Human Instance Segmentation from Video using Detector-based Conditional Random Fields

Vibhav Vineet[1]
vibhav.vineet-2010@brookes.ac.uk
Jonathan Warrell[1]
jwarrell@brookes.ac.uk
L'ubor Ladický[2]
lubor@robots.ox.ac.uk
Philip H. S. Torr[1]
philiptorr@brookes.ac.uk

[1] Oxford Brookes University
Oxford, UK
cms.brookes.ac.uk/research/visiongroup/

[2] Oxford University, Oxford, UK

This work is supported by the EPSRC and the IST Programme of the European Community, under the PASCAL2 Network of Excellence. P. H. S. Torr is in receipt of Royal Society Wolfson Research Merit Award.

Abstract

In this work, we propose a method for instance based human segmentation in images and videos, extending the recent detector-based conditional random field model of Ladicky et.al. Instance based human segmentation involves pixel level labeling of an image, partitioning it into distinct human instances and background. To achieve our goal, we add three new components to their framework. First, we include human parts-based detection potentials to take advantage of the structure present in human instances. Further, in order to generate a consistent segmentation from different human parts, we incorporate shape prior information, which biases the segmentation to characteristic overall human shapes. Also, we enhance the representative power of the energy function by adopting exemplar instance based matching terms, which helps our method to adapt easily to different human sizes and poses. Finally, we extensively evaluate our proposed method on the Buffy dataset with our new segmented ground truth images, and show a substantial improvement over existing CRF methods. These new annotations will be made available for future use as well.

1 Introduction

Ever since computer vision researchers embarked on developing algorithms for complete scene understanding, understanding human activities has been an important goal [5, 17]. More importantly, human activity recognition has found a niche in security and video surveillance and most of the related algorithms [5] depend on recovering human shapes and structures in the images. Further, accurate human segmentation can help in developing better systems for white balancing [16]. Another prominent application area is video object-cutout and paste [2, 15], where human instances from one environment can be segmented, and pasted onto another environment. Thus, knowledge of pixel level human segmentation adds to the potential of a range of applications, and we focus on this problem in our current work.

BMVC 2011 http://dx.doi.org/10.5244/C.25.80

The task of human instance segmentation has been studied extensively at the level of human tracking [21], pose estimation [7] and parts detection [12]. However, the existing methods face several challenges including difficult optimization problems, large variation in the appearance of humans and environments, and the multitude of combinations of different human parts and poses. Traditional segmentation methods [20] fail on this task as they embody only class information, and thus can not distinguish particular instances e.g., pixels on two distinct individuals would all, at best, get the same label "person". Further, the output of detection methods [4] is typically too sparse to generate the pixel-level accuracy necessary for the tasks above.

Over the years, a plethora of methods have been designed to generate consistent human segmentation, detection, parts detection and pose estimation in images and videos. A. H. Vela [22] proposes an iterative grabcut based [18] method to segment humans from images and videos. They start by applying human and face detectors to delineate the search space and generating an initial guess of the color model for the human from the face and skin color. They iteratively apply grabcut in the region to update the color model and refine the output. However, this method builds on the initial guess from the body and face detectors which sometimes fail to locate these parts in complex environments. Another associated disadvantage is the absence of a good set of representative features, and general human model. Several other works propose ways to improve the human body parts detection and pose estimation at reasonable computation cost. Ferrari et.al [7] progressively reduces the search space containing the correct hypothesis of body parts, increasing the complexity in features, thus reducing the overall computational complexity. Snapp et.al. [19] learn a sequence of structured models on a coarse to fine scale of images by pruning the hypotheses at the coarser levels and increasing the feature complexities trying to achieve a better hypothesis at reasonable cost. Although these pose estimation methods provide knowledge about the presence of human at a location, they generally fail to achieve pixel level segmentation.

Dalal and Triggs [4] provide an efficient discriminative classifier based on histogram of gradients features and linear SVMs for detecting humans. Felzenswalb et.al. [6] propose a deformable parts-based model to detect complex structured objects. These detection based methods enclose the objects within a bounding box but they fail to perform pixel level segmentation. The need for features capable of providing a pixel-level semantic labeling of an image led to Textonboost [20] which provides a boosting based discriminative object model, using a combination of features based on texture, location and contextual information. Ladicky et.al. [13] use these features and propose a hierarchical CRF based method to take advantage of combining pixel, super-pixel and region level information. These methods prove instrumental in providing object-class segmentation but fail to achieve an instance based segmentation of objects. Further, several works [8, 11, 14] integrate results from detection and segmentation in CRF framework to achieve complete scene understanding. Ladicky et.al. [14] present a principle way of integrating results from detection and segmentation in a single CRF framework. Their method can be used to recover each individual instance of an object. However, they fail to integrate parts-based detection in their framework.

We propose a method which addresses these problems. Succinctly, our main contributions are, 1) An adaptation of the detector based CRF framework of [14] to the problem of human instance segmentation. We extend the method so as to be able to integrate parts-based detection potentials. 2) We provide general techniques for integrating instance level information for structured objects such as humans into this framework. Specifically, we propose to include a shape prior information which biases the segmentation to typical human shape. Further, we match a test instance with exemplar instance models so as to easily incorporate

sensitivity to different scales, poses and identities. 3) We provide a pixel-level human/non-human segmentation on a set of images to generate training and test images from the TV series, "Buffy: the Vampire Slayer" for our evaluation purpose. We will provide these segmented ground truth images for future research purpose as well. We extensively evaluate our methods on this dataset and show both qualitative and quantitative improvements over a current state-of-the-art method.

We organize our paper as follow: Section 2 outlines the methods adopted, with subsections giving details on individual components. Experimental setup and results are discussed in section 3 and we conclude the paper in section 4.

2 Methods

We begin by outlining how we formulate the problem of human instance segmentation in a conditional random field (CRF) framework in section 2.1, outlining the general form of the energies we use. We then give the details of the separate components of these energies in Sec. 2.2-2.5. Our training and inference techniques are then discussed in Sec. 2.6.

2.1 Problem formulation: Human Instance Segmentation in Video

We formulate the problem of segmenting human instances in video as a labeling problem defined in a conditional random field (CRF) framework. This problem is modeled as an optimization problem where the goal is to achieve the global minimum of an energy or cost function defined over the CRF. At test time, we are given video frames $\mathbf{z}^t$, where $t \in \{1...T\}$ is a time index. Our task is to produce a labeling for each image, $\mathbf{x}^t$, defined across pixels $p = 1...P$, where $\mathbf{x}^t \in \mathcal{L}_t^{|P|}$, $\mathcal{L}_t$ being the label set for image (frame) t. The label sets take the form $\mathcal{L}_t = \{0, 1, ..., L^t\}$, the semantics of which are that 0 implies background, and labels $1, ..., L^t$ correspond to different human instances within the image. The assignment of instances to particular labels is arbitrary within each frame (e.g. the same person need not receive the same label across frames), and the value of L^t is set individually for each frame from separate detector responses, as will be described below.

Model 0: Our approach is based on the detector-CRF framework of [14]. We begin by directly adapting the energy used in [14] to our problem. We will refer to this as 'model 0' (all energy terms are implicitly conditioned on the test images $\mathbf{z}$):

$$E_{\mathrm{M0}}(\mathbf{x}) = \sum_t [E_{\mathrm{pix}}(\mathbf{x}^t) + \sum_{l=1...L^t} \psi_{\mathrm{d}}(\mathbf{x}_{c^t_{d(l)}}, H_{d(l)}, l)] \quad (1)$$

Here, E_{pix} is any CRF energy defined over labelings at the pixel level of a single frame. As in [14], we use an energy based on the Associative Hierarchical CRF (AHCRF) for this term [13], which includes unary potentials and contrast-sensitive Potts pairwise potentials which penalize neighboring pixels taking different labels by a cost dependent on the change in RGB values, as well as potentials over a hierarchy of higher order cliques which reward consistency over larger segments using truncated linear Potts models. We assume that the unary potentials are shared between foreground labels $1...L^t$, and hence are based on a simple binary classifier which says if a pixel belongs to a human or the background. As in [14] we learn this classifier via boosting. The ψ_{d} are the detector potentials of [14]. These assume that we have pre-run a detector over each test image, and generated a set of bounding boxes for proposed humans in the image, which are refined to a set of proposed segments (cliques) using grab-cut. We associate a label l with each of these proposed instances, thus giving us

L^t, and write $c^t_{d(l)}$ for the clique generated by the unique detection d associated with label l at frame t. Further, we have in Eq. 1 $\mathbf{x}_c$ for the labeling of clique c and H_d for the detector response associated with detection d, and we use the same form for the potential as proposed in [14]:

$$\psi_{\mathrm{d}}(\mathbf{x}_{c^t_d}, H_d, l) = \min_{y^t_d=\{0\,1\}} \left(-w_d |c^t_d| \max(0, H_d - H_t) y^t_d + \frac{w_d |c^t_d| \max(0, H_d - H_t) N_d y^t_d}{p_d |c^t_d|} \right) \quad (2)$$

where y^t_d is a latent binary variable, w_d is a weight for all detector potentials, H_t is a threshold on the detector response, $N_d = \sum_{p \in c^t_d} (x^t_p \neq l)$ is the number of 'inconsistent pixels' in the clique (taking values other than l) and p_d is the proportion of inconsistencies allowed.

Model 1: We begin by improving model 0 above in two regards. First, we add parts based terms for the face and body ψ_{f} and ψ_{b} to better model the internal structure of human instances. Second, we add higher order terms ψ_{vid} spanning multiple frames of video to enforce smoothness over homogeneous regions across frames. This gives us the following energy:

$$E_{\mathrm{M1}}(\mathbf{x}) = \sum_t [E_{\mathrm{pix}}(\mathbf{x}^t) + \sum_{l=1\ldots L^t} (\psi_{\mathrm{b}^t_l}(\mathbf{x}^t) + \psi_{\mathrm{f}^t_l}(\mathbf{x}^t))] + \sum_{\mathrm{c} \in \mathrm{C}_{\mathrm{vid}}} \psi_{\mathrm{vid}}(\mathbf{x}, c) \quad (3)$$

Here, ψ_{vid}, which we will call video potentials, are defined over a set of cliques C_{vid} on the full cross-frame labeling $\mathbf{x}^{1\ldots T}$, and their form will be described in Sec. 2.2. Further, we note that ψ_{b} and ψ_{f}, the separate 'body' and 'face' potentials, are substituted in place of the single detection potential per human instance ψ_{d} in Eq. 1. The form of these potentials will be given in detail in Sec. 2.3 and 2.4.

Model 2: Although model 1 incorporates a basic parts model of human instances, it does not reflect any global properties of the instance, such as pose and scale. Recent work has shown that histograms of edge-based features carry much information about these properties [1], and a simple way of building this into our model is to add an extra (higher order) term for each instance ψ_{h} based on its histogram:

$$E_{\mathrm{M2}}(\mathbf{x}) = \sum_t [E_{\mathrm{pix}}(\mathbf{x}^t) + \sum_{l=1\ldots L^t} (\psi_{\mathrm{b}^t_l}(\mathbf{x}^t) + \psi_{\mathrm{f}^t_l}(\mathbf{x}^t)) + \psi_{\mathrm{h}^t_l}(\mathbf{x}^t))] + \sum_{\mathrm{c} \in \mathrm{C}_{\mathrm{vid}}} \psi_{\mathrm{vid}}(\mathbf{x}, c) \quad (4)$$

As will be discussed in Sec. 2.5, these terms will be used to provide a soft constraint that the feature histograms of the instances found should match to known instance histograms in the training data (we will include gradient, texture and color features, see Sec. 3), thus implicitly enforcing consistency over scale/pose. We separate models 1 and 2 above as they require different optimization techniques (Sec. 2.6).

2.2 Video potentials (Model 1)

Our E_{pix} energies include higher-order consistency terms defined over cliques on segments generated at each frame. Multiple segmentations are generated for each frame using the meanshift algorithm [3], varying (spatial and range) parameters, $\mathrm{h_s}, \mathrm{h_r}$. In general, any unsupervised segmentation method can be used to generate these segments. A robust P^n potential is then placed over each proposed segment to softly enforce consistency (see [13])

Our video potentials similarly place soft constraints on consistency, but over 3d spatio-temporal segments defined on the full cross-frame labeling $\mathbf{x}^{1\ldots T}$. We use 3d meanshift to

generate such segments [8, 9, 24], which includes a temporal parameter h_t which can be varied along with the spatial and range parameters h_s, h_r. Let r_{t_i} be the i^{th} segment in frame at t which is assigned to the 3d clique $c \in C_{vid}$. Consistency in segments across frames is maintained, which means the i^{th} segments in frames at t and (t+1), r_{t_i} and $r_{(t+1)_i}$, are assigned to the same 3d clique c. Thus, our 3D mean-shift implementation ensures consistency in the segment ids across all the frames in a shot. We define a P^n Potts potential on these 3d cliques, which enforces foreground/background consistency across segments in different frames. For clique $c \in C_{vid}$, this takes the form:

$$\psi_{\text{vid}}(\mathbf{x}, c) = \min\left(\min_{l' \in \{0,1\}} (k_{vid} \sum_t |N^t_{l'c}|) \frac{1}{p_{vid}} \gamma_{max}, \gamma_{max}\right) \tag{5}$$

where k_{vid} is the consistency prior for the Potts model (the cost incurred for each inconsistent pixel in the clique), p_{vid} is the truncation parameter (controlling rigidity), γ_{max} the truncation value (the maximum cost allowed), and $N^t_{l'c} = \sum_{p \in c_t} [x^t_p > 0 \wedge l' = 0] \vee [x^t_p = 0 \wedge l' = 1]$, i.e. the number of pixels in frame t falling into clique c (c_t) inconsistent with the foreground/background label l' (see [10] for details on the P^n potts model).

2.3 Parts-based detection and segmentation (Model 1)

Model 0 described above (Sec. 2.1) assumes we have a single human detector, whose outputs determine the number of instances assumed present in an image and their proposed regions. The detector potentials in Eq. 2 attempt to enforce consistency over the latter, rewarding the label associated with each detection in the region proposed. In model 1 we extend this framework to incorporate multiple parts per instance, restricting ourselves here to a simple face-body model. We assume we have separately trained face and body detectors, which generate for each image a set of proposal face and body bounding boxes. In general, we now have the problem of grouping these part proposals into instances, which will then be associated with labels $l = 1 \ldots L^t$. We choose a straightforward method: we keep all body detections whose overlap is not too great, and assign each to a different label l; we then consider each face in turn, grouping it with at most one body depending on the overlap between the bounding boxes; any faces not grouped with a body are then rejected as false positives. We denote the faces and bodies remaining by $f \in F$ and $b \in B$, and our grouping process automatically allows us to write f_l and b_l respectively for the face and body associated with label l (where we may have $f_l = f_0$, the null face, if no face is associated with a particular body). Given this notation, our face and body potentials then take the same form as the detector potentials in Eq. 2, using the label associations implied by the grouping process:

$$\psi_{\text{b}^t_l}(\mathbf{x}^t) = \psi_{\text{d}}(\mathbf{x}_{c^t_b}, H_b, l), \qquad \psi_{\text{f}^t_l}(\mathbf{x}^t) = \psi_{\text{d}}(\mathbf{x}_{c^t_f}, H_f, l) \tag{6}$$

where c^t_b and c^t_f are the cliques associated with body b and face f in image t respectively, and H_b, H_f are the detector responses. The null face is automatically given a 0 detector response (and an empty clique), while for all other faces we set H_f to a high constant value and p_d in Eq. 2 to 1, ensuring that pixels are automatically marked with the relevant instance labels in these regions, as the face detector has high accuracy. As in model 0, c^t_b is generated by applying grab-cut to the refine the initial detector bounding box. However, we also incorporate a shape prior, described in Sec. 2.4 to provide a bias towards regions with expected human shapes. For c^t_f the whole bounding box is used, as the face boxes are typically tight.

2.4 Shape Prior (Model 1)

As described in the previous section, we use grabcut to refine the bounding boxes of the body detectors to form the cliques for the detector potentials. Grabcut generally builds a color model for the foreground and background objects based on the bounding box location and iteratively applies a graph cut to extract and refine the segments. However, grab-cut suffers from the innate problem of color not being a discriminative feature. We thus propose to use a human shape prior which biases the segmentation towards typical human shapes for the body segments in our model.

We place a simple grid of cells $j = 1...J$ over $i = 1...I$ body bounding boxes extracted from our training set, and, denoting the binary vector (foreground/background) $\mathbf{v}_{ij}$ as the labeling of the j'th cell of the i'th training box, calculate the prior weights $\pi_j = \sum_i |\mathbf{v}_{ij}|_1 / \max_v \sum_i |\mathbf{v}_{ij}|_1$. where $|.|_1$ is the L_1-norm, i.e. giving the number of positives in the cell. We then adapt the grab-cut energy function to incorporate these prior weights, giving us the energy:

$$E_{\text{grabcut}}(\mathbf{v}_b^t, \theta) = \sum_p (\pi_{j_{bp}} U_1(v_{bp}^t, \theta_1) + U_0(v_{bp}^t, \theta_0)) + \sum_{pq \in \mathcal{N}} V(v_{bp}^t, v_{bq}^t) \tag{7}$$

where v_b^t is a binary vector such that $v_{bp}^t = 1 \Rightarrow p \in c_b^t$, θ_1 and θ_0 are the foreground/background color models of the grab-cut energy, U_1 U_0 the associated unary potentials, $\mathcal{N}$ the neighbor relation, V the grab-cut pairwise potential (which as in the energies of Sec. 2.1 is a contrast-sensitive Potts model), and j_{bp} denotes the cell that pixel p falls in with respect to bounding box b ($\pi_{j_{bp}} = 0$ if p falls outside the box b). As in [18], Eq. 7 can be minimized iteratively with respect to v_b^t and θ, generating the cliques c_b^t for the body potentials in Eq. 6.

2.5 Exemplar histogram matching (Model 2)

In addition to incorporating information on instance shape, in model 2 we also add terms to the energy which encourage the feature histograms of the instance regions in the solution to conform to known *exemplar* histograms from our training set. As has been shown recently [1], histograms of edge-based features are highly discriminative of properties such as pose in humans. By encouraging the instances found to match known exemplars in their histograms of such features (and others, see Sec. 3), we are thus encouraging consistency in the global properties of the instances. We denote the exemplars as $\mathbf{e}^k$, $k = 1...K$, where for each exemplar e_{mn}^k denotes the number of features falling in bin n of the histogram for feature m for exemplar k. Similarly, we write $z_{mp}^t \in \{1...N\}$ for the observed value (bin) of feature m at pixel p in image t, S^{tl} for the set of pixels in image t taking label l, and S_{mn}^{tl} for the subset of pixels of S^{tl} for which $z_{mp}^t = n$. Given this, the histogram matching term in Eq. 4 is:

$$\psi_{\mathrm{h}_\mathrm{l}^\mathrm{t}}(\mathbf{x}^t) = \min_k (\sum_{mn} ||S_{mn}^{tl}| - e_{mn}^k|) \cdot [|S^{tl}| > 0] \tag{8}$$

The form of Eq. 8 implies that a constant cost is paid for the amount by which the entries of the observed histograms of instance l deviate from their expected values given the best matching exemplar $\mathbf{e}^k$. The factor $[|S^{tl}| > 0]$ ensures we do not pay a cost for labels which are absent (i.e. proposal instances from the detectors not supported by other potentials).

2.6 Training and inference

We adopt a piecewise approach to training, which involves training the potentials of each model separately, and then manually tuning weights on the separate potentials (not shown in Eqns. 1, 3 and 4). We assume we have a (set of) training videos including (1) ground truth bounding boxes of human instances, body and face segments, (2) pixel level binary

human/background masks on a large number of frames, and (3) pixel level segmentations of instances in a limited number of frames. Given this training data, the detectors are trained directly on the relevant bounding boxes, as is the shape prior of Sec. 2.4. The unary potentials of all models are trained on the human/background masks, using a boosted classifier over multiple quantized features (see Sec. 3) as in [14]. Finally, all annotated instance level segments are used as exemplars for model 2 (Sec. 2.5), and their histograms extracted over the same quantized features while adding a small smoothing term to prevent 0 entries.

For inference in models 0 and 1, we can directly use the graph-cut based inference techniques of [14]. This is because all the higher-order terms in these models take the form of P^n-Potts potentials, thus giving rise to an energy with submodular α-expansion moves. The video potentials in these models cause the solutions of all video frames to be linked, and we thus build a large graph across all frames in the shot and solve jointly for $\mathbf{x}^{1\ldots T}$. The detections generate a distinct set of labels for each frame (as noted, we do not enforce consistent instance labeling across frames), and we could perform inference by directly running α-expansion in this large label-set across the whole video. However, this is inefficient, since only human/non-human consistency is enforced across frames (see Eq. 5), and so we do not require the instance labels of neighboring frames to solve $\mathbf{x}^t$ for a given frame. Consequently, we begin by performing a single binary graph-cut across the whole video on the labels human/non-human (with all detector P^n potentials fixed to the human label), which is exact since the problem is binary, and then perform separate α-expansions on each frame separately using the instance labels of the frame $\mathcal{L}^t$.

Model 2 can no longer be solved directly by α-expansion, as the terms of Eq. 8 break the submodular properties of the energy. However, this model can be solved by Dual Decomposition (DD) techniques, recently proposed for enforcing similar histogram matching constraints between foreground regions in the context of cosegmentation [23]. As above, we begin by performing a single human/non-human segmentation across all frames, and for each frame individually run α-expansion using the energy of model 1. This gives us an initialization of the instance labeling for each frame $\mathbf{x}^t$, from which we identify the closest matching exemplars $\mathbf{e}^l$ (see Sec. 2.5) for each label l using the L_1-distance between the initial observed and exemplar histograms (unnormalized). For each frame, we then rewrite the problem of minimizing Eq. 4 subject to these fixed associations as an integer program:

$$\min_{\mathbf{x}^t} E_{\mathrm{M2}}(\mathbf{x}^t|\mathbf{e}^{1\ldots L^t}) = \min_{\mathbf{x}^t}(E_{\mathrm{M1}}(\mathbf{x}^t) + \sum_{l=1\ldots L_t} \psi_{\mathrm{h}^t_{\mathrm{l}}}(\mathbf{x}^t|\mathbf{e}^l)) = \min_{\mathbf{x}^t,\mathbf{a}^t}(E_{\mathrm{M1}}(\mathbf{x}^t) + \sum_{l=1\ldots L_t}\sum_{mn}|a^{tl}_{mn}|) \quad (9)$$

subject to:

$$\forall t,l,m,n \;\; a^{tl}_{mn} = |S^{tl}_{mn}| - e^l_{mn}\cdot[|S^{tl}|>0], \quad -N_t \le a^{tl}_{mn} \le N_t; \qquad \forall t,p \;\; x^t_p \in \{0,1\ldots,L^t\} \quad (10)$$

where auxiliary variables $\mathbf{a}$ have been introduced, N_t is the total number of pixels in frame t, and we define $\psi_{\mathrm{h}^t_{\mathrm{l}}}(\mathbf{x}^t|\mathbf{e}^l)$ as in Eq. 8, but where the min over k is replaced by setting $k = k' \; s.t. \; \mathbf{e}^{k'} = \mathbf{e}^l$. An expansion move over label l gives rise to a restricted version of the integer program in Eq. 9, the dual of which can be formed by introducing Lagrange multipliers to relax the first constraint in Eq. 10. This can be efficiently optimized by a subgradient method using the techniques of [23], where each step involves solving a submodular graph-cut. A solution to the primal is not guaranteed by solving the dual (but can be recognized if the duality gap is closed). The expansion moves generated by suitable rounding however typically involve decreases in the energy of similar magnitude to standard (submodular) α-expansion.

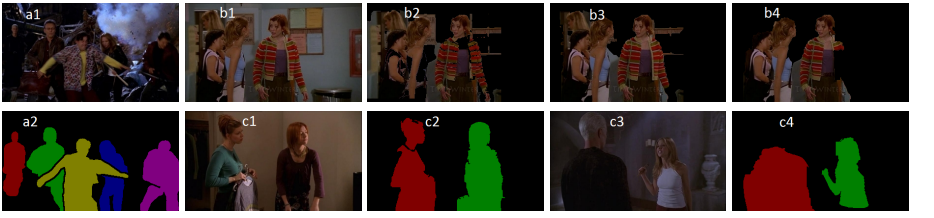

Figure 1: Our method of instance based segmentation is illustrated by images (a1-2), a pair from the set of exemplars, and images (c1-4), two pairs from the output set, with each pair representing the original and the segmented images with each distinct instance receiving a distinct colour. Images (b1-4) highlight the difficulties caused by occlusion in the dataset for the histogram potential; it results in the removal of one human (occluded) as shown in b1-4 representing original image, grabcut, model 1, and model 2 outputs.

U/I	*E4*	*E5*	*E6*	E_{avg}	***RC***	*E4*	*E5*	*E6*	E_{avg}
GC_d	0.454	0.555	0.481	0.496	GC_d	0.589	0.762	0.727	0.693
M0	0.543	0.591	0.615	0.583	*M0*	**0.765**	**0.953**	**0.925**	**0.881**
M1	**0.569**	0.649	0.652	0.623	*M1*	0.725	0.935	0.944	0.868
M2	0.546	**0.653**	**0.673**	**0.624**	*M2*	0.626	0.915	0.884	0.808

PR	*E4*	*E5*	*E6*	E_{avg}	***OA***	*E4*	*E5*	*E6*	E_{avg}
GC_d	0.664	0.672	0.586	0.640	GC_d	0.798	0.822	0.846	0.822
M0	0.651	0.609	0.647	0.636	*M0*	0.8166	0.8087	0.8858	0.837
M1	0.697	0.680	0.678	0.685	*M1*	0.8323	0.8534	0.9013	0.8623
M2	**0.745**	**0.695**	**0.737**	**0.726**	*M2*	**0.8326**	**0.8589**	**0.9158**	**0.8691**

Table 1: Results in tabular form; shown are Union/Intersection (U/I) (top left), Recall (RC) (top right), Precision (PR) (bottom left), and Overall accuracies (OA) (bottom right) of the methods on three different episodes. GC_d: detector + iterative graph cut, M0: base energy+(face and body) detectors without shape priors, M1: M0+shape priors, M2: M1+histogram matching potentials. E4, E5, E6, and E_{avg} correspond to the results on images taken from episodes four, five, six, and average over these three episodes respectively.

3 Experiments

We outline here our experimental setup and report our results on the Buffy dataset. For evaluation purposes, we select a set of 452 images from first two episodes for training and 160 images from next three episodes for test purposes. We provide a pixel-level human/non-human segmentation on these images, with each human instance getting same label/color, generating the ground truth images for training and testing purposes. Figure 2(b) shows some such segmented ground truth images. Further, we select a set of 60 images from the training dataset and provide pixel-level segmentation with each human instance receiving a distinct label/color as shown in figure 1(a). This set of images serves as exemplars for evaluating the histogram matching potential. As a whole, the dataset includes not only images with large variations in appearance and clutter in indoor environments but also wide variation in the scales and poses of human instances, making it challenging for the segmentation task.

Setup: Our method, as discussed earlier, is an unification of unary, pairwise, and higher order (segment, detector, and histogram) potentials in a hierarchical CRF framework. We learn the discriminative pixel-based and segment-based unary potentials as in [13], using a joint-boost method where the classifier is defined on multiple dense features such as color, textons,

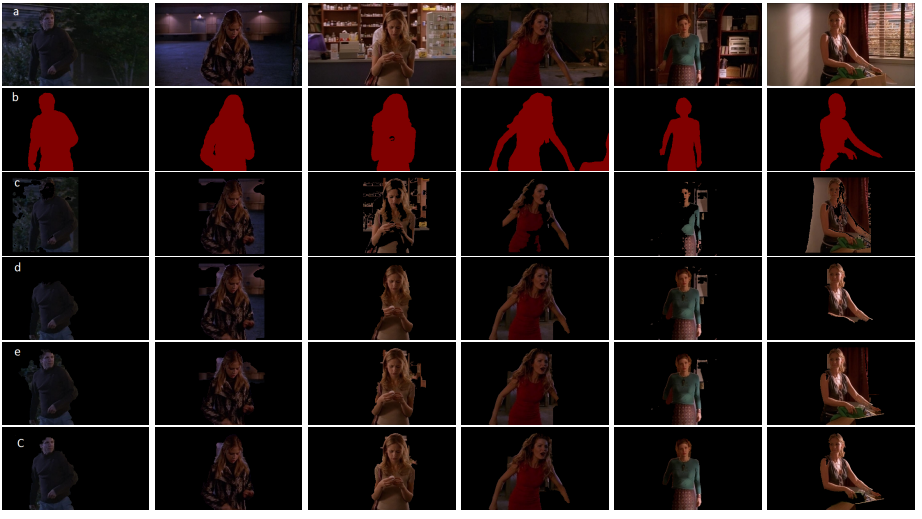

Figure 2: These figures outline the gradual improvement in quality of visual output on using model 1 and model 2. (a) Original images, (b) ground truth images, (c) Detector + iterative graph cuts (GC_d), (d) results after inclusion of detection potential (model 0), (e) results after inclusion of parts detection potentials and shape priors (model 1), (f) with histogram potentials also (model 2).

histograms of oriented gradients (HOG), and pixel location, similar to Textonboost [20]. Our pairwise potential term between a pair of pixels and segments takes the form of a contrast sensitive Potts model. Further, we adjust the parameters for detector and histogram potentials to balance their contributions, and also introduce an extra foreground label not tied to any particular detection to recover instances which are missed by any detector. The α-expansion algorithm is in to convergence for both models 0 and 1, and we make 2 subgradient steps at each dual decomposition expansion for model 2.

3.1 Results

We evaluate the efficiency of our model on the images taken from episodes four (E4), five (E5), and six (E6) of the Buffy dataset, and compare them with the base method, model 0 and iterative graph cuts applied to the detector outputs (GC_d) in Table 1. We also show E_{avg}, the overall result, after averaging the results on all three episodes. Overall accuracy improves by almost 4.0% and 2.5% over GC_d and model 0 methods respectively after inclusion of shape prior (model 1). Further, adoption of histogram potential (model 2) improves the accuracy by 0.68% over model 1 (Table 1 compares the average across episodes, while taking the average across frames yields 0.8315, 0.8597, and 0.8667 for models 0, 1 and 2 respectively). A union/itersection comparison shows an average improvement from 0.583 to 0.623 after using the model 1 and to 0.624 on inclusion of the matching potential as in model 2. Looking at the scores for each episode, we see that while model 2 improves the U/I score for E5 and E6, it does slightly worse for E4. This is possibly caused by the prominence of occlusion due to crowds in episode 4, and absence of sufficient variability in the exemplar instances, capable of matching to occluded human models (see also Fig. 1), which could be improved by adding more annotations. Further, we also achieve a very good precision at moderate recall rate on all episodes, which is expected as the matching potential and grabcut based detection potential complement each other.

Beyond these quantitative comparisons, we highlight the qualitative improvement in

Fig. 2. The visual output has gradually improved on adoption of different models (model 1, model 2). Most importantly, it is notable that the use of shape prior and histogram potential helps in recovering the shapes of humans and recovering background regions, which are mislabeled as foreground by the base model 0. Finally, we show examples of the full output of our method in Fig. 1. Each individual instance of a human in the image receives a different color. However, color in itself is not a representative of any particular human.

4 Conclusion

In this paper, we propose an instance based method for human segmentation in images and videos, which is motivated by a range of applications. We formulate our model as a detector-based CRF, allowing us to use efficient graph-cut algorithms for inference, and achieve promising results on the Buffy dataset with our new annotations.

Beyond its mentioned applications, we believe there are indeed more possible extensions. Introduction of exemplar shape priors along with the exemplar histograms will help in overcoming the ambiguities associated with the histogram matching. These arise because different shapes can have similar histograms of features. Another purposeful extension would be to enforce consistency of an instance across frames in space-time framework. Further, we can potentially identify the poses/identities of humans by using the matched exemplars of model 2. We will also make the segmented dataset and code available for others.

References

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(1):44–58, 2006.

[2] X. Bai, J. Wang, D. Simons, and G. Sapiro. Video snapcut: robust video object cutout using localized classifiers. *ACM Trans. Graph.*, 28(3), 2009.

[3] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR (1)*, pages 886–893, 2005.

[5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, pages 726–733, 2003.

[6] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

[7] V. Ferrari, M. J. Marín-Jiménez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

[8] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, pages 1–8, 2009.

[9] M. Hahn, F. Quronfuleh, C. Wöhler, and F. Kummert. 3d mean-shift tracking of human body parts and recognition of working actions in an industrial environment. In *HBU*, pages 101–112, 2010.

[10] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009.

[11] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Obj cut. In *CVPR (1)*, pages 18–25, 2005.

[12] M. P. Kumar, A. Zisserman, and P. H. S. Torr. Efficient discriminative learning of parts-based models. In *ICCV*, pages 552–559, 2009.

[13] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, pages 739–746, 2009.

[14] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What, where and how many? combining object detectors and crfs. In *ECCV (4)*, pages 424–437, 2010.

[15] Y. Li, J. Sun, and H. Y. Shum. Video object cut and paste. *ACM Trans. Graph.*, 24(3): 595–600, 2005.

[16] J. Nikkanen, T. Gerasimow, and L. Kong. Subjective effects of white-balancing errors in digital photography. In *SPIE*, pages 1–11, 2008.

[17] A. Patron, M. Marszalek, A. Zisserman, and I. D. Reid. High five: Recognising human interactions in tv shows. In *BMVC*, pages 1–11, 2010.

[18] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[19] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV (2)*, pages 406–420, 2010.

[20] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[21] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton. Dynamical binary latent variable models for 3d human pose tracking. In *CVPR*, pages 631–638, 2010.

[22] A. H. Vela. Pose and face recovery via spatio-temporal grabcut human segmentation. In *MS thesis*, 2010.

[23] S. Vicente, V. Kolmogorov, and C. Rother. Cosegmentation revisited: Models and optimization. In *ECCV (2)*, pages 465–479, 2010.

[24] A. Yilmaz. Approximate mean-shift method. *http://server.cs.ucf.edu/ vision/source.html*.