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] School of Technology
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.

1 Human Instance Segmentation

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. [1]. Instance based human segmentation involves pixel level labeling of an image, partitioning it into distinct human instances and background. Human instance segmentation is central to numerous tasks ranging from human activity recognition in security and video surveillance to building better user interface systems as in white balancing and video object-cutout systems.

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.

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. Thus, our final energy function which incorporates parts-specific potential, shape prior information, 3D consistency potential and histogram matching potential as soft constraints along with the previous model (E_{pix}) is,

$$\begin{aligned} E(\mathbf{x}) &= \sum_t [E_{\text{pix}}(\mathbf{x}^t) + \sum_{l=1...L^t} (\psi_{\text{b}_l^t}(\mathbf{x}^t) + \psi_{\text{f}_l^t}(\mathbf{x}^t)) + \psi_{\text{h}_l^t}(\mathbf{x}^t))] \\ &+ \sum_{c \in C_{\text{vid}}} \psi_{\text{vid}}(\mathbf{x}, c) \end{aligned} \quad (1)$$

- Video potential (ψ_{vid}): Our video potentials place soft constraints on consistency, but over 3d spatio-temporal segments defined on the full cross-frame labeling $\mathbf{x}^{1...T}$. We use 3d meanshift to generate such segments [2], which includes a temporal parameter h_t which can be varied along with the spatial and range parameters $\text{h}_\text{s}, \text{h}_\text{r}$.
- Parts-based detection and segmentation ($\psi_{\text{b}_l^t} + \psi_{\text{f}_l^t}$): We extend the model of Ladicky et.al. [1] 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 group these part proposals into instances, which are then associated with labels $l = 1...L^t$.
- Shape prior: Ladicky et.al. [1] use grabcut to refine the bounding boxes of the body detectors to form the cliques for the detector potentials. 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 used in ($\psi_{\text{b}_l^t}$).
- Exemplar histogram matching ($\psi_{\text{h}_l^t}$): In addition to incorporating information on instance shape, 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.

Figure 1: These figures outline the visual output after using our models.(a) Original images, (b) ground truth images, (c) results of our model.

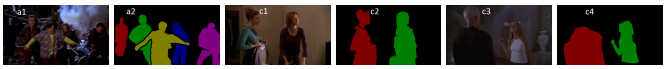

Figure 2: Our method of instance based segmentation is illustrated by images (a1-2), a pair from the set of exemplars, and images (c1-4), two pairs from the output set, with each pair representing the original and the segmented images with each distinct instance receiving a distinct colour.

2 Experiments

We 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 1(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 2(a).

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). Overall accuracy improves by almost 4.0% and 2.5% over GC_d and model 0 methods respectively after inclusion of shape prior. Further, adoption of histogram potential improves the accuracy by 0.68% over model 1. 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. Beyond these quantitative comparisons, we highlight the qualitative improvement in Fig. 1 and Fig. 2.

[1] 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.

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