Intelligent Surveilance Systems

Paul Koppen, Nicholas Piël, Štefan Konečný
University Of Amsterdam

February, 2008

Abstract

Detecting objects in static images without any prior knowledge is considered to be one of the hardest tasks in machine vision. One of the most successful approaches is the use of haar-like features as described by Viola and Jones [12]. This approach is known for its high accuracy and low classification time cost. However, training such a classifier is computationally expensive.

In this paper we will demonstrate that we can construct a full-body classifier from combining detectors that label sub parts of a human body that will outperform each individual detector. We will show an approach which does not require any domain knowledge and operates on single frames, we will improve this design by incorporating background subtraction and knowledge learned from previous frames.

1 Introduction

This paper explores ways to combine the results of different visual detectors into a single high performance classifier. Towards this end we have constructed a human body detection system which achieves detection and false positive rates that outperform the original detectors. This full body detection system is most clearly distinguished from previous approaches in the low number of features it requires to find a human body. Each classification needs no more than two detected regions for the system to be able to completely reconstruct the person their location and size. In other detection systems full bodies are resembled by iteratively growing clusters of body parts, increasing the chance of correct classification.

The notion that different (weak) classifiers produce uncorrelated false positives, although correct regions are collectively detected, is directly exploited by our system. This enables us to find full bodies with only few detected parts. Now running on single images, this notion can be extended to series of images where incorrect detections are also uncorrelated through time.

Two approaches to combine different classifiers have been tested, a rule based system and a relational boundary learner plus the influence of background subtraction as a means to rule out false positives. Knowing that the body part detectors have been trained on complete images, background subtraction is applied in a late stage, on the detected regions, rather than on the input images directly.

1.1 Background

In this project our main goal is to detect people in static images without any prior knowledge. When we restrict ourselves to using just the pixel data we are dealing with arguably the most difficult problem in computer vision. In most recent research, robust people detection combines different features beyond the pixel data from a single frame, such as motion [11], in order to improve results. However, research has also been done into human detection by just taking the pixel data in account. For example, Micilotta et al. [8] detect and assemble humans from individual part detectors with an added skin color detector. Mori et al. [9] use segmentation to detect body parts and reconstruct these to a full body.

In our approach we try to boost individual haar cascade classifiers by making use of the relations between detected regions. This is done either through rules extracted from expert knowlegde, or through a learning system which learns those relations. The developed algorithms were evaluated on 900 sequential images provided by the CASSANDRA [5] project and the ISLA laboratory. The frames were captured from six different realistic video sequences recorded in a real-word setup (train station). From the set of images, (and previous research results) it was decided to

only label people in the foreground. The developed algorithms have been tested on these hand labelled image sequences. The learning system also used a subset of 100 images for training.

1.2 Haar cascades

For our body part classifier we make use of so called haar cascade classifiers [12] [7] This classifier works on grey scale images by utilizing window like features as depicted in figure 1.

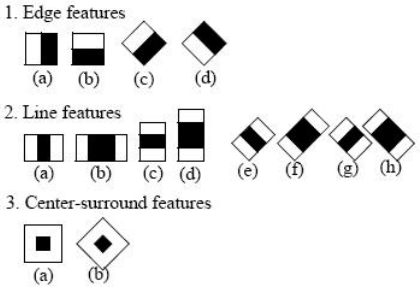

Figure 1: The used haar like features

By scanning through possible positions of these windows in images of various size (scaled by a certain factor), the sum of pixel values covered by black and white region are summed, and subtracted from each other (weighted, to compensate the difference size of the regions). These value are further applied as weak classifiers in an AdaBoost [4] framework, and form strong classifiers which constitute the stages of the final cascade. If the sample is classified as a positive instance, it is processed by a following stage of cascade, otherwise discarded. Thus only instances positively classified by all stages of the cascade are classified positively by the final classifier. The cascade acts as an degenerated, single branch decision tree.

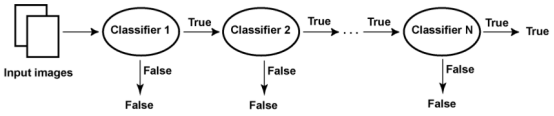

Figure 2: The cascade architecture of weak classifiers

Thanks to the use of simple image features and the cascade method, we are able to quickly discriminate between feasible and infeasible regions and focus our attention to the most promising ones. This results in the possibility of real time classification, which is vital in the surveillance problem domain.

1.3 Initial Experiments

For our classifier we planned to make use of the Haar-cascade detectors as provided by OpenCV [2]. However, initial experiments showed that these detectors performed below our expectations. As we need a relative high 'hit rate' in order to be able to combine the results of different classifiers we tried to improve these initial results in several ways:

1. Peeling of layers of the cascade
2. Tweaking parameters of the haar detectors
3. Applying morphological operations
4. Filtering images
5. Using different cascades

1.3.1 Peeling of layers of the cascade

As explained previously a haarcascade detector is a cascade of classifiers where with each stage it tries to reduce the false positive rate and the detection rate of the previous stage. Thus the main idea is that by peeling of one of the last cascades, we might increase the detection rate at the cost of a higher false positive rate. However, visual examination after this mutation showed that this resulted in uncontrollable and erratic behaviour.

1.3.2 Tweaking the parameters of the haar detectors

A haar detector applies a sliding window over the image in order to be able to detect object at different locations. To make sure that it will also detect objects of different sizes it will grow the window with a certain ratio. When we also scale down the image by a factor 1.3 prior to object detection and set the ratio at which the window size increases to 1.1 we can cut down the time cost of object detection without any loss in performance. Our experiments show that when we set the minimum window size to a region of 10 by 10 pixels we can even filter out some of the false positives.

1.3.3 Applying distortion correction operations

All the haar cascades we where using where trained on people straight in front of the camera. In our test data, the camera is tilted down towards the scene. We suspected that this might negatively influence the performance of our detectors and tried to improve the performance by applying distortion correction operations prior to detection. A result of such an operation can be seen in figure 3

(a) Original image (b) Straigthened image

Figure 3: Straigthening the image

Unfortunately these operations did not result in a better performance of our detectors. It could be that such operations introduce new artefacts which outweigh the positive results. Also, since estimating the correct operations is a hard problem which could be the subject of a whole project by itself we did not spent any more time on this approach.

1.3.4 Image filtering

Since previous attempts [6] to filter the image prior object detection has been shown to be fruitless we did not attempt to waste much time here. However, some quick experiments did show that disabling image histogram normalisation does significantly improve the performance of the haar cascade detectors. Therefor all our results where performed without histogram normalisation.

1.3.5 Using more cascades

Since the obtained results where still somewhat disappointing we experimented with extra body part classifiers. Due to time constraints we were unable to train our own classifiers. (Training requires a large labelled dataset and the training process itself can take up to a week). We were able to obtain extra 'Head Shoulder' detectors provided by S. Korzec [5] and M Castrillo-Santana [1].

While OpenCV also provides some Face detectors and recent experiments [3] show that they outperform the other classifiers, our experiments show that these classifiers are not suited for our domain. We have therefore chosen to exclude them. One can see the resulting list of classifiers in the next table.

CASCADE	AUTHOR	Target
HS_{10-18}	Korzec	Upper body
HS	Castrillo-Santana	Upper body
upperbody	Kruppa et al.	Upper body
lowerbody	Kruppa et al.	Lower body
fullbody	Kruppa et al.	Full body

2 Methodology

We will now continue with explaining how we will improve upon the individual body part detectors. The first section will describe our approach to interpretation and representation of the results of our body part detectors. Section 2.2 discusses our rule based system, which we built from a set of static relational rules after carefully examining combined behaviour of the loose body part detectors. Recognizing the possibility of dynamically learning those rules, section 2.3 will explicate our approach to learning from a train set, and how to use the learned model to classify new data.

2.1 Interpretation and Representation

The body part detectors that are used as a basis for our classifiers are based on cascades of haar-like features trained for a single body part. With an image as input, per detector a set of rectangles is returned designating the areas of specific detected body parts. The task at hand is to find stronger performance hidden in the relations between the different body part detectors. This means that the relations between different rectangles are modelled.

Various approaches to model relational differences are possible. Our methods analyze the relations between no more than two body parts at a time. The relation between two body part regions is calculated over three different geometrical measures; *angle*, *distance*, and *scale*. Angle is the relative direction from one body part to another. The distance is calculated as the Euclidean distance. Scale is the surface area of

one body part with respect to the other. In each measurement, rectangle centres are used as the position. Please notice that the relations must be calculated independent to the image size.

$$\begin{aligned} \text{direction} &= \arctan(\frac{dy}{dx}) \\ \text{Euclidean distance} &= \sqrt{dx^2 + dy^2} \\ \text{scale} &= \frac{h_1}{h_2} \end{aligned}$$

Each detector returns rectangles that hold a fixed width/height ratio specific to the body part. This constancy is used by our systems to recover the assumed full body from any single body part. Since each classification is performed on two body parts, the most reliable of the two is chosen for full body reconstruction.

2.2 Rule based System

Recognizing the original body part detectors' poor performance, the quality of our rule based classifier mainly relies on the high precision resulting from the combination of (a) lower body regions with full body regions, and (b) similar upper bodies. Especially the similarity of all the upper body detectors is remarkable. They perform almost the same, however, since their errors are uncorrelated, when they fire on the same region, we can almost certainly define a complete full body from that.

Prior to application of the rules, a pre-processing step is performed, aimed at combining the results of the different upper body detectors. This is done by splitting, clustering and combining the result of the three chosen upper body detectors.

Using simple clausal rules, candidates for full body (person classification) are constructed. These are then post-processed to average nested or similar regions. Averaging two regions is simply averaging each corresponding corner, extended with a weighting where appropriate.

The four rules are:

1. Construct a person from overlapping full body and lower body regions
2. Construct a person from overlapping full body and correctly aligned upper body regions
3. Construct a person from two overlapping upper body regions, suggesting high probability
4. Construct a person when a lower body is correctly aligned with an upper body

2.3 Learning based Approach

As a dynamic alternative to hard coded boundaries on the body part relations, a learning system is introduced which models the relations using mean and variance, which are derived using a training set. Noticing that each relation is a vector holding three measures, we calculate the entire covariance matrix instead of just the variance in each individual measure, which makes it also a more robust classifier. Using the sample mean and an unbiased estimator for the covariance matrix, we are not bound to presupposing normal distribution. For each combination of two body parts, we model both a positive and a negative distribution. This is useful for classification, which will be discussed later on in this section.

So, using four different body part types, our model exists of six relations represented by 3x1 mean and 3x3 covariance matrix for two sets (positive and negative). The mean and covariance are derived from one training set, accounting for both positive and negative examples in the following manner: For each image (f) in the manually labelled gold set (I), the relation (Rel) of any two body parts (r) belonging to the same identity (p) is added to the positive set (R^+). The relation of any two body parts not belonging to the same identity is added to the negative set (R^-).

2.3.1 Explanation of model

$$\begin{aligned} I &= \{f_1, \ldots, f_m\} && \text{labelled images} \\ f_i &= \{p_1, \ldots, p_n\} && \text{set of identities} \\ p_i &= \{r_1, \ldots, r_o\} && \text{ordered set of regions.} \end{aligned}$$

$$\begin{aligned} x_{fp_i} &= \text{region } i \text{ of person } p \text{ in frame } f \\ R^+_{ab} &= \{Rel(x_{fpa}, x_{fpb}) | a \neq b\} \\ R^-_{ab} &= \{Rel(x_{fpa}, x_{fqb}) | p \neq q, a \neq b\} \end{aligned}$$

Where Rel is the relation as defined in section 2.1

2.3.2 Definition of the Learning system

$$\mu^s_{ab} = \frac{1}{|R^s_{ab}|}\sum_{r\in R^s_{ab}} r$$
$$\Sigma^s_{ab} = \frac{1}{|R^s_{ab}|-1}\sum_{r\in R^s_{ab}} (r-\mu^s_{ab})(r-\mu^s_{ab})^T$$

When classifying new images, the results from the loose body part detectors are compared against the associated positive and negative model. Comparison is performed by calculating the Mahalanobis distance (D) to each of the means. The distance is formally defined as:

$$D(x) = \sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

In our case, we use it as a dissimilarity measure for a relation (r) on two body parts (a and b) to both positive (D^+) and negative (D^-) distributions:

$$D^+(r_{ab}) = \sqrt{(r_{ab}-\mu^+_{ab})^T {\Sigma^+_{ab}}^{-1}(r_{ab}-\mu^+_{ab})}$$
$$D^-(r_{ab}) = \sqrt{(r_{ab}-\mu^-_{ab})^T {\Sigma^-_{ab}}^{-1}(r_{ab}-\mu^-_{ab})}$$

2.3.3 Classification of full bodies given detected body parts

A certainty factor (α) is introduced, which puts a threshold on classifying relations as positives. If the distance to the positive set is α times smaller than the distance to the negative set, then the relation is classified as positive and a full body region is recovered from the two body parts.

Let $B = \{$*full, upper, lower, head*$\}$ be the retrieved set of detected body parts with each of *full, upper, lower, head* $= \{r_1, \ldots, r_n\}$, a collection of detected regions for each type, then

$$F = \{FullBody(r_a, r_b) | r_a \in B_a, r_b \in B_b, a \neq b, \alpha D^+(Rel(r_a, r_b)) < D^-(Rel(r_a, r_b))\}$$

Because multiple body part pairs exist for a single person, multiple largely overlapping full body regions may be generated for a single person. A late fusion which merges similar regions, overcomes this hurdle. Similarity is a value between 0 and 1, which is defined as the overlapping area divided by the union area of the two regions. A threshold value (typically 0.4) is applied to accept or refute similarity. Any two similar full body regions are merged by taking the average position and size.

$$Similarity = \frac{region_a \cap region_b}{region_a \cup region_b}$$

To summarize, the learning system learns from positive and negative examples. It models the training data using mean and covariance to estimate the distribution of body part relations. Classification of full bodies on new sample data is performed by comparing the distances to the positive and negative sets respectively. Similarity is used in a late fusion to combine multiple detected full bodies for single identities. More pre- and post-processing techniques have been investigated and will be presented in the next section.

3 Enhancements

In the previous chapter we focused on working with static images without any prior knowledge. However, on the domain we are working with we obtain an image sequence from a set of static cameras. This provides us with extra information we can choose to exploit in order to improve the performance. Also, since the cameras are static we can try to segment the foreground from the background and only apply our detectors to the foreground part of the image, we will take a look at this approach in section 3.1. Another way to try and lower the false positive rate is exploiting the knowledge that people can not appear or disappear in thin air. Since our images are a sequence of frames we can use the knowledge we have gained in previous frames in our current frame. We will explain exactly how we exploited this in section 3.2.

3.1 Background Subtraction

As our frames are taken by stationary cameras, we may use a method to differentiate between a static background and interacting objects to reduce the false positive rate. One such a method is presented in [10], which is effectively robust against lightning changes, repetitive motion, and tracking though cluttered images. We adopted the concept of "pixel processes" to calculate mean and variance in an online fashion. Rather than using a mixture model of Gaussians, our method stores the mean and variance for each pixel.

So, each colour component for each pixel is modeled over time using the following calculation:

$$\begin{aligned} \mu_{t+1} &= \alpha f_{t+1} + (1-\alpha)\mu_t \\ \Sigma^2_{t+1} &= \alpha(f_{t+1} - \mu_t)^2 + (1-\alpha)\Sigma^2_t \end{aligned}$$

where $f_t + 1$ is the current pixel value and α is a weighting factor that determines the speed at which the distribution's parameters change. This process incrementally builds a model, weighing historical data less and actual data more. A typical value for α is 0.05.

Using this model, each current pixel in the new image can be classified as either background or foreground, depending on how much it fits within the model. If it fits well, then it is highly likely static background. On the other hand, if it does not fit at all (e.g. a multitude of σ away from the mean), then it is likely to be part of a moving object and thus foreground. By matching each pixel against its model, a black and white image is built, where white indicates foreground and black indicates background.

(a) Original image (b) Detected foreground

Figure 4: Background subtraction

This classification can be used to exclude detected regions based on the foreground-background ratio over the detected area. Detected regions that contain too much black area are dropped out of the classifier because it is likely to be falsely labelled as positive.

3.2 Time Coherence

Since the data on which we operate is a sequence we can exploit this to improve our classifier. Remember that we are trying to combine individual body part detectors to a single person. If we assume that we can do this fairly reliable and we assume that a subsequent frame contains the same persons we can use the result of our 'person detector' in the previous frame as some sort of full body detector for the current frame. Thus, when trying to combine body part detectors at f_t we not only use the detectors that fire at f_t but supplement it with the persons detected in f_{t-1}.

4 Evaluation

To compare performance of our combined full body detector to the original body part detectors, it is necessary to first test both on the same domain. Secondly, since our system detects people from body parts, labelled body parts must be accompanied by an ID to collectively define a single person.

4.1 Mr. Tag

None of the datasets provided to us have this specific labelling, nor did any labelling tool allow for this structure of definition. So we built our own labelling tool named 'mr Tag'. This multi-platform program enables you to manually label an astonishing three thousand regions per hour! It loads a series of frames directly from a selected folder. Each time when you skip to the next frame it automatically copies all regions so that you only need to adjust the changed parts, saving precious minutes. Multiple different regions are possible (for us that was; upper body, lower body, full body, and face) and twelve different identities. Each region is marked with the identity number so they can even be preserved through time series. Regions can be easily adjusted by dragging and scaling them with the mouse. With this tool we have been labelling all 900 images and thus identified a total of about 10.000 regions.

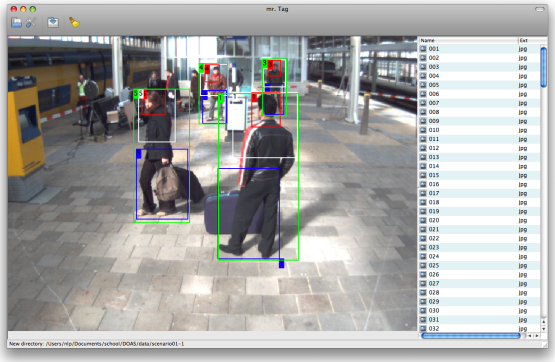

Figure 5: Mr. Tag - The labeling tool

4.2 Performance Assessment

To evaluate the performance of our methods, we should first look at the performance of the individual body part detectors. Assessing the performance of the body part detectors in our domain is done by calculating the True Positive Rate (TPR) and the False Positive Rate (FPR) of each detector. We define these rates as follows:

$$TPR = \frac{TP}{P}$$
$$FPR = \frac{FP}{N}$$

Thus, the TPR is a rate which resembles how much of the Gold set labels we are able to retrieve. We want this to be a as high as possible. FPR defines how specific our classifier is, as it is a ratio of the errors we produce to the maximum amount of errors (N)

We have a hit (TP) when one of our testboxes covers a certain ratio of a goldbox (recall) with a certain precision. For this we use the following formula:

$$Recall = \frac{\text{goldbox} \cap \text{testbox}}{\text{goldbox}}$$
$$Precision = \frac{\text{goldbox} \cap \text{testbox}}{\text{testbox}}$$

When given a set of testboxes, we have to identify for each goldbox which testbox matches that box best, as each goldbox can only have a single testbox attached to it. For this we make use of the overlap ratio:

$$OverlapRatio = \frac{\text{goldbox} \cap \text{testbox}}{\text{testbox} \cup \text{goldbox}}$$

Thus, in order to evaluate the performance, we calculate the *OverlapRatio* of each *testbox* with each *goldbox*, from which we take the combination of boxes that has the highest *OverlapRatio*. From these boxes we calculate the *Recall* and *Precision*. When the *Recall* is above some threshold γ and the precision is higher then some threshold β, we register a $TruePositive$ otherwise a $FalsePositive$. We repeat until we have no more *goldboxes* or *testboxes* then add the respective $FalsePositives$ and continue with the next frame.

5 Results

The developed algorithms were evaluated on 900 sequential images provided by the CASSANDRA [5] project and the ISLA laboratory. The frames were captured from six different realistic video sequences recorded in a real-word setup (train station). From the set of images, (and previous research results) it was decided to only label people in the foreground. The developed algorithms have been tested on these hand labelled image sequences. The learning system also used a subset of 100 images for training.

In the table below one can find the results of the individual body detectors we have used.

CASCADE	TPR	FPR
HS_{10}	**0.58**	0.97
HS_{11}	0.51	0.65
HS_{12}	0.41	0.40
HS_{13}	0.38	0.33
HS_{14}	0.28	0.12
HS_{15}	0.24	0.08
HS_{16}	0.18	0.06
HS_{17}	0.15	0.05
HS_{18}	0.11	0.04
HeadShoulder	0.08	0.09
upperbody	0.12	0.03
lowerbody	0.26	0.06
fullbody	0.17	**0.02**

We can clearly see that HS_{10} has the highest True Positive Rate, however this comes at a huge cost in False Positives. We can also see that the fullbody detector has the lowest FPR but on average the lowerbody seems to be the best of both worlds, having a low FPR but still able to obtain a relative high TPR.

CLASSIFIER	TPR	FPR
Expert Rules	0.41	0.16
Training	0.48	0.29
Training + Time	0.97	0.48

Our results depicted in the table above show that our expert rules are able to obtain a TPR of 0.41 while maintaining a low FPR of just 0.16. The trainer boosts the TPR even more but comes with a high cost in FPR. We can clearly see that by adding time coherence we can dramatically boost the TPR of the classifier. This improvement is extraordinary considering its simplicity.

When we plot all these values we obtain the graph in figure 6. Please notice that we represented the

HS_{10-18} classifier as a ROC curve where we connected the body part connectors HS_{10} - HS_{18} to form a single line.

From this graph we can see that we clearly outperform all individual detectors. Also, we expect that we can drop the FPR even more as our detector was able to classify people in the background observing the scene. Thus lots of these hits were incorrectly identified as a False Positive as we did not label all people in the background.

Unfortunately we where unable to thoroughly test the results of the background subtraction method. However the first visual inspection of its results look promising. And we think this can be a great way to further limit the false positives.

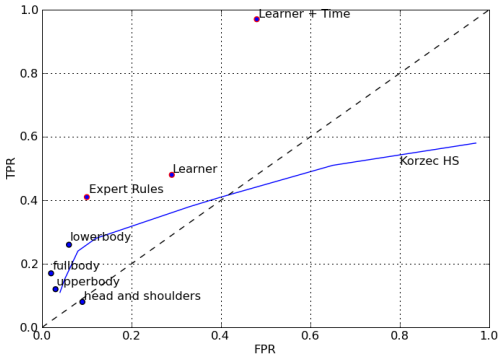

Figure 6: TPR vs FPR. For the various cascades we can clearly see that our classifier outperforms the rest

(a) The individual classifiers fireing

(b) Our expert rules combining them to single persons

Figure 7: Visualizing the results of our people detector vs the part detectors.

5.1 Conclusion

We have presented an approach for accurate human body classification, based on an array of weak classifiers. The approach was used to construct human body detection systems which perform significantly better than each of the single classifiers. The same approach was used in two different settings. One through a rule based system, the other through a simple learning system. In both cases only needing two detected regions to fully recover the entire person in each image.

The systems were designed to operate on still images, which casts the challenge to the domain of hardest applications in computer vision as of today. We have also shown that when we add domain specific information we can boost the performance even more. As our results by applying time coherence show.

We have shown how to robustly evaluate the results of our classifier in section 4. And while our methods do allow for tweaking of various parameters, the restrictions in time, unfortunately did not allow us to optimize these parameters.

Finally, this paper presents a set of detailed experiments on a difficult body detection dataset which has been previously studied. This dataset includes people under a very wide range of conditions including: pose, illumination, and occlusion. Correctly identifying the people in such images is a hard task, especially regarding the performances of all previous classifiers. Nevertheless systems that do detect the right people are extremely valuable, both to the CASSANDRA project and to other areas that use cameras in public areas.

References

[1] M. Castrillón Santana, O. Déniz Suárez, M. Hernández Tejera, and C. Guerra Artal. Encara2: Real-time detection of multiple faces at different resolutions in video streams. *Journal of Visual Communication and Image Representation*, pages 130–140, April 2007.

[2] Intel Corporation. Open source computer vision library.

[3] Modesto Costrill'on-Santana. Face and facial feature detection evaluation. *Third International Conference on Computer Vision Theory and Applications VISAPP08*, 2008.

[4] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.

[5] Sanne Korzec. Classifying the head-shoulder region and orientation in pedestrians. 2007.

[6] Sanne Korzec, Henco Visser, and Marcel Goksun. Detecting humans by combining human part-detectors in an urban setting.

[7] Rainer Lienhart and Jochen Maydt. *An Extended Set of Haar-like features for Rapid Object Detection*, volume 1. IEEE International Conference on Image Processing, 2002.

[8] Antonio Micilotta. Detection and tracking of humans by probalistic body part assembly.

[9] G. Mori, X. Ren, A.A Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. *IEEE Computer Vision and Pattern Recognition*, pages 326–333, 2004.

[10] C. Stauffer and W. E. L Grimson. *Adaptive background mixture models for real-time tracking*. Proceedings IEEE Conference on Computer Vision and Pattern Recognition, 1999.

[11] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[12] Paul Viola and Michael J. Jones. *Rapid Object Detection using a Boosted Cascade of Simple Features*. IEEE Computer Vision and Pattern Recognition, 2001.