

# Automatic Annotation of Humans in Surveillance Video

D.M. Hansen, B.K. Mortensen, P.T. Duizer, J.R. Andersen and T.B. Moeslund  
Laboratory of Computer Vision and Media Technology  
Aalborg University, Denmark

## Abstract

*In this paper we present a system for automatic annotation of humans passing a surveillance camera. Each human has 4 associated annotations: the primary color of the clothing, the height, and focus of attention. The annotation occurs after robust background subtraction based on a Codebook representation. The primary colors of the clothing are estimated by grouping similar pixels according to a body model. The height is estimated based on a 3D mapping using the head and feet. Lastly, the focus of attention is defined as the overall direction of the head, which is estimated using changes in intensity at four different positions. Results show successful detection and hence successful annotation for most test sequences.*

## 1. Introduction

Due to the many potential applications "Automatic Surveillance System" research has received much attention recently [9]. It is generally believed that a successful surveillance system will contain a figure-ground segmentation to detect humans in the images, tracking to maintain temporal coherence, and finally a recognition part to recognize identity, actions etc. [9]. So far no system has been able to successfully do any of the three parts in a robust manner and hence no commercial systems are available. While waiting for robust subsystems to build upon, different successful sub-applications have been built.

In our work we follow this trend and aim at an automatic annotation system for surveillance video. The long term goal is a network of connected cameras, each annotating every person passing by in terms of appearance, size, and behavior/action. In this particular paper we use one camera and limit the annotation to a few informative measures.

Different types of appearance measures exist such as color and style of hair, shirt, pants, shoes, beard, and glasses. In this work we focus on the most fundamental, namely the primary color of the hair, upper body clothing and lower body clothing. Regarding the size of a person

many measures exist, but we focus on the primary one, namely the height of the person. Annotating the behavior/action of a person is a complex task. In this work we limit behavior to the attention of a person defined as the direction of the head. We are not aiming at actual head pose but rather the general direction of attention.

The paper is structured as follows. In Section 2 it is described how humans are segmented and tracked. In Sections 3, 4, and 5 we describe the actual annotation of the three measures: appearance, size, and behavior, respectively. In Section 6 results are presented and finally the paper is concluded in Section 7.

## 2. Segmentation of Humans

Before any annotation can commence, each human has to be segmented from the rest of the image, i.e., figure-ground segmentation. As this is the first step in many systems analyzing humans several approaches exist based on e.g., background subtraction [4, 15], motion [14, 16], appearance [5, 13], and shape [7, 17]. For an overview see [9].

We apply a robust version of background subtraction, known as the Codebook method [6], because it has been shown to operate for ten hours without losing significant selectivity [3]. The method contains three steps: modeling the background, pixel classification and model updating.

Each pixel is modeled as a group of codewords which correspond to the codebook for this particular pixel. Each codeword is a cylinder in RGB-space. In each new frame each pixel is compared to its codebook. If the current pixel value belongs to one of the codewords it is classified as background, otherwise foreground.

The codebooks are built during training and updated at run-time. The training phase is similar to the pixel classification mentioned above except that a foreground pixel results in the construction of a new codeword and a background pixel is used to modify the codeword it belongs to using a standard weighting scheme. The codebooks generated in this way during training will typically fall into three categories:

**Static codebook** For example a pixel representing a road with no shadows or occlusions. Typically only one codeword is used.

**Quasi-static codebook** For example a pixel containing the sky, but sometimes occluded by vegetation due to wind gusts. During training typical two codewords will be constructed for this codebook, one for the sky and one for the vegetation.

**Noisy codebook** One of the above combined with "noise" in the form of a pedestrian, car etc. passing by the pixel or noise due to incorrect segmentation. The result will be an often high number of codewords for this codebook.

To handle the noisy codebooks a temporal filter is applied. It is based on the Max Negative Run-Length (MNRL), which is the longest time interval in which a codeword has not been activated. The filter effectively removes codewords with little support during the training phase, such as passing pedestrians.

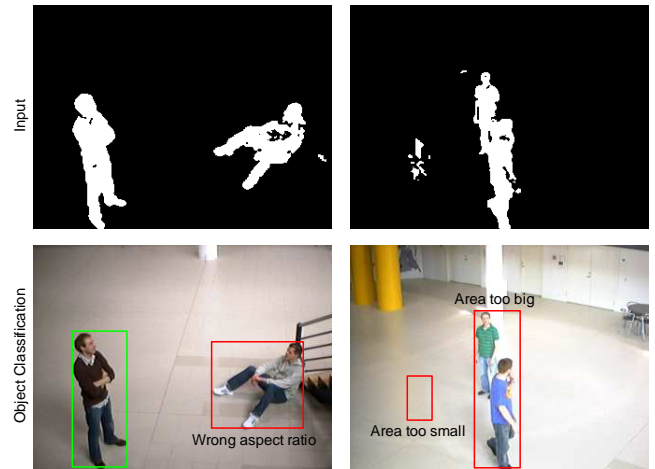
During run-time the activated codewords are updated in two ways. Firstly, as described above, using a standard weighting scheme to ensure updates with respect to slow changes in the scene, e.g., the position of the sun. The second type of update handles fast changes in the scene. Imagine a car enters the scene and is parked inside the scene. Obviously the car will be segmented as a foreground object. If the car stays in the scene it is considered background and new foreground objects passing the car can be correctly classified as foreground objects. This is done by defining a new codeword whenever a pixel is classified as foreground. If this new codeword has support in terms of a small MNRL then it is defined as a codeword and added to the codebook for this pixel. These types of codewords are denoted *non-permanent codewords* and can be removed again if they lack support for some time. A codeword learned during training is denoted a *permanent codeword* and can not be removed. Using the two updates provides a robust figure-ground segmentation for further processing.

## 2.1. Tracking

After having performed figure-ground segmentation we need to filter the output first spatially to ensure that each blob corresponds to one and only one human and second temporally to obtain a track of each human over time.

We assume walking or standing humans and can therefore define an interval of acceptance for the ratio between the height and width of the bounding box. Furthermore, after improving the output using standard filter methods we

introduce size criteria<sup>1</sup>, see Figure 1, together with a proximity criterion which merge correct sized blobs located on top of each other [1].



**Figure 1. An illustration of the size criteria.**

Each time a new track is initiated an ID is assigned. The tracking of IDs is done using temporal filtering with a zero'th order predictor and an Euclidean distance measure. Groups and partial occlusion will result in blobs that are too large or small respectively, which are ignored. This can result in small tracks or tracks not assigned an ID in every frame. To decide whether we want to keep a track or not for further processing we use the same principle as used for the MNRL [1].

## 3. Appearance: Color

Representing a human by colors has been used in tracking for some time. The standard approach is to divide the silhouette into a number of regions (usually three) and represent each region by either a Mixture of Gaussian or a histogram, see e.g., [10, 13] and [9] for an overview. These representations tend to describe the average color in a region. Furthermore, they often use fixed region borders resulting in a long shirt contributing to the colors of both the upper and the lower body parts.

Neither the averaging nor the fixed region borders constitute a problem in tracking. However, for annotation a different method is required, which allows for a detection of the primary color for both the upper and the lower body parts, respectively. Before explaining how this is done we first describe the chosen color space representation.

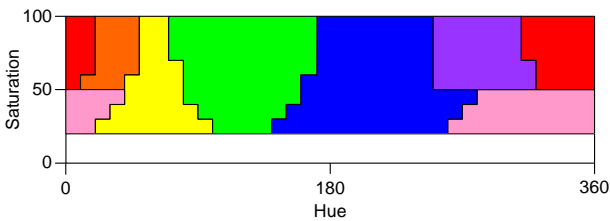
<sup>1</sup>These criteria are defined with respect to the camera's position (calibration), the layout of the scene, and the position of the human in the scene.

### 3.1. Color Space Representation

We use HSV colors since these are more practical for human interpretation<sup>2</sup> in an annotation system and at the same time decouples color and intensity allowing for less lighting dependent annotations.

Even though humans can distinguish between thousands of color shades, they are normally only able to remember 11 basic colors: red, green, yellow, orange, brown, pink, purple, white, gray, and black [2].

In order to convert from HSV color coordinates to these 11 color terms, the HSV-color space has to be subdivided. The hue-saturation color space is divided as shown in Figure 2 [1].



**Figure 2. The hue-saturation space is divided into eight fields [1].**

The hue-saturation space is divided into eight different fields. When the saturation is low, which is the white field in the figure, it is hard to separate different colors. Instead the color term is either black, gray or white, depending on the brightness of the color. The seven other fields represent the rest of the color terms. The brown color is missing in the figure, because it shares the hue and saturation with other fields. When the brightness is low, both orange, yellow and pink look brown. This is used to define the brown color.

### 3.2. Assigning Pixels to Body Parts

Taking the silhouette as input we are interested in first finding the color of the pixels belonging to the different body parts, see Figure 3(b), and then constructing a textual representation.

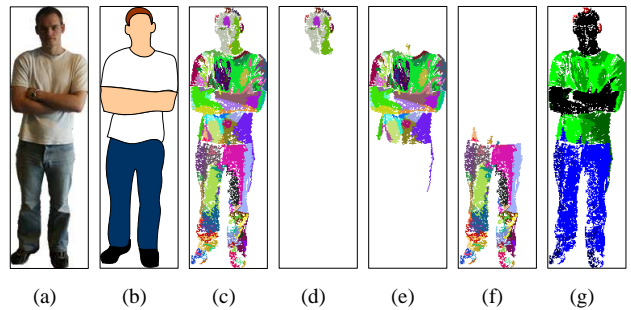
The input color image is segmented using the  $k$ -means clustering algorithm. The number of samples,  $n$ , is equivalent to the number of pixels in the image. The feature space is 3 dimensional, corresponding to the dimension of the color space. Experiments in this work show that better segmentation is achieved using all color components instead of only hue and saturation. The samples are divided into  $N_c$  clusters (20 in this work) using the  $k$ -means algorithm with an euclidian distance measure.

<sup>2</sup>For enquires into a database of annotations.

To compensate for oversegmentation, similar clusters are merged together. Two clusters are merged, if the euclidian distance in the hue-saturation space between their centers is below a predefined threshold.

A cluster can consist of a high number of blobs, which are not necessarily connected. We therefore apply a connected component analysis based on contours [1] to find all the individual blobs. The resulting blobs can be seen in Figure 3(c). For all blobs, the size and the mean color are found. Too small blobs are ignored as are blobs having a color similar to skin. We use a look-up-table in hue-saturation space to detect skin pixels [1].

The different blobs are assigned to one of the body parts based on the assumption that people are upright. We use the vertical position of their centers to assign blobs [12]: Head  $\in [0, 16\%]$ . Upper body  $\in [16, 45\%]$ . Lower body  $\in [45, 100\%]$ . 0% = top of bounding box and 100% = bottom of bounding box. See Figure 3.



**Figure 3. Assignment of blobs to body parts. (a) Input. (b) An ideal color segmentation of a person. (c) Blobs. (d) Head. (e) Upper body. (f) Lower body. (g) Merged blobs. Black represent skin, red the hair, green the shirt and blue the pants.**

The body parts still consist of a large number of blobs. To lower this number blobs are merged in the same manner as for clusters. Only this time, the color threshold is set higher to merge more blobs. Different thresholds can be used in the three body parts. Normally, people wear uniform colored pants, but might wear a multicolored shirt. Therefore, the threshold should be lower for the upper body than the lower body.

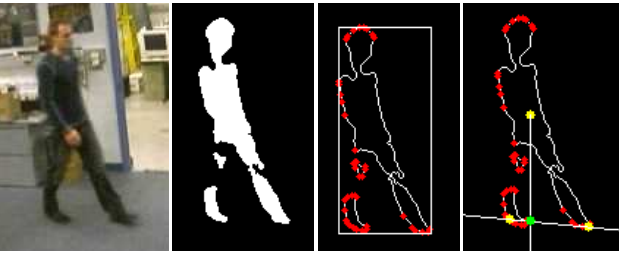
The merged blobs are illustrated in Figure 3(g). In the figure the black color represent skin blobs, which have not been merged. The two red shades are hair, the three green shades are the shirt and the three blue shades are the pants. The largest of the merged blobs is used as output in the textual description, since only the most dominant color is desired. In order to obtain stable color estimates we filter the output for each track by selecting the most often occurring color along a track.

## 4. Size: Height

To estimate the height we assume walking or standing persons. A prerequisite of the method is a calibration between the camera and floor, yielding the projection matrix shown in Equation 1.  $A$  is a matrix containing the intrinsic parameters,  $R$  and  $\vec{t}$  are the extrinsic parameters. The extrinsic parameters are chosen so the  $X$  and  $Y$  world axis are in the ground plane, which is assumed to be a two dimensional plane.

$$H = A [R\vec{t}] = \begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{14} \\ h_{21} & h_{22} & h_{23} & h_{24} \\ h_{31} & h_{32} & h_{33} & h_{34} \end{bmatrix} \quad (1)$$

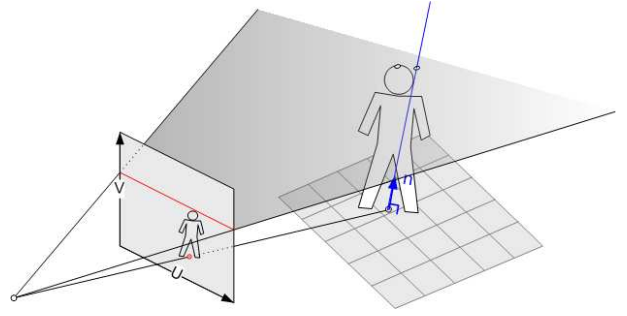
Two image points must be located to calculate the height. One point located on the ground between the person's feet called the ground point and a point at the top of the head. The ground point is estimated using convex hull points. In order to find the ground point, two foot points are located as the two convex hull points closest to the two lower corners of the bounding box, see Figure 4. The ground point is the intersection between a vertical line through the median point of the body and a line defined by the two foot points, see Figure 4. The head point is chosen as the topmost pixel in the body silhouette.



**Figure 4.** The principle used to find the top of the head and ground point.

After locating the ground point and head point, the next step is to map these points into world coordinates. The ground point is assumed to be in the ground plane, hence  $Z = 0$  in world coordinates. As described in [8],  $H$  can be altered to represent a mapping from the image plane to the ground plane (where  $Z = 0$ ) of the world coordinate system. This makes it possible to compute the person's ground plane position in world coordinates. Knowing the ground plane position, it is now possible to extend a line in the direction of the ground plane normal until it intersects a plane extended by the head point. This is illustrated in Figure 5. The height  $Z$  is calculated by Equation 2, where  $u$  and  $v$

are the head point coordinates and  $X$  and  $Y$  are the world coordinates for the ground point.



**Figure 5.** An illustration of the height estimate using the ground point and head point. The height is given by the  $Z$  world coordinate of the intersection point between the normal and the plane extended by the head point.

$$Z = \frac{(h_{21} - vh_{31})X + (h_{22} - vh_{32})Y + h_{24} - vh_{34}}{vh_{33} - h_{23}} \quad (2)$$

The height estimate is stored with the tracking ID for each frame the person is tracked. When a track is ended, e.g., by the person leaving the scene, we have several estimates of the person's height. However, only one height estimate is desired and this is chosen as the median value of all the height estimates. This makes the estimate robust towards outliers, which may be caused by poor segmentation in a few frames of the total tracking period. Furthermore, the method is based only on a couple of matrix operations and a convex hull method making it computationally inexpensive.

## 5. Behavior: Head Direction

The orientation of the person's head is used as an estimate for the person's attention. We first extract the head in an image and then classify the head direction into one of five different categories.

### 5.1. Head Extraction

Before the orientation of the head can be determined, template matching is first applied to extract the head.

The head is extracted from the silhouette of the person. The binary image allows for fast matching with exclusive-or operations. A number of preprocessing steps trim the size of the silhouette to further reduce processing time.

The template consists of a head and torso whose size is changed dynamically for scale invariance. The head and torso is dimensioned based on typical human measures, and the size is determined by an estimate of the person’s shoulder width. The shoulder width is estimated as 90 percent of the minimum distance from the median of the silhouette to either side of the person’s bounding box. This approach provide a reliable measure invariant to typical segmentation errors. Figure 6 shows the steps in the head extraction process.

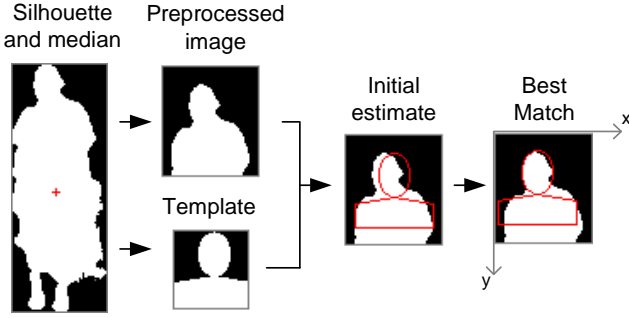


Figure 6. The steps in the head extraction process.

## 5.2. Head Direction Classification

The head is classified into one of five classes, corresponding to an orientation of Right 90° , Right 45° , 0° , Left 45° and Left 90° , where 0° is a person facing the camera. For determining the head orientation at low resolution, a view-based method inspired by [11] is used. The method in based on calculating moments of pixel intensity distribution in subimages of the extracted head. The moment features provide invariance to scaling and robustness toward imperfect segmentation and head extraction. In [11], the head is divided into twelve subimage in a 3 × 4 grid, as depicted in Figure 7(a), and for each subimage a feature value  $W$  is calculated using Equation 3.

$$W = \frac{\text{mean}(\text{subimage}) - \text{mean}(\text{image})}{\text{std}(\text{image})} \quad (3)$$

Only those pixels that are within the silhouette of the person are part of the calculations. Furthermore, any highlights in the head region are removed beforehand, to reduce the influence of face items like eyeglasses or shiny skin. Inspection of these twelve features in the training data showed some features provide very little or no information. The corners had a poorly separated mean value and a high variance. Furthermore, many of the features are highly correlated. To improve this, 40 subimages in a 5 × 8 grid were used to

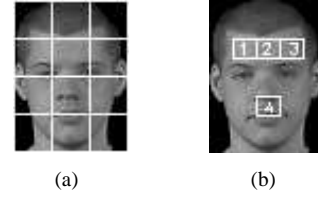


Figure 7. (a) head region divided into twelve subimages. (b) four selected subimages of a 5 × 8 grid.

locate more specific areas in the head, and the four features depicted in Figure 7(b) were selected. The numbers in the subimages corresponds to the numbering of the features, and the plot in Figure 8 depicts their mean value and variance. As seen in the figure, these features are able to separate the classes in the training data, but does also raise the requirements for the training data to be representative. The different grid division gives areas in the head where e.g. the area is skin-only when the head is oriented 0° and hair-only when the head is oriented ±90° .

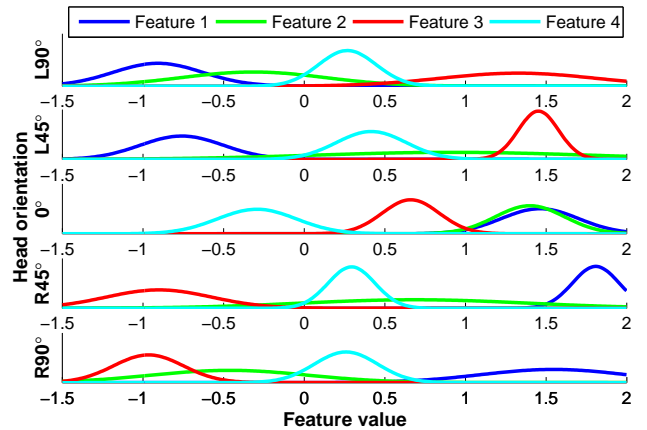


Figure 8. The mean and variance of the four features, calculated from the four selected subimages.

The four-dimensional feature vector is classified using  $k$ -nearest-neighbor using Euclidean distance. A temporal filter is applied to each track selecting the most occurring head orientation class within the last  $X$  frames as the output. Experiments have shown that using  $X = 15$  frames gives the best output and seems to be a reasonable size to filter out sudden head movement. 15 frames correspond to one second meaning that attention is defined as a constant head direction for one second.

## 6. Results

In this section we present test results for the previous sections. All results are recorded with a standard web-camera with a resolution of 320x240 pixels and with a frame rate of 15 Hz.

### 6.1 Segmentation of Humans

111 frames are selected randomly from five different video sequences. 91 frames contained one or more humans. The pixels belonging to a human are found manually to obtain ground truth. On average the false acceptance rate<sup>3</sup> is 0.06% and the false rejection rate<sup>4</sup> is 5.95%. The errors originate especially from foreground camouflage and shadows. When the appearance of a person's clothing is similar to the background then a "hole" can appear in the silhouette leading to false rejections. This, however, seldom affects the rest of the system. Shadows can affect both the estimation of the colors as well as the height estimation. In most cases, however, the temporal filtering solves the problems<sup>5</sup>.

The tracking works very well because occlusions are ignored. If the system is to be used to annotate individuals during occlusions, then better segmentation and tracking is required. Possible solutions can be found in [9].

### 6.2 Appearance: Color

The color descriptor is tested in two test cases. The first case tests the correctness of the annotated color term and the second case tests the variation of the extracted color values for a person walking through the scene.

The purpose of the first test case is to analyze the ability to annotate different colors. Furthermore, the transformation from HSV to the 11 color terms is tested. 33 input images are manually segmented, so only the person and a white background is present. Four people agreed on the color terms yielding ground truth. See examples in Figure 9. The first row is the ground truth for the upper body, while the second row is the annotated color terms.

All shirts are annotated correctly, except for the turquoise-green t-shirt (#6), which is annotated as blue. The reason for wrong assignment is that the HSV value lies in between the green and blue color term, where the border is sharply defined, despite it being rather fuzzy in reality. This problem also exist when humans determine a color and a

<sup>3</sup>The false acceptance rate is defined as the number of background pixels falsely accepted as foreground pixels, divided by the total number of manually annotated background pixels.

<sup>4</sup>The false rejection rate is defined as the number of foreground pixels falsely rejected, divided by the total number of manually annotated foreground pixels.

<sup>5</sup>For more results on the figure-ground segmentation please refer to [3].

suggestion is therefore to represent borderline colors by two different colors.

This test also highlights difficulties when a person is wearing shirt and pants of the same color. For test subject 2, it is difficult to find the border between the upper and lower body. The clusters often separates the upper and lower body, but after the merging of clusters, the upper and lower body can be fused. The annotated color terms are however often correct, because the largest color blobs of both the upper and lower body are still correct - blue in this particular case.

Lastly it was found that problems occur when people are wearing skin-colored clothes. In these situations, the produced color term might only represent the second most dominant color. This is the case with the pants of test subject 10. The pants are classified as skin making the white shoes the largest part of the lower body. The pants are therefore labeled as white. To handle such situations more advanced analysis of the blobs is required.

The second test case tests the color descriptor when connected to the tracking module. Instead of using manually segmented images, the output produced by the tracking module is used directly. The same test recordings as in the first test case are used.

In general, for all test subjects the same issues are observed. Outliers occur mainly because color descriptors are made even though the person is not completely in the scene. Also, misclassifications from the figure-ground segmenting caused by strong shadows lead to false color terms. However, these problems are eliminated (except for the hair where the resolution is too low to produce trustworthy results) by temporal filtering and the correct color terms are found in all cases except a few where blue and green are mixed up as illustrated in Figure 9.

### 6.3 Size: Height

We recorded a video with four people moving in different directions and at different speeds within the scene to test the height estimate. The results are shown in Table 1. A total of 17 tracks or height estimates are performed and the number of frames available for the height estimation is between 33 and 90 (2-6 seconds). The table shows the statistics of each person. Despite the segmentation issues and the merging of two tracks, the poorest height estimate is only 3.4 cm from the actual height.

### 6.4 Behavior: Head Direction

Two types of tests are conducted, a quantitative and a qualitative test.

In the quantitative test eight test subjects are instructed to put their heads in one of the five head positions for a certain period of time by standing still and looking at certain



**Figure 9. 9 test images. The first row of color terms is ground truth. The second row is the output from the system.**

	Person 1	Person 2	Person 3	Person 4
# tracks	4	7	2	4
Height	182.0 cm	192.0 cm	170.0 cm	186.0 cm
Mean	181.4 cm	191.5 cm	171.9 cm	186.0 cm
Std.dev.	1.9 cm	1.9 cm	2.2 cm	0.5 cm
Min	179.7 cm	189.4 cm	170.3 cm	185.7 cm
Max	184.0 cm	193.7 cm	173.4 cm	186.7 cm

**Table 1. The result of estimating the height for four people. # tracks lists the number of tracks for each person in the recording. Height is the actual height of the person.**

Sub.	R90°	R45°	0°	L45°	L90°
1	100%	4.9%	75.8%	0%	100%
2	48.2%	100%	100%	90.1%	100%
3	100%	100%	100%	0%	100%
4	95.8%	45.6%	86.3%	75.3%	100%
5	20.8%	88.3%	100%	100%	100%
6	0%	97.6%	100%	100%	0%
7	100%	86.3%	100%	100%	100%
8	100%	88.1%	100%	100%	100%

Total correct classifications:	80.1%
Classifications in adjacent class(es):	19.9%
Classifications in other classes:	0.0%

**Table 2. The result of estimating head direction for the test subjects.**

markers mounted in the room. Two samples per class per subject is used to train a classifier which is tested on more than 2000 frames.

The results are shown in Table 2. The table shows the percentage of correctly classified frames for a specific class and test subject along with the total correct classifications. The system yields a classification rate of 80.1%. Of the 19.9% misclassifications all were classified within the adjacent class(es), which is  $\pm 45^\circ$  of the correct class. If the five classes are merged into three classes (Right, Front and Left) the result is a 98.5% correct classification rate. The results vary between the test subjects. The poorest classification is for test subject 1 who has hair that sticks up. The hair moves the subimages to an undesired area causing more misclassifications. Test subject 5 and 6 are the same person without glasses and with glasses, respectively. With glasses it is harder to classify Left  $90^\circ$  correctly and it is misclassified as Left  $45^\circ$ . In the qualitative test 40 video sequences (7000+ frames) are used. In each sequence a test person walks towards the camera and changes his/her position in a supervised manner. When the subject is within 9

meters enough head pixels ( $> 10 \times 15$ ) are available for classification. The results depend on the distance to the camera. When the subjects are more than 4.5 meters away 49.0% is correctly classified. 75.4% was correctly classified for subjects closer than 4.5 meters. The total correct classification for the entire walking distance is 62.2%. By merging the five classes into three classes (Right, Front and Left) the far distance, close distance and total correct classification becomes 72.5%, 92.4% and 82.5%, respectively.

In real-life applications a frame-by-frame head direction is not always desirable and annotation of general tendencies in the form of a textual output rather than frame-by-frame output is more feasible, e.g., registration of how many customers are looking at a certain display window. In such applications, a coarse time location of a change in head position is sufficient. The filtered output from the head direction classification have stable periods where the same class is outputted for several frames. If we define a stable period

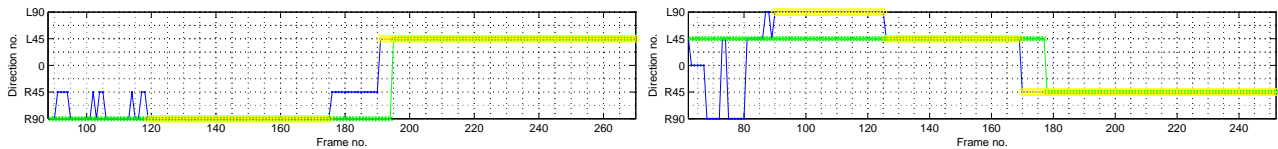


Figure 10. Head direction. Blue: Estimate. Green: Ground truth. Yellow: Stable period.

to be at least one second we can create a textual output. For the left graph in Figure 10, the textual output is {Frame 119-175 looking Right 90°, Frame 191-270 looking Left 45°} using the stable periods and is depicted graphically in the figure by yellow lines<sup>6</sup>. Besides the inaccurate location of the transition in head direction, the textual output is correct.

A problematic issue is shown in Figure 10 (right). In this case, an extra head direction is listed in the textual output. However, it is typically a head direction within 45° of the correct direction and occurs when the subject is located farthest away from the camera or during head direction transition. Hence in the context of for example determining if a customer is looking at the stores on his left or right, the proposed head direction provides a reliable result.

## 7. Conclusions

The paper has presented annotation of persons passing a surveillance camera in terms of primary color of upper and lower body part, the height of the person, and head direction to estimate attention. The annotations are incorporated into a system based on a robust segmentation of individuals.

One limitation of the system is its lack of occlusion handling. For dense areas this will reduce the number of persons that are actual annotated and a more advanced segmentation is required to handle such situation, see [9] for examples. In future work an active camera will be controlled by the segmentation to zoom in on the face for obtaining a quality picture to store along with the other annotations or for further processing, e.g., ID or facial expression recognition.

## References

- [1] J. Andersen, P. Duizer, D. Hansen, and B. Mortensen. Automatic Annotation of Humans in Surveillance Video Recordings. Technical report, Computer Vision and Media Technology, Aalborg University, Denmark, 2006.
- [2] B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley, CA, 1969.
- [3] P. Fihl, R. Corlin, S. Park, T. Moeslund, and M. Trivedi. Tracking of Individuals in Very Long Video Sequences. In

<sup>6</sup>These frame indexes can be converted to 3D view direction using the camera calibration.

- Int. Symposium on Visual Computing*, Lake Tahoe, Nevada, USA, November 6-8 2006.
- [4] M. Heikkila and M. Pietikainen. A Texture-Based Method for Modeling the Background and Detecting Moving Objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4), 2006.
- [5] M. Hu, W. Hu, and T. Tan. Tracking People through Occlusion. In *International Conference on Pattern Recognition*, Cambridge, UK, Aug 2004.
- [6] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis. Real-time Foreground-Background Segmentation using Codebook Model. *Real-time Imaging*, 11(3):167–256, 2005.
- [7] V. Krüger, J. Anderson, and T. Prehn. Probabilistic Model-Based Background Subtraction. In *Scandinavian Conference on Image Analysis*, Joensuu, Finland, Jun 19-22 2005.
- [8] C. Madden and M. Piccardi. Height Measurement as a Session-Based Biometric for People Matching Across Disjoint Camera Views. In *Image and Vision Computing New Zealand*, Dunedin, New Zealand, 28 - 29 Nov 2005.
- [9] T. Moeslund, A. Hilton, and V. Krüger. A Survey of Advances in Vision-Based Human Motion Capture and Analysis. *Journal of Computer Vision and Image Understanding*, 104(2-3), 2006.
- [10] K. Okuma, A. Taleghani, N. Freitas, J. Little, and D. G. Lowe. A Boosted Particle Filter: Multitarget Detection and Tracking. In *European Conference on Computer Vision*, Prague, Czech Republic, May 11-14 2004.
- [11] S. Park and J. Aggarwal. Head Segmentation and Head Orientation in 3D Space for Pose Estimation of Multiple People. In *IEEE proc. Southwest Symposium on Image Analysis and Interpretation (SSIAI 2000)*, Austin, TX, USA, 2000.
- [12] S. Park and J. Aggarwal. Simultaneous tracking of multiple body parts of interacting persons. *Computer Vision and Image Understanding*, 102(1), 2006.
- [13] D. Roth, P. Döubek, and L. Gool. Bayesian Pixel Classification for Human Tracking. In *IEEE Workshop on Motion and Video Computing (MOTION'05)*, Breckenridge, Colorado, Jan 2005.
- [14] H. Sidenbladh. Detecting Human Motion with Support Vector Machines. In *International Conference on Pattern Recognition*, Cambridge, UK, Aug 2004.
- [15] C. Stauffer and W. Grimson. Adaptive Background Mixture Models for Real-Time Tracking. In *Computer Vision and Pattern Recognition*, Santa Barbara, CA, USA, June 1998.
- [16] P. Viola, M. Jones, and D. Snow. Detecting Pedestrians Using Patterns of Motion and Appearance. *International Journal of Computer Vision*, 63(2), 2005.
- [17] B. Wu and R. Nevatia. Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detection. In *International Conference on Computer Vision*, Beijing, China, Oct 15-21 2005.