A Bottom-Up, View-Point Invariant Human Detector

Nicolas Thome, Sébastien Ambellouis
Laboratoire Électronique, Ondes et Signaux pour les Transports, INRETS
20, rue Elisée Reclus BP 317
59666 Villeneuve d'Ascq Cedex, France

Abstract

We propose a bottom-up human detector that can deal with arbitrary poses and viewpoints. Heads, limbs and torsos are individually detected, and an efficient assembly strategy is used to perform the human detection and the part segmentation. Firstly, a topological model is used to represent the structure of the human body, and the topologically equivalent configurations are ranked with additional priors. Promising results prove the approach efficiency for detecting people in low-resolution and compressed images.

1. Introduction

Detecting humans in images and inferring their pose is arguably one of the most challenging problem in computer vision due to large variations in body shape, appearance, clothing, illumination and background clutter.

Top-down approaches use global models of the human body for inferring the part location and/or detecting humans, by minimizing a given model to image criterion. Exemplar-based approaches [1] propose to learn discriminatively the mapping between image features and model parameters. Realizing the difficulties of using 3D models, many researchers have used of 2D assembly of parts, such as cardboard models or pictorial structures [4]. The main shortcoming related to pictorial structures corresponds to the fact that the geometrical relationship between parts is strongly view-dependent. More generally, top-down strategies are sensitive to occlusions, as they attempt to represent the assembly of body parts with a single model.

Bottom-up approaches typically adopt a two-stage strategy: a bottom-up detector is applied on the image to extract candidate parts, then a top-down procedure makes inference about the configuration and finds the best assembly. Mori *et. al.* [7] propose to use Normalized Cuts to find a few salient body parts, and then solve the assembly problem by brute-force search. However, the bottom-up detectors rely on very specific features like focus, that are validated for a set of base-ball images, but that do not seem appropriate in another context. Mikolajczyk *et. al.* [6] propose to represent individual parts by co-occurrences of SIFT features, that are learned from training images using AdaBoost. Although the part recognition rates are impressive, the system is specially dedicated to recognizing faces, upper body and legs learned form frontal and profile viewpoints. Therefore, and it is not clear how well the method will generalize for detecting humans in arbitrary poses and viewpoints. The work of Ren *et. al.* [8] is the most relevant to us. They propose to model body parts as pairs of parallel straight lines, and to solve the assembly problem using Integer Quadratic Programming.

Our contribution for providing a bottom-up human detector is three-fold. Firstly, we build part detectors that are robust to (self)-occlusions and that are computationally efficient (section 2). Secondly, we make use of a topological model for the assembly (section 3.1), providing a view-point invariant human detector. Finally, we incorporate additional priors for ranking the topologically equivalent configurations (section 3.2).

2. Detecting Body Part Candidates

We attempt here at detecting heads, legs, arms and torsos. We want our detectors to have high recall performances, *i.e.* able to detect most of the limbs present in the images. We can, however, tolerate a relatively low precision, as we use the body model assembly presented in section 3 to filter out the spurious detections, and provide a human detector.

2.1 Head Detector

The head is identified by a combination of a face detector and a circle template matching. The face detection is performed by means of the Viola-Jones cascade detector [10]. We use both front and side views detectors, and setup the detection with a maximum sensitivity. In order to identify head facing back the camera, we use a template shape matching approach dedicated to locating circles. We use an algorithm based on the Distance Transform (DT) matching [5]. Let us consider the original image I, and let us denote I_E the binary image obtained by applying a Canny edge detector. Basically, the DT matching consists in computing a dissimilarity between a segmented template T_c representing the prior shape (*e.g.* a circle). Computing the DT of I_E leads to a non-binary image, that we denote I'_E. The template T_c is mapped onto I'_E after different affine transformations. The matching measure $DT_{T_c}(I)$ is determined by the pixel values of I'_E which lie "under" T_c:

$$DT_{T_c}(I) = \frac{1}{|T'_c|} \sum_{t \in T_c} I'_E(t) \qquad (1)$$

where $|T_c|$ denotes the number of edge points in T_c and $I'_E(t)$ denotes the chamfer distance between edge t in T_c and the closest edge in I_E.

2.2 Limbs Detector

Legs and arms are modeled as parallel lines of contrast. This is clearly a simplistic representation, but that remains valid for a wide range of poses and viewpoints, as discussed in section 1. Roughly speaking, there exits two different kinds of strategies to detect pairs of parallel segments. Top-Down strategies [4] process similarly as the method described in section 2.1 for our shape-based head recognition. However, the algorithm requires to exhaustively search for rotations and foreshortening when detecting limbs, and is therefore very time consuming, even with a coarse discretization of the parameter space. Alternatively, some authors take advantage of the structure of the image and directly use the edge map I_E, decompose it into segments, and apply pair-wise constraints based on parallelism to identify the candidates [8]. However, it is difficult in unspecified conditions to detect both segments for parts in I_E, mainly due to clutter, occlusions and loose clothing.

In this work, we propose an intermediate solution. Firstly, we extract a set of straight line segments from I_E. Then, we apply a template matching by searching for parallel segments having a low dissimilarity when computing the chamfer distance DT_G stated in equation 1. The well known non-maxima suppression method is applied to only keep the largest response candidates. This hybrid strategy provides a scale invariant limb detector, that is able to detect arms and legs with a single segment extracted from the edge map, without being time consuming: for a given segment, the parameter space is reduced to one parameter, *i.e.* the width of the limb.

2.3 Detecting Torsos

The torso identification is arguably the most difficult task when using bottom-up detectors, due to strong occlusions by the other parts like arms. Therefore, our experiments prove that is not reasonable to expect detecting torsos by identifying even a single segment, due to these strong self-occlusions. For that reason, we choose to use a parallel line template T_l, and to detect torsos by using the chamfer matching technique stated in equation 1. To make the torso detection view point invariant (its width is sensitively larger for a front view than for a side view), the dissimilarities are computed by varying the aspect ratio of the template. We tolerate torso candidates with a quite large dissimilarity threshold θ, since the template poorly explain the image evidence in case of self-occlusions. We come back to that point in section 3.2.1. In addition, to make the approach computationally efficient, we propose to infer the possible torso dimensions from the detections of the other parts, using anthropometric data [2]. Thus, once the different candidates for the limbs are extracted, we cluster them depending on their width value, that gives a good estimate of the size of the part (contrary to its height, due to foreshortening). Then, we run the mean-shift algorithm [3] for clustering the different limbs into an unspecified number of classes. The center of each class is used to guide the torso detection. Our strategy takes advantage of the previous limbs detection results (less sensitive to the occlusions) to specify the torso search area in the parameter space. It makes the detection more accurate without increasing its computational cost.

3. The Part Assembly Process

Once the different candidates detected, we form a set of configurations by merging "loosely neighboring" parts in the image. For each configuration, we compute its dissimilarity D_H of being a human as follows:

$$D_H = w_g D_g + w_t D_{top} + w_a D_{app} + w_l D_{lg} \qquad (2)$$

D_g is a term dedicated to pruning configurations that are not physically valid. We use a strategy inspired from [7], using anthropometric data [2] for checking the size compatibility between limbs. Formally, we have $D_g = 0$ if the configuration is physically valid and $D_g = \infty$ otherwise.

3.1 Topological Human Body Model

D_{top} corresponds to a topological matching between the part assembly and a model of the human skeleton. This score is estimated by using a graph matching strategy inspired from the "shock graphs" [9]. Thus, we generate off-line a topological model graph, denoted G_M, representing the human body structure, *i.e.* the connections between parts. For each formed configuration in the image, we generate on-line a graph from the part assembly, denoted G_I. Matching the part assembly to the topological model is formulated as the problem of estimating the dissimilarity between the structures of the graphs G_I and G_M. For that purpose, we compute a Topological Vector Signature (TSV) for the roots of the two graphs, denoted χ_M^0 and χ_I^0, as detailed in [9]. The dissimilarity D_{top} between G_I and G_M is estimated by computing the euclidean distance between the TSVs of the two roots:

$$D_{top} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\chi_M^0(i) - \chi_I^0(i)\right]^2} \tag{3}$$

where $N = \max\left(\Delta(G_M), \Delta(G_I)\right)$, $\Delta(G_M)$ and $\Delta(G_I)$ being the maximal degree of G_M and G_I, respectively.

The major strength of the graph matching consists in only using the topological information for performing the correspondence, *i.e.* the connection between limbs, leading to a **viewpoint invariant human detector**. Thresholding D_{top} to D_{top}^T makes it possible to seek for assemblies having a structure in accordance to the human body connections and to efficiently filter out spurious configurations. Note that if no configuration scores below D_{top}^T, the detection process concludes that the image does not contain any human. In that case, we have $D_H = \infty$. We insist here on the fact that our topological matching scheme is much more robust to missing parts than top-down approaches trying to determine the model parameters that best explain the image evidence. Indeed, although we use a single model for modeling the part assembly, we can take advantage of the strong background of the shock graphs: for example, theoretical properties have been derived regarding the robustness of the TSV to minor perturbations (see [9]) such as noise, insertion/suppression of node, *etc.*

3.2 Incorporating additional Priors

When a human is effectively present in the image, there are usually many assemblies that are topologically equivalent, and that have to be ranked with another strategy. There are, indeed, dependencies among the body parts that cannot be captured by a tree. The terms D_{app} and D_{lg} of equation 2 are defined for that purpose.

D_{app} is a term that encodes prior about symmetry in clothing and support assemblies for which the appearance of left and right limbs is similar. Let us consider H_l and H_r the color histograms of two left and right detected limbs. The dissimilarity between H_l and H_r is determined using the χ^2 distance:

$$D_{app} = \chi^2(H_l, H_r) = \frac{1}{2}\sum_{i=1}^{B}\frac{\left[H_l(i) - H_r(i)\right]^2}{H_l(i) + H_r(i)} \tag{4}$$

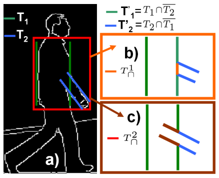
where B denotes the number of bins of the histograms.

3.2.1 Occlusion-sensitive Image Likelihood

When multiple body parts fit the same image region, the independent parts models poorly explain the overall image evidence. Let us consider two parts L_1 and L_2 that overlap in the image plane. Both parts have been detected using the chamfer distances for templates T_1 and T_2, respectively. In the overlap area, only one template model is valid for predicting the image data. The last term of equation 2, $D_{lg}(L_1, L_2)$, corresponds to a more global reasoning about the configuration, dedicated to estimating a combined image likelihood of the assembly, by explicitly taking into account self-occlusions:

$$D_{lg}(L_1, L_2) = DT_{T_1} + DT_{T_2} + min_{i\in\{1;2\}}DT_{\cap}^i \tag{5}$$

Figure 1 illustrates the occlusion-sensitive likelihood formulation, and details the terms of equation 5. Minimizing $DT_{\cap}^i$, $i \in \{1;2\}$, consists in identifying which part partially occludes the other in the overlap area. Figure 1a) illustrates a given binary edge image and two overlapping templates T_1 and T_2 for the torso (in green) and the arm (in blue), respectively. Both possibilities (the torso occluding the arm and *vice-versa*) are shown in figure 1b) and 1c), respectively. In this example, we expect our occlusion-sensitive likelihood to score the configuration c) with the smallest distance $DT_{\cap}^2$, as the arm is actually occluding the torso.



$$DT_{T_1} = \frac{\sum_{t\in T_1} I'_E(t)}{|T_1|} \qquad DT_{T_2} = \frac{\sum_{t\in T_2} I'_E(t)}{|T_2|}$$

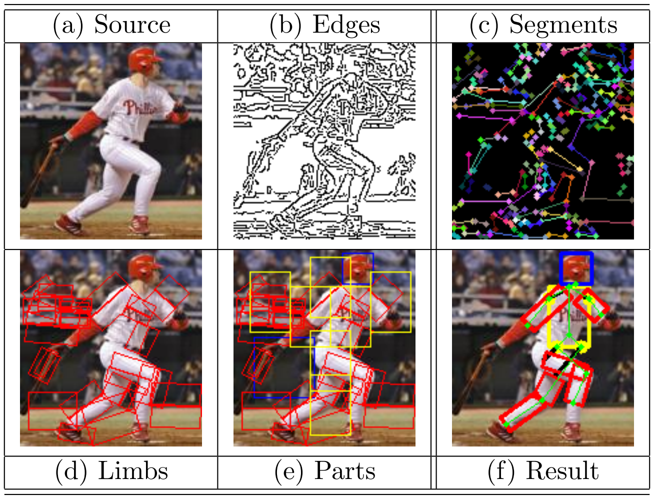
$$DT_{\cap}^1 = \frac{\sum_{t\in T_{\cap}^1} I'_E(t)}{|T_{\cap}^1|} \qquad DT_{\cap}^2 = \frac{\sum_{t\in T_{\cap}^2} I'_E(t)}{|T_{\cap}^2|}$$

Figure 1. Occlusion-sensitive Likelihood

4 Results

We present here results illustrating our approach ability to detect people by the combination of its visible parts, and the possibility to segment the limbs in the image.

Figure 2 illustrates the result for an upright standing pose, with a strong amount of background clutter. From



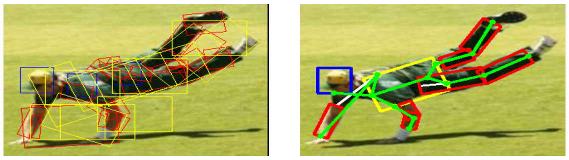
(a) Source (b) Edges (c) Segments

(d) Limbs (e) Parts (f) Result

Figure 2. Body Part Detection and Assembly

the raw image (figure 2a)), we compute the edge map I_E, and extract a set of straight line segments(figure 2b)). The limb candidates are illustrated with red rectangles (figure 2c)). As argued in section 2.2, we can notice that many correct limb hypotheses are generated although a single segment is extracted from I_E (*e.g* upper arms, right lower leg). Figure 2d) illustrates the results for the overall set of bottom-up detectors, *i.e.* limbs, heads (blue), and torsos (yellow). There are many false positives. Indeed, the parallel line model is itself not very distinctive, and our detector happily fires on background regions. Moreover, we set our torso and head detectors with a maximum sensitivity, to overcome our simplistic part representation and the auto-occlusions between limbs. The model assembly makes it possible to efficiently filter out the spurious detections, by incorporating the *prior* knowledge detailed in section 3, and to detect that a person is present in the image. The configuration with the largest likelihood with respect to equation 2 is presented in figure 2e). In this example, we can notice that there are strong auto-occlusions from the upper arms to the torso, making the occlusion-dependent likelihood formulated in section 3.2.1 efficient for modeling the assembly. In addition, the symmetry in clothing prior is here efficient for removing assemblies that include "background parts".

Figure 3 illustrates the ability of the approach to detect someone plunging on the ground. Indeed, neither the part detection nor their assembly make any assumption about the pose of the person. Thus, the part labeling and segmentation is efficient for any viewpoint, making the approach applicable in more general settings than pedestrian detectors. In addition, we claim that our human detector is robust to strong image degradations, because the part segmentation only use contrast features. For example, the images shown in figure 3a) and 3b) have a small resolution (136 × 147 and 184 × 147, respectively), are compressed (JPEG), there is motion blur, *etc.* It proves that our generic part detection scheme combined with a modeling of the articulated human structure can successfully achieve human detection in various pose and view points, and remains efficient in complex situations.



(a) Part detection (b) Best Assembly

Figure 3. Pose-Invariant Part Assembly

5 Conclusion and Future Works

We present a bottom-up approach for detecting humans in images with arbitrary poses. The main direction for future works consists in incorporating the temporal aspect of the image sequence for providing a video-based human detector.

References

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *PAMI*, 28(1), jan 2006.

[2] N. A. anthropometric data of children. http://ovrt.nist.gov/projects/anthrokids/, 1977.

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

[4] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61(1):55–79, 2005.

[5] D. M. Gavrila. A bayesian, exemplar-based approach to hierarchical shape matching. *PAMI*, 29(8):1408–1421, 2007.

[6] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV*, volume I, pages 69–81, 2004.

[7] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR*, volume 2, pages 326–333, 2004.

[8] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *CVPR*, volume 1, pages 824–831, 2005.

[9] A. Shokoufandeh, D. Macrini, S. Dickinson, K. Siddiqi, and S. W. Zucker. Indexing hierarchical structures using graph spectra. *IEEE PAMI*, 2005.

[10] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.