Part-based Clothing Segmentation for Person Retrieval

Michael Weber[1] Martin Bäuml[1]
[1]Institute for Anthropomatics
Karlsruhe Institute of Technology
Adenauerring 2, 76131 Karlsruhe, Germany
michael.weber@student.kit.edu baeuml@kit.edu

Rainer Stiefelhagen[1,2]
[2]Fraunhofer Institute of Optronics,
System Technologies and Image Exploitation
Fraunhoferstr. 1, 76131 Karlsruhe, Germany
rainer.stiefelhagen@kit.edu

Abstract

Recent advances have shown that clothing appearance provides important features for person re-identification and retrieval in surveillance and multimedia data. However, the regions from which such features are extracted are usually only very crudely segmented, due to the difficulty of segmenting highly articulated entities such as persons. In order to overcome the problem of unconstrained poses, we propose a segmentation approach based on a large number of part detectors. Our approach is able to separately segment a person's upper and lower clothing regions, taking into account the person's body pose. We evaluate our approach on the task of character retrieval on a new challenging data set and present promising results.

1. Introduction

In this paper we address the problem of clothing segmentation for person retrieval. We focus on character/actor retrieval in multimedia data, specifically in TV series. Character/actor retrieval allows users to quickly search for scenes with their favorite actors or characters within a TV episode, a full series or even their complete media collections. However, our segmentation approach is general enough to be applied to other application domains as well. In fact, our method does not even require video as input but works fine on still images.

Clothing-based person retrieval is a promising but challenging approach to person retrieval in multimedia data. Approaches that solely rely on facial features fail when faces cannot be observed due to occlusion, extreme non-frontal poses or too small resolutions. Clothing information can help to overcome those problems, but, for a good representation of a person's clothes, the respective regions have to be segmented first. This stands in contrast to the fact, that usually a person detector or tracker only provides a rough location estimation (e.g. a bounding box) of a person.

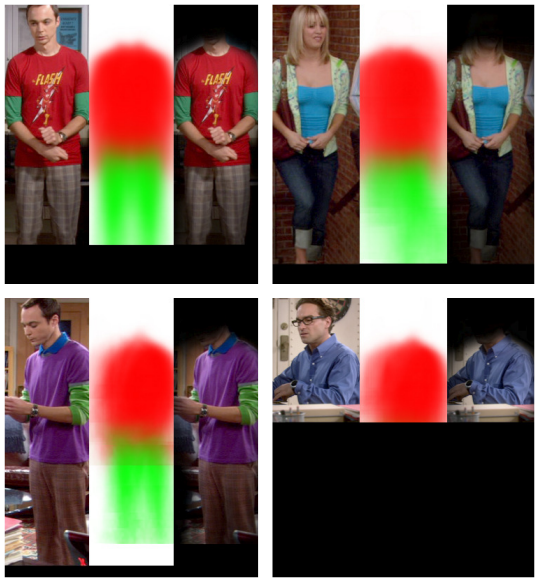


Figure 1: Segmentation masks and results. For each person from left to right: 1. cropped bounding box around the person, 2. computed segmentation mask for the person's pose (red denotes the upper clothing region, green the lower clothing region), 3. segmented clothing regions.

We propose a segmentation approach which employs a large number of pose specific part detectors. For each of the parts, we learn a segmentation mask in order to segment the clothing of a person in the respective pose part. Using a part-based segmentation approach will help to deal with occlusions and the non-rigid nature of the human body. We combine the individual part segmentations to obtain a full segmentation of the person's clothing in order to compute descriptors for re-identification and retrieval. The segmentation information is also used to adjust the influence of specific clothing part descriptors depending on their actual visibility for a given instance.

The remainder of this paper is structured as follows: Sec-

tion 2 discusses related work. In Section 3 we present our approach for learning, refining and applying our part segmentors. In Section 4 we describe how we use the segmented clothing regions to perform person retrieval and in Section 5 we present results on a data set of still images of two episodes from the sitcom *The Big Bang Theory*.

2. Related work

Person re-identification and retrieval in both multimedia and surveillance data have received growing interest over the last years. Approaches can be divided into those based on biometric features such as facial features (e.g. [2, 3, 4, 8]), those based on clothing information alone (e.g. [5, 10, 11, 14, 18]), and hybrid approaches (e.g. [1, 9, 16, 17]).

Most approaches which employ clothing information focus on improving the features, and use rather crude segmentation methods. If video from static cameras is available, background segmentation or difference images can be used to segment the persons [10, 18]. However, in unconstrained multimedia data, one often has to deal with moving cameras, quickly changing backgrounds and large foreground objects (i.e. close views on the actors) which leads to failing background segmentation.

In the case of still images, the most popular method is to detect faces first, and then use a fixed region relative to the detected face, such as a rectangle in some relative distance beneath the face [1, 12]. Song et al. [16] additionally adjust the position of the box in order to minimize the overlapping of clothing box regions of different persons. Also, just using the full person bounding box as region for the description can suffice to perform some re-identification [11, 14]. Sivic et al. [15] fit bounding boxes to hair, face and body regions based on a pictorial structure model.

Gallagher et al. [9] pre-segment the image into superpixels, and then use superpixels from the region beneath a face detection to acquire a more accurate clothing segmentation. If multiple images of the same person are available, they improve the segmentation by co-segmenting multiple regions simultaneously.

Our approach is probably most similar to the segmentation step of the Implicit Shape Model (ISM) [13]. In contrast to the ISM, our masks are based on discriminatively trained part detectors instead of clustered local features. Furthermore, the part detectors we employ are trained to detect specific body part configurations, whereas the ISM parts are clustered based on their (visual) similarity in appearance space alone, possibly mixing different poses in configuration space.

3. Part-based clothing segmentation

Mask-based segmentation for non-rigid entities is difficult since both the viewpoint and the actual pose of the object can influence the visible outline of an entity (here a person). We deal with this by using a large number of local part detectors that are trained to detect specific joint configurations within a local region. We build upon the poselet approach by Bourdev and Malik [7]. Poselets are such part detectors that discriminatively detect specific joint configurations and have been successfully applied for person detection [6, 7]. For each of these given part detectors, we train a specific segmentation mask which segments the clothing of a person in the respective patch of a pose. Given these masks, we apply the segmentation to an unknown image by first detecting the relevant poselets, and then combining the part masks to a full segmentation of the person's clothing regions.

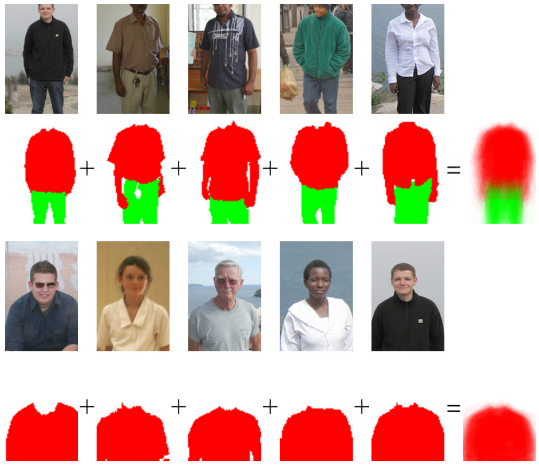


Figure 2: Creating masks from multiple training images. The red regions denote the upper clothing, the green regions the lower clothing.

3.1. Mask training

Based on the way the poselet detectors are trained, we can make the assumption that in all detections of a specific poselet the same joint is approximately at the same position.

For creating the part segmentation masks, we use the region labeled images from the H3D data set [7]. This data set provides images labeled with 15 types of person-related regions e.g. upper clothes, face, background or occlusion. We use these region labels to build two binary images for each person in the data set – one containing the upper and the other containing the lower clothing regions. We then create probability masks for each of the poselets using multiple detections from the corresponding detector. Two masks are created for each poselet, one resulting from all detections upper clothes binary images and the other resulting from the lower clothes binary images.

The training images for the pose-specific mask are obtained by running the poselet detectors on the H3D data set.

To limit the influence of weak detections, we discard all poselet detections with a detection score $s < 1$. From the remaining detections, we use the first n occurrences with the highest detection scores, where n controls the trade-off between quantity and quality of the used occurrences. We empirically determined $n = 50$ which in our experience provides a reasonable trade-off for mask creation.

For some poselets, there are less than n detections on the training data set. This usually happens if the underlying pose is underrepresented in H3D. For those poselets, we do not create segmentation masks. They are ignored in the further segmentation process because we determined experimentally that they actually degrade the segmentation performance. This is due to the fact that anomalies have a too large influence on the segmentation mask if the number of samples is too small.

Given the binary masks of the *good* detection occurrences, we train a 64×96 pixel mask for a specific poselet i by properly resizing the labeled binary masks and then calculating the weighted average

$$p_i(x,y) = \frac{\sum_k b_k(x,y) \cdot s_k}{\sum_k s_k} \quad , \tag{1}$$

where s_k is the detection score – i.e. the distance to hyperplane of the support vector machine classifier – of the k-th poselet detection and $b_k(x,y)$ denotes pixel (x,y) of the corresponding binary mask in the training set. See Figure 2 for an illustration of the mask training.

3.2. Mask refinement

The training so far was based on unsupervised poselet detections. However, this poses a problem if there are false detections among the ones used for training the masks. To reduce the influence of such bad detections on the final mask, we remove those masks from the training set b_k that deviate too much from the trained mask p_i, as determined by the distance

$$d(b_k, p_i) = \sum_{x,y} |b_k(x,y) - p_i(x,y)|. \tag{2}$$

All training images with a distance larger than a threshold θ are discarded. The mask is then re-trained with the remaining images.

3.3. Combination of part segmentations

For segmenting a person's clothes in an unknown image, we assume that we roughly know where the person is located in the image. This can, for example, be achieved by using the poselet based person detector [7], but basically any person detector that outputs a bounding box around the person is suitable. In our experiments we will use hand labeled bounding boxes in order to simulate a perfect person detector, thus also taking those person instances into account which would be missed for example by the poselet person detector.

Figure 3: Combination of part segmentations. After running the part detectors on the image, the corresponding part masks are combined by computing the weighted average of the masks at each pixel. Note that usually there are much more than the four depicted poselet detections.

We run all poselet part detectors on the image and record their detections. Then, we find those poselet detections that match the person detection (i.e. the poselet's vote to the person's center is closer to the detection bounding box center than a given threshold). Both steps integrate nicely with the use of the poselet person detector since all results from the person detection step can actually be re-used.

We can now combine the individual masks to obtain a full segmentation of the person's upper and lower clothing. Similar to the training, the masks are weighted by the detection confidence in order to give more weight to masks where the detector is confident that it found the correct pose. While detection scores from different poselets are technically not comparable (the poselets' classifier is a SVM and their confidence is given by the distance to the hyperplane), we found in practice that they still provide valid weights for the individual masks.

The combination of the masks to the segmentation $p(x,y)$ is done according to the following:

$$p(x,y) = \frac{\sum_{k=1}^{|H_{x,y}|} p_k(x - x_{0k}, y - y_{0k}) \cdot s_k}{|H_{x,y}|} \quad , \tag{3}$$

where $H_{x,y}$ denotes the set of all poselet detections cover-

ing pixel (x, y), s_k is the score of poselet detection k and p_k is the trained segmentation mask (cf. Equation 2) for this poselet. x_{0k} and y_{0k} denote the detection location of poselet k. The normalization factor $|H_{x,y}|$ assures that pixels do not get high segmentation confidences simply because of many detections. See Figure 3 and Algorithm 1 for an illustration of the combination procedure.

Algorithm 1 Segmentation Process

```
compute poselets on patch
select relevant poselets by position
prob_img = zeros(patchsize)
occurrence_img = zeros(patchsize)
for p in poselets do
    mask = resize(get_mask(p.Id))
    mask *= p.detection_score
    for x,y = 1 to patchsize do
        prob_img((x,y)+p.position) += mask(x,y)
        occurrence_img((x,y) + p.position) += 1
    end for
end for
prob_img /= occurrence_img
```

4. Person retrieval

In order to demonstrate what can be achieved simply by a better clothing segmentation, we use only simple color features to describe the segmented regions. Given a database of images or videos, we first compute the segmentation masks for the upper and lower body regions of each person instance in the database. We then compute two RGB histograms ($32\times32\times32$ bins each) for each person, one for the segmented upper and one for the segmented lower clothing region. The segmentation masks weight each pixel's contribution to the histogram.

A query instance of a person can now be used to find other occurrences of this person in the database by computing the distance between the input descriptors and the descriptors of each person in the database. The results are then ranked according to the distance and reported to the user.

In order to compare two descriptors H_i and H_j we use the following distance function:

$$d(H_i, H_j) = w_u \cdot d_h(H_{iu}, H_{ju}) + w_l \cdot d_h(H_{il}, H_{jl}) \quad , \quad (4)$$

where d_h is any histogram distance function (in our experiments we use the Bhattacharyya distance) and H_{iu}, H_{il} are the histogram descriptors for upper and lower clothes of person i. The weights w_u and w_l describe the relative expressiveness of the respective clothing region and are determined by

$$w_r = min(\frac{1}{a_1} \cdot \sum_{x,y} p_{1r}(x, y),\ \frac{1}{a_2} \cdot \sum_{x,y} p_{2r}(x, y)) \quad , \quad (5)$$

where $r \in \{u, l\}$, $p_{ir}(x, y)$ is the probability mask of person i and region r, and a_i denotes the area of the bounding box of person i. The weights are normalized such that $w_u + w_l = 1$. This effectively weights a region's influence on the final distance based on the minimum support it has in either region, and thus helps to mitigate problems caused by reduced visibility of one of the regions, e.g by occlusions. For example, when the legs are (partly) occluded in either query or candidate image, their descriptors contribute less to the overall distance.

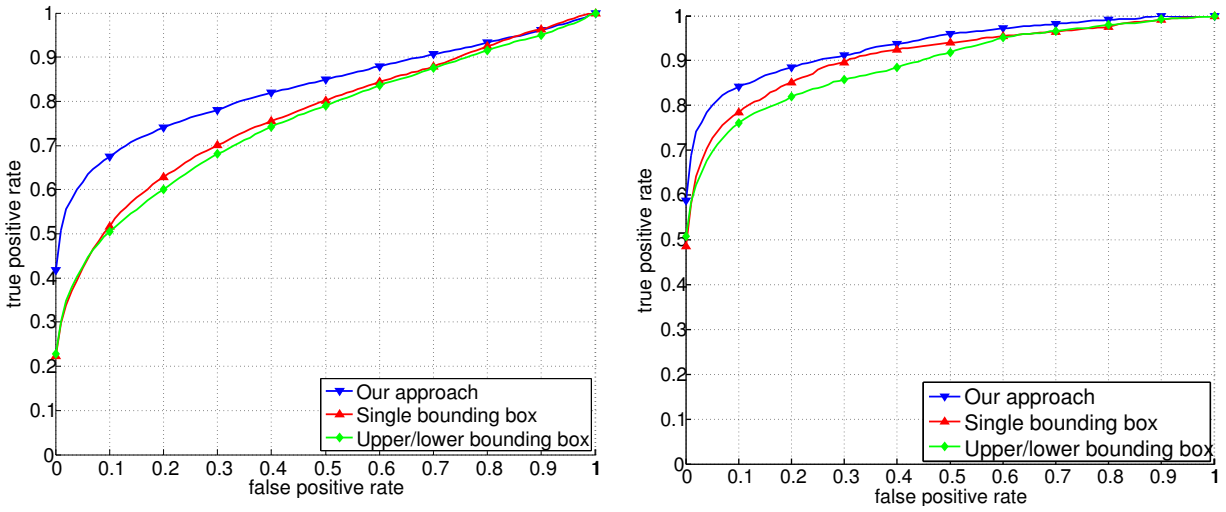
Figure 5: Some images from the Big Bang Theory data set. The data set contains people in different poses, seen from different viewpoints and with varying occlusions.

5. Evaluation

For the evaluation of our approach, we collected a new data set for person detection and retrieval in multimedia data based on the sitcom *The Big Bang Theory (BBT)*[1]. The data set consists of images from episodes one and three of season one. In both episodes every 250-th frame was selected, what is approximately one image every 10 seconds. In total, this resulted in 132 images with 228 person occurrences for episode one, and 128 images with 198 person occurrences in episode three.

For each person occurrence, we labeled a rough bounding box and their identity. Obviously, the same person can look very different in two images if wearing different clothes. Therefore, we also labeled the current clothing configuration of each person. Many characters in the BBT data set occur several times, with multiple clothing configurations each. For details see Table 1. The data set is quite

[1] We will make the data set available for research purposes on our website `http://cvhci.anthropomatik.kit.edu/projects/pri`



(a) (b)

Figure 4: Evaluation results for (a) episode S1E1 and (b) episode S1E3. Our segmentation approach improves the retrieval performance on both episodes.

challenging due to the large number of different poses and view angles. Also, quite often parts of a person are occluded (e.g. the legs in close-up shots). See Figure 5 for some sample images from the data set.

Episode	# Images	# Labels	# Persons	# Cloth. Config.
S1E1	132	228	10	12
S1E3	128	198	13	23

Table 1: Statistics for the two episodes of the Big Bang Theory data set.

5.1. Baseline methods

We compare our approach against two baseline methods which use the labeled bounding boxes as segmentation. As features to describe the clothing regions we use the same features as for describing the segmented clothing regions in our approach, that is RGB color histograms with $32 \times 32 \times 32$ bins. For the first baseline method (*single bounding box*), we compute the histogram over all pixels in the annotated bounding box. The histograms of two person occurrences are compared by the Bhattacharyya distance.

In the second method (*upper/lower bounding box*), we divide the bounding box horizontally into two boxes of the same size. The upper box represents the upper clothing region, the lower box the lower clothing region. As descriptor for the second method, we compute one RGB histogram for each box, similar to our approach with the two segmented clothing regions. We compare descriptors for two persons by calculating the average of the Bhattacharyya distances between the histograms of the corresponding regions.

5.2. Experimental results

We evaluate our approach on the two episodes of the *Big Bang Theory* data set. We perform character retrieval using each of the labeled persons as query image. All other persons in the episode form the candidate set. The goal of the retrieval is to find all other occurrences of the same person in the same clothing configuration. Obviously, we cannot expect for a fully clothing-based approach to also find occurrences of the same person but in different clothes. The results are reported as average over all possible queries in terms of True Positive Rate (TPR) vs. False Positive Rate (FPR).

The retrieval results of both episodes can be found in Figure 4. Our segmentation approach clearly helps to improve the retrieval performance on both episodes. For episode S1E1 we achieved an initial average TPR of 42% with our approach compared to 23% and 23% for the two baseline methods. As expected, the more accurate segmentation of the clothing regions improves the quality of the descriptors significantly. For episode S1E3 we achieved an initial TPR of 58% vs. 50% and 48%. For this episode, the retrieval performance is generally better which is most likely due to the fact that in this episode there are longer scenes where some characters largely remain in a similar pose (Penny and

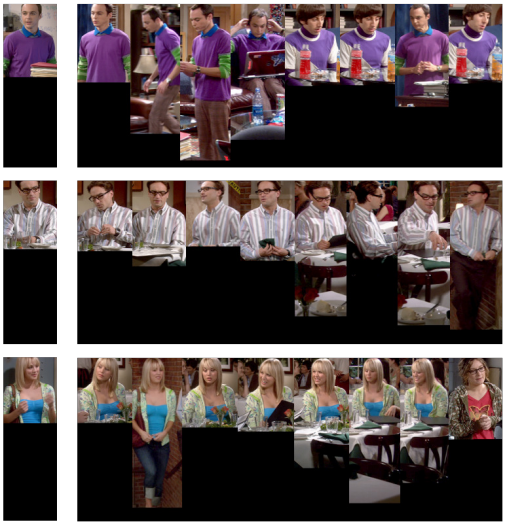


Figure 6: Multiple queries (left image) and the top 8 results. Query and candidate images can be in quite different poses (e.g. standing vs. sitting). Some problems remain: The query image in the first row contains some red from a book, which is too special to be removed by the segmentation mask. Therefore, some false results of Howard (first row, columns 5, 6 and 8) with the likewise purple shirt and a red bottle are ranked quite high. Such false positives could however very likely be removed by (a) better descriptors of the clothing region and (b) a hybrid approach which also takes facial features into account.

Leonard sitting in the restaurant). Some sample queries and their first ranked results can be found in Figure 6.

6. Conclusion

We presented a novel clothing segmentation approach that can deal with unconstrained poses and occlusions of persons. This is achieved by employing a large number of pose-specific part detectors for each of which we learn one segmentation mask. The masks' quality is further improved by filtering out bad training images in a refinement step. Relevant clothing regions in a new image are segmented by first detecting the parts, and then applying the masks at the detection locations. We have shown that a good segmentation obtained by our approach leads to increased person retrieval performance on a challenging multimedia data set.

7. Acknowledgments

We thank Lubomir Bourdev for making the H3D dataset and his poselet part detectors available. This research has been partially funded by the German Federal Ministry of Education and Research (BMBF) under project PaGeVi.

References

[1] D. Anguelov, K.-C. Lee, S. B. Gokturk, and B. Sumengen. Contextual Identity Recognition in Personal Photo Albums. In *CVPR*, 2007.

[2] N. Apostoloff and A. Zisserman. Who are you? Real-time person identification. In *BMVC*, 2007.

[3] O. Arandjelovic and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. In *CVPR*, 2005.

[4] M. Bäuml, K. Bernardin, M. Fischer, H. Ekenel, and R. Stiefelhagen. Multi-Pose Face Recognition for Person Retrieval in Camera Networks. In *AVSS*, 2010.

[5] A. Borràs, F. Tous, J. Llados, and M. Vanrell. High-Level Clothes Description Based on Colour-Texture and Structural Features. *Pattern Recognition and Image Analysis*, pages 108–116, 2003.

[6] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting People Using Mutually Consistent Poselet Activations. In *ECCV*, 2010.

[7] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009.

[8] M. Fischer, H. K. Ekenel, and R. Stiefelhagen. Person re-identification in TV series using robust face recognition and user feedback. *Multimedia Tools and Applications*, 2010.

[9] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *CVPR*, 2008.

[10] N. Gheissari, T. B. Sebastian, and R. Hartley. Person Reidentification Using Spatiotemporal Appearance. *CVPR*, 2006.

[11] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *ECCV*, 2008.

[12] G. Jaffré and P. Joly. Improvement of a Person Labelling Method Using Extracted Knowledge on Costume. *Computer Analysis of Images and Patterns*, pages 489–497, 2005.

[13] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77(1):259–289, 2008.

[14] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *Pattern Recognition*, 36:1997–2006, 2003.

[15] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: video shot retrieval for face sets. In *CIVR*, 2005.

[16] Y. Song and T. Leung. Context-Aided Human Recognition-Clustering. *ECCV*, 2006.

[17] D. a. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. *Workshop on Applications of Computer Vision*, 2009.

[18] K. Yoon, D. Harwood, and L. Davis. Appearance-based person recognition using color/path-length profile. *Journal of Visual Communication and Image Representation*, 17:605–622, 2006.