# Human detection based on a probabilistic assembly of robust part detectors

Krystian Mikolajczyk, Cordelia Schmid, Andrew Zisserman

# Human detection based on a probabilistic assembly of robust part detectors

K. Mikolajczyk[1] C. Schmid[2] A. Zisserman[1]

(1) Dept. of Engineering Science (2) INRIA Rhône-Alpes
Oxford, OX1 3PJ 38330 Montbonnot
United Kingdom France
km,az@robots.ox.ac.uk schmid@inrialpes.fr

**Abstract.** We describe a novel method for human detection in single images which can detect full bodies as well as close-up views in the presence of clutter and occlusion. Humans are modeled as flexible assemblies of parts, and robust part detection is the key to the approach. The parts are represented by co-occurrences of local features which captures the spatial layout of the part's appearance. Feature selection and the part detectors are learnt from training images using AdaBoost.
The detection algorithm is very efficient as (i) all part detectors use the same initial features, (ii) a coarse-to-fine cascade approach is used for part detection, (iii) a part assembly strategy reduces the number of spurious detections and the search space. The results outperform existing human detectors.

## 1 Introduction

Human detection is important for a wide range of applications, such as video surveillance and content-based image and video processing. It is a challenging task due to the various appearances that a human body can have. In a general context, as for example in feature films, people occur in a great variety of activities, scales, viewpoints and illuminations. We cannot rely on simplifying assumptions such as non-occlusion or similar pose. Of course, for certain applications, such as pedestrian detection, some simplifying assumptions lead to much better results, and in this case reliable detection algorithms exist. For example, SVM classifiers have been learnt for entire pedestrians [14] and also for rigidly connected assemblies of sub-images [13]. Matching shape templates with the Chamfer distance has also been successfully used for pedestrian detection [1, 5].

There is a healthy line of research that has developed human detectors based on an assembly of body parts. Forsyth and Fleck [4] introduced body plans for finding people in general configurations. Ioffe and Forsyth [6] then assembled body parts with projected classifiers or sampling. However, [4, 6] rely on simplistic body part detectors – the parts are modelled as bar-shaped segments and pairs of parallel edges are extracted. This body part detector fails in the presence of clutter and loose clothing. Similarly, Felzenszwalb and Huttenlocher [2] show

that dynamic programing can be used to group body plans efficiently, but simplistic colour-based part detectors are applied. An improvement on body part detection is given in Ronfard *et al.* [17] where SVMs are trained for each body part. An improvement on the modelling of body part relations is given in Sigal *et al.* [21], where these are represented by a conditional probability distribution. However, these relations are defined in 3D, and multiple simultaneous images are required for detection.

In this paper we present a robust approach to part detection and combine parts with a joint probabilistic body model. The parts include a larger local context [7] than in previous part-based work [4, 17] and they therefore capture more characteristic features. They are however sufficiently local (cf. previous work on pedestrian detectors [14]) to allow for occlusion as well as for the detection of close-up views. We introduce new features which represent the shape better than the Haar wavelets [14], yet are simple enough to be efficiently computed. Our approach has been inspired by recent progress in feature extraction [10, 18–20], learning classifiers [15, 22] and joint probabilistic modelling [3].

Our contribution is three-fold. Firstly, we have developed a robust part detector. The detector is robust to partial occlusion due to the use of local features. The features are local orientations of gradient and Laplacian based filters. The spatial layout of the features, together with their probabilistic co-occurrence, captures the appearance of the part and its distinctiveness. Furthermore, the features with the highest occurrence and co-occurrence probabilities are learnt using AdaBoost. The resulting part detector gives face detection results comparable to state of the art detectors [8, 22] and is sufficiently general to successfully deal with other body parts. Secondly, the human detection results are significantly improved by computing a likelihood score for the assembly of body parts. The score takes into account the appearance of the parts and their relative position. Thirdly, the approach is very efficient since (i) all part detectors use the same initial features, (ii) a coarse-to-fine cascade approach successively reduces the search space, (iii) an assembly strategy reduces the number of spurious detections.

The paper is structured as follows. We introduce the body model in section 2. We then present the robust part detector in section 3, and the detection algorithm in section 4. Experimental results are given in section 5.

## 2 Body model

In this section we overview the body model which is a probabilistic assembly of a set of body parts. The joint likelihood model which assembles these parts is described in section 2.1. The body parts used in the model are given in section 2.2, and geometric relations between the parts in section 2.3.

### 2.1 Joint likelihood body model

Our classification decision is based on two types of observations, which correspond to the body part appearance and relative positions of the parts. The ap-

pearance is represented by features $F$ and the body part relations by geometric parameters $R$. The form of a Bayesian decision for body $B$ is:

$$\frac{p(B|\mathcal{R},\mathcal{F})}{p(non\ B|\mathcal{R},\mathcal{F})} = \frac{p(\mathcal{R}|\mathcal{F},B)}{p(\mathcal{R}|\mathcal{F},non\ B)} \cdot \frac{p(\mathcal{F}|B)}{p(\mathcal{F}|non\ B)} \cdot \frac{p(B)}{p(non\ B)} \tag{1}$$

The first term of this expression is the probability ratio that body parts are related by geometric parameters measured from the image. The second term is the probability ratio that the observed set of features $F$ belong to a body:

$$\frac{p(\mathcal{F}|B)}{p(\mathcal{F}|non\ B)} = \prod_{f\in\mathcal{F}} \frac{p(f,\mathbf{x}_f|B)}{p(f,\mathbf{x}_f|non\ B)}$$

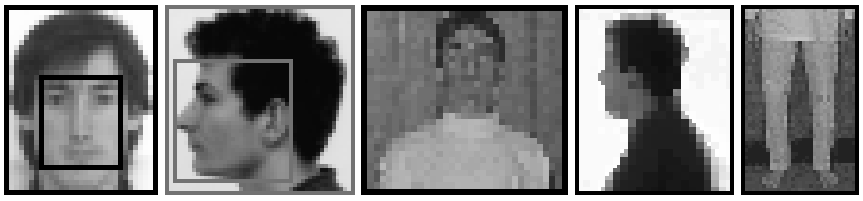
This set consists of a number of local features $f$ and their locations $x_f$ in a local coordinate system attached to the body. The third term of (1) is a prior probability of body and non-body occurence in images. It is usually assumed constant and used to control the false alarm rate.

Individual body part detectors are based on appearance (features and their locations) and provide a set of candidates for body parts. This is discussed in section 3. Given a set of candidate parts the probability of the assembly (or a sub-assembly) is computed according to (1). For example, suppose that a head $H$ is detected based on the appearance, i.e. $p(\mathcal{F}|H)/p(\mathcal{F}|non\ H)$ is above threshold, then the probability that an upper body (U) is present can be computed from the joint likelihood of the upper-body/head sub-assembly $p(U,H)$. Moreover, a joint likelihood can be computed for more than two parts. In this way we can build a body structure by starting with one part and adding the confidence provided by other body part detectors. Implementation details are given in section 4.

### 2.2 Body parts

In the current implementation we use 7 different body parts as shown in Figure 1. There are separate parts for a frontal head (a bounding rectangle which includes the hair), and face alone. Similarly there is a profile head part and a profile face part.

Each body part is detected separately, as described in section 3 based on its likelihood ratio.



(a) (b) (c) (d) (e)

**Fig. 1.** Body parts. (a) Frontal head and face (inner frame). (b) Profile head and face (inner frame). (c) Frontal upper body. (d) Profile upper body. (e) Legs.

### 2.3 Body geometric relations

The probability of a false positive for an individual detector is higher then for several detectors with a constraint on geometric relations between parts. The geometric relationship between the parts is here represented by a Gaussian $G(x_1 - x_2, y_1 - y_2, \sigma_1/\sigma_2)$ depending on their relative position and relative scale. $\sigma_1$ and $\sigma_2$ correspond to the scales (sizes) at which two body parts are detected. These parameters are learnt from training data. The size of a human head can vary with respect to the eyes/mouth distance. Similarly, the scale and the relative location between other body parts can vary for people. Figure 2(b) shows the Gaussian estimated for the head location with respect to the face location. Figure 2(c-d) shows the geometric relations for other body parts. We need to estimate only one Gaussian relation between two body parts, since the Gaussian function in the inverse direction can be obtained by appropriately inverting the parameters. Note that each of the detectors allows for some variation in pose. For example, the legs training data covers different possible appearance of the lower body part.

(a) (b) (c) (d) (e) (f)

**Fig. 2.** Gaussian geometric relations between body parts. (a) Frontal face location. (b) Frontal head location with respect to the face location. (c) Profile location. (d) Profile head location with respect to the profile location. (e) Profile upper body location with respect to the head. (f) Frontal upper body location with respect to the head location, and legs with respect to the upper body.

## 3 Body part detector

In this section we present the detection approach for individual body parts. In sections 3.1 and 3.2 we describe the low-level features and the object representation. Section 3.3 explains the classifiers obtained from the features and the learning algorithm.

### 3.1 Orientation features

An object's appearance is represented by orientation-based features and local groupings of these features. This choice is motivated by the excellent performance of SIFT descriptors [10, 12] which are local histograms of gradient orientations. SIFT descriptors are robust to small translation and rotation, and this is built into our approach in a similar way.

*Orientation features.* Our features are the dominant orientation over a neighbourhood and are computed at different scales. Here we use 5 scale levels and a 3-by-3 neighbourhood. Orientation is either based on first or second derivatives.

In the case of first derivatives, we extract the gradient orientation. This orientation is quantized into 4 directions, corresponding to horizontal, vertical and two diagonal orientations. Note that we do not distinguish between positive and negative orientations. We then determine the score for each of the orientations using the gradient magnitude. The dominant direction is the one which obtains the best score. If the score is below a threshold, it is set to zero. Figure 3(b) shows the gradient image and Figure 3(c) displays the dominant gradient orientations where each of the 5 values is represented by a different gray-level value. Note the groups of dominant orientations on different parts of the objects.

A human face can be represented at a very coarse image resolution as a collection of dark blobs. An excellent blob detector is the Laplacian operator [9]. We use this filter to detect complementary features like blobs and ridges. We compute the Laplacian $(d_{xx} + d_{yy})$ and the orientation of the second derivatives $(\arctan(d_{yy}/d_{xx}))$. We are interested in dark blobs therefore we discard the negative Laplacian responses, since they appear on bright blobs. Figure 3(d) shows the positive Laplacian responses. Similarly to the gradient features we select the dominant orientation. Second derivatives are symmetrical therefore their responses on ridges of different diagonal orientations are the same. Consequently there are 3 possible orientations represented by this feature. Figure 3(e) displays the dominant second derivative orientations where each orientation is represented by a different gray-level value.

(a) (b) (c) (d) (e)

**Fig. 3.** Orientation features. (a) Head image. (b) Gradient image. (c) Dominant gradient orientations. (d) Positive Laplacian responses. (e) Dominant orientations of the second derivatives.

*Feature groups.* Since a single orientation has a small discriminatory power, we group neighbouring orientations into larger features. The technique described below was successfully applied to face detection [11, 19]. We use two different combinations of local orientations. The first one combines 3 neighbouring orientations in a horizontal direction and the second one combines 3 orientations in a vertical direction. Figure 4(a) shows the triplets of orientations. A single integer value is assigned to each possible combination of 3 orientations. The

number of possible values is therefore $v_{max} = 5^3 = 125$ for the gradient and $v_{max} = 4^3 = 64$ for the Laplacian. More than 3 orientations in a group significantly increase the number of possible combinations and poorly generalize. In summary, at a given scale there are four different feature group types $v_t$: horizontal and vertical groups for gradient orientations and horizontal and vertical groups for the Laplacian.

### 3.2 Object representation

The location of a feature group on the object is very important as we expect a given orientation to appear more frequently at a particular location and less frequently at the other locations. The location is specified in a local coordinate system attached to the object (Figure 4(b)). To make the features robust to small shifts in location and to reduce the number of possible feature values we quantize the locations into a $5 \times 5$ grid (Figure 4(c)).

In the following we will use the notation $(x, y, v_t)$ to refer to a feature group of type $v_t$ at the grid location $(x, y)$. For simplicity we will refer to this as a *feature* $(x, y, v_t)$.

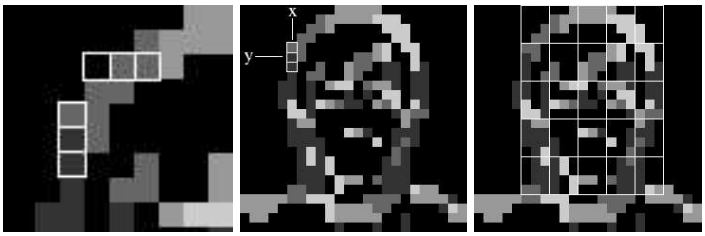



**Fig. 4.** Local groups of features. (a) Two groups of local orientations. (b) Location of the feature on the object. (c) Grid of quantized locations.

### 3.3 Classifiers

To build a reliable detector we need a powerful classifier. Such classifiers can be formed by a linear combination of weak classifiers, and trained with a learning algorithm to excellent classification results at a small computational cost [8, 22]. In the following we explain the form of our weak classifiers.

*Weak classifiers.* The above described features are used to build a set of classifiers. A weak classifier is the log likelihood ratio of the probability of feature occurrence on the object with respect to the probability of feature occurrence on the non-object:

$$h_{f_a} = \ln\left(\frac{p(f_a|object)}{p(f_a|non\ object)}\right)$$

where $f_a$ is a single feature $(x, y, v_t)$. Intuitively, some features occur frequently together on object but randomly together on non-object. Therefore, a better weak classifier using joint probability between two features is

$$h_{f_{ab}} = \ln(\frac{p(f_a, f_b|object)}{p(f_a, f_b|non\ object)}) \tag{2}$$

where $f_a, f_b$ is a pair of features, which simultaneously occur on the object. The probabilities $p(f_a|object)$ and $p(f_a, f_b|object)$ and the corresponding probabilities for non-object can be estimated using multidimensional histograms of feature occurrences. Each bin in the histogram corresponds to one feature value. The probabilities are estimated by counting the feature occurrence on positive and negative examples. Some features do not appear at a particular object location which indicates a zero probability. To avoid a very large or infinite value of a weak classifier we smooth the predictions as suggested in [15].

*Strong classifiers.* A strong classifier is a linear combination of M weak classifiers

$$H_M(x_i) = \sum_{m=0}^{M} h_{f_m}(x_i)$$

where $x_i$ is an example and the class label is $sign[H(x_i)]$. The weak classifiers $h_{f_a}$ and $h_{f_{ab}}$ are combined using the real version of AdaBoost as proposed in [8, 15]. The error function used to evaluate the classifiers is

$$E(H_M) = \sum_i \exp[-y_i H_M(x_i)] \tag{3}$$

where $y_i$ is a class label $[-1, 1]$ for a given training example $x_i$.

A strong classifier is trained separately for each of the four feature types $v_t$. This is motivated by the efficiency of the cascade approach. One feature type at one scale only has to be computed at a time for each cascade level. We compute features at different scales, therefore the number of strong classifiers is the number of feature types times the number of scales. The initial number of strong classifiers is therefore 20 (4 feature types at 5 scales). The number of weak classifiers used by AdaBoost depends on the scale of features and can vary from 16 to 5000.

*Cascade of classifiers.* The strong classifiers are used to build a cascade of classifiers for detection. The cascade starts with the best of the fastest strong classifiers. In this case the fastest classifiers are computed on the lowest scale level and the best one corresponds to that with the lowest classification error (equation 3). Next, we evaluate all the pairs and the following classifier in the cascade is the one which leads to the best classification results. If the improvement is insignificant we discard the classifier. The number of classifiers in a cascade is therefore different for each body part. The coarse-to-fine cascade strategy leads to a fast detection. The features are computed and evaluated for an input window only

if the output of the previous classifier in the cascade is larger than a threshold. The thresholds are automatically chosen during training as the minimum classifier responses on the positive training data. The output of each detector is a log likelihood map given by the sum of all strong classifiers

$$D(x_i) = \sum_{c=1}^{C} H_c(x_i)$$

where $C$ is the number of strong classifiers selected for the cascade. The location of the detected object is given by a local maximum in the log likelihood map. The actual value of the local maximum is used as a confidence measure for the detection. Note that the windows classified as an object have to be evaluated by all the classifiers in the cascade. The algorithm selected 8 strong classifiers out of 20 initial for each of the face detectors and 8 classifiers for each of the head detectors (4 feature types at 2 scales). The upper body and legs detectors use 4 classifiers selected out of 20 (two feature types of gradient orientations at 2 scales).

## 4 Detection system

In this section we describe the detection system, that is how we find the individual parts and how we assemble them. Detection proceeds in three stages: first, individual features are detected across the image at multiple scales; second, individual parts are detected based on these features; third, bodies are detected based on assemblies of these parts.

*Individual part detector.* To deal with humans at different scales, the detection starts by building a scale-space pyramid by sampling the input image with the scale factor of 1.2. We then estimate the dominant orientations and compute the groups of orientations as described in section 3.1. For the profile detection we compute a mirror feature representation. The estimated horizontal and vertical orientations remain the same, only the diagonal orientations for gradient features have to be inverted. Thus, for a relatively low computational cost we are able to use the same classifiers for left and right profile views. A window of a fixed size ($20 \times 20$) is evaluated at each location and each scale level of the feature image. We incorporate the feature location within the window into the feature value. This is computed only once for all the part detectors, since we use the same grid of locations for all body parts. The feature value is used as an index in a look-up table of weights estimated by AdaBoost. Each look-up table of a body part corresponds to one strong classifier. The number of look-up tables is therefore different for each body part detector. The output of the detector is a number of log likelihood maps corresponding to the number of body parts and scales. The local maxima of the log likelihoods indicate the candidates for a body part. To detect different parts individually we threshold the confidence measure. A threshold is associated with each part and is used to make the final classification decision. However, better results are obtained by combining the responses of different part detectors and then thresholding the joint likelihood.

*Joint body part detector.* Given the locations and magnitudes of local maxima provided by individual detectors we use the likelihood model described in section 2.1 to combine the detection results. We start with a candidate detected with the highest confidence and larger than a threshold. This candidate is classified as a body part. We search and evaluate the candidates in the neighbourhood given by the Gaussian model of geometric relations between two parts.

For example, suppose that a head (H) is detected. This means that the log likelihood ratio

$$D_H = \log \frac{p(\mathcal{F}|H)}{p(\mathcal{F}|non\ H)}$$

is above threshold. We can then use the position $(x_H, y_H)$ and scale $\sigma_H$ of the detected head to determine a confidence measure that there is an upper body (U) at $(x, y)$ with scale $\sigma$. In detail $G(x_H - x, y_H - y, \sigma_H/\sigma)$ is used to weight the computed $D_U$ (where $D_U$ is defined in a similar manner to $D_H$ above). The final score is

$$D_{U|H}(x, y, \sigma) = D_U(x, y, \sigma) + G(x_H - x, y_H - y, \sigma_H/\sigma) D_H(x_H, y_H, \sigma_H) \quad (4)$$

and the upper body is detected if this score is above threshold.

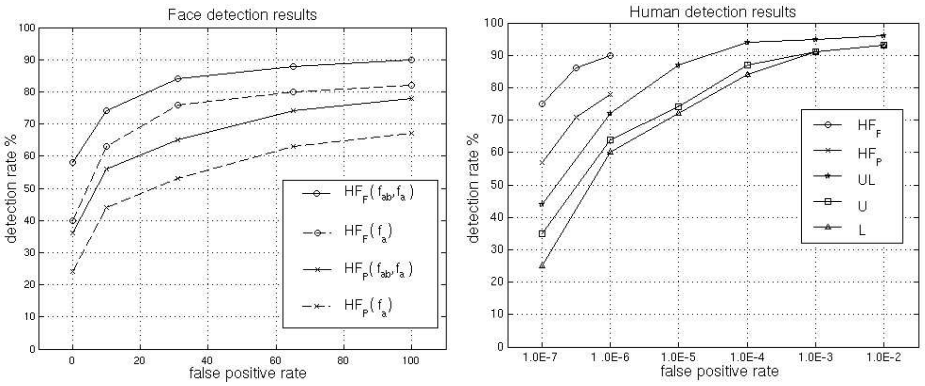
A confidence measure can also be computed for more than two parts; e.g. for an upper body, head and legs ($L$) sub-assembly $D_{L|U,H} = D_L + G(R_{L|U}) D_{U|H}$. If this score is higher than a threshold we accept this candidate as the body part and remove the closely overlapping neighbours. We can set the decision threshold higher than for the individual detectors since the confidence for body part candidates is increased with the high confidence of the other body parts. Given the new body part location we continue searching for the next one. There are usually few candidates to evaluate in the neighbourhood given by the Gaussian model.

The current implementation does not start to build the model from legs since this detector has obtained the largest classification error (cf. equation 3) and the legs are not allowed to be present alone for a body. In most of the body examples the highest log likelihood is obtained either by a face or by a head.

## 5 Experiments

### 5.1 Training data

Each body part detector was trained separately on a different training set. Approximately 800 faces were used to train the frontal face detector and 500 faces for the profile detector. The frontal views were aligned by eyes and mouth and the profiles by eyebrow and chin. For each face example we add 2 in-plane-rotation faces at -10 and 10 degrees. To train the frontal upper body/leg model we used 250/300 images of the MIT pedestrian data base [14]. 200 images for training the profile upper body model were collected from the Internet. The initial classifiers were trained on 100K negatives example obtained from 500 images. We then selected for each body part 4000 non-object examples detected with initial classifiers. The selected examples were then used to retrain the classifiers with AdaBoost.




(a) (b)

| # of false detections | 10 | 31 | 65 | 95 |
|---|---|---|---|---|
| Viola-Jones | 78.3% | 85.2% | 89.8% | 90.8% |
| Rowley-Baluja-Kanade | 83.2% | 86.0% | - | 89.2% |
| Schneiderman-Kanade | - | - | 94.4% | - |
| our approach | 75% | 85% | 89% | 90% |

(c)

**Fig. 5.** (a) ROC curves for face detectors. $HF_F(f_a)$ are the results for combined frontal head $H$ and face $F$ detector using single features $f_a$. $HF_F(f_{ab}, f_a)$ are the results for the detector using both single features and feature pairs. Similarly for the profile detector $HF_P$. (b) ROC curves for head/face, upper body and legs detectors. The results for head/face are converted from $HF(f_{ab}, f_a)$ displayed in figure (a). $U$ are the results for individual upper body detector and $L$ for the individual legs detector. $U|L$ are the results for upper body detector combined with legs detector. (c) Face detection results compared to state of the art approaches.

### 5.2 Detection results

*Face.* The MIT-CMU test set is used to test the performance of our face detectors. There are 125 images with 481 frontal views and 208 images with 347 profiles. The combined head-face models for frontal and profile faces were used in this test. Figure 5(a) shows the face detection results. The best results were obtained with the frontal face detector using a combination of simple features and feature pairs. We obtain a detection rate of 89% for only 65 false positives. These results are comparable with state of the art detectors (see figure 5(c)). They can be considered excellent given that the same approach/features are used for all human parts. Compared to the classifiers using only single features the gain is approximately 10%. A similar difference can be observed for the profile detectors. The performance of the profile detector is not as good as the frontal one. The distinctive features for profiles are located on the object boundaries, therefore the background has a large influence on the profile appearance. Moreover the test data contains many faces with half profile views and with in-plane-rotation of more then 30 degrees. Our detector uses a single model for profiles and currently we do not explicitly deal with in plane rotations. The detection rate of 75% with only 65 false positives is still good and is the only quantitative result reported on profile detection, apart from [19].

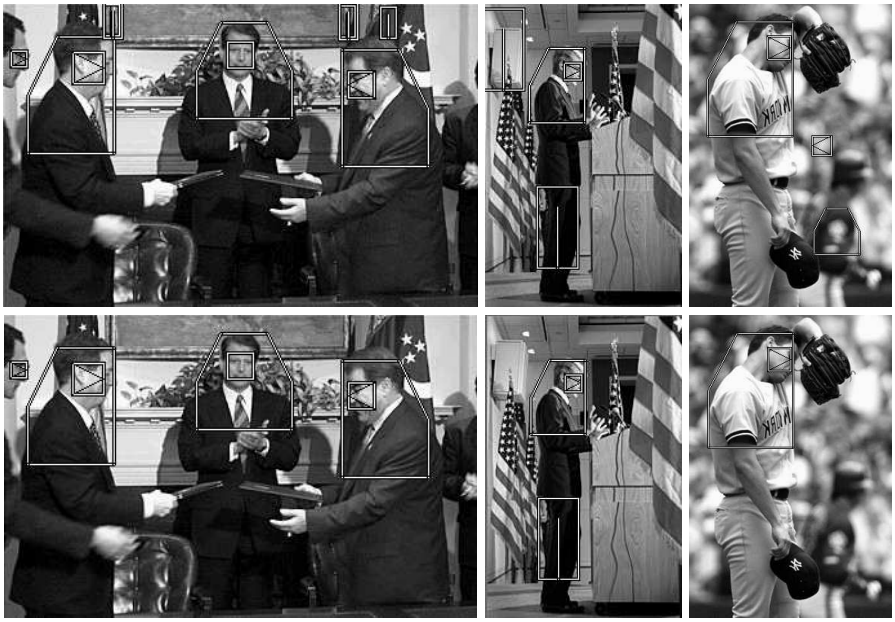


**Fig. 6.** Results for human detection. Top row: individual body part detection. Bottom row: detection with the joint likelihood model. The joint likelihood model significantly improves the detection results.

*Human.* To test the upper body and legs detector we use 400 images of the MIT pedestrian database which were not used in training. 200 images containing no pedestrians were used to estimate the false positive rate. There are 10800K windows evaluated for the negative examples. The false positive rate is defined as the number of false detections per inspected window. There are $10800K/200 = 54000$ inspected windows per image. Figure 5 (b) shows the detection results for the head/face, the frontal view of the upper body part and legs as well as the joint upper body/legs model. The results for head/face are converted from figure 5(a) and displayed on 5(b) for comparison. The best results are obtained for frontal head/face with the joint model. The result for the upper body and the legs are similar. For a low false positive rate the joint upper-body/legs detector is about 15% better than the individual upper-body and legs detectors. We obtain a detection rate of 87% with the false positive rate of 1:100000, which corresponds to one false positive per 1.8 images. This performance is better than the ones reported for pedestrian detection in [13, 14]. Note that an exact comparison is not possible, since only the number of images selected for the training/test is given. In addition, our approach performs well for general configurations in the presence of occlusion and partial visibility, see figures 6 and 7.

Figure 6 illustrates the gain obtained by the joint likelihood model. The top row shows the results of the individual detectors and the bottom row the combined results. The improvement can be observed clearly. The false positives disappear and the uncertain detections are correctly classified. Some other examples are shown in Figure 7.

**Fig. 7.** Human detection with the joint model. Top row: images from movies "Run Lola Run" and "Groundhog Day". Bottom row: images from MIT-CMU database.

## 6 Conclusions

In this paper we have presented a human detector based on a probabilistic assembly of robust part detectors. The key point of our approach is the robust part detector which takes into account recent advances in feature extraction and classification, and uses local context. Our features are distinctive due to encoded orientations of first and second derivatives and are robust to small translations in location and scale. They efficiently capture the shape and can therefore be used to represent any object. The joint probabilities of feature co-occurrence are used to improve the feature representation. AdaBoost learning automatically selects the best single and pairs of features. The joint likelihood of body parts further improves the results. Furthermore, our approach is efficient, as we use the same same features for all parts and a coarse-to-fine cascade of classifiers. The multi-scale evaluation of a $640 \times 480$ image takes less than 10 seconds on a 2GHz P4 machine.

A possible extension is to include more part detectors, as for example an arm model. We also plan to learn more than one lower body detector. If the training examples are too different, the appearance cannot be captured by the same model. We should then automatically divide the training images in sub-sets and learn a detector for each sub-set. Furthermore, we can use motion consistency in a video to improve the detection performance in the manner of [11].

### Acknowledgements

Funding for this work was provided by an INRIA postdoctoral fellowship and EC Project CogViSys.

## References


1. P. Felzenszwalb. Learning models for object recognition. In *Proc. of the CVPR, Hawaii, USA*, pp. 1056-1062, 2001.
2. P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *Proc. of the CVPR, Hilton Head Island, USA*, pp. 66–75, 2000.
3. R. Fergus, P. Perona, and A.Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. of the CVPR, Madison, USA*, pp. 264–271, 2003.
4. D. Forsyth and M. Fleck. Body plans. In *Proc. of the CVPR, Puerto Rico, USA*, pp. 678–683, 1997.
5. D. M. Gavrila. Pedestrian detection from a moving vehicle. In *Proc. of the ECCV, Dublin, Ireland*, pp. 37–49, 2000.
6. S. Ioffe and D. Forsyth. Probabilistic methods for finding people. *International Journal of Computer Vision*, 43(1):45–68, 2001.
7. H. Kruppa and B. Schiele. Using local context to improve face detection. In *Proc. of the BMVC, Norwich, England*, pp. 3–12, 2003.
8. S. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum. Statistical learning of multi-view face detection. In *Proc. of the ECCV, Copenhagen, Denmark*, pp. 67–81, 2002.
9. T. Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch - a method for focus-of-attention. *International Journal of Computer Vision*, 11(3):283–318, 1993.
10. D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the ICCV, Kerkyra, Greece*, pp. 1150–1157, 1999.
11. K. Mikolajczyk, R. Choudhury, and C. Schmid. Face detection in a video sequence - a temporal approach. In *Proc. of the CVPR, Hawaii, USA*, pp. 96–101, 2001.
12. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proc. of the CVPR, Madison, USA*, pp. 257–263, 2003.
13. A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. IEEE *Transactions on PAMI*, 23(4):349–361, 2001.
14. C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000.
15. Y. S. R. E. Shapire. Improving bossting algorithm using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.
16. D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. In *Proc. of the CVPR, Madison, USA*, pp. 467–474, 2003.
17. R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *Proc. of the ECCV, Copenhagen, Denmark*, pp. 700–714, 2002.
18. H. Schneiderman. Learning statistical structure for object detection. In *Proc. of the CAIP, Groningen, Netherlands.*, pp. 434–441, 2003.
19. H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *Proc. of the CVPR, Hilton Head Island, USA*, pp. 746–751, 2000.
20. H. Sidenbladh and M. Black. Learning image statistics for bayesian tracking. In *Proc. of the ICCV, Vancouver, Canada*, pp. 709–716, 2001.
21. L. Sigal, M. Isard, B. H. Sigelman, and M. J. Black. Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In *Proc. of the NIPS, Vancouver, Canada*, 2003.
22. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of the CVPR, Hawaii, USA*, pp. 511–518, 2001.