Detection and Tracking of Humans by Probabilistic Body Part Assembly

Antonio S. Micilotta Eng Jon Ong Richard Bowden
CVSSP, University of Surrey, Guildford, UK
amicilotta@airpost.net {e.ong, r.bowden}@surrey.ac.uk

Abstract

This paper presents a probabilistic framework of assembling detected human body parts into a full 2D human configuration. The face, torso, legs and hands are detected in cluttered scenes using boosted body part detectors trained by AdaBoost. Body configurations are assembled from the detected parts using RANSAC, and a coarse heuristic is applied to eliminate obvious outliers. An *a priori* mixture model of upper-body configurations is used to provide a pose likelihood for each configuration. A joint-likelihood model is then determined by combining the pose, part detector and corresponding skin model likelihoods. The assembly with the highest likelihood is selected by RANSAC, and the elbow positions are inferred. This paper also illustrates the combination of skin colour likelihood and detection likelihood to further reduce false hand and face detections.

1 Introduction

Our goal is to robustly estimate the location and approximate 2D pose of humans in real world cluttered scenes. This is a challenging task as the shape and appearance of the human figure is highly variable. The problem is further compounded as people wear a variety of clothes, and skin tone varies with race. We have implemented the Adaboost object detection method [16], and have created body part detectors for the face, torso, legs and hands. These are then assembled via RANSAC [4] in real-time.

Several approaches for the visual recognition of humans have been explored. Detection of pedestrians has been proposed by Elzein et al. [2] where a wavelet-based approach and multistage template matching is used to detect pedestrians for a 'smart' car application. The Chamfer System [5] performs shape-based pedestrian detection: a hierarchy of pedestrian templates is matched with distance-transformed images in a tree traversal process. The method locks onto desired objects in a coarse-to-fine manner.

On a lower level, human detection has also been achieved by detecting individual body parts, and assembling them into a human figure. Ioffe and Forsyth [7] make use of a parallel edge segment detector to locate body parts, and assemble them into a 'body plan' using a pre-defined top level classifier. This methodology has shortcomings in the presence of background clutter and loose clothing. Similarly, Felzenszwalb and Huttenlocher [3] use rectangular colour-based part detectors, and assemble detected parts into a body plan using pictorial structures. Ronfard et al.[11] use detectors trained by dedicated Support Vector Machines (SVM) where a feature set consists of a Gaussian filter image

and 1st and 2nd derivatives. Haar wavelets are used by Mohan et al. [10] to represent candidate regions and SVMs to classify the patterns. They then combine the detector outputs using another SVM. Roberts et al. [12] have created probabilistic region templates for the head, torso and limbs. Likelihood ratios for individual parts are learned from the dissimilarity of the foreground and adjacent background distributions. The process is computationally expensive, and the greatest likelihoods occur where the foreground and background hold similar colour distributions. Mikolajczyk et al. [9] model humans as flexible combinations of boosted face, torso and leg detectors. Parts are represented by the co-occurrence of orientation features based on 1st and 2nd derivatives. The procedure is computationally expensive, and 'robust part detection is the key to the approach' [9].

Detection of faces in colour images has also been explored. A method of face tracking for video communications is described in [15], where the objective is to centre the face in the image. This is achieved using blink detection and colour histogram matching. Hsu et al. [6] detect skin regions over an image and search for the co-occurrence of eyes and a mouth in these regions. The co-occurrence of facial features was also used by Cristinacce et al. [1] using a Pairwise Reinforcement of Feature Responses in a similar fashion to that of Mikolajczyk [9].

Our approach is novel in that it uses RANSAC to combine appearance, colour and structural cues with a strong prior on pose configuration to detect structures of humans within images. This paper is set out as follows: A basic discussion of AdaBoost applied to object detection is presented in Section 2. Our first contribution illustrates how the occurrence of colour associated with hand and face detections can be used to increase the likelihood of true detections (Section 2.2). Our chief contribution offers a methodology of assembling all part detections using RANSAC, a heuristic, and an *a priori* mixture model of upper-body configurations (Section 3). From the final model, the elbow positions are then inferred from a secondary prior. Finally, results are shown, and conclusions drawn.

2 Boosted Body Parts Detectors

Boosting is a general method that can be used for improving the accuracy of a given learning algorithm [14]. More specifically, it is based on the principle that a highly accurate or 'strong' classifier can be produced through the linear combination of many inaccurate or 'weak' classifiers. The efficiency of the final classifier is increased further by organising the weak classifiers into a collection of cascaded layers. This design consists of a set of layers with an increasing number of weak classifiers, where each layer acts as a non-body-part rejector with increasing complexity. An input image is first passed to the simplest top layer for consideration, and is only moved to the next layer if it is classified as true by the current layer. The reader is directed to [16] for a detailed discussion of AdaBoost cascades.

2.1 Detection of Body Parts

Using AdaBoost as mentioned above, we separately trained four different body part detectors using their respective image databases. The face, torso and hand training images were sized at 20x20 pixels, and the legs at 20x40 pixels. Examples are shown in Figure 1.

Firstly, we provide definitions of the different body part detectors, each of which takes the form of a cascaded strong classifier. In order to detect a specific body part in

Figure 1: Training image examples of the four body parts - face, torso, legs and hands

a bounding box, we firstly offset all the weak classifiers belonging to the detector to that location. A positive or negative detection is then computed by combining weak classifier outputs in strong-classifier layers. Each detector returns a score for part detection, which is then normalised to produce a likelihood, defined as L_F, L_T, L_L and L_H respectively. Use of this notation can be found in Section 3.3. The performance of the detectors is shown in Figure 4 (a).

2.2 Exploiting Colour Cues for Reduced False Detections

Since detections are performed in gray scale, it would be advantageous to exploit colour cues to act as a-priori constraints to initially reduce the most obvious false detections. This is especially useful if the colour is fairly consistent across different instances of the objects. Here, we find that the hands and face benefit from this constraint.

A weak skin colour model consisting of a single gaussian in the Hue-Saturation colour space has been created from a large selection of natural images. Given a novel image, a skin-colour likelihood map can be generated using this Gaussian model. As previously mentioned, face detection/tracking on colour images has received considerable attention - a large difference however is that they first segment skin colour regions, and conduct face-like feature searches in these segmented regions. Our approach is the reverse: The hand and face detectors are applied across the entire image, and provide a bounding box for each detection. The skin probability map is used to determine the median skin colour likelihood for each bounding box. Should this probability fall below a weak threshold (ie. 6 standard deviations), the detection is rejected. Our motivation is twofold: firstly, skin-colour segmentation can be unreliable, possibly leaving blank areas in the skin-colured regions. Furthermore, facial cavities are ignored, and fingers are thinned. Such partially segmented body parts would fail to meet the detection requirements of the associated part detectors, as the training images are cropped from larger natural images. Secondly, even if the segmentation were clean, restricting the search space to these regions is naive as a generic skin filter is not guaranteed to segment all skin-like objects.

Figure 2 demonstrates this method with the use of the face detector. Figure 2 (a) shows all face detections – it is evident that several false detections occur over ambiguous textures. With the aid of the above-mentioned generic skin model, many of these false detections can be eliminated, as shown in Figure 2 (b). The resulting detection/skin joint-likelihood image for all detections is provided in Figure 2 (c) – these likelihoods come into play when determining the joint-likelihood model in Section 3.3. The effect of the inclusion of colour is shown by the comparison of the two face detection performance curves of Figure 4 (a).

(a) (b) (c)

Figure 2: Illustration of reduction of false detections using colour cues

3 Human Body Assembly

The methods described in the previous sections provide the detected body parts needed to construct a human model. To ensure that most of the body parts are detected, fewer layers in the cascade are selected, resulting in a larger number of false detections. In order to determine a final body configuration, a four step process is followed:

1. RANSAC is used to assemble random body configurations, each consisting of a head, a torso, a set of legs, and a pair of hands. A weak heuristic is then applied to each configuration to eliminate obvious outliers (3.1).

2. Each remaining configuration is compared to an *a priori* mixture model of upper-body configurations, yielding a measure of fitness for the upper body pose (3.2).

3. A resultant likelihood for each configuration is obtained by combining the likelihood determined by the prior model with those of the body part detectors and corresponding skin colour (if applicable). The configuration with the highest likelihood is voted for by RANSAC to represent the detected human (3.3).

4. With a chosen body configuration, the elbow positions are then inferred (3.4)

3.1 Coarse Elimination of Poor Body Configurations

An image with several human figures and dense background clutter can produce a large number of separate part detections. RANSAC selects subsets of detections that represent body configurations, however testing all these configurations would still be computationally expensive – a coarse heuristic is therefore employed to discard unlikely configurations.

Rules of the heuristic are designed according to a generic human model, and include a reference length measurement. Referring to Figure 3, Da Vinci's Vitruvian Man subdivides the human figure into eight lengths, each of which is equal to the length of the head, measured from the top of the skull to the chin. For the purpose of this paper, this length is referred to as a *skeletal unit length*. The head can be further subdivided into 3 lengths, a,b and c – a typical face detection occupies b and c, thereby allowing us to approximate the body unit length.

Comparing the ROC curves of Figure 4 (a) it is evident that the face detector is the most robust. For this reason, the face detector forms the base for every body configuration.

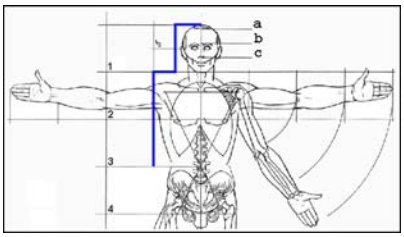


Figure 3: Vitruvian Man

The skeletal unit length and centre position of a selected face is determined, and form the parameters that assist in solving a body configuration. The rules of the heuristic are set out in the following order, with *x* and *y* referring to horizontal and vertical directions:

1. A torso is added to the model if: its centre x position lies within the face width; the torso scale is approximately 3 $\times$ face scale ($\pm$ 0.5); the face centre lies within the detected torso region.
2. A set of legs is added if: its centre x position lies within the face width; the top y position lies within 4 $\times$ body unit lengths ($\pm$ 0.5); the leg scale is approximately 2 $\times$ face scale ($\pm$ 0.5).
3. A pair of hands are added if: both hands are less that 4 $\times$ body unit lengths from the face; the hand scale $\approx$ face scale ($\pm$ 0.2).

False hand detections form the bottleneck in the system as a large number are accepted by the heuristic. The configurations that are passed by the heuristic are then compared to an *a priori* mixture model of upper-body configurations to obtain a likelihood for the upper body pose (see equation 3), which plays an important role in eliminating false hand detections as awkward hand poses yield a low likelihood.

3.2 Prior Data for Pose Likelihood

In this second step, we use an *a priori* mixture model of upper-body configurations to estimate the optimal upper body pose. Each body configuration obtained by the above-mentioned selection process provides the position of 8 points, namely the four corners of the torso detector, the chin and brow of the face detector, and the hands. These 8 x, y coordinates are concatenated to form a feature vector $\mathbf{Y} \in \mathbb{R}^{16}$.

An *a priori* model ϕ of upper-body configurations was built from approximately 4500 hand labelled representative examples ($\in \mathbb{R}^{16}$ as above) from image sequences of subjects performing various articulated motions. It is obvious that the manifold on which the data set lies is unlikely to be linear, and we therefore use a Gaussian Mixture Model (GMM) to represent the training set. The optimum number of components is said to be the number for which further increases do not produce significant gain in overall cost. This optimum number of components k is chosen to be the point of inflection of the cost function, constructed from k-means. k 16x16 covariance matrices $Cov_{\phi,k}$ are formed from data set ϕ, where $Cov_{\phi,k} = \frac{1}{N_k - 1}(\phi_i - \boldsymbol{\mu}_{\phi,k})(\phi_i - \boldsymbol{\mu}_{\phi,k})^T$, and $\boldsymbol{\mu}_{\phi,k}$ is the mean of each component of the GMM. A measure of how well a feature vector $\mathbf{Y}$ (or body configuration) fits the prior data set can now be determined.

Firstly, the Mahalanobis distances $md_{\phi,k}$ from $\mathbf{Y}$ to each component of the GMM are determined. However, due to the lack of variance in the shoulder and hip components of the feature vectors that created the prior, $Cov_{\phi,k}$ is often singular. $md_{\phi,k}$ can therefore not be determined as per $md_{\phi,k} = (\mathbf{Y} - \boldsymbol{\mu}_{\phi,k})Cov_{\phi,k}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_{\phi,k})^T$ as $Cov_{\phi,k}$ cannot be inverted. Instead, we obtain $md_{\phi,k}$ from the eigenvectors and eigenvalues $\mathbf{P}_{\phi,k}$ and $\mathbf{b}_{\phi,k}$ of $Cov_{\phi,k}$. Firstly, we project $\mathbf{Y}$ into the eigenvector domain to obtain $\mathbf{Y}_{\phi,k}$ as follows:

$$\mathbf{Y}_{\phi,k} = \mathbf{P}_{\phi,k}^T(\mathbf{Y} - \boldsymbol{\mu}_{\phi,k}) \quad (\mathbf{Y}_{\phi,k} \in \mathbb{R}^{16}) \tag{1}$$

With $\mathbf{Y}_{\phi,k}$ and $\mathbf{b}_{\phi,k}$ in column vector form, the distance from $\mathbf{Y}$ to each GMM component can now be determined as:

$$md_{\phi,k}(\mathbf{Y}) = (\mathbf{Y}_{\phi,k})^T.\mathbf{b}_{\phi,k}^{(-1/2)} \tag{2}$$

where $\mathbf{b}_{\phi,k}^{(-1/2)}$ represents the pseudo inverse (the square root and reciprocal of each component) of $\mathbf{b}_{\phi,k}$. The final pose likelihood can then be obtained from the weighted sum of likelihoods of each component:

$$L_Y = \sum_{i=1}^{k} \frac{N_i}{N} \left[\left(2\pi^{\frac{d}{2}} |Cov_{\phi,i}|^{\frac{1}{2}} \right)^{-1} exp(-\frac{1}{2} md_{\phi,i}^2) \right] \tag{3}$$

3.3 Final Configuration Selection

At this point in the algorithm, a coarse skin model with a low threshold has been used to reduce obvious false face and hand detections (Section 2.2). This skin model is far too general for use in robust segmentation, but does however allow us to discard pixels that are outliers. From the inliers that remain, a new model is learned, creating a refined, user specific skin model. A skin colour likelihood for the hands is determined using this model as the skin colour of the face and hands is assumed to be similar. These skin likelihoods contribute to the body joint-likelihood model for a body configuration.

The nine determined likelihoods, namely the mixture model (L_Y), face (L_F), torso (L_T), legs (L_L), left hand (L_{LH}), left hand skin (L_{LHS}), right hand (L_{RH}) and right hand skin (L_{RHS}) are combined to provide an overall body configuration likelihood, L_{BC}.

$$L_{BCi} = L_{Yi}.L_{Fi}.L_{Ti}.L_{Li}.L_{LHi}.L_{LHSi}.L_{RHi}.L_{RHSi} \tag{4}$$

The configuration with the greatest likelihood is selected to represent the final configuration.

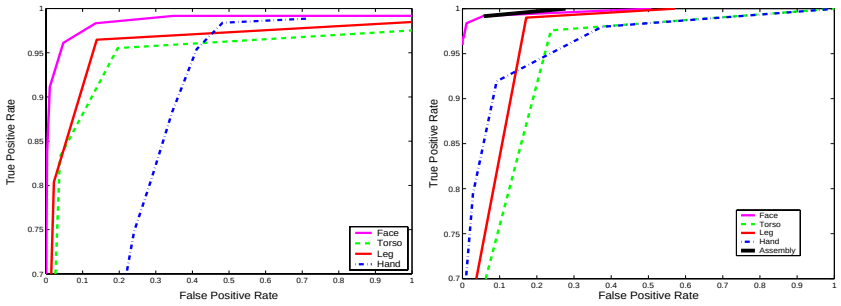
3.4 Estimation of Elbow Positions

A second *a priori* model ψ was built from the same image sequences of Section 3.2, however, these labelled examples also include the elbow positions ($\in \mathbb{R}^{20}$). The final body configuration $\mathbf{Y} \in \mathbb{R}^{16}$ of 3.2 and the *a priori* model ψ are used to construct a new feature vector $\mathbf{Y^R} \in \mathbb{R}^{20}$ that includes predicted elbow positions. These elbow positions are extracted and included in the final representative model. Details of this approach are given in [8].

4 Results

Comparison of the different part detectors is a difficult task. The most obvious problem is that each part is of different scale, and we would therefore expect a larger number of false hand detections than false torso detections. In addition, a fair test would make use of an image database in which all the aforementioned body parts exist – however we have been unsuccessful in finding such a database, and we have made use of three databases. Our in-house face database consists of colour images containing 500 faces, and is similar in size to the MIT-CMU face database (507 faces). The torso and leg detectors were tested on 460 (of 900) images of the MIT pedestrian database, while the hand detector was tested on a colour image database containing 400 hands. Figure 4(a) shows the detection performance of the detectors applied to their respective test datasets, where layers from the classifier are removed to increase the number of detections. In this research, detection is considered true if at least 75% of its bounding box encloses the groundtruthed body part. In addition, we do not merge overlapping false detections as in [13].

We have plotted two curves for the face detector to show the advantage of including colour. The face detector proves to be the most robust of the detectors. This is an intuitive result as the face is a self contained region, while the other body parts are disguised with background clutter and and have a greater variable appearance. Due to the high variability of hand shape, we expect the hand detector to offer the poorest performance.



(a) (b)

Figure 4: (a) Performance on test databases (b) Performance on a video sequence

Making use of the ROC curves plotted for each detector, the desired number of layers was chosen such that the probability of detecting all objects was no less than 80%. This decision naturally came with the trade-off of an increased number of false detections. Figure 5 (a) shows all detections from the body part detectors applied to a set of images sized at 1024x768 pixels. The false detections are rapidly eliminated by applying RANSAC and the heuristic, as shown in Figure 5 (b). Figure 5 (c) highlights the greatest body configuration likelihood as determined by the joint-likelihood model, and also overlays the final body pose with the elbow positions. The total number of detections is indicated beneath each image, and is naturally dependent on the image contents. This is evident in the second image where the store's display window contains suit jackets, thereby producing numerous false detections. The entire process from detection to assembly takes approximately 5 seconds on a P4.

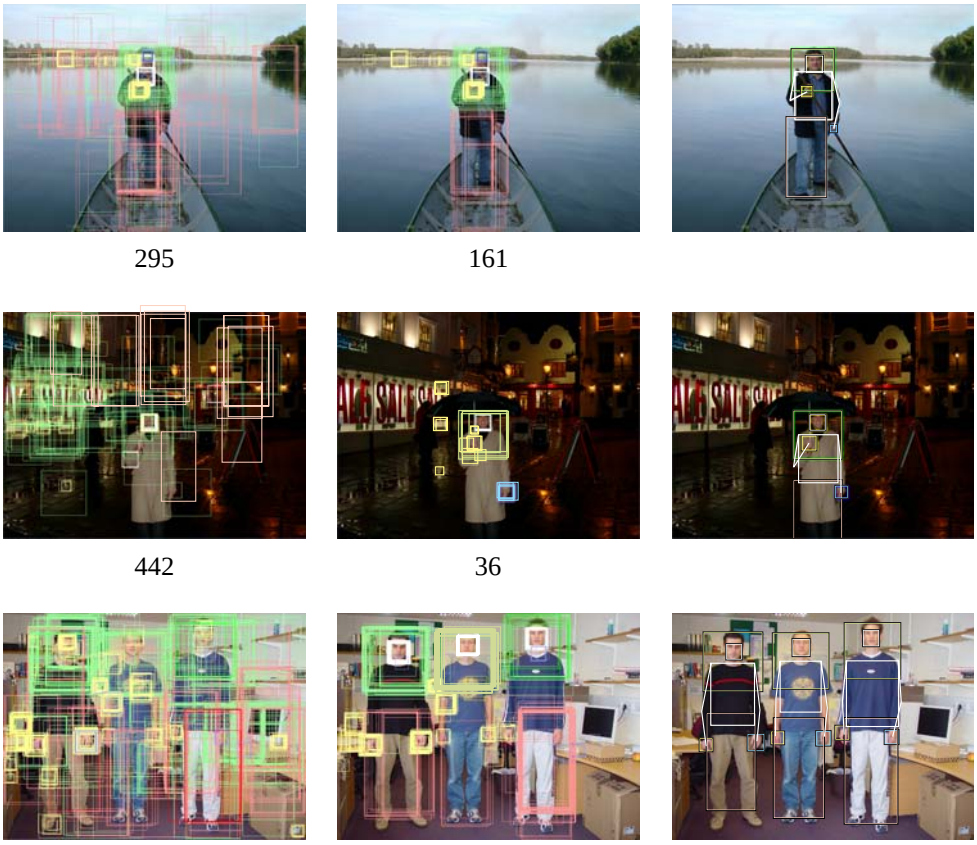


Figure 5: (a) All detections (b) Reduced detections (c) Final assembly
(Number of detections shown beneath the image)

Extending this work to video sequences allows the detectors to be applied in a tracking framework. This localises the search space in subsequent frames, thereby reducing the number of false detections and improving speed performance. An initial face detection is conducted as before, with consequent body part detections limited by the heuristic proximity rules as defined in Section 3.1. Subsequent position and scale variations of each detector are governed by prior detections. Should a body part fail to be detected, the search region for the corresponding detector is increased linearly and the scale is adjusted by a Gaussian drift term until the detector recovers.

Figure 6 illustrates the assembly and elbow prediction of a subject walking into an office and performing hand gestures. The scene is particularly complex, and has wood furniture and cork pin boards that have a similar colour to skin. Furthermore, the sub-

ject is wearing beige trousers, offering very little contrast to the background of a cream wall and filing cabinets, making background segmentation poor. Our assembly system overcomes these difficulties, and with the use of temporal information, operates at 8 frames/sec (frames sized at 640x480). This is a considerable improvement compared to the static image case.

A corresponding performance curve for this sequence is provided in Figure 4 (b). To maintain consistency with the performance curves of Figure 4 (a), each frame of this sequence was treated as a discrete image, with the search space encompassing the entire image. Here the hand detector also makes use of a foreground likelihood and offers similar performance to the torso and leg detectors. The purpose of using a sequence was to offer a full subject such that the performance of the assembly method could be evaluated. As expected, the assembly curve supersedes the others, illustrating the robust false part elimination by the assembly methodology.

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 6: Body part assembly and elbow prediction from a video sequence

5 Conclusions

We have extended an existing boosting technique for face detection to build three additional body part detectors. Due to the variability of these body parts, their detection performance is lower, and a technique was developed to eliminate false detections. Knowledge of Da Vinci's human figure schematic enabled the design of a coarse body configuration heuristic. By combining this heuristic with RANSAC and an *a priori* mixture model of upper-body configurations, we are able to assemble detections into accurate configurations, and estimate elbow positions to complete the upper body pose. When this approach is applied to a video sequence, exploitation of temporal data reduces the false detection rate of all the detectors, and improves speed performance dramatically.

References

[1] D. Cristinacce, T. Cootes, and I. Scott. A multi-stage approach to facial feature detection. In *Proc. of BMVC*, pages 231–240, 2004.

[2] H. Elzein, S. Lakshmanan, and P. Watta. A motion and shape-based pedestrian detection algorithm. In *Proc. of IEEE Intelligent Vehicles Symposium*, pages 500–504, 2003.

[3] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient matching of pictorial structures. In *Proc. of CVPR*, volume 2, pages 66 – 73, 2000.

[4] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In *Comm. of the ACM*, volume 24, pages 381–395, 1981.

[5] D. M. Gavrila and V. Philomin. Real-time object detection for "smart" vehicles. In *Proc. of ICCV*, volume 1, pages 87–93, 1999.

[6] Rein-Lien Hsu, M. Abdel-Mottaleb, and A.K. Jain. Face detection in color images. *IEEE Transactions on PAMI*, 24(5):696–706, May 2002.

[7] S. Ioffe and D. Forsyth. Probabilistic methods for finding people. *International Journal of Computer Vision*, 43(1):45–68, 2001.

[8] A.S. Micilotta and R. Bowden. View-based location and tracking of body parts for visual interaction. In *Proc. of BMVC*, volume 2, pages 849–858, 2004.

[9] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust body part detectors. In *Proc. of ECCV*, volume 1, pages 69–82, 2004.

[10] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on PAMI*, 23(4):349–361, April 2001.

[11] B. Triggs R. Ronfard, C. Schmid. Learning to parse pictures of people. In *Proc. of ECCV*, volume 4, pages 700–707, 2002.

[12] T. Roberts, S. McKenna, and I. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. In *Proc. of ECCV*, pages 291–303, 2004.

[13] H.A. Rowley, S. Baluja, and T.Kanade. Neural network-based face detection. *IEEE Transactions on PAMI*, 20(1):23–38, January 1998.

[14] R. Schapire. The boosting approach to machine learning: An overview. In *Proc. of MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

[15] K. Schwerdt and J. L. Crowley. Robust face tracking using color. In *Proc. of AFGR*, pages 90 – 95, 2000.

[16] P. Viola and M. Jones. Robust real-time object detection. In *Proc. of IEEE Workshop on Statistical and Computational Theories of Vision*, pages 1–25, 2001.