Pedestrian tracking based on colour and spatial information

Florian H. Seitner*and Brian C. Lovell†

*Pattern Recognition and Image Processing Group, Vienna University of Technology, AUSTRIA
†National ICT Australia and School of ITEE, The University of Queensland, AUSTRALIA
{seitner, lovell}@itee.uq.edu.au

Abstract

This paper describes a tracking with appearance modelling system for pedestrians. A cascade of boosted classifiers and Haar-like rectangular features [6, 12] are used for the pedestrian detection. Statistical modelling in the HSV colour space is used for adaptive background modelling and subtraction, where the use of circular statistics for hue is proposed. By using the background model in combination with the detector, the system extracts a feature vector based on colour statistics and the spatial information. Circular [9] and linear statistics are applied on the extracted features to robustly track the pedestrians and other moving objects through the scene. An adaptive appearance model copes with partial or full occlusions and addresses the problem of missing or wrong detections in single frames.

Keywords*: tracking, background segmentation, appearance model, HSV, circular statistic, Haar-like features*

1. Introduction

The proposed tracking system uses a background segmentation algorithm in combination with an object classifier to quickly find pedestrians in each video frame. After detection of a possible pedestrian, the moving object is subdivided into three zones (head, upper body, and lower body) and the colour and spatial properties of each part which form the basic appearance model in this system are extracted. The colour information is analyzed in the HSV (Hue, Saturation and Value) colour space which provides a natural means of colour representation. The HSV colour model describes each colour by one angular (Hue) and two linear values (Saturation and Value). Although HSV has been applied to a wide range of applications like motion analysis, background modelling, and image retrieval, often its mixed topological nature of linear and circular domains is not appropriately taken into account. For example, it is clear that the mean of angles 359° and 1° is not 180° like the arithmetic mean would yield — it should be 0°. Furthermore, twins born one minute before and one minute after midnight are born only two minutes apart — not 23 hours and 58 minutes. Therefore, important definitions of circular statistics are given in Section 2 and used in this work to accurate process directional hue data.

Section 3 describes how color distributions can be approximated by parametric descriptions and how the adaptive background model distinguishes between foreground and background. The detector (Section 4) uses a cascade of boosted classifiers and Haar-like features to describe pedestrians in a highly efficient way. Section 5 describes how tracking features can be extracted by using the obtained pedestrian detections and the foreground data. The structure of the adaptive appearance model is described in Section 6 and results and conclusions of this work are given in Section 7 and Section 8.

2. Applied circular statistics

The algebraic structure of the line and the circle are different and therefore adequate methods of circular data analysis as discussed in [9] must be used when working with directional data. In contrast to the linear domain only one operation, the addition modulo 2π is available in the circular domain. Due to the fact that the circle is a closed curve, its natural periodicity must be taken into account. As described in [8] a set of N angular estimates can be represented by N unit phasors with arguments equal to the corresponding angular estimates. The mean angle $\hat{\mu}_p$ is then given by the argument of the phasor sum and this value is *independent* of the choice of origin. The general definitions of circular mean and variance based on this phasor sum are of the following form.

Definition 1 *Circular Sample Mean and Sample Variance: Let* $\{\hat{\alpha}(k)\}, \hat{\alpha} : \mathbb{Z} \longmapsto \mathbb{R}$ *be a set of* N *observations of a random variable in the circular domain* $[0, P)$. *Then the circular sample mean* $\hat{\mu}_p$ *and the circular sample variance* $\hat{V}_p$ *are defined by*

$$\hat{\mu}_p = \frac{P}{2\pi}\left(\left(\arg\left[\sum_{k=0}^{N-1} e^{\frac{j2\pi\hat{\alpha}(k)}{P}}\right]\right)\right)_{2\pi} \tag{1}$$

and

$$\hat{V}_p = \frac{P^2}{4\pi^2}\left[1 - \frac{1}{N}\left|\sum_{k=0}^{N-1} e^{\frac{j2\pi\hat{\alpha}(k)}{P}}\right|\right] \tag{2}$$

where $((\;))_{2\pi}$ *denotes reduction modulo* 2π *onto* $[0, 2\pi)$.

The circular variance $\hat{V}_p, \hat{V}_p \in \left[0, \frac{P^2}{4\pi^2}\right]$ cannot be compared directly with its linear equivalent σ^2 which lies in the domain $[0, \infty)$. However by using the relationship between the normal distribution on the circle (wrapped normal, [9]) and the normal distribution on the line a circular standard deviation in the range $[0, \infty)$ can be defined like

$$\hat{\sigma}_p = \sqrt{-2log_n(1 - \frac{4\pi^2}{P^2}\hat{V}_p)}. \tag{3}$$

Therefore when using statistical definitions in the context of hue values, we always refer to the above definitions from circular statistics.

3. Background model

An adaptive background model is used for background subtruction and motion-based foreground selection. A parametric model like used by Francois *et al.* [2] for real-time segmentation of video streams is used. The model operates on the HSV colour space since it clearly separates chromatic and intensity information which makes it suitable for both intensity and colour measurements. Each colour channel of a background reference pixel is modelled as a single and separate distribution since we use a static camera sequences and assume that each pixel of the background can be represented as a single colour (single model). A model based on mixtures of multiple distributions as used in [10] would also cope with multi-model backgrounds but is computationally more expensive and would bring no advantage in a single model background.

Since intensity and saturation are aligned in the linear domain, Gaussian distributions characterized by a mean μ and a variance σ are used for modeling those two channels of a pixel. Note that colours are not Gaussian distributed [11] but as shown in numerous works can be well approximated by the standard normal distribution [2, 13]. The probability of observing a saturation value S at a pixel with a reference distribution $N(\mu_s, \sigma_s)$ for the saturation is therefore given by

$$P(S) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{(S-\mu_s)^2}{2\sigma_s^2}}. \tag{4}$$

The probability for the intensity is calculated identically. Better suited distributions like the Beta distribution for the saturation and the Rayleigh distribution for the intensity would probably provide a more accurate description than the Gaussian for values near the extremes. By using the Beta distribution, the probability for the saturation is

$$P(S) = \begin{cases} \frac{\Gamma(\eta_s+\gamma_s)}{\Gamma(\eta_s)\Gamma(\gamma_s)} S^{\gamma_s-1}(1-S)^{\eta_s-1} & , 0 \leq S \leq 1 \\ 0 & elsewhere, \end{cases} \tag{5}$$

where Γ is the gamma function and $\eta_s > 0, \gamma_s > 0$ are the parameters of the Beta distribution for the saturation. The natural finite domain $S \in [0, 1]$ of the Beta function is a significant advantage. The intensity V for a pixel could be more appropriately described by a Rayleigh distribution which has a natural semi-finite domain $V \in [0, \infty)$ and can be written as

$$P(V) = \begin{cases} \left(\frac{V}{\sigma_v^2}\right) e^{-\frac{V^2}{2\sigma_v^2}} & , V \geq 0 \\ 0 & elsewhere. \end{cases} \tag{6}$$

Nevertheless, most cameras have only a limited working range and do not cope well near the extremes. Since we are mainly interested in an accurate modelling of pixels within the camera working range, the simpler Gaussian distribution is appropriate for saturation and value components. However, for the hue component the Gaussian distribution is inappropriate since circular data behaves quite differently from linear data. Here a von Mises distribution (named after the Austrian mathematician Richard von Mises) which is the circular equivalent of the Gaussian distribution is an adequate density function. The probability for a hue value H described by a von Mises distribution can be written as

$$P(H) = \frac{1}{2\pi I_0(\kappa_h)} e^{\kappa_h \cos(H - \hat{\mu}_p)} \tag{7}$$

where the concentration κ_h and the mean direction $\hat{\mu}_p$ characterize the von Mises distribution. I_0 is the modified Bessel function of the first kind of order 0. Note that the von Mises distribution degenerates to a uniform distribution for $\kappa_h = 0$. Similar in shape to the von Mises distribution is the wrapped Gaussian [9]. This distribution of the form

$$P(H) = \frac{1}{\sqrt{2\pi}\sigma_h} \sum_{k=-\infty}^{\infty} e^{-\frac{(H+2\pi k)^2}{2\sigma_h^2}} \tag{8}$$

where H is $N(0, \sigma_h)$, wraps the ordinary normal distribution around a circle. The resulting distribution is circular. A comparison between the von Mises and the wrapped normal distribution is done in [7, page 67]. It is shown that the wrapped normal distribution is a very accurate approximation to the von Mises distribution for moderate SNR (signal-to-noise ratio). For a color distribution described by a mean μ and a standard deviation σ, the SNR can be defined as

$$SNR = \frac{\mu}{\sigma}. \tag{9}$$

For a unimodal, static background we can assume that background changes occur slowly and signal distortions produced by, say, camera sensors are in a moderate range. Therefore the wrapped normal distribution is an appropriate simpler alternative to the von Mises distribution for modelling the hue values and is used in this system.

Each pixel of the background is therefore described by two Gaussians $N(\mu_s, \sigma_s)$ and $N(\mu_v, \sigma_v)$ for saturation and intensity and one wrapped Gaussian $\tilde{N}(\mu_h, \sigma_h)$. Initially, the means of all three colour channels of the reference distribution are set to the corresponding values of the pixels in the first frame. The variance for each background distribution is set to a minimal variance value of $\sigma_{min} > 0$. After the initialization the model is continually performing two main tasks. First the background mask is generated by comparing the reference distributions and the current frame. Secondly the distributions of our background model is updated by using the current frame information.

Background generation The model decides if a pixel with value $I = [H, S, V]'$ belongs to the background by thresholding the distance between the three colour channels and the means of the correspondent colour distributions $\mu = [\mu_h, \mu_s, \mu_v]'$ in the background model. A distance measurement defined by a distance δ_{hsv} is performed in all three colour channels independently

$$\delta_{hsv}(I, \mu) = \begin{pmatrix} \delta_h \\ \delta_s \\ \delta_v \end{pmatrix} = \left| \begin{pmatrix} \measuredangle(H, \mu_h) \\ S - \mu_s \\ V - \mu_v \end{pmatrix} \right|. \tag{10}$$

The circular domain of the hue is taken into account when computing the hue difference δ_h. We make the simplified assumption that the colour channels are independent from each other to reduce computational complexity. Porikli observed in [10] that this assumption degrades the quality of the results only minimally. If for a pixel at position x the difference for one of the channels is larger than a foreground threshold $\varepsilon_{\{h,s,v\}}(x)$ the pixel is marked as foreground $F(x) = 1$, otherwise it is labelled as background $F(x) = 0$.

$$F(x) = \|\delta_{hsv}(I(x), \mu(x)) < \varepsilon(x)\|_\infty \tag{11}$$

The threshold $\varepsilon_{\{h,s,v\}}$ depends on the variance of the corresponding colour channel

$$\varepsilon_{\{h,s,v\}}(x) = 2\sigma_{\{h,s,v\}}(x). \tag{12}$$

The range of 2σ is equivalent to a 95.5% confidence interval for a standard normal distribution. Since colour information is not Gaussian distributed [11] we can still expect each colour value to lie in the interval $[\mu - 2\sigma, \mu + 2\sigma]$ with a confidence of at least 75 percent by applying Tchebychev's Inequality theorem.

For reliable computation of the hue difference we have to test if the saturation value $S(x)$ of the frame pixel or the mean $\mu_s(x)$ of the reference distribution is close or equal to 0. Pixels with saturation equal to 0 are in the achromatic range of the HSV colour space. In this range the pixel lies on the central line of gray values and its hue information is meaningless and not usable as a distance measure. We define a pixel as achromatic if its saturation lies below a saturation threshold $\epsilon_{achr_s} = 0.2$. According to this we only use the reliable channels of the frame pixel and the reference distributions for comparison and distinguish between four cases:

1. if $S < \epsilon_{achr_s}$ and $\mu_s < \epsilon_{achr_s}$, check $\delta_v < \varepsilon_v$.

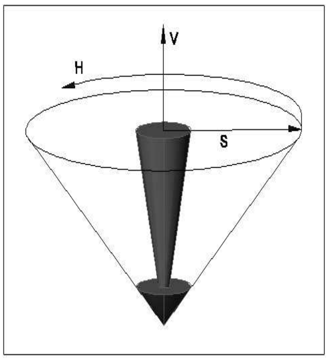

Figure 1. Separation of HSV colour space

2. if $S < \epsilon_{achr_s}$ and $\mu_s > \epsilon_{achr_s}$, check $\delta_v < \varepsilon_v$ and $|S - \frac{\mu_s}{\mu_v}| < \varepsilon_s$.
3. if $S > \epsilon_{achr_s}$ and $\mu_s < \epsilon_{achr_s}$, check $\delta_v < \varepsilon_v$ and $|\frac{S}{V} - \mu_S| < \varepsilon_s$.
4. if $S > \epsilon_{achr_s}$ and $\mu_s > \epsilon_{achr_s}$, check $\delta_v < \varepsilon_v$, $\delta_h < \varepsilon_h$ and $|S\cos(H, \mu_h)| < \varepsilon_s$.

In the first case no useful colour information is available and therefore only the intensity is used to measure the distance between the pixel and colour distribution. As shown in [2] the saturation can be scaled by the intensity to reflect the uncertainty of the colour information for lower intensity values (Case 2 and 3). The scaling is done when a saturation value is low and only partly reliable. In Case 4 the reference pixel as well as its reference distributions lie in the chromatic range and all channels can be used as distance measures. In this case the saturation is projected on the mean hue direction.

For low brightness values the saturation component is unreliable. Therefore an additional threshold $\epsilon_{achr_v} = 0.2$ for the intensity is used in our model. We add all pixels with an intensity $V < \epsilon_{achr_v}$ to Case 1 since this intensity range of the HSV colour space represents the nearly black pixels with strong achromatic properties. In Figure 1 the separation of the HSV colour space is shown. The gray region represents the range where only the intensity is used. The rest of the cone is the colour range where also the saturation and the hue are used for comparison. Note that instead of using a saturation threshold for deciding if a hue value is useful, another possibility would be to weight each hue value by its corresponding saturation as shown by Hanbury et al. [4].

Update background model After the pixels in the current frame have been labelled as foreground or background, the colour distributions of all reference pixels are updated by

$$\mu(t) = [1 - \alpha]\mu(t-1) + \alpha I \tag{13}$$

and

$$\sigma^2(t) = [1 - \alpha]\sigma^2(t-1) + \alpha[\mu(t) - I]^2. \tag{14}$$

Here α is the learning rate which defines how quickly old frames are forgotten. A minimal standard derivation σ_{min} is introduced, which prevents the decreasing of the standard deviation $\sigma(t)$ below a minimal value. This is useful in a long period of time when the background remains constant. As in the background generation step, only those channels of a pixels are updated which contain useful information to support the update. In the case where a reference pixel and a frame pixel are both in the achromatic range (Case 1) and no useful colour information is available, only its intensity distribution is updated. Note that the update of directional values according to equations (13) and (14) can be regarded as building the circular mean for grouped angular data with two sample points $\mu(t-1)$ and I and corresponding frequencies $[1 - \alpha]$ and α. No corrections for grouping is necessary but the range of $[0, 2\pi]$ has to be considered.

4. Detector

In this system a detector searches for pedestrians in single video frames. Therefore each frame is broken up into multiple sub-images and a classifier decides if the window contains a pedestrian.

As basic features for classification, a set of static *Haar-like rectangle* features [12] as shown in Figure 2 is used. This kind of features can quickly be computed by using *integral images* and builds the basic structure of each of our classifiers. By using an adapted version of the AdaBoost algorithm [3] we construct a *cascade of boosted classifiers* for quickly detect pedestrians. Boosting is widely used in the field of pattern classification and is the idea of letting multiple and simple (=weak) classifiers decide a classification task by a majority vote. For boosting any kind of learning algorithm like SVMs, neural networks, or decision trees can be used as a weak classifier.

The detector was trained with 1500 images of pedestrians from different angles. All training

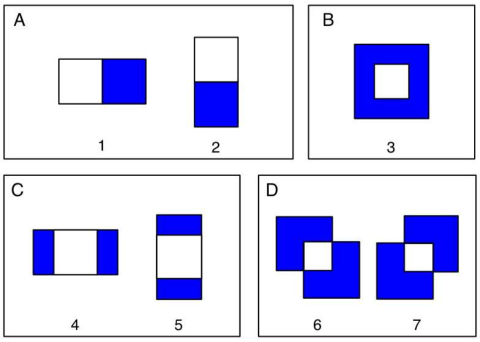

Figure 2. Features used in the detector

samples were manually extracted from multiple video sequences. The negative training samples were created by selecting random regions in images not containing any pedestrians (bootstrapping, [1]). The final detector cascade consists of 13 boosted classifiers. Since only one detector for multiple views of a pedestrian is used, the boosted classifiers in the cascade are quite complex with the smallest one containing 35 features and the largest with 257.

5. Tracking features

Feature Vector The effectiveness of the tracking process depends strongly on the choice of the tracking features. Our tracker uses the detector to get basic spatial information of a possible object and augments additional information by using the background model. The detector provides a set of K detection windows $d_k, k \in \{1..K\}$ in the current frame. Each window is defined by a size and a position. We firstly divide each detection window into three individual body zones (head, upper body, lower body) by using a fixed height ratio $r = (r_{head}, r_{ub}, r_{lb})' = [\frac{1}{4}, \frac{3}{8}, \frac{3}{8}]'$. Next each part is processed until it mainly contains a connected region of foreground pixels. After the processing the colour information is extracted by building the colour histograms and the means of each body part. The feature vector for each body part has the form

$$f_{\{head,ub,lb\}}(d_k) = \begin{pmatrix} P_k \\ S_k \\ H_{k,\{h,s,v\}} \\ \mu_{k,\{h,s,v\}} \\ \sigma_{k,\{h,s,v\}} \end{pmatrix} \tag{15}$$

where P_k is the position, S_k the size and $H_{k,\{h,s,v\}}, \mu_{k,\{h,s,v\}}$ and $\sigma_{k,\{h,s,v\}}$ the histograms,

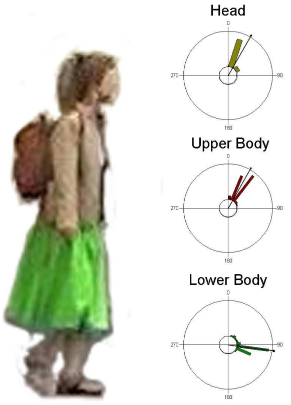

Figure 3. Hue colour features

means and variances for all three colour channels. In Figure 3 the statistical hue informations like the circular histogram and circular mean of a person are shown. For deciding if a hue or saturation value represents useful information and should therefore be included in a histogram, the same rules as in the background segmentation are applied . For generating the hue histogram all hue values are additionally weighted by their corresponding saturation.

In addition to using colour as an additionally tracking feature for following a pedestrian, it would also be possible to search for a person with known clothes on a multi-camera system. This could be highly interesting for automatically finding persons like a lost child or a robber after a bank robbery.

Distance Measures Since our feature vector contains spatial as well as colour information, different distance measures are used to calculate the distances between two of its elements. For measuring a spatial difference D_{spat} between to points $P_i, P_j \in \mathbb{R}^2$ the Euclidean distance is used

$$D_{spat}(P_i, P_j) = \|(P_i, P_j)\|_2 . \tag{16}$$

The difference D_{hist} between two colour histograms H_i and H_j is computed by using the Bhattacharyya distance

$$D_{hist}(H_i, H_j) = 1 - \sum_{k=1}^{B} \sqrt{H_i(k)H_j(k)} \tag{17}$$

where B is the number of histogram bins. In our work it was set to $B = 10$. The distance D_{dist} between a reference distribution with mean μ_i and standard deviation σ_i and a second distribution with mean μ_j is computed by

$$D_{dist}(\mu_i, \sigma_i, \mu_j) = \frac{|\mu_i - \mu_j|}{2\sigma_i}. \qquad (18)$$

After the distances for all elements of the feature vectors are computed independently with the corresponding distance functions, all distance values are tested against border conditions like a maximal position difference $\Delta_{velocity}$ or a maximal scale difference Δ_{scale}. This validation rejects all objects which do not provide enough features to support a robust tracking. After validation all distances are normalized, multiplied by a weight factor and totalized to an overall sum D.

The tracking features are for assigning detections to objects in the appearance model but are also used by the detector to find multiple detections of the same object in a frame. All sub-windows which were passed to the detector and classified as pedestrian windows $d'_{k'}, k' \in \{1..K'\}$, are tested for their similarity to each other. Similar detections with $D(d'_i, d'_j) < \xi_{merge}, i, j \in \{1..K'\}$ are grouped together to a single detection $d_k, k \in \{1..K\}$ with $K \leq K'$. Here ξ_{merge} is a similarity threshold. For each resulting object window d_k the number of windows M_k which were merged together is remembered, since it reflects the certainty of the detection to be a pedestrian.

6. Appearance model

The adaptive appearance model (AAM) used in this work has to address multiple tasks. A major problem the AAM should cope with is the stable handling of occlusions. In the case where an object is partly or completely occluded the AAM should be able to predict its current position and size. This is done by using the information about the velocity, direction, size and position of the occluded object, which was collected during the previous frames. Since a missing object detection due to the detector can be regarded as a complete occlusion, also this special case can be addressed by the appearance model. Additionally non-pedestrian objects which are incorrectly classified as persons by the detector can be filtered by the AAM. They normally occur only briefly or stay stationary for long periods of time and therefore can be distinguished from 'real' pedestrians.

Model states The appearance model can be divided into multiple steps. At the beginning the distances $D(o_j, d_k)$ between all existing objects $o_j, j \in \{1..N\}$ of the AAM and the detections $d_k, k \in \{1..K\}$ in the current frame are calculated. Here N and K represent the numbers of objects in the appearance model and the number of detections in the current frame. An existing object o_j is assigned to a detection if the distance $D(o_j, d_k)$ is smaller than a similarity threshold ξ_{sim}

$$assign(o_j, d_k) = \begin{cases} 1 & \text{if } D(o_j, d_k) \leq \xi_{sim} \\ 0 & otherwise. \end{cases} \qquad (19)$$

All objects in the AAM which could be assigned to a detection, are updated with the new information. For objects without a corresponding detection in the current frame, the position and size are updated according to the model assumptions. Additionally the appearance model creates new objects for all detections which could not be assigned to an existing object. When an object is created it gets an initial life value α_0 and is regarded as a possible pedestrian candidate. If a detection d_k is assigned to the object o_j, the life value α_j of the object is increased by a life bonus α_{bonus} like

$$\alpha_j = \alpha_j + min(\alpha_{max}, \alpha_{bonus} M_k). \qquad (20)$$

Here α_{max} is the maximal life bonus an object can receive per frame. The term M_k is the number of multiple detections during the detection phase which were merged together to the single detection d_k. A detection which is confirmed by multiple detection windows is therefore regarded as higher probability to be a pedestrian than a single window detection. If the life value α_j of an object rises above a threshold ξ_{acc} the object is accepted as a pedestrian and the object can be post-labelled as pedestrians in all the previous frames. If the life span decreases below a threshold $\xi_{destroy}$ it is removed from the appearance model.

In Figure 4 the different phases in the lifetime of an object are shown. The red rectangle on the left side represents the phase of the object creation until it is accepted as a pedestrian. The green rectangle shows the phase where the object is regarded as a pedestrian. Note that the objects in all frames of the first rectangle are post-labelled as pedestrians. The red rectangle on the right side represents the time when the tracker loses the object because it has left the field of view or is occluded for too long a time. Finally the object is removed from the object pool.

Collision areas Before the appearance model compares the new detections with the existing ob-

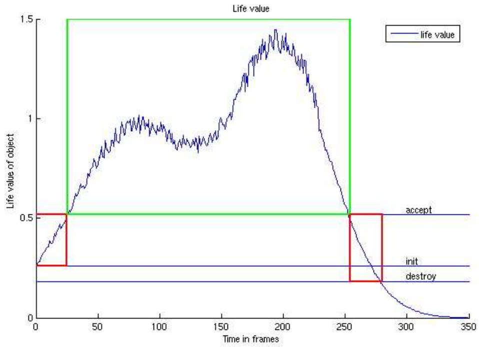

Figure 4. Life span of an object

jects in the model, it tests for collision areas. We define a collision area as a region in the frame, where at least two existing objects of the appearance model are close together and occlusions may occur. An object in a collision area still is distinguishable from the background, but since it can be partly occluded by other objects, its colour information is not as reliable as it would be outside of a collision area. Therefore, in the regions of multiple object occurrence we use a stronger similarity threshold ξ_{sim} to reject the wrongly assigned detections. After a detection has been assigned to an AAM object in a collision area, a careful update of the object information is necessary. Due to partly or complete occlusions and mixed colour histograms of colliding objects, no clear extraction of the colour information of a single object is possible. Therefore no update of the colour tracking features is done and no new objects are added to the appearance model in these regions.

7. Results

The detector was tested on 3 manually labelled test sequence where each sequence contained around 400 frames. The detection rate was near $d \approx 0.87$ with a false positive rate of $f \approx 10^{-4}$. This implies an average detection rate for each boosted classifier in the cascade of $d_{avg} = d^{\frac{1}{13}} = 0.989$ and a false positive rate of $f_{avg} = f^{\frac{1}{13}} = 0.49$. This is a lower detection rate compared to similar detection systems [12] but since we used only one detector for multiple views the results are still quite good. Additionally the appearance model filters most of the false positives and provides an approximation of position if a pedestrian is not detected.

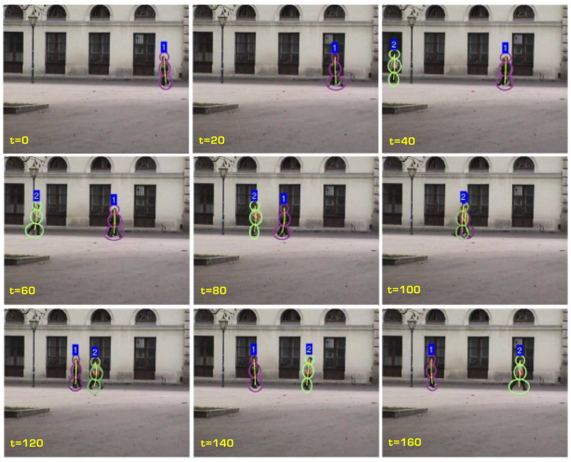

Figure 5. Tracking with occlusion

The appearance model performed very well in all three test sequences and could cope with multiple persons, occlusions, and missing or wrong detections. To test how well the appearance model copes with erroneous detections, we lowered the classification threshold of the detector progressively. By doing so the detector classifies more sub-windows as pedestrians which leads to a higher number of competing detection windows. This gives us the possibility of testing the quality of the tracking features and how well they distinguish the various objects.

The tracker performed quite well until we reached a false positive rate of $f \approx 25.10^{-4}$ (15 wrong detections per frame). No wrong pedestrian was accepted or real pedestrian lost below this threshold. Above this false positive rate, incorrectly accepted pedestrians increased rapidly. In two test sequences two and three wrong pedestrians were accepted. In the third test sequences three wrong pedestrians were accepted and furthermore one of three pedestrians was lost at the beginning, found again later, and tracked as a different pedestrian. The lost pedestrian is partly due to the large number of competing detection windows, but also because of the failing background model which needs a few frames at the beginning to initialize the background distributions correctly. The object validation according to certain border conditions for tracking features proves to be highly effective. Many wrong detections are already rejected during the feature extraction.

To test how good the AAM copes with occlusions we raised the classification threshold. Therefore, fewer pedestrian detections were registered and the

AAM had to cope with multiple frames where object detections are not provided by the detector. The AAM provides good approximations if an object is occluded for not longer than a time range of 15 to 25 frames. Since the color features of the pedestrians remain nearly constant, objects could be robustly found and reassigned after an occlusion.

The background model itself seems to perform well. It separates foreground and background well enough to extract the tracking features of moving objects. Since all test sequences are created with static cameras, one distribution per colour channel is sufficient for modeling the colour distribution.

8. Conclusion

In this paper we address the circular nature of the hue in the HSV colour space and provide accurate density functions for modelling color distributions. Furthermore, an adaptive background model is used in combination with an object detector to quickly locate pedestrians in video frames and to extract their colour and spatial information. An appearance model uses this information to robustly track the objects through the scene. This system demonstrates, how a person can quickly be subdivided into multiple body parts and a feature vector consisting of spatial as well as colour information can be extracted. Here a direct comparison of how well the use of appropriate circular statistics improves the processing of directional colour data provides interesting possibilities for future investigations.

In addition to using colour only as a tracking feature, the subdivision into multiple body parts provide an additional possibility. Multiple video sources like different video cameras can be used for searching for pedestrians according to a specified colour scheme for hair and clothing.

Analyzing the current results, the systems still seems to contain several possibilities for improvements which should be addressed in future work. A separation of the training samples into different views angles as in [5] and using multiple and view specialized detectors will probably improve the detection rate of the detector.

To additionally improve the background model, shadow removal techniques as well as mixtures of Gaussian distributions for multi-model backgrounds [10] can be used. An improved foreground segmentation would increase the quality of the tracking features and make the prediction of the object location and size more reliable.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience Publication, 2000.

[2] A. R. Francois and G. G. Medioni. Adaptive color background modeling for real-time segmentation of video streams. In *Proceedings of the International Conference on Imaging Science, Systems, and Technology*, pages 227–232, 1999.

[3] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. In *The Annals of Statistics*, volume 28, pages 337–374, October 2001.

[4] A. Hanbury and J. Serra. A 3d-polar coordinate colour representation suitable for image analysis. Technical Report PRIP-TR-077, PRIP, TU Vienna, Vienna, 2002.

[5] M. Jones and P. Viola. Fast multi-view face detection. Technical Report TR2003-096, Mitsubishi Electric Research Laboratories (MERL), June 2003.

[6] R. Lienhart and J. Maydt. An extended set of Haar-like features for rapid object detection. In *IEEE ICIP2002*, volume 1, pages 900–903, Sept. 2002.

[7] B. C. Lovell. *Techniques for Non-Stationary Spectral Analysis*. PhD thesis, University of Queensland, 1991. Brisbane.

[8] B. C. Lovell, P. J. Kootsookos, and R. C. Williamson. The circular nature of discrete-time frequency estimates. In *IEEE International Conference on ASSP*, pages 3369–3372, Toronto, May 1991.

[9] K. V. Mardia. *Statistics of directional data*. Academic Press, London, 1972.

[10] F. Porikli and O. Tuzel. Human body tracking by adaptive background models and mean-shift analysis. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, March 2003.

[11] N. Sebe and M. S. Lew. A maximum likelihood investigation into color indexing. In *Proceedings Visual Interface 2000*, pages 101–106, 2000.

[12] P. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 511–518, December 12-14 2001. Kauai, Hawaii.

[13] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.