nodes in the graph for video v at time t.
- Let $\mathcal{G}^{(v,t)} = \{\mathcal{V}^{(v,t)}, \mathcal{E}^{(v,t)}\}$ denote the nodes and edges for the undirected graph of frame t in video v.
- Let $\mathcal{V}^{(v,t)} = \{1, \cdots, S^{(v,t)}\}$ denote the set of superpixel nodes for frame t in video v.
- Let $\mathcal{E}^{(v,t)} = \{(i,j) : i,j \in \mathcal{V}^{(v,t)}$ and i,j are adjacent superpixels in frame t in video $v\}$.
- Let $\mathcal{X}^{(v,t)} = \{\mathcal{X}_{\mathcal{V}}^{(v,t)}, \mathcal{X}_{\mathcal{E}}^{(v,t)}\}$ be the set of features in frame t in video v where
  - $\mathcal{X}_{\mathcal{V}}^{(v,t)}$ is the set of node features $\{\mathbf{x}_s^{(t)} \in \mathbb{R}^{D_n} : s \in \mathcal{V}^{(v,t)}\}$ for frame t in video v.
  - $\mathcal{X}_{\mathcal{E}}^{(v,t)}$ is the set of edge features $\{\mathbf{x}_{ij}^{(t)} \in \mathbb{R}^{D_e} : (i,j) \in \mathcal{E}^{(v,t)}\}$ for frame t in video v.
- Let $\mathcal{X}_{\mathcal{T}}^{(v,t,t-1)}$ be the set of temporal features $\{\mathbf{x}_{ab}^{(t,t-1)} \in \mathbb{R}^{D_{temp}} : a \in \mathcal{V}^{(v,t)}, b \in \mathcal{V}^{(v,t-1)}\}$ between adjacent frames $t, t-1$ in video v.
- Let $\mathcal{Y}^{(v,t)} = \{\mathbf{y}_s^{(v,t)} \in \{0,1\}^L, s \in \mathcal{V}^{(v,t)} : \sum_{l=1}^{L} y_{sl}^{(v,t)} = 1\}$ be the set of labels for the nodes in frame t in video v.

D_n, D_e, D_{temp} denote the dimensions of the node, edge, and temporal features, respectively, and L denotes the number of labels. In the rest of this paper, the superscripts "v", "node", and "edge" are omitted for clarity, but the meaning should be clear from context. The superscript t is also omitted except when describing interactions between frames.

The STRF model is shown in Figure 2. The top two layers correspond to a conditional restricted Boltzmann machine (CRBM) [24] with the (virtual) visible nodes colored orange and the hidden nodes colored green. The bottom two layers correspond to a CRF with temporal potentials. Note that if we consider the model at time t only and ignore the previous frames, we revert to the GLOC model from [14]. We now describe the components of the STRF model.

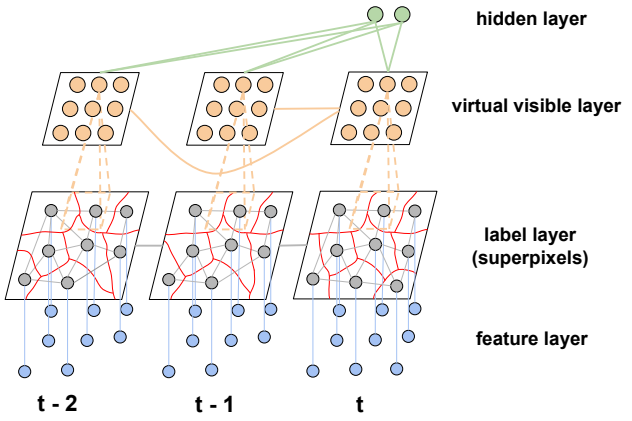


Figure 2. **High level view of the STRF model.** The model is shown for the current frame at time t and two previous frames. The dashed lines indicate the virtual pooling between the visible units of the CRBM and the superpixel label nodes. Parts of this model will be shown in more detail in subsequent figures.

3.1. RBM

The restricted Boltzmann machine (RBM) [23] is a generative model in which the nodes are arranged in a bipartite graph, consisting of a hidden layer and visible layer, as shown in Figure 3(a). In our model, superpixels are used as the base image representation, but superpixels can vary in shape and number from frame to frame. In order to map between superpixels and the fixed size grid of the RBM, we follow a *virtual pooling* approach from [14]. The pooling is shown in Figure 2, as the dashed line between the (virtual) visible layer and label layer. The projection matrix between the superpixels and the fixed grid of the RBM is defined as

$$p_{rs} = \frac{Area(Region(s) \cap Region(r))}{Area(Region(r))}, \qquad (1)$$

where r is the index for the visible units in the RBM and s is the index for superpixels. $Region(r)$ and $Region(s)$ refer to the pixels corresponding to the visible unit r and superpixel s, respectively. The energy between the label nodes and the hidden nodes for an image is defined as

$$E_{\text{rbm}}(\mathcal{Y}, \mathbf{h}) = -\sum_{r=1}^{R^2}\sum_{l=1}^{L}\sum_{k=1}^{K} \bar{y}_{rl} W_{rlk} h_k - \sum_{k=1}^{K} b_k h_k - \sum_{r=1}^{R^2}\sum_{l=1}^{L} c_{rl}\bar{y}_{rl}, \qquad (2)$$

where the virtual visible nodes $\bar{y}_{rl} = \sum_{s=1}^{S} p_{rs} y_{sl}$ are deterministically mapped from the label layer by multiplying with the projection matrix $\{p\}$ from Equation (1). In addition, there are R^2 multinomial visible units, L labels, and K hidden units. $\mathbf{W} \in \mathbb{R}^{R^2 \times L \times K}$ represent the pairwise weights between the hidden units h and the visible units y, and b, c represent the biases for the hidden units and multinomial visible units, respectively. The model parameters W, b, c are trained using contrastive divergence [9].

3.1.1 CRBM

While the RBM can be used to model the label shape within a particular frame of video, it may be inefficient at modeling temporal dependencies in the video. The conditional restricted Boltzmann machine (CRBM)[24] is an extension of the RBM that uses previous frames in a video to act as a dynamic bias for the hidden units in the current frame. The CRBM energy at time t is defined as:

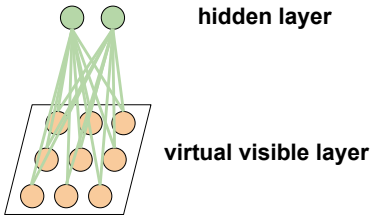
$$E_{\text{crbm}}\left(\mathcal{Y}^{(t,<t)}, \mathbf{h}^{(t)}\right) = E_{\text{rbm}}\left(\mathcal{Y}^{(t)}, \mathbf{h}^{(t)}\right) - \sum_{w=1}^{W}\sum_{r=1}^{R^2}\sum_{l=1}^{L}\sum_{k=1}^{K} \bar{y}_{rl}^{(t-w)} B_{wrlk} h_k^{(t)} - \sum_{q=1}^{Q^2}\sum_{w=1}^{W}\sum_{r=1}^{R^2}\sum_{l=1}^{L} \bar{y}_{qrl}^{(t-w)} A_{qwrl} \bar{y}_{rl}^{(t)}, \qquad (3)$$

which includes the RBM energy $E_{\text{rbm}}\left(\mathcal{Y}^{(t)}, \mathbf{h}\right)$ defined earlier in Equation (2). The W frames before the current frame t act as the "history", which is always conditioned on at time t. Following the notation in [24], $\mathcal{Y}^{(<t)}$ refers to the labels of the W previous frames before the current frame. $A \in \mathbb{R}^{Q^2 \times W \times R^2 \times L}$ represent the weights between visible units in the history to the current visible units at time t and $B \in \mathbb{R}^{W \times R^2 \times L \times K}$ represent the weights between visible units in the history to the hidden units. Note that there is a dense connection between the hidden units h and the visible layer at each time step. If each time step is considered independently, this corresponds to an RBM, as shown in more detail in Figure 3(a).

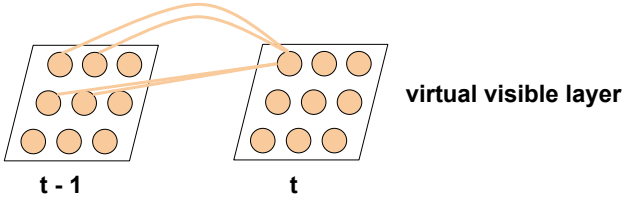
The hidden units h are densely connected to visible units $\bar{y}$ in both the current frame and in the history because the hidden units are meant to model changes in object shape across time. However, connections between visible units in the history and visible units in the current frame act more as temporal smoothing and so the interactions are likely to be more local. Thus, each visible unit $\bar{y}_{rl}^{(t)}$ at time t is only connected to a local neighborhood Q of visible units in previous frames. Figure 3(b) shows this local modeling for a single visible unit. By modeling only the local interactions between visible units (instead of using a dense connection), we also significantly reduce the number of parameters.

The main differences between the usage of the CRBM in our model compared to its original usage [24] are:

- Our CRBM is used within a discriminative framework for labeling. It is not meant to generate realistic data,



(a) Hidden layer to visible layer connections at each time step.



(b) Visible layer connections from t to $t-1$, shown for a single visible unit. The visible unit at time t in the upper-left corner is connected only to a local neighborhood of size Q from the previous frame.

Figure 3. Components of the CRBM.

but rather to complement the local modeling provided by the CRF and help improve labeling performance.

- Our CRBM models the label shape across time, and does not model the observed features directly (which is the case in the original usage of the CRBM).
- In our model, the visible units at time t frame are connected to a local neighborhood (of size Q) of the visible units in the history. In contrast, the original CRBM has a dense connection between the visible units at time t and the visible units in the history.

3.2. CRF

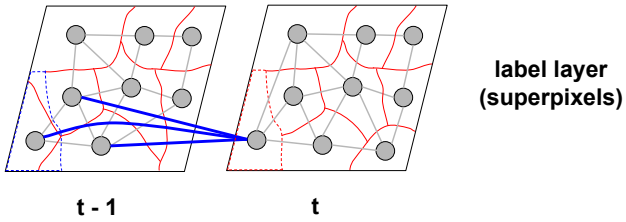
The conditional random field (CRF) [15] is a discriminative model which is used as both a baseline and a component for our later models. For the task of semantic labeling, a variant of the CRF, called the spatial CRF (SCRF) was found to outperform the CRF empirically, as described in [14]. The SCRF overlays an $N \times N$ grid on top of the image and learns a different set of node weights for each grid position. The energy of the SCRF is defined as

$$E_{\text{scrf}}(\mathcal{Y}, \mathcal{X}) = E_{\text{node}}\left(\mathcal{Y}, \mathcal{X}_{\mathcal{V}}\right) + E_{\text{edge}}\left(\mathcal{Y}, \mathcal{X}_{\mathcal{E}}\right), \quad (4)$$

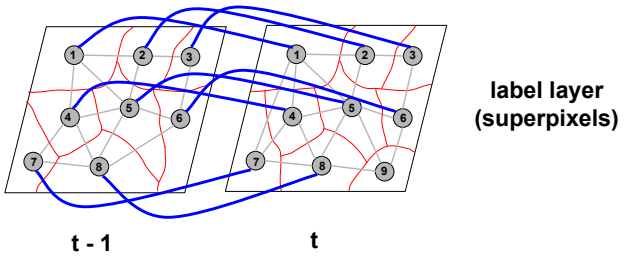
$$E_{\text{node}}\left(\mathcal{Y}, \mathcal{X}_{\mathcal{V}}\right) = -\sum_{s \in \mathcal{V}} \sum_{l=1}^{L} y_{sl} \sum_{n=1}^{N^2} p_{sn} \sum_{d=1}^{D_n} \Gamma_{ndl} x_{sd}, \quad (5)$$

$$E_{\text{edge}}\left(\mathcal{Y}, \mathcal{X}_{\mathcal{E}}\right) = -\sum_{(i,j) \in \mathcal{E}} \sum_{l,l'=1}^{L} \sum_{e=1}^{D_e} y_{il} y_{jl'} \Psi_{ll'e} x_{ije}, \quad (6)$$

where $\Gamma \in \mathbb{R}^{N^2 \times D_n \times L}$ represent the node weights and $\Psi \in \mathbb{R}^{L \times L \times D_e}$ represent the edge weights. A projection matrix



(a) **Position Smoothness.** Temporal potential incorporating position between frames t and $t-1$, shown for a single label node.



(b) **Superpixel Smoothness.** Temporal potential incorporating the TSP ID between frames t and $t-1$.

Figure 4. Temporal potentials used in the temporal SCRF.

$\{p\}$ is used to map between the superpixels and the grid, in a similar way to the virtual pooling in the RBM[1]. We use mean-field approximate inference [20] along with LBFGS optimization from minFunc [1] for learning the weights.

3.2.1 Temporal SCRF

One way to extend a traditional CRF for labeling in videos is to incorporate temporal potentials, which has been applied to the labeling task [28, 6, 29]. In our model, temporal potentials look only at the previous frame and are used to encourage smoothing between adjacent frames in a video in much the same way that edge potentials encourage spatial smoothing within an image. Two types of temporal potentials are used:

1) Position smoothness: This potential encourages a consistent labeling between superpixels in adjacent frames that are approximately in the same position and have similar appearance. The energy is defined as

$$E_{\text{tpot1}}\left(\mathcal{Y}^{(t,t-1)}, \mathcal{X}_{\mathcal{T}}^{(t,t-1)}\right) =$$

$$-\sum_{a \in \mathcal{V}^{(t)}} \sum_{b \in Int(V^{(t-1)}, a)} \sum_{l,l'=1}^{L} \sum_{e=1}^{D_{temp}} y_{al}^{(t)} y_{bl'}^{(t-1)} \Phi_{ll'e} x_{abe}^{(t,t-1)},$$

where $\Phi \in \mathbb{R}^{L \times L \times D_{temp}}$ represent the temporal weights, and $Int(\mathcal{V}^{(t-1)}, a)$ refers to superpixels in frame $t-1$ that intersect with superpixel a in the current frame. Thus, only

[1]Note that this projection matrix can be different from the one used by the RBM in Equation (2).

superpixels that intersect with superpixel a in the previous frame are included in this potential. Figure 4(a) shows the connections for this temporal potential for a single superpixel node. The figure shows the superpixel in the lower-left corner at time t and its projection at time $t-1$ (shown in dotted blue lines). At time $t-1$, there are three superpixels that are intersected by the dotted blue lines. Thus, there are connections from these three superpixels at time $t-1$ to the superpixel at time t, shown by the solid blue lines.

2) Superpixel smoothness: Temporal superpixels (TSP) [3] are used to segment the frames in a video. They have the desirable property of maintaining their position on an object through time. For example, a TSP on a person's cheek will stay "stuck" to the cheek as long as the person's pose does not change significantly (i.e. the person does not move their head). For our task, these TSPs have been found empirically to be very pure in the sense that a TSP tends to remain a single label for most of its lifetime. The following temporal potential is used to encourage consistent labeling between the same TSPs in adjacent frames,

$$E_{\text{tpot2}}\left(\mathcal{Y}^{(t,t-1)}\right) = - \sum_{a \in \mathcal{Y}^{(t)}} \sum_{b \in \mathcal{Y}^{(t-1)}} \sum_{l,l'=1}^{L} y_{al}^{(t)} y_{bl'}^{(t-1)} \Pi_{ll'} \left[a = b\right],$$

where $\Pi \in \mathbb{R}^{L \times L}$ represent the temporal weights and $[a = b]$ denotes indicator notation checking whether superpixel a is equal (i.e. has the same TSP ID) to superpixel b. Figure 4(b) shows the connections of this temporal potential at time t and $t-1$. Note that superpixels 1-8 exist at both time t and $t-1$, and thus there is a connection (indicated by blue lines) between a superpixel at time $t-1$ to its corresponding superpixel at time t. However, superpixel 9 is "created" at time t and therefore there is no connection from the previous frame.

Incorporating these temporal potentials, the energy for the temporal SCRF model is defined as

$$E_{\text{tscrf}}(\mathcal{Y}^{(t,t-1)}, \mathcal{X}^{(t,t-1)}) = E_{\text{scrf}}\left(\mathcal{Y}^{(t)}, \mathcal{X}^{(t)}\right) + E_{\text{tpot1}}\left(\mathcal{Y}^{(t,t-1)}, \mathcal{X}_{\mathcal{T}}^{(t,t-1)}\right) + E_{\text{tpot2}}\left(\mathcal{Y}^{(t,t-1)}\right), \quad (7)$$

where the SCRF energy defined in Equation (4) is simply augmented by the temporal potentials.

Inference. Inference in the temporal SCRF is similar to inference in the SCRF. There is not much additional cost for inference because the labels for the previous frame are assumed fixed, and thus the temporal potentials only need to be computed once. For the first frame (time $t = 1$), the SCRF is used for inference. Afterward the temporal potentials are computed from the previous frame and then included as an additional set of potentials for the label nodes in the current frame. Additional details about inference can be found in the supplementary material.

Learning. The temporal SCRF is learned using piecewise learning in which scalar parameters κ_1, κ_2 are used to weight the contribution of the two temporal potentials, respectively. 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.

3.3. Shape-Time Random Field

The STRF model is a combination of the temporal SCRF and CRBM components defined earlier. The conditional distribution and energy at time t are defined as

$$P_{\text{strf}}(\mathcal{Y}^{(t)} | \mathcal{Y}^{(<t)}, \mathcal{X}^{(t,t-1)}) \propto \sum_{\mathbf{h}^{(\mathbf{t})}} \exp\left(-E_{\text{strf}}(\mathcal{Y}^{(t,<t)}, \mathcal{X}^{(t,t-1)}, \mathbf{h}^{(\mathbf{t})})\right),$$

$$E_{\text{strf}}\left(\mathcal{Y}^{(t,<t)}, \mathcal{X}^{(t,t-1)}, \mathbf{h}^{(t)}\right) = E_{\text{tscrf}}\left(\mathcal{Y}^{(t,t-1)}, \mathcal{X}^{(t,t-1)}\right) + E_{\text{crbm}}\left(\mathcal{Y}^{(t,<t)}, \mathbf{h}^{(t)}\right),$$

where $\mathcal{Y}^{(<t)}$ refers to the labels in the history, which is assumed fixed at time t. The model is shown in Figure 2. The CRBM (top two layers) provides a dynamic bias for the hidden units, based on previous history, to help with the temporal SCRF label classification (bottom two layers).

Inference. We adopt a feed-forward inference procedure in which inference of the labels $\mathcal{Y}^{(t)}$ at time t involves only a history of the previous W frames. This approach is computationally efficient since the history is fixed at time t, and so the only latent variables are the hidden units of the CRBM and the labels. During inference, the first W frames are computed using the GLOC [14] model. Afterward, inference proceeds in a sliding window fashion as the history is used to compute CRBM potentials that augment the existing potentials from the temporal SCRF. A mean-field approximation for inference of the label nodes is used in which we alternately sample between the hidden units and labels at time t (more detail in the supplementary material).

Learning. The STRF model is learned using piecewise learning in which the temporal SCRF and CRBM components are each 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.

4. Data

Our models are evaluated on videos from the YouTube Faces Database (YFDB) [30], which is a large database of "real world" videos of faces found on YouTube, and not taken from a controlled, laboratory environment. Videos

from YFDB contain a wide variety of motions, hair/skin shapes, lighting conditions and occlusions, making them challenging to label. An example of a video and its corresponding labeling is shown in Figure 1.

Aligning an object in an image into a canonical position as a pre-processing step has been found to be helpful for tasks such as face recognition [10]. For the face videos from YFDB, we tried several alignment approaches: (1) a pre-learned deep funnel using the method of [11], and (2) a pre-learned SIFT-congealed funnel using the approach of [10] and (3) the YFDB-provided alignment. Both funnels were pre-learned on LFW images[2] and alignment was performed for each video frame independently.

However, these approaches generally result in an unstable, coarse alignment. In many cases there are significant scale differences between frames and other transformation instabilities. Therefore, we resorted to a simpler approach to avoid using an unstable alignment. We used the output of the Viola Jones face detector [25], but fixed the height and width of the detected face box to the mean height and width of the detected face boxes for all frames in the video. Then, for each frame, a bounding box for the face is cropped out from the center of the face detection (provided by YFDB), using the dimensions of the mean width and height for the video. Following the process of LFW [13], the bounding box is expanded by a factor of 2.2 in each direction and then resized to 250 $\times$ 250 pixels. This simple fix tends to produce a stable, temporally smooth set of frames. Afterward, we use temporal superpixels (TSP) [3] to segment the video frames (there are about 300-400 superpixels per frame).

Features. The same set of features is used as in [12, 14]. The node features are:

- **Color**: Normalized histogram over 64 bins generated by running K-means over pixels in LAB space. Each pixel is assigned to its closest centroid and a normalized histogram is computed using all the pixel assignments within a superpixel.
- **Texture**: Normalized histogram over 64 textons which are generated according to [17]. Each pixel is assigned to a texton and a normalized histogram is computed for all the pixel assignments within a superpixel.

The following edge features are computed between a pair of adjacent superpixels:

- **Probability of Boundary (Pb)** [18]: Sum of Pb values between adjacent superpixels.
- **Color**: L2-distance between color histograms for adjacent superpixels.
- **Texture**: Chi-squared distance between texture histograms for adjacent superpixels.

[2]The SIFT-congealed funnel was used to attain the image alignments for the previous GLOC [14] experiments.

These edge features are also used for the temporal potentials between adjacent frames, except for the Pb feature. It is unclear how to incorporate the Pb feature, which is defined spatially within a frame, in this temporal manner.

5. Experiments

In our experiments we chose 50 videos randomly from YFDB and labeled a "chunk" of 11 consecutive frames per video. We manually labeled all chunks into hair, skin, or background regions, resulting in a total of 550 labeled frames. The labeled data is then divided into 5 disjoint sets for use in cross validation. For each split, 3 of the folds are used for training, 1 for validation and 1 for testing. There is only one instance of each person in the 50 videos, so the same person is never used for training and testing. Below, we describe the progression of models from the baseline SCRF to the STRF model, and show results in Table 1.

- **SCRF**. Since each training split contains only 330 images, an additional 500 labeled images are added from the Part Labels Database[3]. This database contains hair, skin, background labeled images from LFW [13].
- **SCRF + Temporal**. Temporal potentials are added into the SCRF by learning tradeoff parameters κ_1, κ_2 from the validation fold.
- **SCRF + RBM (GLOC)** [14]. This model is trained piecewise in which the SCRF and RBM components are trained separately and then combined together using a tradeoff parameter λ found from validation. We also trained a joint GLOC model[4] using default parameters (again adding 500 training images to each training fold) but this did not perform as well as the piecewise GLOC model. It is possible that GLOC may be sensitive to the choice of hyperparameters, which may have contributed to this drop in performance.
- **SCRF + RBM + Temporal**. Temporal potentials are added to GLOC, using the same tradeoff parameters $\lambda, \kappa_1, \kappa_2$, discussed earlier.
- **SCRF + CRBM**. A tradeoff parameter λ is used to weight the CRBM and SCRF components.
- **STRF**. The temporal potentials and CRBM components are combined by reusing the same parameters from previous models.

Specific parameters in our experiments are: $K = 400, R = 32, N = 16, Q = 3$. Parameters such as W can vary for each cross validation split depending on which values performed best on the validation fold. For each video frame, STRF takes about 0.28 (sec) for inference on an Intel i7.

Results. Table 1 shows the results of cross-validation for the following metrics (with respect to superpixels).

[3]vis-www.cs.umass.edu/lfw/part_labels/

[4]Code from vis-www.cs.umass.edu/GLOC/index.html

Model	Error Reduction	Overall Accuracy	Hair	Skin	BG	Category Average
SCRF	0	0.902 ± 0.005	0.629 ± 0.047	***0.891 ± 0.025***	***0.958 ± 0.005***	0.826 ± 0.009
SCRF + Temp	***0.034 ± 0.034***	0.905 ± 0.006	***0.698 ± 0.038***	***0.878 ± 0.028***	***0.953 ± 0.006***	***0.843 ± 0.007***
SCRF + RBM [14]	0.028 ± 0.025	0.905 ± 0.006	0.608 ± 0.038	***0.900 ± 0.023***	***0.963 ± 0.003***	0.824 ± 0.008
SCRF + RBM + Temp	***0.096 ± 0.018***	***0.911 ± 0.006***	0.655 ± 0.027	**0.901 ± 0.022**	**0.964 ± 0.004**	0.840 ± 0.006
SCRF + CRBM	0.036 ± 0.015	0.905 ± 0.005	0.632 ± 0.046	***0.894 ± 0.023***	***0.961 ± 0.004***	0.829 ± 0.010
STRF	**0.123 ± 0.025**	**0.914 ± 0.006**	**0.720 ± 0.039**	***0.889 ± 0.025***	***0.959 ± 0.004***	**0.856 ± 0.010**

Table 1. **Labeling performance.** All metrics are with respect to superpixels. For each model, the mean and standard error of the mean (SEM) are given for each metric (from cross-validation). For each metric, the result in **blue** indicates the best performing model and results in ***italics*** indicate models with performances not statistically significantly different than the best model at the $p = 0.05$ level as measured by a two-sided paired t-test. Numbers in regular typeface indicate results that are significantly different from the best model.

- **Error reduction**: Error reduction in overall superpixel accuracy with respect to the SCRF.
- **Overall accuracy**: Number of superpixels classified correctly divided by total number of superpixels.
- **Category-specific accuracy**: For each class, the number of superpixels classified correctly divided by the total number of superpixels.
- **Category average**: Average of the category-specific accuracies.

The mean and standard error of the mean (SEM) for these metrics are reported for all models. In addition, we computed two-sided paired t-tests for STRF compared with all other models. With the exception of the SCRF + Temporal and SCRF + RBM + Temporal models, STRF results in significant improvements over other models for the following metrics: error reduction, overall accuracy, hair accuracy, and category average. We note that for these metrics, STRF still outperformed the SCRF + Temporal and SCRF + RBM + Temporal models in terms of the mean scores. For the skin and background classes, the top performing model, SCRF + RBM + Temporal, is not significantly different than the other models.

Qualitative results are shown in Figure 5 for two video clips (more qualitative results are in the supplementary material). In the first case, the SCRF guesses can vary significantly from frame to frame, possibly due to a lack of temporal smoothing or a global shape prior. SCRF + RBM results show more consistency but still contain errors. STRF, which incorporates both temporal and shape dependencies, results in the best overall label shape and consistency. In the second case, 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.

In some cases, STRF may be propagating errors from previous frames. It is possible that information from future frames may be helpful in mitigating the effects of this error propagation. We can revise revise our inference procedure 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.

Voting. A "voting" approach may be used as a post-processing step after a model such as STRF generates its label guesses. The majority label guess of a TSP is used as the label guess in all frames covered by the TSP, which may help to smooth the labels using information across the video. This "voting" approach used on the STRF guesses results in a small 0.07% improvement in overall superpixel accuracy over STRF. The disadvantage of this approach is that it may require inference on the entire video depending on the coverage of the TSPs.

CRBM Filters. Some of the learned weights (or filters) from the CRBM are shown in Figure 6. Each row in the figure corresponds to the filters for a particular hidden unit. The history weights B are shown to the left of the white line and the corresponding pairwise weights W are shown to the right. Note that in Figure 2 there are two previous time steps used as history, but the filters in Figure 6 show three previous time steps. The history weights seem to learn some of the pose and overall label shape of the corresponding pairwise weights.

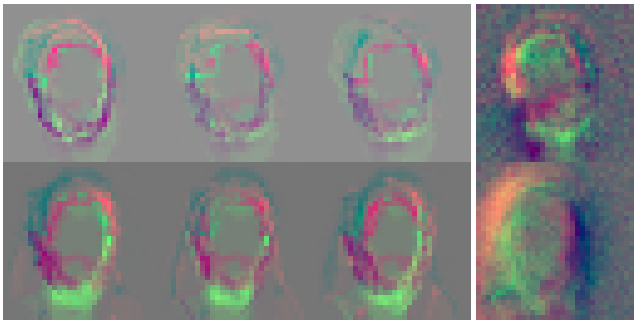


Figure 6. **Sample of learned history weights B and pairwise weights W.** History weights B are shown to the left of the white line and corresponding pairwise weights W are shown to the right. Each row corresponds to the $\{B, W\}$ weights of a particular hidden unit in the CRBM. The strength of hair weights is shown in red and the strength of skin weights is shown in green.

Discussion. The task of labeling face regions in videos is challenging due to the variety of hair and skin appearances

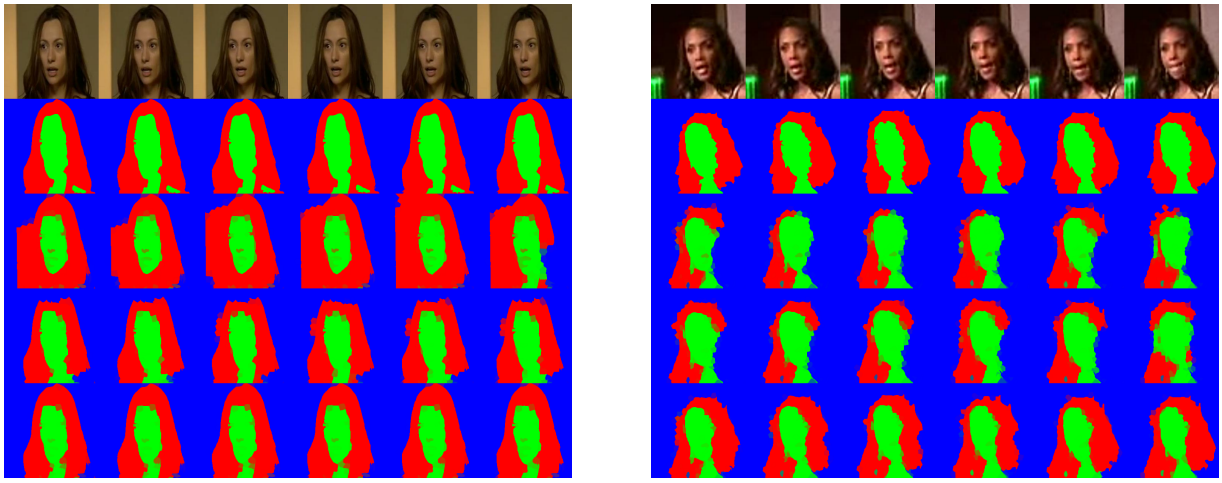


Figure 5. **Qualitative results**. We show two cases where STRF outperforms baselines. In both cases, every other frame from a labeled chunk is shown. The rows correspond to 1) original video frames, 2) ground truth, 3) SCRF, 4) SCRF + RBM, and 5) STRF .

and shapes, complex motions of faces, as well as difficult lighting conditions and occlusions. For this task, we presented the Shape-Time Random Field (STRF) which incorporates both shape and temporal dependencies into a discriminative framework for semantic labeling in video. We discussed efficient inference and learning techniques using STRF and demonstrated both quantitative and qualitative improvements over competitive baseline models.

Acknowledgments. We thank Marwan Mattar for helpful discussions and we thank our reviewers for their constructive feedback.

References

[1] www.di.ens.fr/~mschmidt/Software/minFunc.html.

[2] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *CVPR Workshop on Perceptual Organization in Computer Vision*, 2004.

[3] J. Chang, D. Wei, and J. W. F. III. A Video Representation Using Temporal Superpixels. In *CVPR*, 2013.

[4] S. M. A. Eslami, N. Heess, and J. Winn. The shape Boltzmann machine: A strong model of object shape. In *CVPR*, 2012.

[5] S. M. A. Eslami and C. K. I. Williams. A generative model for parts-based object segmentation. In *NIPS*, 2012.

[6] G. Floros and B. Leibe. Joint 2d-3d temporally consistent semantic segmentation of street scenes. In *CVPR*, 2012.

[7] X. He, R. Zemel, and M. Carreira-Perpinán. Multiscale conditional random fields for image labeling. In *CVPR*, 2004.

[8] X. He, R. Zemel, and D. Ray. Learning and incorporating top-down cues in image segmentation. In *ECCV*, 2006.

[9] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[10] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007.

[11] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller. Learning to align from scratch. In *NIPS*, 2012.

[12] G. B. Huang, M. Narayana, and E. Learned-Miller. Towards unconstrained face recognition. In *CVPR Workshop on Perceptual Organization in Computer Vision*, 2008.

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

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

[15] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[16] Y. Li, D. Tarlow, and R. Zemel. Exploring compositional high order pattern potentials for structured output learning. In *CVPR*, 2013.

[17] J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *ICCV*, 1999.

[18] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using brightness and texture. In *NIPS*, 2002.

[19] R. Salakhutdinov and G. Hinton. Deep Boltzmann machines. In *AISTATS*, 2009.

[20] L. Saul, T. Jaakkola, and M. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.

[21] C. Scheffler, J. Odobez, and R. Marconi. Joint adaptive colour modelling and skin, hair and clothing segmentation using coherent probabilistic index maps. In *BMVC*, 2011.

[22] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 2009.

[23] P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1:194–281, 1986.

[24] G. W. Taylor, G. E. Hinton, and S. Roweis. Modeling human motion using binary latent variables. In *NIPS*, 2006.

[25] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 2004.

[26] N. Wang, H. Ai, and S. Lao. A compositional exemplar-based model for hair segmentation. In *ACCV*, 2011.

[27] N. Wang, H. Ai, and F. Tang. What are good parts for hair shape modeling? In *CVPR*, 2012.

[28] Y. Wang and Q. Ji. A dynamic conditional random field model for object segmentation in image sequences. In *CVPR*, 2005.

[29] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In *ECCV ,European Conference on Computer Vision*, 2008.

[30] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011.