21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Automated person segmentation in videos

Chetan Bhole
University of Rochester
bhole@cs.rochester.edu

Christopher Pal
Ecole Polytechnique de Montreal
christopher.pal@polymtl.ca

Abstract

This paper deals with automatically segmenting a person from challenging videos using a pose detector. A state of the art pose detector is used to detect the pose of a person from a frame in the video sequence. The pose is 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. We also show how the pose results can be improved by our model.

1. Introduction

Detection and segmentation of people in videos is an important problem in computer vision. Segmenting humans can help improve activity recognition systems and thus be useful in video surveillance, sports activity analysis etc. We focus on the problem of automatically segmenting out a full body person in challenging videos. We make available the labelings that were used for evaluation to the research community.

Bi et. al. [1] use motion, color and stereo vision to segment out people from static background video sequences. Rodriguez et. al. [11] detect and segment humans using instances of a codebook to estimate the location and postures with the examples being of humans usually being upright walking or standing. Vineet et. al. [12] 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. [5] use HOG-based detection, face detection and skin color models to use on Grabcut where temporal information is stored as Gaussian mixture models. Wang et. al. [13] segment humans from static images with a joint pose and segmentation model that they solve using dual decomposition. Gulshan et. al. [4] use large amount of training data to learn their segmentation model. Niebels et. al. [10] use a shape prior based mostly on pedestrians to obtain motion volumes. Our technique is more similar to that by Kohli et. al. [6]. They however 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. Also compared to previous mentioned work, we try not to make use of face detectors because many challenging videos contain people of interest where the faces are not frontal views or could be too small or blurred due to motion. Skin detectors don't help when people are completely dressed. Besides this, challenging videos have people moving towards or away from the camera leading to scale changes.

Pose detection[14] on single images has got more reliable though detecting the lower limbs is still challenging compared to the upper torso. We make no assumption about the motion of camera and both the person (foreground) and background can be moving. We learn the smoothness parameters and try to provide robustness even though the pose results might not be reliable. We use optical flow [8] connections between image sequences and we focus on full body segmentations in this work. Our model segments one full body person from the video sequence. The same procedure can be potentially applied repeatedly to segment more people in the same video.

2. Model description

The system is explained with the help of Fig. 1. (1) The video sequence is broken down in batches (in our case of 15 frames) for computational reasons and to allow segmentation in very long sequences. Since there can be variations in lighting with time, it makes sense to learn different sets of parameters for each batch instead of global parameters for the entire sequence. (2) We apply a pose detector on each of the frames and select the frame that has the best pose score (highest confidence) as our key frame to initiate segmentation and hence the

978-4-9906441-1-6 ©2012 IAPR

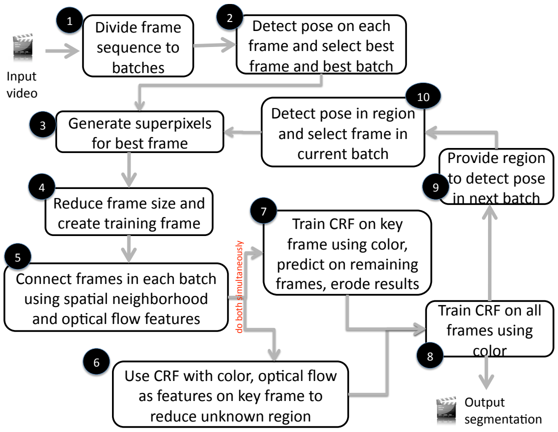
Input video
1
Divide frame sequence to batches
2
Detect pose on each frame and select best frame and best batch
10
Detect pose in region and select frame in current batch
3
Generate superpixels for best frame
9
Provide region to detect pose in next batch
4
Reduce frame size and create training frame
7
Train CRF on key frame using color, predict on remaining frames, erode results
5
Connect frames in each batch using spatial neighborhood and optical flow features
do both simultaneously
Train CRF on all frames using color
8
6
Use CRF with color, optical flow as features on key frame to reduce unknown region
Output segmentation

Figure 1. The process pipeline used.

batch that contains that key frame. It should be noted that even state of the art pose detectors [14] (trained on limited static images) do not provide reliable poses for all frames and the detector fails on many frames in challenging videos. The pose detector result in the key frame is able to give us a reasonable start point to begin segmentation and provides us with a location feature. (3) We generate superpixels [9] for the key frame and (4) select all superpixels that cover the pose stick as foreground training data. This enables us to extract the location of the various body parts. A more liberal region is selected that will contain the full body of the person of interest (shaded gray in Fig. 3(b)) and acts as the unknown region that will be estimated by our model. The remaining region is used as background training data (shaded black in Fig. 3(b)). (5) The current batch of frames is connected together to form a 3 dimensional graph. The frame sequence in the batch is connected together temporally using dense optical flow [8] in the forward and backward direction [3] as shown in Fig. 2(a). These connections better model self occlusion changes. (7) A CRF trained on only color on the key frame is used to predict segmentations in the frames (formed as a 3D graph) in the current batch. (6) Another CRF is trained on the key frame using color and optical flow features. Optical flow features are useful because the pixels in a given body part will tend to have the same motion. We however observed that using optical flow as features tends to include surrounding background regions and hence use it only for the key frame initial segmentation. (8) These segmentations from (6) and (7) are combined (foreground eroded and localized) and becomes training data for our final CRF model. We refine the results by removing segments far from the pose in the key frame. The final CRF predicts the segmentation for the entire sequence in the current batch. (9) We then make use of the the segmentation in the end frame as seeds to narrow down regions in the form of windows in the next batch. (10) The pose detector is run only in this windows in the frames of the next batch. This acts as a tracking step to be able to detect the same person if there are multiple people in the same video. The pose with the best score is selected as the key frame in the new batch and the process is repeated as above from (3).

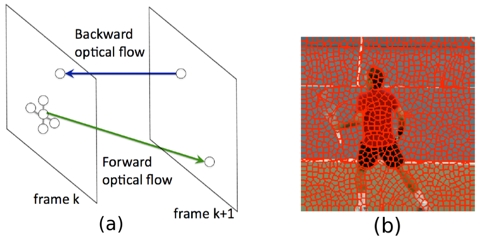

Figure 2. (a) Forward and backward optical flow connections between frames. (b) Superpixels generated on a key frame.

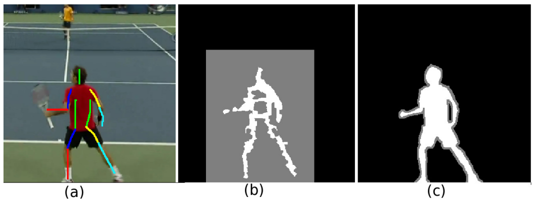

Figure 3. (a) Detected pose, (b) training image used by crf model and (c) evaluation ground truth.

2.1. CRF model

The CRF model is given by:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\exp(-\sum_{p}\sum_{c}\sum_{b}\lambda_{cb}f(x_p,y_p) - \sum_{p}\sum_{c}\sum_{b}\gamma_{cb}g(x_p,y_p) - \alpha\sum_{pq\in N_s}s(x_p,x_q,y_p,y_q) - \beta\sum_{pq\in N_t}t(x_p,x_q,y_p,y_q))$$

where $\mathbf{y}$ denotes the label nodes that can either be the foreground human segmentation or background, $\mathbf{x}$ denote the features uses in this case color features. $Z(\mathbf{x})$ is the normalization constant. p and q stand for pixels or nodes in the graph. The binned uv color-scale is modeled as a binary feature f and each bin has parameter λ_{cb} (per class c per bin b). Similarly, g is the binned optical flow

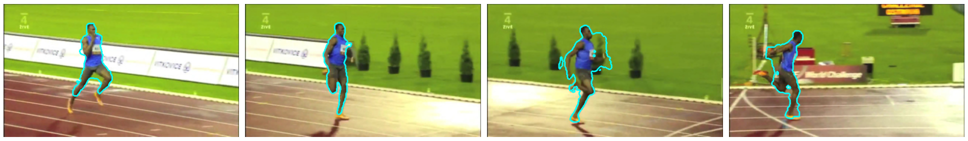

Figure 4. Person running segmentation example.

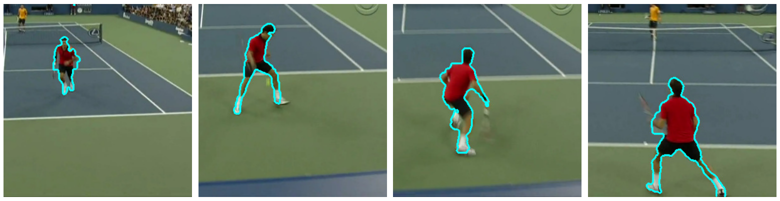

Figure 5. Person playing tennis segmentation example.

feature function and the corresponding weights are γ_{cb}. The spatial neighborhood (N_s) pairwise feature s computes the value (ratio of gradient and expected gradient over entire frame) [2] with the spatial pairwise parameter being α. The temporal neighborhood (N_t) feature t computes the gradient in the temporal neighborhood and β is the temporal pairwise parameter. The temporal neighborhood is restricted to the forward and backward optical flow connections. The features are turned off if not used for the specific CRF. All the above parameters are learnt using the log likelihood and gradient descent and graph cuts [7] is used to perform inference. Note that the inference is performed over the entire graph including the unknown nodes during parameter learning. The motivation behind not using a local color model is that as the pose detector can miss parts of the body, our model can recover the errors due to similarity and/or symmetry of the color distribution of clothes or skin. Using strong location information as in [6] does not let the model recover from pose mistakes.

2.2. Improving pose estimation

The state of the art pose detector fails in many frames of the videos. By segmenting the frame sequence, we can use the segmentation to improve the poor pose estimates. We blur out the region around the segmentation provided by our model and run the pose detector again. If the pose estimate originally predicted any body joint outside the body, the segmentation reduces the probability of that joint being outside the body or segmentation thereby improving the overall pose. An example of this improvement is seen in Fig. 7 where the left image is the initial pose detection result and the right image is the new pose after using segmentation predicted by our model.

3. Data, Evaluation and Results

We segment people from 5 different videos, 3 of them are long sequence videos and 2 of them are shorter. The following really makes these videos challenging. The first long video has a person running quickly and hence there is a lot of motion blur. The second video of a tennis player has the person's scale and orientation changing drastically in just a few frames. The third has the person skiing and is mostly not upright. The first short video has a person performing a difficult dance and the second also has a person dancing except that he has paint over his body and skin detectors fail. Also note that the color distribution for his clothes or body is close to that of the surrounding in both these videos. Note that for the two short videos, we selected the initial frame for pose detection manually instead of it being selected automatically as done for the other videos. This is because the pose scores and estimates for almost all frames were unreliable. In general, our method does well when reliable poses are detected and does poorly due to unreliable poses due to self occlusion or very similar foreground and background. The computational cost of our model is the cost of applying the pose detector on each frame and using a 3D CRF for segmentation on the batches of frames.

The pixel accuracies (number of pixels correctly classified for a class over total ground truth pixels of that class) and average class accuracies (average pixel accuracies over foreground and background) of the segmentations are computed using every 5th (10th) frame for the short (long) videos and manual ground truth segmentations. The results are shown in Table 1. FPA stands for foreground pixel accuracy, BPA stands for background accuracy and ACA stand for average class accuracy.

CP thanks the FQRNT, NSERC and Google for financial support.

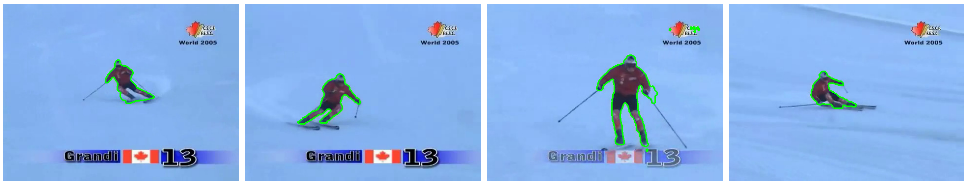

Figure 6. Skiing segmentation example.

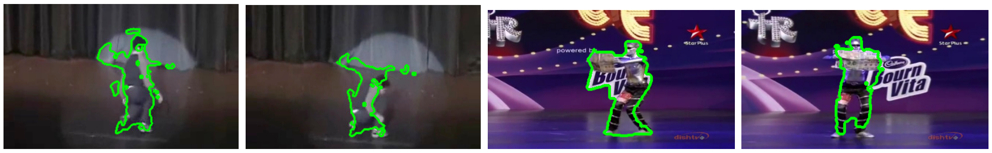

Figure 8. Two dance videos segmentation examples.

Table 1. Segmentation accuracies.

Video	No frames	FPA	BPA	ACA
Running	300	91.14	97.81	94.47
Tennis	101	95.7	99.56	97.63
Skiing	201	91.24	99.86	95.55
Dance-1	36	82.45	96.81	89.63
Dance-2	46	89.71	97.84	93.77

4. Conclusion

We show how a pose detector along with color, optical flow and location features can be used to segment a person in challenging videos with the help of a CRF and without the need for any additional training data. We conclude that combining the segmentation and pose information using multiple frames improves person segmentation accuracy.

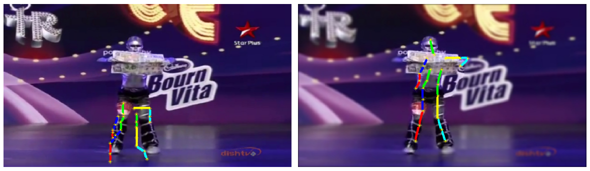

Figure 7. Improvement of pose using segmentation. Zoom for clarity.

References

[1] S. Bi and D. Liang. Human segmentation in a complex situation based on properties of the human visual system. In *Intelligent Control and Automation. WCICA 2006.*, volume 2, pages 9587 –9590, 2006.

[2] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive gmmrf model. In *ECCV*, pages 428–441, 2004.

[3] A. Chen and J. Corso. Propagating multi-class pixel labels throughout video frames. In *Proceedings of WNYIPW Workshop*, 2010.

[4] V. Gulshan, V. Lempitsky, and A. Zisserman. Humanising grabcut: Learning to segment humans using the kinect. In *IEEE Workshop, ICCV*, 2011.

[5] A. Hernandez, M. Reyes, S. Escalera, and P. Radeva. Spatio-temporal grabcut human segmentation for face and pose recovery. In *AMFG10*, pages 33–40, 2010.

[6] P. Kohli, J. Rihan, M. Bray, and P. Torr. Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *IJCV*, 79(3):285–298, 2008.

[7] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts. *IEEE Transactions on PAMI*, 26:65–81, 2004.

[8] C. Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis.* PhD thesis, MIT, 2009.

[9] G. Mori. Guiding model search using segmentation. In *ICCV*, pages 1417–1423, 2005.

[10] J. C. Niebles, B. Han, and L. Fei-Fei. Efficient extraction of human motion volumes by tracking. In *CVPR*, pages 655–662, 2010.

[11] M. Rodriguez and M. Shah. Detecting and segmenting humans in crowded scenes. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 353–356. ACM, 2007.

[12] V. Vineet, J. Warrell, L. Ladicky, and P. Torr. Human instance segmentation from video using detector-based conditional random fields. In *Proceedings of the BMVC*, pages 80.1–80.11. BMVA Press, 2011.

[13] H. Wang and D. Koller. Multi-level inference by relaxed dual decomposition for human pose segmentation. In *Proceedings of CVPR*, 2011.

[14] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392, 2011.