Detection of Cast Shadows in Surveillance Applications

Søren Gylling Erbou[1], Helge B.D. Sørensen[2], Bjarne Stage[3]

[1] Informatics and Mathematical Modelling, Technical University of Denmark, 2800 Kgs. Lyngby.
sge@imm.dtu.dk

[2] Ørsted•DTU, Technical University of Denmark, 2800 Kgs. Lyngby.
hbs@oersted.dtu.dk

[3] Danish Defence Research Establishment, Ryvangs Allé 1, 2100 Kbh. Ø.
bs@ddre.dk

Abstract

Cast shadows from moving objects reduce the general ability of robust classification and tracking of these objects, in outdoor surveillance applications. A method for segmentation of cast shadows is proposed, combining statistical features with a new similarity feature, derived from a physics-based model. The new method is compared to a reference method, and found to improve performance significantly, based on a test set of real-world examples.

1 Introduction

The introduction of digital video cameras, and recent advances in computer technology, make it possible to apply (semi-)automated processing steps to reduce the amount of data presented to an operator in a surveillance application. This way the amount of trivial tasks are reduced, and the operator can focus on a correct and immediate interpretation of the activities in a scene.

The Danish Defence Research Establishment (DDRE) is currently focusing part of it's research on implementing a system for automated video surveillance. The main objectives of the DDRE are to gain general knowledge in this area, and eventually implement an automated surveillance application that is capable of detecting, tracking and classifying moving objects of interest.

At this point the DDRE has carried out some initial studies in testing and implementing parts of the W^4-system [4] for automated video surveillance. The W^4-system effectively detects moving objects, tracks them through simple occlusions (blocking of the view), classifies them and performs an analysis of their behavior. One limitation of W^4 is that the tracking, classification and analysis of objects fails when large parts of the moving objects are actually cast shadows.

Distinguishing between cast shadows and self shadows is crucial for the further analysis of moving objects in a surveillance application. Self shadows occur when parts of an object are not illuminated directly, but only by diffuse lighting. Cast shadows occur when the shadow of an object is cast onto background areas, cf. figure 1. The latter are a

major concern in today's automated surveillance systems because they make shape-based classification of objects very difficult.

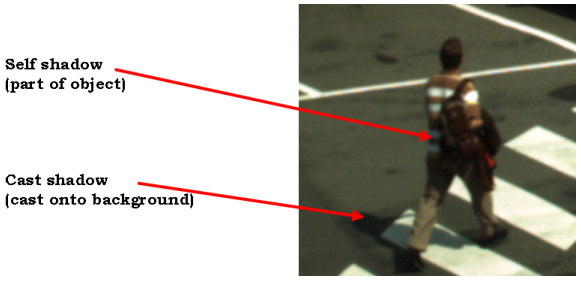

Figure 1: *Types of shadows. Self shadow is shadow on the object itself, a person in this case. Cast shadow is the shadow cast onto the background.*

In [9] Prati *et al.* give a comparative evaluation of the most important methods up until 2001. They conclude that the more general situations a system is designed to handle, the less assumptions should be made, and if the scene is noisy, a statistical approach is preferable to a deterministic model. In [5], Hsieh *et al.* focus on removing cast shadows from pedestrians using a statistical model combined with spatial assumptions. Only situations with pedestrians in an upright posture are handled and the cast shadows are assumed to touch their feet. Javed *et al.* [6] make no spatial assumptions of posture or composition prior to a statistical modelling of shadows, based on a correlation of the derivatives for regions of similar pixels.

In [7] Nadimi *et al.* apply a number of steps in a physics-based shadow detection algorithm. No spatial assumptions are made, but other assumptions makes it less suitable for some types of weather. Furthermore several threshold dependent parameters should be optimized. Finlayson *et al.* [3] use a physics-based approach to derive an illumination invariant, therefore shadow free, gray-scale image of an RGB image. From this image the original RGB image, without shadows, is derived. Finlayson's approach is aimed at shadow elimination in general in images obtained with a color calibrated standard digital camera [2],[3].

The rest of this paper consists of three sections, in section 2 existing methods for shadow handling are described in more detail, leading to a new combined method for segmentation of cast shadows. In section 3 the experimental results are presented, and section 4 is the conclusion.

2 Methods

The statistical approach suggested by Javed *et al.* [6] is implemented as a reference, because it makes no spatial assumptions and has the least number parameters to tune. The physics-based method suggested by Finlayson *et al.* is elegant, but not previously applied in surveillance applications. The new similarity feature proposed in this work is based on the ideas og Finlayson *et al.* Combining Javed's method with the new similarity feature, a new approach for handling cast shadows in surveillance applications is suggested.

2.1 Statistical Approach

Javed *et al.* [6] use a statistical approach for segmenting foreground pixels darker than a reference image (pixel-candidates) into cast shadow, self shadow and object pixels darker than the background. A K-means approximation of the EM-algorithm is used to perform unsupervised color segmentation of the pixel candidates. Each pixel candidate is assigned to one of the K existing Gaussian distributions if the Mahalanobis distance is below a certain threshold. If above this threshold a new distribution is added with it's mean equal to the pixel value. All distributions are assumed to have the same fixed covariance matrix $\Sigma = \sigma^2\mathbf{I}$, where σ^2 is a fixed variance of the colors and $\mathbf{I}$ is the identity matrix. After a pixel candidate is assigned to a distribution, the distribution mean is updated as follows:

$$\mu_{n+1} = \mu_n + \frac{1}{n+1}(x_{n+1} - \mu_n), \tag{1}$$

where x is the color vector of the pixel and μ_n is the mean of the Gaussian before the n+1th pixel is added to the distribution. Using a connected component analysis the spatially disconnected segments are divided into multiple connected segments. Smaller segments are then merged with the largest neighboring segment using region merging. Then each segment is assumed to belong to one of the three classes, cast shadow, self shadow or part of the object darker than the background image. To determine which of the segments are cast shadows, the textures of the segments are compared to the texture of the corresponding background regions. Because the illumination in a cast shadow can be very different from the background the gradient direction is used:

$$\theta = \arctan\frac{f_y}{f_x}, \tag{2}$$

where θ is the gradient direction and f_y and f_x are the vertical and horizontal derivatives respectively. If the correlation is more than a certain threshold, the region is considered a cast shadow. Otherwise it is either self shadow or dark part of the object. This method is considered as a state-of-the-art method in surveillance applications but still faces fundamental problems concerning some very context dependent parameters.

2.2 Physics-based Approach

The physics-based approach suggested by Finlayson *et al.* [3] derives an illumination invariant grayscale image from an RGB-image.

The color of a pixel in an image depends on the illumination, the surface reflection and the camera sensors. Denoting the spectral power distribution of the illumination $E(\lambda)$, the surface spectral reflection function $S(\lambda)$, and the camera sensor sensitivity functions $Q_k(\lambda)$ $(k = R, G, B)$, the RGB color ρ_k at a pixel can be described as an integral over the visible wavelengths λ:

$$\rho_k = \int E(\lambda)S(\lambda)Q_k(\lambda)d\lambda \quad , \quad k = \{R, G, B\}. \tag{3}$$

This description assumes no shading and distant lighting and camera placement. If the camera sensitivity functions $Q_k(\lambda)$ are furthermore assumed to be narrow-band, they can be modelled by Dirac delta functions $Q_k(\lambda) = q_k\delta(\lambda - \lambda_k)$, where q_k is the strength of the sensor. Substituting this into (3) reveals:

$$\rho_k = E(\lambda)S(\lambda)q_k \quad , \quad k = \{R, G, B\}. \tag{4}$$

Lighting is approximated using Planck's law:

$$E(\lambda, T) = Ic_1\lambda^{-5}\left(e^{\frac{c_2}{T\lambda}} - 1\right)^{-1}, \tag{5}$$

where I is the intensity of the incident light, T is the color temperature, and c_1 and c_2 are equal to $3.74183 \cdot 10^{-16} Wm^2$ and $1.4388 \cdot 10^{-2} Km$ respectively. Daylight is very near to the Planckian locus. The illumination temperature of the sun is in the range from $2500K$ to $10000K$ (red through white to blue). For the visible spectrum (400-700nm) the exponential term of (5) is somewhat larger than 1. This is Wien's approximation [6]:

$$E(\lambda, T) \simeq Ic_1\lambda^{-5}e^{-\frac{c_2}{T\lambda}}. \tag{6}$$

If the surface is Lambertian (perfectly diffuse reflection) shading can be modelled as the cosine of the angle between the incident light $\mathbf{a}$ and the surface normal $\mathbf{n}$. This reveals the following narrow-band sensor response equation:

$$\rho_k = (\mathbf{a} \cdot \mathbf{n})Ic_1\lambda^{-5}e^{-\frac{c_2}{T\lambda}}S(\lambda)q_k \quad , \quad k = \{R, G, B\}. \tag{7}$$

Defining band-ratio chromaticities r_k remove intensity and shading variables:

$$r_k = \frac{\rho_k}{\rho_G} \quad , \quad k = \{R, B\}. \tag{8}$$

Taking the natural logarithm (ln) of (8) isolates the temperature:

$$\begin{aligned} r'_k \equiv ln(r_k) &= ln(s_k/s_G) + (e_k - e_G)/T \quad , \quad k = \{R, B\}, && (9)\\ s_k &= \lambda^{-5}S(\lambda)q_k, && (10)\\ e_k &= -c_2/\lambda_k. && (11) \end{aligned}$$

For every pixel the vector $(r'_R, r'_B)^T$ is formed as a constant vector plus a vector $(e_R - e_G, e_B - e_G)^T$ times the inverse color temperature. As the color temperature changes, pixel values are constrained to a straight line in 2D log-chromaticity space, since (9) is the equation for a line. By projecting the 2D color into the direction orthogonal to the vector $(e_R - e_G, e_B - e_G)^T$, the pixel value only depends on the surface reflectance and not temperature hence illumination:

$$\begin{aligned} r'_R - \frac{e_R - e_G}{e_B - e_G}r'_B &= ln(s_R/s_G) - \frac{e_R - e_G}{e_B - e_G}ln(s_B/s_G),\\ &= f(s_R, s_G, s_B). && (12) \end{aligned}$$

Applying (12) to all pixels reveals the illumination invariant image $gs(x, y)$:

$$gs(x, y) = a_1 r'_R(x, y) + a_2 r'_B(x, y), \tag{13}$$

where the constant vector $a = (a_1, a_2)^T$ is orthogonal to $(e_R - e_G, e_B - e_G)^T$, determined by the camera sensitivity functions only (12)(11), and scaled to unit length:

$$\begin{aligned} a &= \frac{a'}{\|a'\|},\\ a' &= \begin{pmatrix} 1 \\ -\frac{e_R - e_G}{e_B - e_G} \end{pmatrix}. && (14) \end{aligned}$$

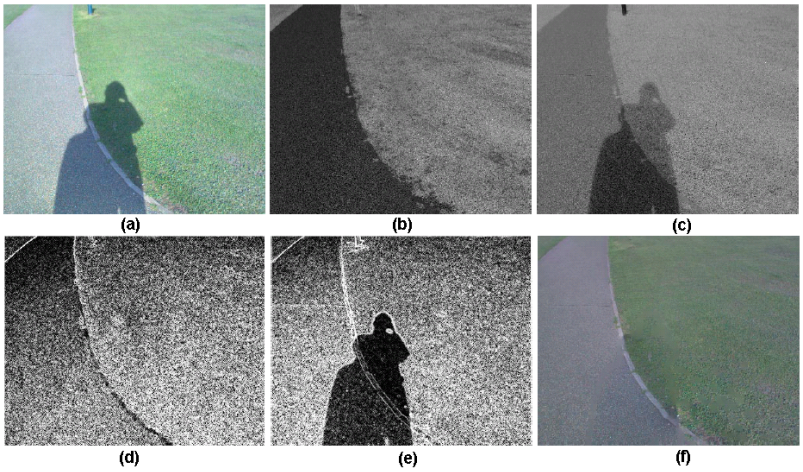

Figure 2: *Finlayson's approach to shadow removal [3]. (a): Original image. (b) Illumination invariant grayscale image. (c): Grayscale of original image. (d): Edge map for invariant image. (e): Edge map for non-invariant image. (f): Recovered shadow-free image.*

Figure 2(b) shows an example of an illumination invariant grayscale image, where edges due to shadows are not visible. Figure 2(a) and 2(c) show the original image, and the normal grayscale image.

If the sensor functions of the camera, and thereby λ_k of (11), are unknown, [2] and [3] outline a procedure for camera color calibration. The invariant direction is estimated by comparing a number of images taken during the day with changing illumination. Daylight is assumed to be Planckian with varying temperature. Each image contains different standard color patches from the Macbeth Color Chart.

The shadow edges are detected by comparing the gradient of each channel in the original log image, $\nabla\rho'(x,y)$, with the gradient of the illumination invariant image, $\nabla gs(x,y)$, cf. figure 2(d) and 2(e). The idea is that if the gradient in $\rho'(x,y)$ is high, while it is low in $gs(x,y)$, the edge is most likely to be a shadow edge. The following threshold function reveals a gradient image of the log response where gradients due to shadows are eliminated (set to zero):

$$S(\nabla\rho'(x,y),\nabla gs(x,y)) = \begin{cases} 0 & \text{if } \|\nabla\rho'(x,y)\| > \mathrm{t}_1 \\ & \text{and } \|\nabla gs(x,y)\| < \mathrm{t}_2 \\ \nabla\rho'(x,y) & \text{otherwise,} \end{cases} \tag{15}$$

where t_1 and t_2 are context dependent thresholds. By integrating S a log response image without shadows is recovered. This corresponds to solving the following Poisson equation:

$$\nabla^2 q'(x,y) = \nabla \cdot S(\nabla\rho'(x,y),\nabla gs(x,y)), \tag{16}$$

where ∇^2 is the Laplacian and q' is the log of the image without shadows. The gradient image of S equals the Laplacian of q' for each color band. Assuming Neumann boundary conditions ($\nabla q' = 0$ for boundary normals), q' can be solved uniquely up to an additive constant using the cosine transform [10]. When exponentiating q' to arrive at the shadow free image q the unknown constant becomes multiplicative. For the colors to appear "realistic" in each band, the mean of the top 5-percentile of pixels is mapped to maximum of the RGB image. In this way the unknown constants are fixed, and a shadow free image q is derived, cf. figure 2(f).

The major drawback of this method is reported to be defining the shadow edges. It turns out that using a robust edge detection algorithm (e.g. Canny or SUSAN [3]) and setting the thresholds are crucial factors. Furthermore a morphological opening is applied on the binary edge map to thicken the shadow edges and thereby improve the suppression of shadow gradients before the re-integration step.

Despite all of the assumptions and difficulties reported the method shows good results on the images shown in [2],[3]. It should be noted that the gradient images and thresholds are very context dependent. However, even when the method performs poorly it still attenuates the shadows. This is often the case for shadows with diffuse edges. Therefore the method is interesting in conjunction with surveillance tasks, where the artifacts introduced by the imperfect shadow edge detection and the re-integration are not crucial.

Due to assumptions in the model, and in the derivation of the shadow free RGB image, the method is far from perfect, but shadows are attenuated significantly. The method has not been applied in a surveillance application yet.

2.3 New Similarity Feature

It was found that the illumination invariant image is sensitive to the limited dynamic range in the video sequences of the camera used (8 bit) and to the spectral sensor functions of the camera not being delta functions. Because of this, determining edges due to shadows in a robust way becomes very difficult. Finlayson *et al.* also reports this to be the major drawback of the method [3].

Instead of only using the illumination-invariant image to determine edges due to shadows, other information should also be used. An important observation to make is that a foreground mask is available from the background model in a surveillance application. This can be used to eliminate artifacts from false shadow edges outside the foreground mask, and should be exploited in the detection of shadow edges.

A dilated version of the edges of the foreground mask is used to determine which gradients to suppress in the gradient image of the illumination invariant image, before reconstructing the "shadow-free" image. Figure 3(a) shows an image and a version of it, figure 3(b), that is reconstructed without suppressing any gradients. Therefore the two images are similar. Figure 3(c) shows the mask used for suppressing gradients, and figure 3(d) shows the corresponding reconstructed image.

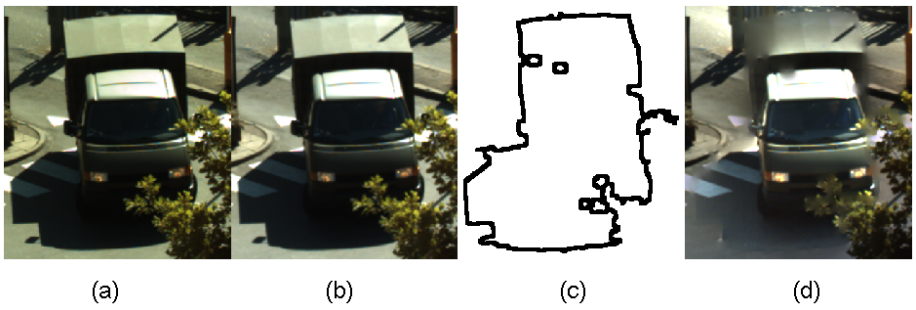

Figure 3: *Reconstruction of an image. (a): Original image. (b): Reconstructed image without suppressed gradients. (c): Suggested mask for suppressing gradients. (d): Reconstructed image with suppressed gradients.*

Both shadow and object gradients are suppressed, but figure 3(d) still clearly contains additional information that can be exploited in the segmentation of cast shadows.

The new similarity feature compares corresponding pixels of the reconstructed image and the background image, for every color segmented region:

$$CS = \frac{1}{\hat{\sigma}^2_{R,BG}(K-1)} \sum_{i=1}^{K} (R_i - BG_i)^2, \tag{17}$$

where CS is the similarity feature of a region, K is the number of pixels of the region times the three colorbands, R and BG are the intensity values of the i'th pixels in the reconstructed image and the background image, respectively. $\hat{\sigma}^2_{R,BG}$ is a variance normalization factor, which is the estimated variance between all pixels in a background image, BG, and all pixels in a reconstructed image, R, of a new frame containing no foreground objects.

Performing a variance normalization of CS makes it a relative measure of similarity that, ideally, only contains variation due to the region not being cast shadow, and not contains variation due to the experimental setup and the complex processing of the images. The estimate of the variance is based only on one sequence since it was difficult to obtain sequences, without foreground objects, that were static while an entire background model was estimated. It is therefore a rough estimate.

The CS measures a normalized mean value of squared differences between regions in the reconstructed foreground image, cf. figure 3(d), and corresponding regions in the background image. If the reconstructed image contains shadow regions along the border of the foreground mask, cf. figure 3(c), these shadow regions are attenuated in the reconstructed image, making them more similar to the background image. This is the key observation that the enhanced similarity feature, CS, is based on. Therefore a large value of CS corresponds to little similarity, which indicates that the region is part of the object. Small values of CS indicate high similarity, i.e. the region is then part of a cast shadow.

It is emphasized that CS only supplies useful information when the shadow edges are actually part of the edge of the foreground mask. In some cases it will not supply any additional information, e.g. when edges due to objects instead of shadows are suppressed. This will tend to smear neighboring background and object regions, for which reason it is suggested only to apply the CS in cases where the correlation threshold, described in 2.1, does not produce confident results. This corresponds to introducing a *reject class* for the correlation feature.

Figure 4 shows the suggested enhanced classification of color segmented regions. The

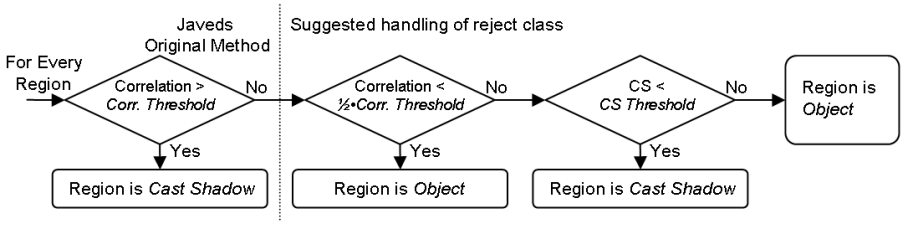

Figure 4: *Flowchart illustrating the enhanced classification of color regions. The enhanced similarity feature, (CS), classifies all regions that the correlation feature assign to a reject class (k·Corr. threshold < Correlation < Corr. threshold ⇒ reject class, 0<k<1).*

left part corresponds to the classification originally suggested by Javed, using a simple correlation threshold. The enhanced classification introduces a reject class if the correlation lies in an interval between k and 1 times the *Correlation threshold* introduced by Javed [6]. k should lie in the interval $[0; 1]$, and is empirically chosen to be 0.5 in this framework. If

the regions in the reject class have a CS larger than the *CS threshold* they are classified as object regions. Otherwise they are classified as cast shadow regions.

3 Data and Results

The camera used for data acquisition is a state-of-the-art industry digital video camera (SVS-204CFCL) with a resolution of 1024x768 pixels. The frame rate currently available is 20 fps., with a dynamic range of 8 bits, and with colors obtained through standard Bayer filtering. A typical scene for a surveillance application is chosen where the typical moving objects are vehicles, people and bicycles.

A kernel-based background model is used to segment foreground objects [1]. Only one frame of an object is used in the data set to avoid stochastic dependence between samples. 18 foreground objects are used in a manual optimization of model parameters and 72 foreground objects are used for validation and comparison of methods [1]. The main performance parameter used is the overall accuracy (AC), defined as the ratio of correctly classified pixels and the total number of pixels that are shadow candidates. True positives (TP) are defined as the proportion of correctly classified object pixels, and true negatives (TN) as the proportion of cast shadow pixels correctly classified.

A color calibration of the camera was performed to determine the the optimal angle of projection in the log-chromaticity space (39.4°). This angle corresponded well with the angle obtained from the spectral sensitivity functions of the camera.

As a reference Javed's statistical segmentation of shadow candidates is used. This is compared to the new method using the new similarity feature. In the optimization of model parameters of the two methods different values for the region merging criteria were found to be optimal. In the reference method more regions were merged into larger regions, making it hard to obtain a performance better than mediocre, because some regions contained both shadow- and object pixels and was classified as a whole. Due to the new similarity feature, the optimal merging parameter was found to produce more and therefore smaller regions to classify, making the method less susceptible to regions containing both types of pixels. Figure 5 compares the classification using the reference method and the enhanced method on the example of figure 3.

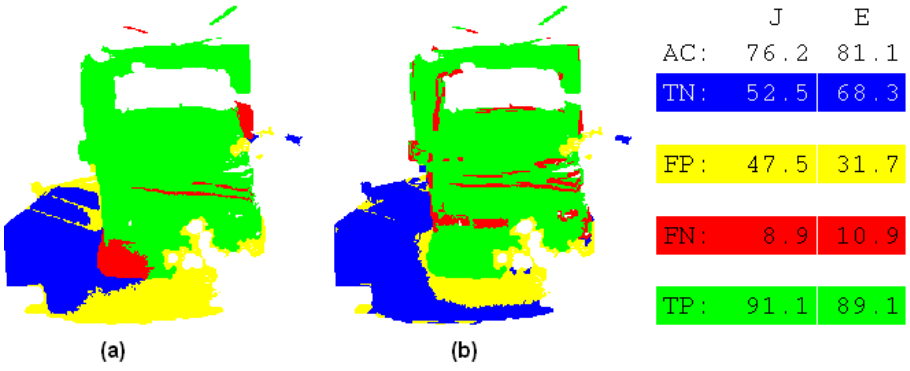

Figure 5: *Classification (%), AC=Accuracy, TN=True cast shadow pixels, FP=False object pixels, FN=False cast shadow pixels, TP=True object pixels. (a): Reference method (*J*). (b): Enhanced method (*E*) applying the new similarity feature.*

Table 1 shows the mean and std. of the absolute performance measures, based on the

test set, for the two methods.

Method	AC	TP	TN
Javed (J) - Mean (Std.) [%]	64.9 (17.8)	63.4 (30.0)	64.7 (33.4)
Enhanced (E) - Mean (Std.) [%]	69.2 (13.7)	69.7 (18.3)	66.0 (23.9)

Table 1: *Absolute performance of the two methods (*J *and* E*) based on the test set of 72 examples. Mean values and standard deviations are shown. AC=Accuracy, TP=True object pixels, TN=True cast shadow pixels.*

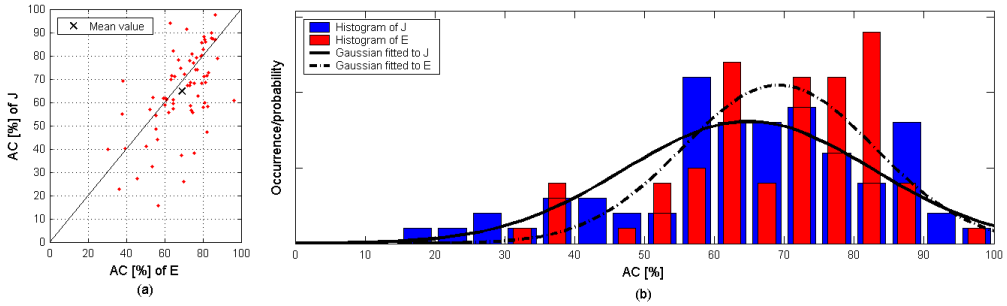

Figure 6: *Comparison of performance. (a): Accuracy of Javed's method (*J*), as a function of accuracy of enhanced method (*E*), based on the test set. (b): Histograms and fitted Gaussians of* J *and* E*, based on the test set.*

Figure 6 illustrates some of the results from table 1. There is a trend that examples with a higher AC in (E), are improved more than the examples with decreased AC, are decreased. This gives rise to the higher mean values, and indicates that fewer examples tend to have much better AC, while more examples tend to have slightly decreased AC.

A paired t-test is applied to determine if there is any significant difference, at a 5% level, in the mean values of the performance measures of the two methods. Table 2 shows the results.

Paired t-test, H_0: $\mu_E - \mu_J = 0$	AC	TP	TN
Difference in mean value ($E - J$)	1 (0.009)	1 (0.020)	0 (0.326)
Lower confidence bound [%]	1.31	1.28	−3.42

Table 2: *Statistical comparison of the absolute measures, AC=Accuracy, TP=True object pixels, TN=True cast shadow pixels. Row 1:* 0 *denote that the mean value cannot be rejected to be equal at a* 5% *level, and* 1 *that the difference of the means is significantly positive. p-values are shown in parentheses. Row 2: Lower confidence bounds for the differences in mean values for the absolute measures, at a* 95% *confidence level.*

0 denotes that the means cannot be rejected to be equal at a 5% level, and 1 that the difference of the means is significantly positive. The p-values are shown in parentheses. The conclusion to make from the test is that the new method (E) produces significantly better accuracy (AC) and is better at classifying object pixels correctly (TP), than the reference method J.

The lower confidence bounds of the difference in mean values, at a 95% confidence level, are shown in the second row of figure 2. They show that the difference in true mean values of the AC and TP for method E, are likely to be at least 1.3% above those of method J.

4 Conclusion

An enhanced method for shadow removal is suggested, based on a new similarity feature derived from a physics-based model. The new method significantly improves the mean accuracy at a 5% significance level, compared to the reference method.

The new similarity feature is only applied when the correlation feature of the reference method is uncertain, ensuring that the spatial assumption does not degrade performance, when compared to the reference method.

The final conclusion therefore is, that the suggested enhanced method for shadow removal, on average is better than the state-of-the-art method suggested by Javed. The enhanced method is also more robust, since it tends to improve the accuracy substantially, for examples where the reference method tends to fail completely.

Combining Javed's statistical-based method with some of the physics-based ideas of Finlayson, and a new similarity feature, therefore reveals a better and more robust algorithm for segmentation of cast shadows from moving objects.

The use of the illumination invariant image, as suggested by Finlayson, might be able to improve the performance even more, but requires a larger dynamic range than the 8 bits currently available with the present camera.

References

[1] Erbou, SG. "Segmentation of Cast Shadows from Moving Objects". M.Sc. Thesis, Ørsted•DTU, Technical University of Denmark, October 2004.

[2] Finlayson, GD., Hordley, SD. "Color Constancy at a Pixel". Journal of the Optical Society of America A, Vol.18 no. 2, pp.253-264, 2001.

[3] Finlayson, GD., Hordley, SD., Drew, MS. "Removing Shadows from Images". European Conference on Computer Vision (ECCV), part IV, p823-836, 2002.

[4] Haritaoglu, I., Harwood, D., Davis, LS. "W^4: Real-Time Surveillance of People and Their Activities". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.8, pp. 809-830, August 2000.

[5] Hsieh, JW., Hu, WF., Chang, CJ., Chen, YS. "Shadow elimination for effective moving object detection by Gaussian shadow modeling". Image and Vision Computing 21, pp.505-516, 2003.

[6] Javed, O., Shah, M. "Tracking And Object Classification For Automated Surveillance". European Conference on Computer Vision (ECCV), part IV, p343-357, 2002.

[7] Nadimi, S., Bhanu, B. "Moving Shadow Detection Using a Physics-based Approach". IEEE Proceedings of Pattern Recognition. Vol. 2, pp. 701-704, 2002.

[8] Park, S., Aggarwal, JK. "Segmentation and Tracking of Interacting Human Body Parts under Occlusion and Shadowing". IEEE Proceedings of the Workshop on Motion and Video Computing. pp. 105-111, 2002.

[9] Prati, A., Mikic, I., Trivedi, MM., Cucchiara, R. "Detecting Moving shadows: Algorithms and Evaluation". IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 25, No. 7, pp. 918-923, 2003.

[10] Press, WH., Teukolsky, SA., Vetterling, WT., Flannery, BP. "Numerical Recipes in C: The Art of Scientific Computing". Cambridge University Press. 2nd ed. 1992.