Finding people in repeated shots of the same scene

Josef Sivic[1] C. Lawrence Zitnick[2] Richard Szeliski[2]

[1] University of Oxford [2] Microsoft Research

Abstract

The goal of this work is to find all occurrences of a particular person in a sequence of photographs taken over a short period of time. For identification, we assume each individual's hair and clothing stays the same throughout the sequence. Even with these assumptions, the task remains challenging as people can move around, change their pose and scale, and partially occlude each other.

We propose a two stage method. First, individuals are identified by clustering frontal face detections using color clothing information. Second, a color based pictorial structure model is used to find occurrences of each person in images where their frontal face detection was missed.

Two extensions improving the pictorial structure detections are also described. In the first extension, we obtain a better clothing segmentation to improve the accuracy of the clothing color model. In the second extension, we simultaneously consider multiple detection hypotheses of all people potentially present in the shot.

Our results show that people can be re-detected in images where they do not face the camera. Results are presented on several sequences from a personal photo collection.

1 Introduction

The goal of this work is to find all occurrences of a particular person in a sequence of photographs for use in labeling and browsing. We assume these photographs were taken within a short period of time, such as a few minutes. This is a challenging task since people can move around, change their pose and scale, and occlude each other. An example of such a sequence is shown in figure 1.

One approach to this problem is to find people using face detection [12, 13, 19, 21] and recognition [24]. While these techniques can typically find many people within a scene, some people will go undetected due to them not facing the camera or occlusion. In these cases, as we will explore in this paper, other cues can be used for detection such as hair or clothing.

We restrict our problem by assuming that: people are roughly in an upright position; their head and torso are visible; background colors are stable across shots; people are distinguishable by the color of their clothing; and each person is facing the camera (and is detected by the face detector) in at least one shot of the sequence.

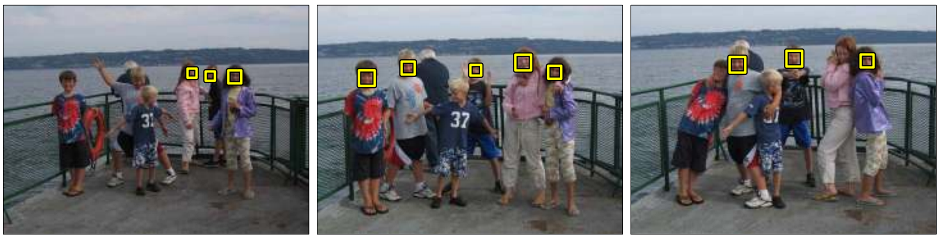


Figure 1: A sequence of repeated shots with people moving around and occluding each other (3 of 5 photos shown). Face detections are overlaid in yellow. Although the face detector is quite successful, some faces are missed, usually due to the fact that they are not facing the camera.

To detect people across images, even when they are not facing the camera, we propose a two stage algorithm. In section 3, we describe the first stage in which we identify people by grouping together all face detections belonging to the same individual. In the second stage, occurrences for each person are found for cases in which face detections are missing. As discussed in section 4, this is accomplished for each individual using a color-based pictorial structure model created from the people groupings found in the first stage. Finally, two extensions for improving the detections of the pictorial structure models are described in section 5. The first extension is a method for segmenting partially overlapping people to obtain more accurate color models. The second extension handles cases in which individual detections may be ambiguous by simultaneously considering multiple detection hypotheses for all people potentially present in the shot. Results of our algorithm are presented for several sequences of photos with varying conditions and numbers of people in section 6.

2 Related work

Clothing colors have been used before to recognize people with detected frontal faces in a personal photo collection [20, 23]. A torso color statistic is computed from a rectangular region below a frontal face detection and people are matched on the clothing colors in addition to the face similarity. In this work we go beyond face detections and re-detect people in photographs where their face detection is missing.

The task we carry out is different from pedestrian detection [2, 10, 22] as we do not require the full body to be visible. Moreover, we build an appearance model for each person rather than using generic (e.g. edge based) person templates.

Another set of approaches for detecting people is to model a person as a collection of parts [7, 11, 13, 14, 17]. In particular, we build on the approaches of Felzenszwalb and Huttenlocher [7] and Ramanan *et al.* [17], where the human body is modelled as an articulated pictorial structure [9] with a single color appearance model for each part. In [17], impressive human detection results are shown on videos with mostly side views of a single running or walking person. A two part pictorial structure model (torso and head) has also been used for detecting particular characters in TV video [5]. In contrast, personal photo collections often contain images of multiple people and we aim to identify the same people across a set of photos. Moreover, people wear clothing with multiple colors and we pay attention to this by explicitly modelling parts using mixtures of colors.

Clothing colors have been also used to identify people in surveillance videos [4, 15],

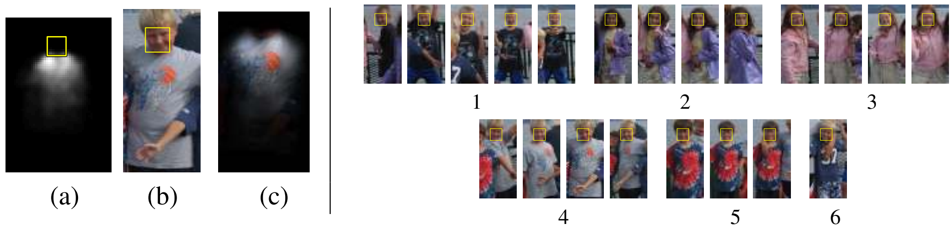


Figure 2: Left: (a) Spatially weighted mask shown relative to a face detection. Brightness indicates weight. This mask is used to weight contributions to a color histogram describing each person's clothing. (b) Cut out of a person with face detection superimposed. (c) Weight mask superimposed over the image. Right: (1)-(6) The clusters correctly computed for the six different people found in the photos shown in figure 1.

where the regions of interest (people detections) are usually given by background subtraction.

3 Grouping initial face detections

Our first goal is to find the different people present in the sequence. This can be achieved by first detecting (frontal) faces [12, 21] and then grouping them into sets belonging to the same person based on their clothing colors.

A color histogram is created for the clothing corresponding to each face detection. The histogram is computed from a region below the face detection using the spatially weighted mask shown in figure 2. The mask and its relative position to the detected face were obtained as an average of 35 hand-labelled clothing masks from about 30 photos of people with detected faces. 16 bins are used for each RGB channel resulting in a histogram with 4,096 bins.

We group the histograms belonging to the same individual using a single-linkage hierarchical agglomerative clustering algorithm [3], where the similarity between two histograms is measured using the χ^2 distance [16]. This simple clustering algorithm works very well in practice and correctly discovers the different people present in a photo sequence as demonstrated in figure 2.

Face similarity is not currently used as a distance function in our system, mostly in order to focus our research on exploiting clothing information. However, incorporating both face and clothing information could be used for this stage in the future.

4 Detecting people using pictorial structures

The frontal face detector usually does not find all occurrences of a person in the photo sequence. This can be due to a high detection threshold (the face detector is tuned to return very few false positives) or because the person does not face the camera or the face is not visible. Within this section, our goal is to find people whenever they are visible in the photo.

We model a person as a pictorial structure [7, 17] with three rectangular parts corresponding to hair, face and torso regions as shown in figure 3(a). We choose not to model the limbs explicitly as they might not be visible are or hard to distinguish from the torso and background in many shots.

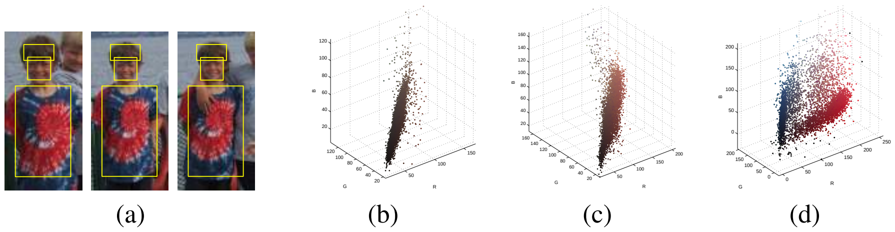


Figure 3: (a) Rectangular parts corresponding to hair, face/skin, and torso for three different image cutouts that had a frontal face detected. (b)-(d) Samples from a Gaussian mixture model in RGB space learned for hair, face/skin and torso respectively. The color of each point corresponds to its actual RGB value. Note how in the case of the torso (d) the Gaussian mixture model captures the two dominant colors (blue and red).

Using the pictorial structure model, the posterior probability of a particular arrangement of parts, $l_1, \dots, l_n$, given an image I is decomposed as

$$P(l_1, \dots, l_n | I) \propto P(I | l_1, \dots, l_n) P(l_1, \dots, l_n), \tag{1}$$

where $P(I|l_1, \dots, l_n)$ is the image appearance likelihood and $P(l_1, \dots, l_n)$ is the prior probability of the spatial configuration. Assuming parts do not overlap, the appearance likelihood is factorized as $\prod_{i=1}^{n} P(I|l_i)$. In this work we use a 'star' [1, 8] pictorial structure (tree of depth one), i.e. the locations of parts are independent given the location of the root (reference) part l_r, $P(l_1, \dots, l_n, l_r) = P(l_r) \prod_{i=1}^{n} P(l_i|l_r)$. Assuming $P(l_r)$ to be uniform, the posterior probability given by (1) can be written as

$$P(l_1, \dots, l_n, l_r | I) \propto \prod_{i=1}^{n} P(I|l_i) \prod_{i=1}^{n} P(l_i|l_r). \tag{2}$$

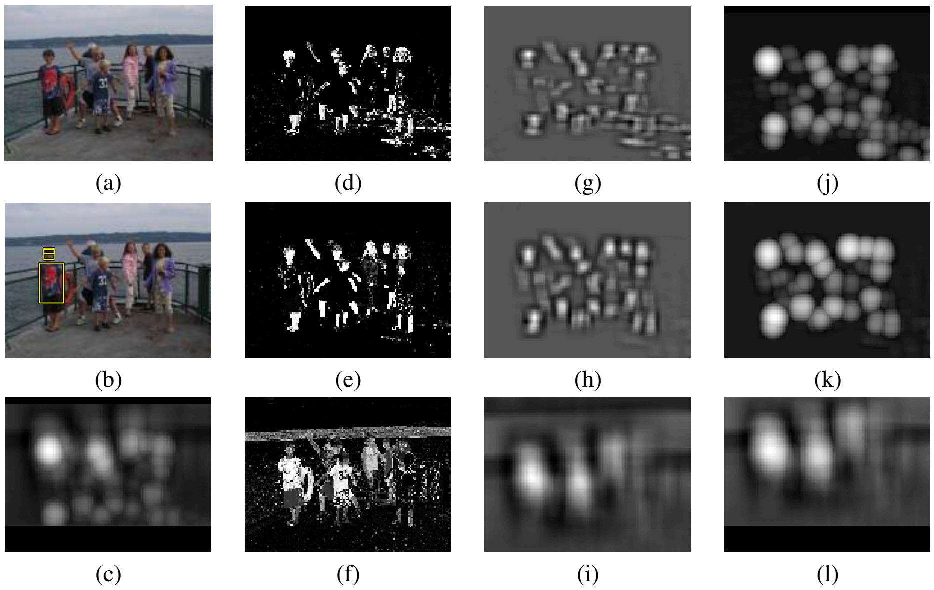
4.1 Modelling the appearance of each part

Given a face detection, three rectangular parts covering the hair, face and torso are instantiated in the image as shown in figure 3(a). The position and size of each part relative to the face detection is learned from manually labelled parts in a set of photos.

The appearance of each part is modelled as a Gaussian mixture model (GMM) with $K = 5$ components in RGB color space. To reduce the potential overlap with the background, the color model is learned from pixels inside a rectangle that is 20% smaller for each part. The GMM for each part is learned from all occurrences of a particular person as found by the clustering algorithm in section 3. Figure 3 shows an example of proposed part rectangles and learned color Gaussian mixture models. Note that using a mixture model is important here as it can capture the multiple colors of a person's clothing. In addition to the color models for each person, we learn a background color model (again a GMM in RGB space) across all images in the sequence.

4.2 Computing part likelihoods

Given a foreground color model for part p_j and a common background color model, the aim is to classify each pixel x as belonging to the part p_j or the background. This is



(a) (d) (g) (j)

(b) (e) (h) (k)

(c) (f) (i) (l)

Figure 4: (a) Image 1 of the sequence in figure 1. The goal is to detect the person shown in figure 3(a). In this image, the face of this person was not detected by the frontal face detector. (b) The configuration of parts with the minimum cost defined by equation (4). (c) The negative cost of the location of the virtual reference part (c.f. equation (6)). Note how the largest peak corresponds to the face of the correct person. (d)-(f) Foreground color posterior probability for hair, face/skin and torso respectively, computed according to equation (3). (g)-(i) Part log-likelihood for hair, face/skin and torso respectively using the rectangular box model described in section 4.2. (j)-(l) Distance transform of part log-likelihood images (g)-(i). Note that the distance transform images are shifted by offset to the virtual reference position. See text for more details. In (c)-(l) brightness indicates value.

achieved by computing the posterior probability that pixel x belongs to part p_j as (assuming equal priors)

$$P(p_j|x) = \frac{P(x|p_j)}{P(x|p_j) + P(x|bg)}, \tag{3}$$

where $P(x|p_j)$ and $P(x|bg)$ are likelihoods under the part p_j and background color GMM respectively. Examples of posterior probability images are shown in figure 4(d)-(f).

Similarly to [6], a matching score $\rho_i(l_i)$ for each part i is computed by convolving the posterior probability image with a rectangular 'center-surround' box filter. The aspect ratio of the box filter is fixed (and different) for each part with the surround negative area being 20% larger than the positive central area. Each filter is normalized such that both positive and negative parts sum to one. Note that the matching score ranges between -1 and +1 with +1 corresponding to the ideal match. The likelihood of part i at particular location l_i is defined as $\exp(\rho_i(l_i))$. In practice, the exponentiation is never performed explicitly as all the computations are done in negative log space. To detect occurrences of a person at different scales the box filter is applied at several scales. Note that the negative ('surround') part of the filter is important to ensure a maximum filter response at the correct scale. Examples of the part log-likelihood images are shown in figure 4(g)-(i).

4.3 Prior shape probability

The location of each part is described by its x, y position in the image. The shape probability $P(l_i|l_r)$ of part i relative to the reference part r is modelled as a 2D Gaussian distribution with mean μ_i and covariance Σ_i, $\mathcal{N}(l_i - l_r; \mu_i, \Sigma_i)$. The mean μ_i is learned from labelled training data, while the covariance Σ_i is diagonal and set manually. Empirically, the covariances estimated from the labelled data are too small and result in a very rigid pictorial structure. The reason might be that the labelled data is based on people found facing the camera (with frontal face detections); during detection, we wish to detect people in more varied poses. Note that the reference part r is treated as virtual (i.e. its covariance is not modelled) and corresponds to the middle of the face. Note also that the relative scale of parts is fixed.

4.4 Finding the MAP configuration

Our goal now is to find the maximum a posteriori (MAP) configuration of parts $(l_1^*, \ldots, l_n^*, l_r^*)$ maximizing the posterior probability given by equation (2). Working in negative log space, this is equivalent to minimizing the matching cost

$$\min_{l_1,\ldots,l_n,l_r} \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{i=1}^{n} d_{ir}(l_i, l_r) \right), \tag{4}$$

where $m_i(l_i) = -\rho_i(l_i)$ is the negative matching score of part i at location l_i defined in section 4.2 and

$$d_{ir}(l_i, l_r) = (l_i - l_r - \mu_i)^\top \Sigma_i^{-1} (l_i - l_r - \mu_i) \tag{5}$$

is the Mahalanobis distance between the location of part l_i and the reference part l_r. The Mahalanobis distance describes the deformation cost for placing part i at location l_i given the location l_r of the reference part. Intuitively, the cost increases when part i is placed away from the ideal location $l_r + \mu_i$.

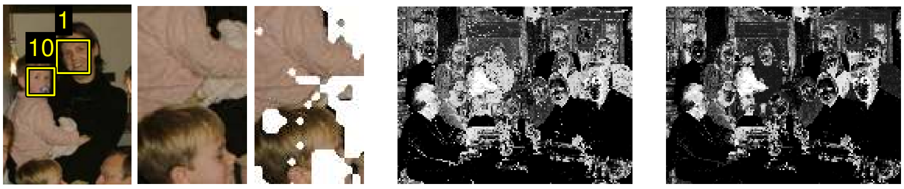
In a naive implementation of the 'star' pictorial structure, finding the MAP configuration by minimizing the cost (4) requires $O(nh^2)$ operations, where n is the number of parts and h is the number of possible locations for each part. In [7], the authors show how to find the minimal cost configuration in $O(nh)$ time using the generalized distance transform. In more detail, the pictorial structure matching cost equation (4) can be rewritten as

$$l_r^* = \arg\min_{l_r} \left[\sum_{i=1}^{n} \min_{l_i} \left(m_i(l_i) + d_{ir}(l_i, l_r) \right) \right], \tag{6}$$

i.e. the cost of placing a reference part at position l_r is the sum of the minimum matching costs for each part l_i, but relative to the position l_r. The minimization of

$$D_i(l_r) = \min_{l_i} \left(m_i(l_i) + d_{ir}(l_i, l_r) \right) \tag{7}$$

inside the sum in (6) can be computed for all h pixel locations l_r in $O(h)$ time using the 2D generalized distance transform described in [7]. One can think of $D_i(l_r)$ as an image where a pixel value at location l_r denotes the minimum matching cost of part i over the whole image, given that the reference part r is located at l_r. Examples of the distance transform images are shown in figures 4(j)-(l). Intuitively, the value of the distance transform $D_i(l_r)$



(a) (b) (c)

Figure 5: **Improving torso color models.** (a) Left: Close-up of person 10 and person 1 in image 2 of figure 7 with face detections overlaid. (a) Middle: A cut-out of the initial torso region proposed below face detection of person 10. (a) Right: Masked torso region obtained using the segmentation algorithm of section 5.1. Note how most of the dark regions of person 1 are removed. (b)-(c) Posterior probability image (eq. 3) of the torso color model of person 10 in image 3 (of figure 7) before (b) and after (c) improving the torso color model. Brightness indicates value. Note how the areas of dark clothing of person 1 have low probability in (c).

is low at locations with a low part matching cost m_i (high matching score ρ_i), shifted by the part offset μ_i, and increases with the distance from the low cost locations. The total cost of placing a reference part r at location l_r is computed according to equation (6), which is essentially a sum of the distance transform images $D_i(l_r)$. An example of the total (negative) cost image is shown in figure 4(c). Note how the highest peak in the final (negative) cost corresponds to the face of the correct person to be detected. The corresponding MAP configuration of parts is shown in figure 4(b).

5 Extensions

Two extensions of the pictorial structure detection algorithm are described next.

5.1 Improving the torso color models

Due to partial occlusion, a torso color mixture model of one person can contain clothing colors of another person. Consider for example figure 5, where the torso region proposed for the person with the pink shirt contains dark colors from the shirt of the other person. The solution we propose is to compute a more accurate segmentation mask so that clothing color models are computed only from pixels belonging to a particular person. This is achieved by allowing multiple color mixture models to 'compete' for image pixels in regions potentially belonging to multiple people.

In more detail, consider a single person j in a set of photos. We start by computing the torso color GMM from the entire torso region as described in section 4.1. Next, we determine the set A of people that have significant overlap with the torso region of person j. We then classify each pixel x in the torso region of person j as belonging to person j, the background or a person from the set A. In particular, the classification is performed by assigning each pixel to the color GMM with the highest likelihood. Finally, the color model of person j is recomputed only from pixels classified as belonging to person j. This procedure is performed for all people which have torso regions that significantly overlap ($\geq$ 10%) in at least one of the photographs. The improvement in the color model is illustrated in figure 5.

Note that this procedure can be applied iteratively [18], i.e. the new color model can be used to further refine the classification of the torso region pixels. Experimentally, we have found one iteration to be sufficient.

5.2 Considering multiple detection hypotheses

In section 4, the detection algorithm finds the part configuration for the pictorial structure model with the lowest matching cost for each person. It is possible that several good matches might exist within an image for a particular person. Typically, these ambiguous matches occur when other people in the photo have similarly colored clothes. To overcome this problem, we consider multiple hypotheses for each person. The match scores across people can then be compared.

In particular, we find multiple detection hypotheses by detecting local maxima in the cost function (4) for each person at all scales. Next we exclude all hypotheses below the detection threshold and hypotheses with the reference part falling within an existing face detection. Finally, the remaining detection hypotheses are assigned using a greedy procedure. We iteratively find the best detection, among all the undetected people, and remove all detection hypotheses with a nearby reference part. Each person is considered only once per image, i.e. if a person is detected, all of their other hypotheses are discarded.

An alternative to the greedy assignment procedure would be exploring all possible combinations of detection hypotheses of people present in the image, maximizing the sum of their detection scores. While this might be feasible for a small number of people and detection hypotheses, in general this would be intractable.

6 Results

In this section, we present results of the proposed algorithm applied to photo sequences from a personal photo collection. The original high resolution images (up to 2,048 × 1,536) pixels were downsampled to one fifth of their original size before processing. To cope with the scale changes present in the data, the pictorial structure detection algorithm was run over four scales. The current Matlab implementation of the pictorial structure detection takes a few seconds per scale on a 2GHz machine.

The algorithm was tested on 11 photo sequences of 2-7 photographs containing 1-14 people wearing a range of clothing and hairstyles. In total there are 113 frontal face detections (with no false positive detections) and 59 occurrences of people with missed frontal face detections. 53 missed detections (out of 59) are correctly filled-in by our proposed algorithm (with no false positive detections).

Figure 6 shows the pictorial structure detection results for the sequence of figure 1. Figure 7 shows a challenging example containing fourteen people and demonstrates the benefit of our algorithm's extensions described in sections 5.1 and 5.2.

7 Conclusions and future work

We have described a fully automatic algorithm, which given a sequence of repeated photographs of the same scene (i) discovers the different people in the photos by clustering frontal face detections and (ii) re-detects people not facing the camera using clothing and hair. Successful detections were shown on challenging indoor and outdoor photo sequences with multiple people wearing a range of clothing and hair-styles.

There are several possibilities for extending this work. Along with color information, clothing and hair texture could be used. The proposed pictorial structure detections could also be verified using additional tests considering spatial layout of clothing colors. Finally,

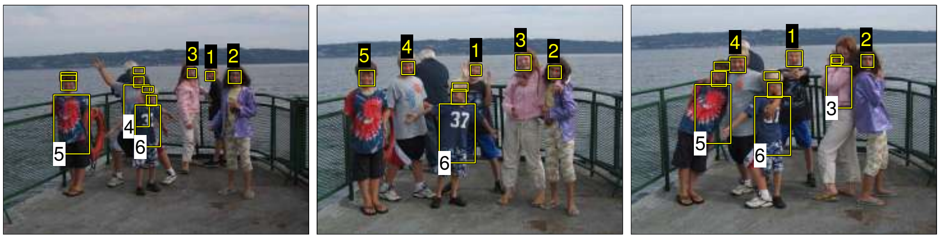


Figure 6: **Re-detecting people in photographs. Example I.** Detected people in the image sequence of figure 1. People are labelled with numbers according to the clusters shown in figure 2. Pictorial structure detections are labelled with dark numbers on white background while the original frontal face detections are labelled with yellow numbers on dark background. In this image sequence, all people with missing face detections are re-detected correctly despite some facing away from camera (e.g. person 6 in image 3). Note how person 4 and person 6 are correctly detected in image 1 despite partially occluding each other. Note also how the pictorial structure model deforms to cope with the slanted pose of person 5 in image 3.

face recognition could be used to distinguish cases where people are wearing identical or similar clothing.

References

[1] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *Proc. CVPR*, pages I:10–17, 2005.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, pages 886–893, 2005.

[3] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2001.

[4] A. Elgammal and L. S. Davis. Probabilistic framework for segmenting people under occlusion. In *Proc. CVPR*, pages 145–152, 2001.

[5] M. Everingham. Person recognition with a two part model. EC project CogViSys meeting, July 2003.

[6] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *Proc. CVPR*, pages 2066–2073, 2000.

[7] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[8] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *Proc. CVPR*, pages I:380–387, 2005.

[9] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computer*, c-22(1):67–92, Jan 1973.

[10] D. Gavrila and V. Philomin. Real-time object detection for "smart" vehicles. In *Proc. ICCV*, pages 87–93, 1999.

[11] S. Ioffe and D. A. Forsyth. Probabilistic methods for finding people. *IJCV*, 43:45–68, 2001.

[12] S. Z. Li and Z. Q. Zhang. Floatboost learning and statistical face detection. *IEEE PAMI*, 26(9):1112–1123, 2004.

[13] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proc. ECCV*, pages I:69–82, May 2004.

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

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

[16] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.

[17] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proc. CVPR*, pages 1:271–278, 2005.

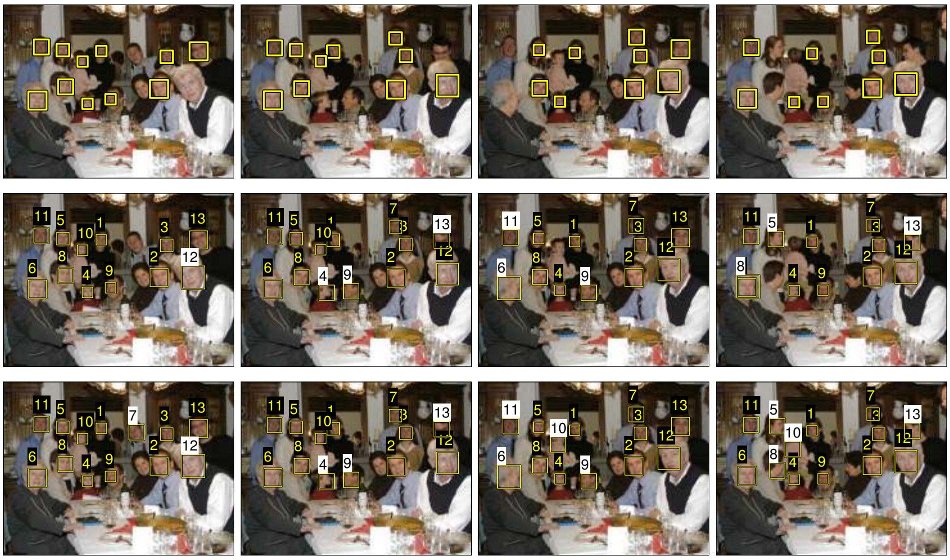


Figure 7: **Re-detecting people in photographs. Example II.** First row: Four (out of five) photographs of the same scene with face detections overlaid. The face detector finds 13 out of 14 people in at least one of the photos, i.e. one person is missed in all photos. The clustering algorithm correctly groups the detected faces into 13 groups (not shown). Second row: Detections using the pictorial structure model without applying the extensions of sections 5.1 and 5.2. Only the face/skin part of the pictorial structure model is shown for each detection. The original face detections are shown with yellow labels on a dark background. The pictorial structure model detections are shown with dark labels on a white background. 16 frontal face detections in all 5 photos are missed by the frontal face detector mainly due to people not facing the camera. 12 out these 16 detections are correctly filled in by the pictorial structure model. Note for example person 6 in image 3 or person 5 and person 13 in image 4. Three people are missed (person 10 in images 3 and 4 and person 7 in image 1) and one person is detected incorrectly (person 8 in image 4 is incorrectly detected over person 6). Third row: The failed cases are corrected by improving torso color models (section 5.1) and considering multiple detection hypotheses for each person (section 5.2). Detections of person 10 in images 3 and 4 are still inaccurate as the person has pink clothing very similar to their skin color. Detections of person 10 in images 3 and 4 and person 8 in image 4 are still below the global detection threshold used to produce the quantitative results in section 6.

[18] C. Rother, V. Kolmogorov, and A. Blake. ”GrabCut”: Interactive foreground extraction using iterated graph cuts. In *Proc. ACM SIGGRAPH*, pages 309–313, 2004.

[19] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *Proc. CVPR*, pages 1746–1759, 2000.

[20] B. Suh and B. B. Bederson. Semi-automatic image annotation using event and torso identification. Technical report, Computer Science Department, University of Maryland, 2004.

[21] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, pages 511–518, 2001.

[22] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. ICCV*, pages 734–741, 2003.

[23] L. Zhang, L. Chen, M. Li, and H. Zhang. Automated annotation of human faces in family albums. In *MULTIMEDIA ’03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 355–358, 2003.

[24] W. Zhao, R. Challappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35:399–458, 2003.