Detecting Perceptual Color Changes from Sequential Images for Scene Surveillance

Mika RAUTIAINEN†, Timo OJALA†, and Hannu KAUNISKANGAS†, Nonmembers

SUMMARY This paper proposes a methodology for detecting matte-surfaced objects on a scene using color information and spatial thresholding. First, a difference image is obtained via a pixel-wise comparison of the color content of a ‘clean’ reference image and a sample image. Then, spatial thresholding of the difference image is performed to extract any objects of interest, followed by morphological post-processing to remove pixel noise. We study the applicability of two alternate color spaces (HSV, CIE Lab) for computing the difference image. Similarly, we employ two spatial thresholding methods, which determine the global threshold from the local spatial properties of the difference image. We demonstrate the performance of the proposed approach in scene surveillance, where the objective is to monitor a shipping dock for the appearance of needless objects such as cardboard boxes. In order to analyze the robustness of the approach, the experiment includes three different types of scenes categorized as ‘easy,’ ‘moderate,’ and ‘difficult,’ based on properties such as heterogeneity of the background, existence of shadows and illumination changes, and reflectivity and chroma properties of the objects. The experimental results show that relatively good recognition accuracy is achieved on ‘easy’ and ‘moderate’ scenes, whereas ‘difficult’ scenes remain a challenge for future work.

key words: visual surveillance, image differencing, thresholding

1. Introduction

A scene surveillance system should give a clear indication of the changed objects in a given scene. In many cases intensity information is not sufficient to detect inanimate scene changes, which is demonstrated in Fig. 1. Particularly, traditional change detection methods have difficulties in recognizing static matte surfaced objects from a background of similar intensity. Previously proposed techniques for detecting changes between two consequent images have been based on pixelwise or local neighborhood image differencing, for the purpose of intruder detection [7]–[9] and vehicle tracking [6]. Many of the developed object detection systems incorporate motion information to precisely locate the actual objects of interest from noise [1]–[3]. Makarov’s [4] approach for scene surveillance adapts to slow changes in background by dynamically updating the background information. To overcome the problem of changing illumination conditions, methods such as Shading Model [11] and Circular Shift Moments [10] have been presented. These techniques work rather well when intensity differences are sufficiently large allowing detection of actual changes in a scene. Young et al. [5] compared some illumination compensation and frame differencing techniques. Rosin [12] presented multiple change detection methods based on thresholding difference images. Paschos and Valavanis [13] used color and texture information in automated surveillance system based on scene segmentation, demonstrating the advantages of xyY and HIS color spaces over RGB color space in finding chromaticity changes for the purpose of wetland monitoring.

In this study we extend two spatial thresholding methods originally developed for detecting intensity differences into detecting relevant color changes between a reference image and a sample image.

2. Methodology

The proposed approach comprises of following steps: reference image acquisition, sample image acquisition, calculation of the difference image, spatial thresholding of the difference image into a binary image, morphological post-processing of the binary image to remove pixel noise, and scene change detection from the post-processed binary image.

The acquisition of the reference image and the sample images is described shortly in experiments. The difference image is computed in two alternate color spaces, CIE Lab and HSV. The HSV color space is obtained from the RGB color space as described by...
The CIE Lab color space was defined by CIE in 1976 as a non-linear transformation from XYZ-tristimulus values for the purpose of presenting object surface colors. In this study, XYZ-values are calculated from RGB pixel values using following transform matrix\cite{16}:

\[
\begin{bmatrix}
X \\
Y \\
Z
\end{bmatrix} =
\begin{bmatrix}
0.431 & 0.342 & 0.178 \\
0.222 & 0.707 & 0.071 \\
0.020 & 0.130 & 0.939
\end{bmatrix}
\begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
\] (1)

Next, CIE Lab color coordinates are obtained from XYZ-values using:

\[
L = 116 \times f \left( \frac{Y}{Y_n} \right) - 16 \\
a = 500 \times \left( f \left( \frac{X}{X_n} \right) - f \left( \frac{Y}{Y_n} \right) \right) \\
b = 200 \times \left( f \left( \frac{Y}{Y_n} \right) - f \left( \frac{Z}{Z_n} \right) \right)
\] (2)

where

\[
X_n = 0.31006 \\
Y_n = 0.31616 \\
Z_n = 0.37378
\]

\[
f(t) = \begin{cases} 
  t^{1/3}, & \text{if } t > 0.008856 \\
  7.787t + \frac{16}{116}, & \text{otherwise}
\end{cases}
\] (3)

The \((X_n, Y_n, Z_n)\) values represent the CIE standard illuminant C coordinate points in the XYZ color space.

Given a reference image \(R\) and a sample image \(T\), the pixel-wise chromaticity difference in the CIE Lab color space is defined as:

\[
D_{ab} = \sqrt{(\Delta a)^2 + (\Delta b)^2}
\] (4)

where

\[
\Delta a = a_R - a_T \\
\Delta b = b_R - b_T
\]

\(a, b\) = color channels

The pixel-wise chromaticity difference in HSV color space is defined as:

\[
D_{HSV} = \sqrt{(\Delta S_x)^2 + (\Delta S_y)^2 + (\Delta V)^2}
\] (5)

where

\[
\Delta S_x = S_R \cos(H_R) - S_T \cos(H_T) \\
\Delta S_y = S_R \sin(H_R) - S_T \sin(H_T) \\
H = \text{hue} \\
S = \text{saturation}
\]

Luminosity/lightness can be incorporated as follows:

\[
D_{Lab} = \sqrt{(\Delta a)^2 + (\Delta b)^2 + (\Delta L)^2}
\] (6)

\[
D_{HSV} = \sqrt{(\Delta S_x)^2 + (\Delta S_y)^2 + (\Delta V)^2}
\] (7)

where

\[
\Delta L = L_R - L_T \\
\Delta V = V_R - V_T
\]

We study two alternate methods for spatial thresholding of the difference image, i.e. the global threshold is determined from the local spatial properties of the difference image. Rosin\cite{12} compared four approaches to threshold intensity images by modeling either signal or noise characteristics of the image using either spatial properties or intensity distribution. Rosin concluded that thresholds based on spatial properties of the signal or noise gave better results. We expect this to be the case in detecting color changes as well, as the commonly appearing matte surfaced objects tend to produce uniform changes in the difference image. Consequently, we employ two thresholding methods, of which the other is based on modeling the spatial properties of the signal and the other on modeling the spatial properties of the noise.

The first threshold is based on modeling the spatial properties of the signal using a stable number of regions. Even though the location, size, and the number of the regions of change in the difference image are unknown, we could expect that they remain roughly constant over a wide range of threshold values, while down at the noise level minor changes in threshold can considerably alter the number of regions. This suggests that if a range of threshold values providing a stable number of regions of change is determined, then these regions are unlikely to exist due noise, hence a value from this range will provide a useful threshold\cite{12,14}.

Rosin and Ellis found that instead of counting the exact number of regions for each candidate threshold value, the difference image’s Euler number could be used, providing almost identical results\cite{14}. The Euler number quantifies local connectivity of the difference image, thus characterizing the number of regions. The advantage of the Euler number over region counting is that it can be computed locally in a single raster scan over the image.

Euler number \(E\) is calculated using the critical points of the difference image:

\[
E = C_{MIN} + C_{MAX} - C_{SADDLE}
\] (8)

where

\[
C_{MIN} = \text{the number of minima} \\
in \text{the difference image} \\
C_{MAX} = \text{the number of maxima} \\
in \text{the difference image} \\
C_{SADDLE} = \text{the number of saddles} \\
in \text{the difference image}
\]

The critical points in the difference image are deter-
mined using three rules:

1. A pixel holds a critical point if the intensity is smaller than its local surroundings (minima).
2. A pixel holds a critical point if the intensity is larger than its local surroundings (maxima).
3. A pixel holds a critical point if the intensity is both maximum and minimum among its local orthogonal surroundings (saddle).

Figure 2 illustrates the physical meaning of critical points in a three-dimensional surface, where \( f(x, y) \) denotes the color difference at pixel \((x, y)\).

The global threshold is determined by examining the behaviour of the Euler number as a function of the threshold (see the Euler curve in Fig.3). A range of stable threshold values corresponds to a plateau in the Euler curve, suggesting that detecting the flat part of the curve would provide the best threshold. However, Rosin [12] found that the Euler number varies slowly within the stable range, due to the noisy, fragmented nature of images, and suggested that the Euler curve should be regarded as a decaying exponential instead.

At low threshold values the Euler number changes rapidly with the threshold, as there are many regions and holes caused by noise. At high threshold values there are few regions and the Euler number is stable, eventually converging to zero, when there are no regions remaining. Consequently, a suitable partition point between signal and noise is the “corner” of the Euler curve, which corresponds to the point, where the curve has maximum deviation from the straight line connecting the two end points of the curve. In “corner” detection the straight line starts from the highest positive peak that sides the curve’s lowest point (Euler number becomes negative, when there are more holes than connected regions in the difference image). The ‘corner’ indicates the point of stability in thresholding and corresponding difference value is selected as the global threshold (later referred to as the ‘Euler threshold’).

The second threshold is based on the assumption that the background noise in the difference image is white, hence its spatial distribution over the image will be random. For spatial data there are many measures of randomness, which are often based on the assumption that the observations (differences) follow a Poisson distribution [17]. The goal is to find a global threshold that minimizes the amount of spatially random noise, also known as shot noise, in the thresholded image. As a result of this, spatial ‘clumpiness’ is maximized producing uniform color changes in the thresholded image. This is accomplished by selecting the threshold (later referred to as the ‘Poisson threshold’), which maximizes relative variance:

![Fig. 2](image1.png)

**Fig. 2** Critical points in a topology of a two-dimensional difference image.

![Fig. 3](image2.png)

**Fig. 3** Euler number as a function of threshold and determination of the optimal threshold.

![Fig. 4](image3.png)

**Fig. 4** Six spatial neighbourhoods thresholded at three different values. Top row has the lowest relative variance, and bottom row the highest, indicating most suitable threshold.

![Fig. 5](image4.png)

**Fig. 5** An example of the color difference image with background noise (image enhanced).
where $s^2$ and $\bar{x}$ correspond to the sample variance and sample mean of the thresholded pixels’ distribution in local windows of $32 \times 32$ pixels in size. Although the test of relative variance is sensitive to the window size and point density, it works adequately well if $\bar{x}$ is sufficiently large.

Figure 4 demonstrates the use of Poisson threshold, showing six spatial neighbourhoods, each thresholded at three different values. Relative variance is highest in the lowest row indicating the most suitable threshold that maximizes uniform regions.

Figure 5 shows a histogram-equalized difference image to illustrate spatially random noise. Poisson threshold tries to minimize the amount of shot noise in the resulting binary image.

After thresholding, obtained binary image is subjected to morphological post-processing: the closing operator removes holes in objects and the opening operator removes individual noise pixels. Any objects in the post-processed binary image correspond to scene changes between the reference image and the sample image.

3. Experiment in Scene Surveillance

We demonstrate the performance of the proposed approach in scene surveillance, where the objective is to monitor a shipping dock of a premise for the appearance of needless objects such as cardboard boxes, paper and foam. This type of litter on a shipping dock can lead to a hazard such as fire, due to sabotage or accident.

In order to analyze the robustness of the approach, scenes from three different shipping docks categorized as ‘easy,’ ‘moderate,’ and ‘difficult’ were included. Reference images of each dock are shown in Fig.6.

The categorization is based on various background scene and appearing object properties such as the heterogeneity of the background, existence of shadows and illumination changes, and reflectivity and chroma properties of the objects. Whereas the ‘difficult’ dock had heterogeneous and complicated background with shadows, the ‘easy’ scene had uniform matte color surfaces, both in appearing objects and in the background.

The image data included 98 images in total: 28, 31, and 39 images for the ‘easy,’ ‘moderate,’ and ‘difficult’ scenes, respectively. For each scene the imagery included a reference image from a ‘clean’ dock and sample images with a varying quantity of objects of different color and shape on the dock. The imaging took place outdoors, hence the images were subject to illumination changes due to clouds. The images were taken using Olympus Camedia C-1400L camera with an image resolution of $640 \times 512$ pixels.

A prototype of the scene surveillance application was developed in Matlab-environment. The user can control various parameters such as the selection of the color space and the thresholding method. Given a reference and a sample image, the system determines the ‘filthiness’ of the scene, which corresponds to the percentage of the area of observed changes to the total area of the scene. If ‘filthiness’ exceeds a user-defined threshold, an alarm is generated to the user to clean up the dock for safety reasons. Four different combinations of a color space and a threshold were engaged: ab-chroma with Euler threshold (henceforth denoted as $ab + e$), Lab with Euler threshold ($Lab + e$), Lab with Poisson threshold ($Lab + p$) and HSV with Poisson threshold ($HSV + p$). Regarding difference measures, Eq. (4) was used with ($ab + e$), Eq. (6) with ($Lab + e$) and ($Lab + p$) and Eq. (7) with ($HSV + p$).

To quantitatively compare the performances of the four selected methods, we created a ground truth data by manually marking all object pixels in the sample images (Fig.7). The quality of the detection result is measured by comparing it to the ground truth image,
Fig. 7  The ground truth image (left) and detected object pixels (right) when no post-processing is done. In the right hand side image, pixels marked in black correspond to correctly detected object pixels, while those marked in grey correspond to background pixels that were erroneously detected as object pixels.

Fig. 8  Detection results for ‘easy’ scene without (left) and with post-processing (right).

pixel by pixel, and computing two statistics: the proportion of correctly detected object pixels of all object pixels \( (T_p) \) and the proportion of erroneously detected background pixels of all detected pixels \( (F_p) \). Let \( O \) denote the set of detected object pixels, \( GT \) the set of ground truth pixels, and \(|X|\) the amount of elements in set \( X \). Then \( T_p \) and \( F_p \) are defined as:

\[
T_p = \frac{|O \cap GT|}{|GT|}
\]

\[
F_p = \frac{|O \cap GT|}{|O|}
\]

In ideal case \( T_p \) equals to 1 and \( F_p \) to 0. In the case of successful detection of object pixels (\( T_p \) approaches 1), increasing value of \( F_p \) indicates an underestimated threshold.

Figure 8 shows individual detection results in \((F_p, T_p)\) coordinate space for all methods for the ‘easy’ scene, without and with the morphological post-processing. The average of the \((T_p, F_p)\) values of each method is also illustrated. As expected, morphological post-processing improves detection, which is indicated by smaller \( F_p \) and higher \( T_p \) values. Therefore, further results will include morphological post-processing.

To summarize the detection results over the sets of sample images (28, 31, and 39 images for the ‘easy,’ ‘moderate,’ and ‘difficult’ scenes, respectively), we compute the averages and standard deviations of \( T_p \) and \( F_p \) for each category (Table 1). Figure 9 illustrates the averages in a two-dimensional plot for each scene category. Detection results of individual samples of the methods with highest \( T_p \) are also included.

The results show that the best proportion of true positives is achieved with the HSV color space and Poisson threshold \((HSV + p)\), but proportion of false positives is clearly higher in comparison to other methods in the case of ‘moderate’ scene. Large standard de-
Table 1 Detection results.

<table>
<thead>
<tr>
<th>Method</th>
<th>'Easy'</th>
<th>'Moderate'</th>
<th>'Difficult'</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T_p$</td>
<td>$F_p$</td>
<td>$T_p$</td>
</tr>
<tr>
<td>$ab+e$</td>
<td>0.53</td>
<td>0.11</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>avg</td>
<td>std</td>
<td>avg</td>
</tr>
<tr>
<td></td>
<td>0.21</td>
<td>0.10</td>
<td>0.29</td>
</tr>
<tr>
<td>$Lab+e$</td>
<td>0.70</td>
<td>0.11</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>avg</td>
<td>std</td>
<td>avg</td>
</tr>
<tr>
<td></td>
<td>0.18</td>
<td>0.12</td>
<td>0.30</td>
</tr>
<tr>
<td>$Lab+p$</td>
<td>0.68</td>
<td>0.10</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>avg</td>
<td>std</td>
<td>avg</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>0.10</td>
<td>0.28</td>
</tr>
<tr>
<td>$HSV+p$</td>
<td>0.82</td>
<td>0.12</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>avg</td>
<td>std</td>
<td>avg</td>
</tr>
<tr>
<td></td>
<td>0.12</td>
<td>0.06</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Figure 9 Detection results for each scene category and $T_p$, $F_p$ and standard deviation of $F_p$ in ‘easy’ and ‘moderate’ scenes for all methods.

Fig. 9 Figure 9. Detection results for each scene category and $T_p$, $F_p$ and standard deviation of $F_p$ in ‘easy’, ‘moderate’ and ‘difficult’ scenes for all methods.

Table 1 Detection results.

The system succeeded in detecting objects that had equal luminance but different chroma compared to the background, which underlines the usefulness of color information. However, problems occurred in detecting changes that had also equal chrominance with the background, e.g., white objects on a grey background. This can be circumvented by using both chrominance and luminance information (Eqs. (6) and (7)), which unfortunately makes the system vulnerable to environmental variations of $F_p$ indicate that every method had severe difficulties with the ‘difficult’ scene. Considerably better proportion of true positives for the ‘difficult’ scene with $HSV+p$ in comparison to other methods can be explained with the tendency of $HSV+p$ to underestimate the threshold.

Figure 9 also shows $T_p$, $F_p$ and the standard deviation of $F_p$ in ‘easy’ and ‘moderate’ scenes (‘difficult’ is excluded due to bad performance). From the diagram it is clearly visible that the high $T_p$ value for $HSV+p$ and $Lab+p$ in ‘moderate’ scenes comes with high $F_p$ and high standard deviation of $F_p$. This indicates that these methods are less reliable than $Lab+e$ in detecting the actual amount of color change. $T_p/F_p$ scattering in ‘moderate’ scene becomes evident from Fig. 9 as well.

The system succeeded in detecting objects that had equal luminance but different chroma compared to the background, which underlines the usefulness of color information. However, problems occurred in detecting changes that had also equal chrominance with the background, e.g., white objects on a grey background. This can be circumvented by using both chrominance and luminance information (Eqs. (6) and (7)), which unfortunately makes the system vulnerable to environmental
changes such as shadows and illumination.

Few interesting remarks can be made of the relative performance of color spaces. HSV color space has larger noise variation in its hue channel than the CIE Lab’s channels a and b. Consequently, CIE Lab handles color changes better at lower levels of object luminance and is less vulnerable to window reflections. Additionally, CIE Lab produces more uniform surfaces, which facilitates more efficient detection. These properties are illustrated in Fig. 10.

4. Conclusions and Future Work

A prototype for color change based scene surveillance system was developed in this study. The system was successful in detecting color changes in ‘easy’ and ‘moderate’ scenes, for which the best proportion of true positive pixels was achieved in HSV color space using thresholding based on assuming that the spatial noise follows a Poisson distribution. However, this combination was also more prone to false detections. For ‘easy’ scenes $HSV+p$ was the best method, but $Lab+e$ proved to be most stable when considering both ‘easy’ and ‘moderate’ scenes. The system had difficulties in detecting changes in the ‘difficult’ scene, due to the rather complicated structure of the background. Performance for difficult scenes could be improved by including a region of interest (ROI) selection, which would reduce the observed scene’s background complexity. Difficulties caused by the strong perspective structure of the scene can be bypassed by positioning the camera perpendicular to the observed scene and by avoiding large local variances in scene depths. Also, objects of color similar to the background caused problems, which can be addressed by using luminance information in addition to color. However, this makes the system vulnerable to illumination changes and shades. Future work includes a more thorough evaluation of the system’s robustness against changes in daytimes and lighting conditions.

Acknowledgments

The financial support provided by the National Technology Agency of Finland and the Academy of Finland is gratefully acknowledged.

Note

The images and the ground truth data used in this study can be downloaded from http://www.mediateam.oulu.fi.

References


Mika Rautiainen received his M.Sc. degree (Information Engineering) from the University of Oulu, Finland, in 2001. He is currently studying towards his Dr.Tech degree in the Department of Electrical Engineering of the University of Oulu. His research interests include image and video understanding, content-based image and video retrieval and processing.

Timo Ojala received his M.Sc. (Electrical Engineering) and Dr.Tech. degrees from the Department of Electrical Engineering, University of Oulu, Finland, in 1992 and 1997, respectively. His research interests include pattern recognition, content-based multimedia retrieval, and digital image and video processing.

Hannu Kauniskangas received his M.Sc. (Electrical Engineering) and Dr. Tech. degrees in 1993 and 1999, respectively from the University of Oulu, Finland. His interests cover content-based multimedia retrieval, digital image and multimedia processing.