Image and Vision Computing 51 (2016) 58–68

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

# Fully automatic person segmentation in unconstrained video using spatio-temporal conditional random fields ☆

Chetan Bhole[a,*], Christopher Pal[b]

[a] University of Rochester, Rochester 14620, USA
[b] Université de Montréal, Montréal, Canada

ARTICLE INFO

*Article history:*
Received 30 July 2014
Received in revised form 5 August 2015
Accepted 24 April 2016
Available online 2 May 2016

*Keywords:*
Person segmentation
Video segmentation
Conditional random field
Optical flow
Fully automatic

ABSTRACT

The segmentation of objects and people in particular is an important problem in computer vision. In this paper, we focus on automatically segmenting a person from challenging video sequences in which we place no constraint on camera viewpoint, camera motion or the movements of a person in the scene. Our approach uses the most confident predictions from a pose detector as a form of anchor or keyframe stick figure prediction which helps guide the segmentation of other more challenging frames in the video. Since even state of the art pose detectors are unreliable on many frames –especially given that we are interested in segmentations with no camera or motion constraints –only the poses or stick figure predictions for frames with the highest confidence in a localized temporal region anchor further processing. The stick figure predictions within confident keyframes are used to extract color, position and optical flow features. Multiple conditional random fields (CRFs) are used to process blocks of video in batches, using a two dimensional CRF for detailed keyframe segmentation as well as 3D CRFs for propagating segmentations to the entire sequence of frames belonging to batches. Location information derived from the pose is also used to refine the results. Importantly, no hand labeled training data is required by our method. We discuss the use of a continuity method that reuses learnt parameters between batches of frames and show how pose predictions can also be improved by our model. We provide an extensive evaluation of our approach, comparing it with a variety of alternative grab cut based methods and a prior state of the art method. We also release our evaluation data to the community to facilitate further experiments. We find that our approach yields state of the art qualitative and quantitative performance compared to prior work and more heuristic alternative approaches.



## 1. Introduction

Video segmentation is an important problem in computer vision. One is often interested in decomposing a video into segmentations that correspond to the different semantic objects found in a scene. In our work here, we focus on segmenting people in video. Video editing and video surveillance applications represent two prominent areas among many others where the automated segmentation of a person can enable key applications. We are interested here in segmenting people under challenging conditions where there are no constraints on the camera or the person's motion. Furthermore, we also wish to deal with real world video that may contain motion blur, drastic scale changes, people in non-upright positions, and clothed people with arbitrarily colored materials as well as compression artifacts in the videos. To automatically segment a person from these types of challenging videos we anchor our approach through using keyframes derived from human pose detection techniques. We use the fairly state of the art pose detector presented in Yang and Ramanan [1] to detect the pose of a person from a frame in the video sequence. The stick figures derived from poses are then used to extract color and optical flow features to train a conditional random field to provide segmentation on multiple frames. Location from the pose is used to refine the results. No additional training data is required by the method for a new video. We also show how pose results can also be improved by our model. We discuss the use of a continuity method that detects large discontinuities of segmentations between batches and the reuse of learnt parameters if the discontinuities are large.

The accuracy of full body pose detection in static images has greatly improved in recent years. However, detecting the lower limbs is still challenging compared to the upper torso. In this work, we

☆ This paper has been recommended for acceptance by Vladimir Pavlovic.

* Corresponding author. Tel.: +1 716 2138593.
*E-mail address:* bhole@cs.rochester.edu (C. Bhole).

http://dx.doi.org/10.1016/j.imavis.2016.04.007
0262-8856/

focus on full body segmentations and make no assumption about the camera and person motion. As such, both a person in the foreground as well as the background may appear to change dramatically due to camera motion. We use the optical flow technique of Liu [2] to define the connectivity structure of a CRF and adjust smoothness parameters using a cross-validation approach that allows us to eliminate the need to obtain hand labeled data for each new video sequence. Our current model segments one full body person from a video sequence; however, the same procedure we use here could fairly easily be adapted and applied repeatedly to segment additional people in the same video. Our experimental results indicate that even though the pose results we use to seed our technique are in general unreliable, we are able to obtain a dramatic increase in robustness over other techniques including prior state of the art work on this problem and our own formulations of strong baseline techniques. Our other baselines are based on more heuristic applications of the grab cut technique of Rother et al. [3] applied in a straightforward but intelligent way to automate the task of video segmentation.

Our work here builds on our preliminary results in Bhole and Pal [4]. Our principal contribution is the creation of a completely automated segmentation technique for humans in video. We use a novel dynamically constructed CRF architecture which is defined based on optical flow analysis and we use a cross-validation technique for learning spatial and temporal smoothing parameters. We provide extensive experiments, exhaustively comparing our method with previous work by Niebles et al. [5] as well as a variety of alternative grab cut methods. Furthermore, we also make available a new segmentation evaluation dataset consisting of painstakingly labeled ground truth maps. We believe that the lack of such data being available the community has been an important factor, limiting the amount work in this area and it is our hope that the release of this data will help remedy the situation. We also include human annotated pose estimates allowing a set of pose experiments to be performed by others. Our method provides both better qualitative and higher quantitative segmentation performance compared to current state of the art methods.

## 2. Related work

Consider first the task of general object or region of interest segmentation in video. Previous work such as that of Chen and Corso [6], Grundmann et al. [7], and Xu and Corso [8] has used motion in video to detect motion segments or regions. Some work has used weakly labeled videos to aid segmentation as in Rochan et al. [9]. Some methods also use the clustering of spatio-temporal cues to extract objects. It is important to consider the effects of camera motion, as well as the motion of foreground and background objects and scene elements on the difficulty of the task that an algorithm must solve. When all three of these aspects of the scene are in motion it can very quickly become significantly harder to segment video into desired regions. Many previous algorithms are not applicable to our goal here as they make strong assumptions about these properties. For example, many previous algorithms for video segmentation either use the assumption of a static camera or global constraints assuming that camera motion was linear. However, some work such as that of Bhat et al. [10] go so far as to leverage state of the art structure from motion and 3D reconstruction for segmentation and video editing.

Some approaches like Jojic and Frey [11] model (non-occluding) segments in a single frame in a layered manner with the layers stacked one over the other to generate a visible scene. They introduce flexible sprites which are masks of contiguous moving segments in which the pixels share the same rigid motion and formulate the problem to maximize the log likelihood of the video solving using unsupervised techniques. Chuang et al. [12], on the other hand, model the video as a weighted combination of the foreground and background maximizing the log likelihood of the foreground, background and weights given the data.

Many of the solutions that use graphical models involve the use of ground-truth provided for keyframes of the same video they test on. Previous approaches set the parameters of the random fields manually while some more recent approaches learn parameters. Li et al. [13] use graph-cuts for inference given the user input ground-truth for some keyframes in the same video. No learning of the MRF parameters was done and the parameters were set to some predetermined values. In this method, initially, 3D segmentation on a 3D MRF is done using the graph-cuts on regions of pixels instead of individual pixels. Since 3D segmentation is more global, sometimes it produces poor results. They then use a local region 2D MRF for 2D segmentation for local regions that give error labels. The user manually selects these regions. This significantly improves performance. Zhang and Ji [14] construct a 3D CRF model and use loopy belief propagation for inference. The local potential is modeled as the logistic output of the 3 layer neural network. Also, they train the CRF with only 3 consecutive frames for every train/test case.

We discuss here the segmentation methods that specifically target person segmentation. Yin et al. [15] concentrate on trying to recongize the upper torso of a person from the rest of the background in a video chat application. Their earlier work [16] used a pair of cameras to get a stereo view and their later work [15] tries to achieve the same performance using monocular video data. The background can be cluttered with motion and define the foreground object using the depth cue with only a single sequence of frames. They use the notion of motons similar to textons in Malik et al. [17]. They take the change of a pixel along time as the motion gradient and change of spatial location of pixel values as the shape gradient to form a motion. Like TextonBoost in Shotton et al. [18], they combine a moton with a location rectangular mask and come up with a feature that counts the difference in the number of different moton pixels located at different locations. The features are fed to a random forest classifier and a CRF to label pixels.

Bi and Liang [19] use an algorithm with motion, color and stereo vision (depth cues) to segment out people from static background video sequences. Rodriguez and Shah [20] detect and segment humans using instances of a codebook to estimate the location and postures using a voting algorithm. Only upright walking or standing humans are used. Vineet et al. [21] use part detection, shape priors and exemplars in a conditional random field (CRF) framework to segment humans on each frame of the sequence separately. Hernandez et al. [22] use HOG-based detection, face detection and skin color models on Grabcut where temporal information is stored as Gaussian mixture models. Gulshan et al. [23] use large amount of training data from the kinect to learn their segmentation model. Wang and Koller [24] segment humans from static images with a joint pose and segmentation model that they solve using dual decomposition.

Niebles et al. [5] use a sequence of an upright human detector, upright human pose template and contour extraction and propagation (shape priors) along with a graphical model to capture temporal information and obtain motion volumes. Unlike our work that uses a state-of-the-art pose detector [1], their method does not use any part-based articulated models and needs the human to be upright. Our technique is more similar to that of Kohli et al. [25]. They also use a pose detector and CRF. However, they don't learn the smoothness or location parameters. Also they segment each frame separately while we segment a joint sequence of frames. We don't concentrate on the 3D orientation and only work with the results obtained from the pose detector. A lot of previous work makes strong assumptions while segmenting people. For example, many segmentation methods work only for pedestrians or upright people, request user interaction in first frame or make use of face or skin detectors. We do not wish to rely on face detectors because many challenging videos contain people of interest where the faces are not frontal views, are too