People Counting In Complex Scenarios[1]

T. Schlögl[2], B. Wachmann[3], H. Bischof[4], W. Kropatsch[5]

[2]Advanced Computer Vision GmbH - ACV, Tech Gate Vienna, Donaucitystraße 1, A-1220 Vienna, Austria, {thomas.schloegl@acv.ac.at}
[3]Siemens AG Österreich, Programm- und Systementwicklung, Graz, Austria
[4]Institute for Computer Graphics and Vision, Graz University of Technology, Austria
[5]Pattern Recognition and and Image Processing Group, Vienna University of Technology, Austria

Abstract:
This paper addresses vision-based people counting under the assumption of an oblique stationary camera setup without any usage of tracking. Our approach is based on motion detection and blob analysis. The estimation of the number of people in the scene is done by accumulating the estimation values of each blob. The blob specific estimation values are computed using four input values: the minimum and maximum number of people which may be present in the blob, the blob area and a predicted number number of people which is computed from the temporal change at the blob location. We have included shadow elimination using colour invariants which are obtained from the measurement of coloured object reflectance assuming a Gaussian colour model. The method was evaluated in single camera scenarios for various sequences taken in halls of shopping centres and railway stations.

1 Introduction

Surveillance of crowded scenarios and their analysis in terms of number of people is gaining importance, for example in shopping areas, public transport, etc. There are numerous approaches to the problem which are either vision or non-vision based.

Non-vision based approaches are characterised by the fact that no camera devices are used and no image processing is performed. Such techniques use for example infrared sensors where passing people interrupt a beam or pressure sensors which sense the footsteps of people. These approaches have the advantage of being insensitive to some crucial environmental factors such as light but imply the constraint that their application is rather limited to count people at a clearly defined line, e. g. to count the passengers boarding and alighting a vehicle via doors. There are a few commercial companies which offer such products.

[1] This work has been carried out within the K plus Competence Center ADVANCED COMPUTER VISION and was funded from the K plus program.

Vision based approaches vary significantly depending on the acquisition geometry. Methods using pure overhead cameras avoid the problem of occlusions but have a rather restricted view and are limited to count at narrow corridors and doors, similarly to non-vision based approaches.

Due to the increased affordability of standard video surveillance systems the application of these systems has been enhanced during the recent years. Public places become more and more CCTV controlled and the utilisation of the installed infrastructure for added value applications as people counting is a logical effort. Hence, there is the challenging task to develop people counting algorithms which can cope with oblique camera setup and occlusions of people. Due to this fact most of the commercial products rely on overhead cameras where people do typically not occlude each other.

Most vision-based people counting systems for oblique camera geometry rely on tracking. These methods have been studied intensively and have been proved to work reliably in scenes with limited people density [1], [2]. However, for scenarios where a high density of people leads to frequent occlusions, tracking is not reliable. In crowded scenes for examples at underground platforms, the tracking and counting performance depends strongly on the clothing and the movement of the people.

There are few approaches for non-tracking-based people counting systems which make use of neural network estimators [3], [4]. Methods related to the latter, use significant features extracted by basic image processing which are fed into trained neural networks in order to yield an estimate of the number of people in the viewing area. The accuracy of those systems depends strongly on the training set of the neural network and on the choice of the feature set.

This paper presents a people counting method which is based on motion blob analysis, shadow-elimination, and the temporal change of the number of people residing at a certain image location. The method was successfully tested in single camera scenarios for various sequences taken in halls of shopping centres and railway stations.

Beside the given overview of other approaches to people counting, this paper provides a detailed description of our people counting system (section 2) and the evaluation using sample video sequences (section 3).

2 System Description

Our approach aims at people counting under an oblique stationary single camera geometry and is based on the combination of motion detection and blob analysis rather than tracking. Omitting tracking provides the ability of faster processing because there is no need to compute hypothesis and/or features.

People are modelled by the height H and shoulder width W. For both parameters minimum (H_{min}, W_{min}), maximum (H_{max}, W_{max}) and typical (H_{typ}, W_{typ}) values are considered. Arms are neglected, since they cover only a small image area in contrast to torso and legs. We restrict our system to people who have their feet on the ground and who stand or walk in an upright manner. The scene is calibrated with a simple calibration scheme which provides the local scaling, i. e. pixel to object size ratio, for every pixel of the scene which refers to the ground plane.

The following description of all processing steps is illustrated in Figure 1.

1. **Frame acquisition**: It is assumed that frame capture is performed at regular time intervals which are typically at the range of 6 to 12 frames per second.

2. **Motion detection**: Intensity profile analysis [5] is used to classify every pixel x of a single frame of time step n as moving, stationary or background pixel. Standard adaptive background and threshold models are computed on RGB-colour images.

3. **Shadow elimination**: Since shadows can increase the blob size significantly, every stationary and moving pixel is checked whether it belongs to a shadow or not. Shadow detection is done pixel based using invariant colour properties which are obtained from the measurement of coloured object reflectance assuming a Gaussian colour model presented by Geusebroek [7]. Geusebroek showed that the first three coefficients $\hat{E}$, $\hat{E}_{\lambda}$ and $\hat{E}_{\lambda\lambda}$ of the second order development of the observed spectral energy distribution approximately can be linked to RGB-components according to Equation 1.

$$\begin{pmatrix} \hat{E} \\ \hat{E}_{\lambda} \\ \hat{E}_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.60 & 0.17 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{1}$$

Due to our target scenarios, we use colour invariants which were derived under the assumptions of equal energy, but uneven illumination of matte, dull surfaces. Such an irreducible set of fundamental colour invariants is determined by Equation 2 [7].

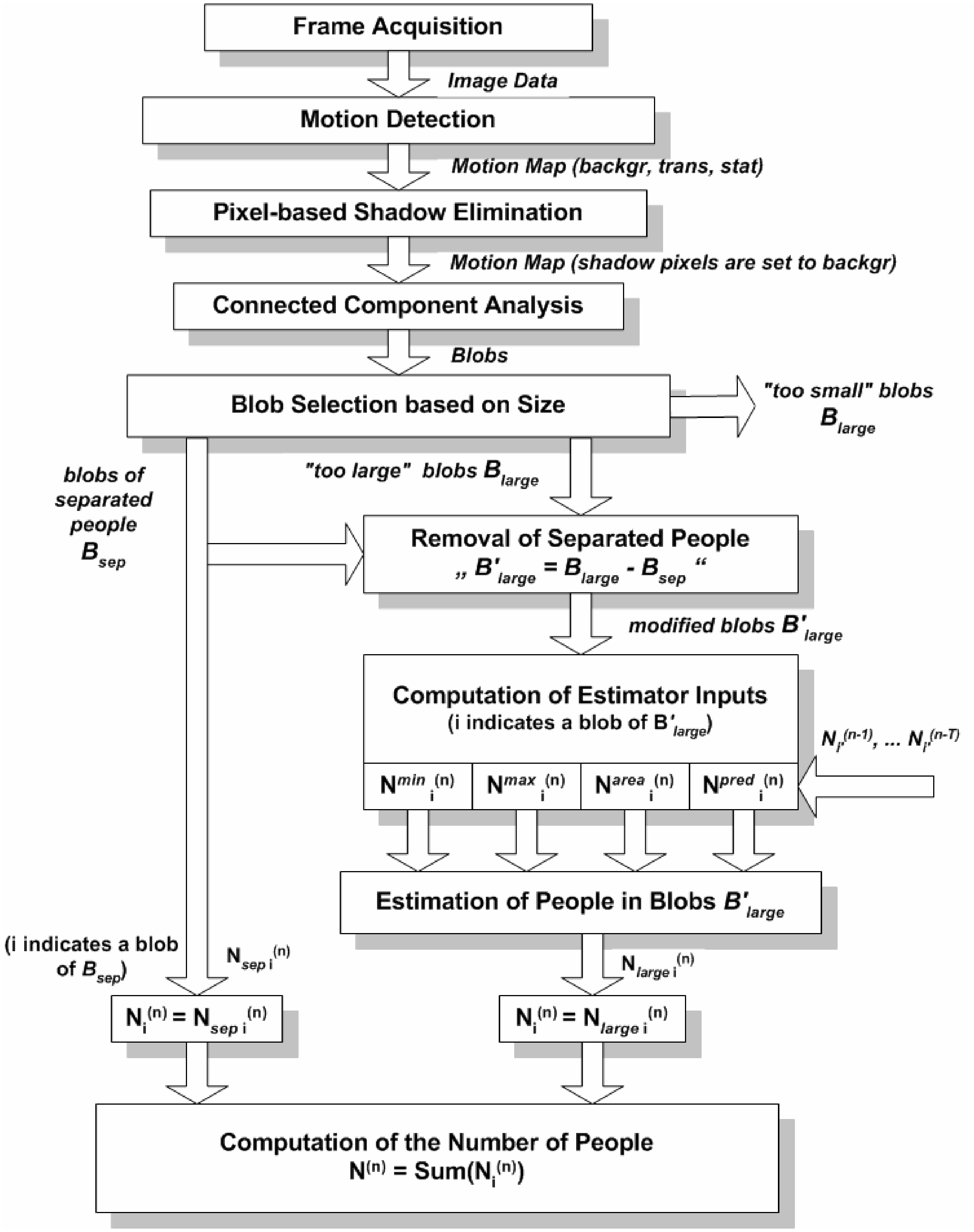
Frame Acquisition
Image Data
Motion Detection
Motion Map (backgr, trans, stat)
Pixel-based Shadow Elimination
Motion Map (shadow pixels are set to backgr)
Connected Component Analysis
Blobs
Blob Selection based on Size
"too small" blobs B_{large}
blobs of separated people B_{sep}
"too large" blobs B_{large}
Removal of Separated People
„ $B'_{large} = B_{large} - B_{sep}$ "
modified blobs B'_{large}
Computation of Estimator Inputs
(i indicates a blob of B'_{large})
$N_i^{(n-1)}$, ... $N_i^{(n-T)}$
$N^{min}{}_i^{(n)}$
$N^{max}{}_i^{(n)}$
$N^{area}{}_i^{(n)}$
$N^{pred}{}_i^{(n)}$
Estimation of People in Blobs B'_{large}
(i indicates a blob of B_{sep})
$N_{sep\,i}^{(n)}$
$N_{large\,i}^{(n)}$
$N_i^{(n)} = N_{sep\,i}^{(n)}$
$N_i^{(n)} = N_{large\,i}^{(n)}$
Computation of the Number of People
$N^{(n)} = Sum(N_i^{(n)})$

Figure 1: Principle of the estimation process

$$C_{\lambda^m x^n} = \frac{\partial^n}{\partial x^n}\left\{\frac{E_{\lambda^m}}{E}\right\} \quad \text{for } m \geq 1, n \geq 0 \tag{2}$$

To increase performance we selected a single invariant (e. g. m=2, n=0) and computed the values of the invariant for every background pixel and for every pixel of the actual frame, i. e. $C_{\lambda\lambda}{}^{background}(x)$ and $C_{\lambda\lambda}{}^{image}(x)$ respectively. For every shadow pixel *x* which does not belong to a person, the absolute difference $\left|C_{\lambda\lambda}{}^{image}(x) - C_{\lambda\lambda}{}^{background}(x)\right|$ is small. Thus a shadow can be detected through thresholding. Hence, we eliminate every pixel which was classified as moving or stationary if it is part of the shadow.

4. **Connected Component Analysis**: Subsequently, all non-shadow pixels which are classified as stationary or moving form blobs based on their connectedness.

5. **Blob filtering**: Due to our very simple people model, blob selection through object size is performed rapidly. All blobs obtained from step 4 are selected based on the local scaling of the image position of the base line and the size of the bounding box. Blobs which are too small to contain a standing person of the minimum height are removed from any further analysis. Blobs which are too large to contain a single person of maximum height are marked. Blobs which meet the people model perfectly are assumed to contain exactly one person.

6. **Estimation of the number of people**: Generally, blobs containing stationary and moving objects are processed in an identical manner. Assuming, $N^{(n)}$ denotes the estimate of the number of people in the scene at time step *n* and *i* indicates every single motion blob $b_i^{(n)}$. Then $N^{(n)}$ is computed by summing up the estimation values $N_i^{(n)}$ of all blobs. $N_i^{(n)}$ is computed based on two different kinds of inputs: (a) information from the corresponding blob $b_i^{(n)}$ and (b) information which is accumulated within the bounding box of $b_i^{(n)}$ from a sliding time window over the time steps *n-k* to *n-1* during the system run.

   a. Information taken from every single blob $b_i^{(n)}$ of the frame of the actual time step *n* is used to compute a frame-based estimate $N^{f}{}_{i}{}^{(n)}$ and comprises

      i. Minimum number of people $N^{min}{}_{i}{}^{(n)}$ which is present in blob $b_i^{(n)}$

      ii. Maximum number of people $N^{max}{}_{i}{}^{(n)}$ which is present in blob $b_i^{(n)}$

      iii. Ratio $N^{area}{}_{i}{}^{(n)}$ between the area of blob $b_i^{(n)}$ and the area of an average sized person which is approximated by the corresponding rectangle w_{typ}xh_{typ} at the blob position.

If $N^{area}{}_i{}^{(n)} \in [N^{min}{}_i{}^{(n)}, N^{max}{}_i{}^{(n)}]$ then $N^f{}_i{}^{(n)} = N^{area}{}_i{}^{(n)}$. Otherwise $N^f{}_i{}^{(n)}$ is equal to the closest value of the interval $[N^{min}{}_i{}^{(n)}, N^{max}{}_i{}^{(n)}]$.

b. In the case of temporarily occluded people, the blob area may shrink considerably and $N_i^{(n)}$ will be decreased. Thus, the number of people within the bounding box of a blob $b_i^{(n)}$ during the previous time steps n-k to n-1, ($k \in \mathrm{Z}_+$) is taken into account. This is done by defining a map $m^{(n)}$ having the same size as the input frame according to $m^{(n)}(x) = \sum_{\forall b_i^{(n)}} s_i^{(n)}(x)$. For every pixel x which is not part of a blob, $s_i^{(n)}(x)$ is set to 0. If pixel x is part of blob $b_i^{(n)}$, $s_i^{(n)}(x) = N_i^{(n)} \big/ A_i^{blob(n)}$. This setting ensures that the sum over the bounding box of any arbitrary rectangle region results in the number of people present in this region[2]. Naturally, the overall sum of $m^{(n)}(x)$ is equal to $N^{(n)}$. The map $m^{(n)}(x)$ is transformed to a so-called integral image [6] and stored in a memory buffer. This structure provides rapid access to the number of people which are present in the bounding box of blob $b_i^{(n)}$ during the preceding frames. Within the sliding window a trend is indicated and yields a prediction value by extrapolation $N^{pred}{}_i{}^{(n)}$.

Eventually, the estimates $N^f{}_i{}^{(n)}$ and $N^{pred}{}_i{}^{(n)}$ are merged. If $N^f{}_i{}^{(n)} \geq N^{pred}{}_i{}^{(n)}$ then $N_i{}^{(n)} = N^f{}_i{}^{(n)}$. If $N^f{}_i{}^{(n)} < N^{pred}{}_i{}^{(n)}$ then $N_i{}^{(n)} = N^{pred}{}_i{}^{(n)}$.

3 Experiments

We tested our method with video sequences captured in shopping malls and entrance halls of railway stations. We present the results of sample video sequences captured in a shopping mall (video 1) and the entrance hall (video 2) of the railway station Vienna West.

All images were processed at the size 360x288 pixels using RGB-colour information for motion detection and shadow elimination. Video 1 contains 454 frames and video 2 contains 1767 frames.

The people density in video 1 is low to medium, i. e. persons in the scene appear often separated and occlusions by other persons occur rarely. The viewing angle in the scene is about 30 degrees. The accuracy is computed numerically and shown in Figure 2. The system output is

[2] Note that the sum might be a fractional number. Assuming a scene where just one person is walking translationally across the scene. Then the sum over all pixels of the bounding box of the moving blob $b_i^{(n)}$ computed for an earlier time step will be less than 1, e. g. 0.9, depending on the blob shape and the walking speed.

identical to the real number of people for 75 percent of all frames. The accuracy of ±1 individual is provided for about 95 percent of the frames. Since all points of the evaluation plot are located close to the optimum line, the system output can be used as reliable estimate of the number of people in the scene.

In video 2 the range of the viewing angle from the camera to every point of the ground ranges from about 45 degrees at the bottom to 25 degrees at the upper part of the image. In contrast to the scene in video 1, occlusions occur frequently in the area near the escalator. People are often hidden. As the accuracy plot in Figure 2 shows, the system has the tendency to underestimate the number of people.

The evaluation plot shows that the system can still provide good qualitative output of the people count.

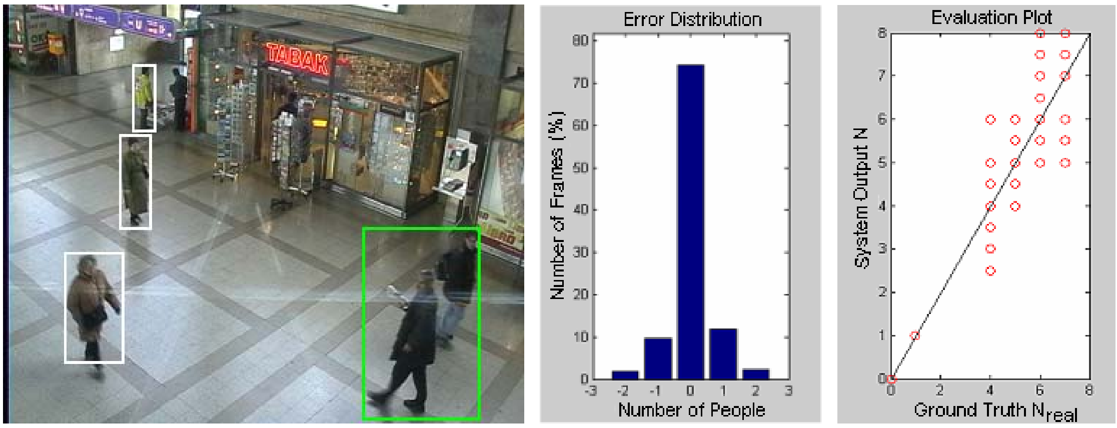

Figure 2: Blobs, error distribution and evaluation plot of the sample video ‘Shopping Mall’

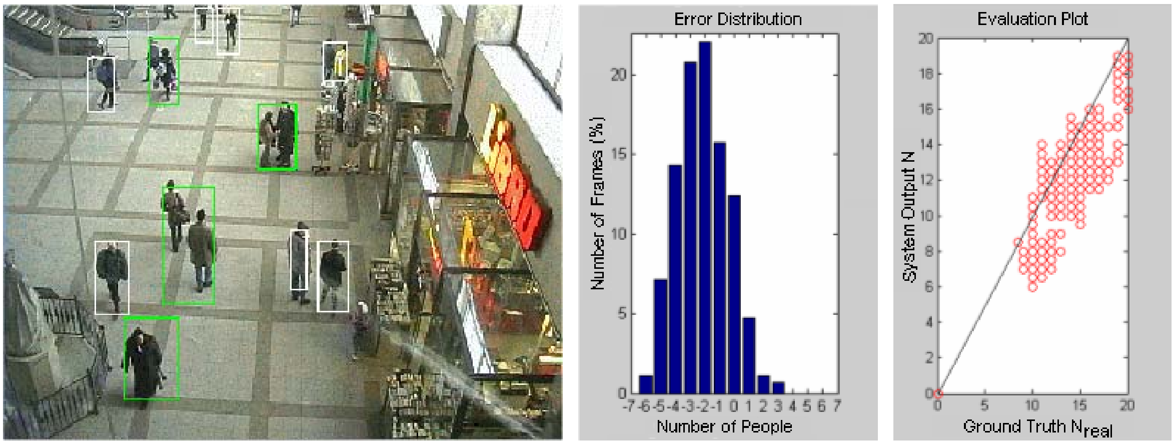

Figure 3: Blobs, error distribution and evaluation plot of the sample video ‘Entrance Hall’

4 Conclusion

We have presented an integral system for people counting which makes use of basic blob analysis and shadow elimination based on colour invariants. The method has been proven to deliver a reliable estimate of the number of people in a complex scene by performing tests for various sequences taken in halls of shopping centres and railway stations. In future work, we will focus on modelling the shape description of the blob in order to refine the estimate by shape segmentation for blobs possibly containing multiple individuals.

5 References

[1] M. Isard and A. Blake, "*Contour tracking by stochastic propagation of conditional density*", proceedings European Conf. on Computer Vision, 1996, Cambridge, UK, pp. 343-356.

[2] C. Tomasi and T. Kanade, "*Detection and tracking of point features*", Tech. Rept. CMU-CS-91132, Carnegie Mellon University School of Computer Science, 1991.

[3] C. S. Regazzoni, A. Tesei, *Distributed data fusion for real-time crowding estimation*, Signal Processing 53, pp. 47-63, 1996.

[4] S.-Y. Cho, T. W. S. Chow, C.-T. Leung, *A Neural-Based Crowd Estimation by Hybrid Global Learning Algorithm*, IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics, Vol. 29, No. 4, August 1999, pp. 535-541.

[5] Collins, Lipton, Kanade, Fujiyoshi, Duggins, Tsin, Tolliver, Enomoto, and Hasegawa, *A System for Video Surveillance and Monitoring: VSAM Final Report*, Technical Report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, May 2000.

[6] P. Viola, M. J. Jones, *Robust Real-time Object Detection*, CRL 2001/01, Cambridge Research Laboratory, Technical Report Series, February 2001.

[7] J.-M. Geusebroek, B. v. d. Boomgaard, A. W. M. Smeulders, H. Geerts, *Color Invariance*, IEEE Trans. Pattern. Anal. Machine Intell., November, 2001.