EVALUATION OF SHADOW CLASSIFICATION TECHNIQUES FOR OBJECT DETECTION AND TRACKING

John-paul R. Renno, James Orwell, Graeme A. Jones

Digital Imaging Research Centre [DIRC], Kingston University
Penrhyn Road, Kingston-Upon-Thames, KT1 2EE, UK.
{ j.renno, j.orwell, g.jones }@kingston.ac.uk

ABSTRACT

In a football stadium environment with multiple overhead floodlights, many protruding shadows can be observed originating from each of the targets. To successfully track individual targets, it is essential to achieve an accurate representation of the foreground. Many of the existing techniques are sensitive to shadows, falsely classifying shadows as foreground. This work presents four different techniques associated with shadow classification. Three of the classifier's originate from the review material whilst the fourth is a novel application of a real-time implementation of the *k-nearest neighbour* algorithm to shadow identification. To assess the performance for each of the classifiers four quantitative evaluation metrics are proposed. Using each of the evaluation metrics, we will discuss the performance of each classifier's segmentation results as well as assess their impact on the tracking performances.

1. INTRODUCTION

Detection of moving objects is essential for automatic monitoring of human activity. A common method for extracting the moving or or *foreground* regions from a video sequence is known as *background subtraction* [1, 2, 3]. This technique subtracts the incoming video frames from a reference image acquired during a period of inactivity and optionally updated over time. The resulting pixel or region differences are classified to detect the presence of moving objects in the scene.

Shadows cause problems for moving target detection and tracking. The appearance of neighbouring background is changed, to the extent that it can be falsely classified as foreground. Thus, measurements of moving objects are less reliable: this may affect the performance of object segmentation, classification, and estimation of position. These problems increase when there are many point light sources, *e.g.* a floodlit stadium. Each light source produces a distinct shadow formation at the base of player (see Figure 1). The underlying motivation of this work is the identification and removal of these shadows, to improve player tracking performance.

In the next section we review relevant previous work. In Section 3 we introduce the existing vision system and describe the various shadow identification techniques. In Section 4 the methods for evaluating these techniques are discussed. In Section 5 the results are presented and discussed. Section 6 provides a brief conclusion of the paper.

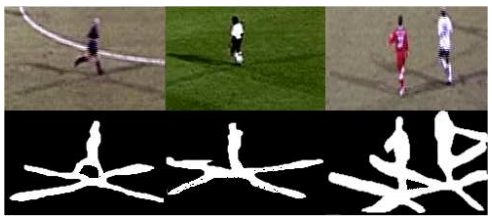

Fig. 1. Demonstration of the impact of multiple object shadows upon target segmentation.

2. REVIEW

Several authors have proposed methods to identify shadows in outdoor environments. Cucchiara *et al* [4] classify pixels into foreground/background using the HSV (Hue, Saturation and Value) colour space, since in this space chromaticity and luminosity components can be easily decoupled. This decoupling exploits the assumption that an area cast into shadow often results in a significant decrease in intensity whilst maintaining a similar chromaticity [1]. Thus, their classification criteria for shadows are: 1) that the *hue* and *saturation* components of the surface's colour should not change significantly and 2) that the *value* component should decrease. One unsolved problem is the specification of the procedure for selecting the appropriate classification thresholds.

McKenna *et al* [1] use similar assumptions, and define a background model with Gaussian distributions for each of

This work forms part of the INMOVE project, supported by the European Commission IST 2001-37422.

the pixel channel chromaticity values. This enables confidence measures to be generated after the background subtraction process, based on the probabilities of a particular pixel value belonging to each distribution in the model. If a pixel vale is classified as foreground using the intensity distribution, and background using the chromaticity model, then, overall, it is classified as shadow. In addition, a third classification method is designed to distinguish between shadows and darker objects that are of a similar colour to the background, using gradient and texture information as the discriminant. One difficulty is that the edge of a shadow will manifest a gradient, just as the edge of a dark object does. To overcome this, both background models are recursively updated: however, they are susceptible to sudden environmental changes; and, in the case of the third classification method, computationally expensive.

Horprasert *et al* [3] separate chromaticity from brightness by defining their own colour model. During a period of scene inactivity, a statistically generated 4-tuple background model is learned. The model components comprise: a pixel RGB mean and variance; and chromaticity and brightness distortion components. Then, a further period of statistical learning is required, to estimate appropriate thresholds for the foreground, background and shadow classes. That entails the construction of normalised histograms for chromaticity and brightness, then choosing the threshold to obtain an assumed 'detection rate'. A significant advantage of this technique is the automatic determination of threshold values. The disadvantages are that the background model is not adaptive, the detection rate needs to be known, and it is computationally expensive

3. PROPOSED VISION SYSTEM

It is desirable for a foreground detection process to be robust both to environmental changes *and* shadows. The proposed method satisfies this requirement in two stages, described below. The first stage is foreground detection: objects *and* their shadows are detected against an adaptive background model. The second stage classifies these foreground regions into object and shadow. Four different shadow classifiers are implemented and evaluated. An overview of the system can be seen in figure 2.

3.1. Stage 1: Foreground Detection

For reliable foreground detection in a changing environment, a background model must adapt to these changes. Examples of common changes include the amount of direct sunlight, wind-blown trees and periodically rotating advertising boards. Given this requirement, an adaptive - multiple background subtraction technique based upon the work of Stauffer and Grimson [2] was used. Their technique models each pixel's RGB value with a mixture of Gaussians. Each of the model Gaussians has a foreground or background classification assigned to it depending upon its persistence and variance. Foreground classification can then be performed by determining a new pixels membership to each of the Gaussian models in the mixture.

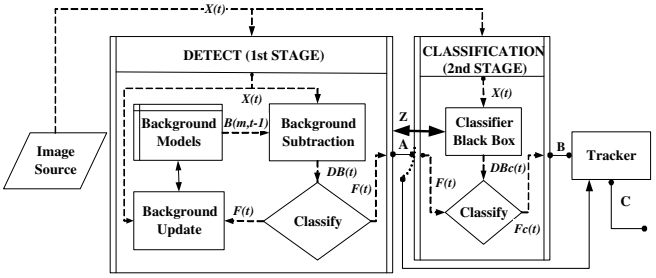

Fig. 2. Stages 1 and 2 of the vision system architecture.

3.2. Stage 2: Shadow Classification

In this section we describe each of the four shadow classifiers. Three of these are implementations of the techniques discussed in Section 2. The fourth is a supervised *k-nearest neighbour* technique. Each of the classifiers performs shadow analysis for pixels previously classed as foreground during stage 1 of the detection process.

3.2.1. Classifier A : HSV colour space conversion

This classifier is based upon the technique for classifying shadows proposed by Cucchiara et al [4]. We make use of their shadow classification and apply it to the pixels detected in the foreground segmentation. This lighter classification reduces the computational time required to process each frame. For each pixel flagged in the foreground mask, the corresponding background and foreground pixels are transformed into the HSV colour space. Each of the pixels is then classified as a shadow if the following condition evaluates as true.

$$\alpha \leq \frac{F_{(x,y)}.V}{B_{(x,y)}.V} \leq \beta \wedge \left|F_{(x,y)}.H - B_{(x,y)}.H\right| \leq \tau_h \wedge \left|F_{(x,y)}.S - B_{(x,y)}.S\right| \leq \tau_s$$

The selection of the appropriate values for the classification thresholds α, β, τ_h, & τ_s are set empirically for a specific video sequence.

3.2.2. Classifier B:Normalised RGB (chromaticity) model

This classifier is based upon the technique for classifying shadows proposed by McKenna *et al* [1]. We make use of their chromaticity model to separately generate a second

background that is representative of the pixels chromaticity. Since each pixel in the background is modelled by a Gaussian distribution, the classification criteria is a statistical probability. The current image is converted into the normalised RGB colour space and then background subtraction is performed on the pixels that have been previously classified as foreground. Any pixel not previously classified as foreground has its background model updated. A foreground pixel is classified as a shadow if both its normalised Red and Green pixel differences are within three standard deviations of their background model mean values.

3.2.3. Classifier C : Brightness & Chromaticity Distortion

This classifier (*BCD model*) is based upon the technique for classifying shadows proposed by Horprasert *et al* [3]. It is a implementation of the published technique, adapted to comply with the architecture presented in Figure 1. Thus, only shadow classifications are used, so that background subtraction is only appled to those pixels previously detected as foreground. After this background subtraction the result is decomposed into the brightness ($\widehat{\alpha}$) and chromaticity ($\widehat{CD}$) distortions and subsequently classified as a shadow if the following condition evaluates as true.

$$\widehat{CD} < \tau_{CD} \ \wedge \ \tau_{\alpha_{lo}} < \widehat{\alpha} < 0$$

The classification threshold τ_{CD} is determined by the algorithm during the statistical learning process. $\tau_{\alpha_{lo}}$ is specified by the authors.

3.2.4. Classifier D : k-nearest neighbour

This classifier is based upon the *k-nearest neighbour (k-nn)* algorithm and was developed using the concept of a fixed search space discussed by Nene and Nayer [5]. Fixed search spaces made it possible to implement the knn algorithm for real-time performance; hence, shadow identification is a novel application of the *k-nn* algorithm. Initialisation of the classifier is achieved by providing supervised training data that is comprised of RGB values which represent the foreground and shadow classes. Classification is performed upon any pixel previously classified as foreground: a pixel's class is defined as the most commonly occurring class from the k elements of the training set that are nearest in RGB space.

4. EVALUATION

There are several possible evaluation methodologies. We concentrate on evaluation methods that presuppose the limited availability of ground-truth data. This enables quantitative performance measures and therefore direct comparisons between the proposed methods; also, differences between real and ground-truth segmentations provide interesting insights into algorithm performance. In this section we briefly describe the ground-truth data format and the quantitative evaluation metrics that use it.

4.1. The Ground Truth

The ground truth is generated manually for every target within each frame of the video sequence. The characteristics maintained in the ground-truth are the objects bounding-box position and id. The bounding box represents the image plane coordinates that encapsulate the targets. Examples of the ground truth can be seen in figure 3.

Fig. 3. Target postures with the ground-truth overlayed.

4.2. The Metrics

We use four evaluation metrics, each directly comparing the segmentation result with the ground truth data. The first two metrics determine how much of the segmentation is correctly and incorrectly classified. They are formally known as the detection rate (Dr) and the false positive rate (FPr) [6] and defined as

$$\mathbf{Dr}_{(t)} = \frac{N^{tp}_{(t)}}{N^{tp}_{(t)} + N^{fn}_{(t)}} \quad , \quad \mathbf{FPr}_{(t)} = \frac{N^{fp}_{(t)}}{N^{fp}_{(t)} + N^{tn}_{(t)}}$$

where N^{tp} and N^{tn} are the numbers of pixels in the segmentation supported by the ground-truth, whilst N^{fp} and N^{fn} are the pixels in the segmentation not supported by the ground-truth. The third metric uses N^{tp} and N^{fp} to compute the signal to noise *SNR* ratio of the segmentation.

$$\mathbf{SNR}_{(t)} = 20 \log_{10}\left(N^{tp}_{(t)} / N^{fp}_{(t)}\right)$$

The fourth and final evaluation metric gives a insight into the performance of the vision system by evaluating the results obtained from tracking. We define the tracker error ($TErr$) as

$$\mathbf{TErr}_{(t)} = \frac{1}{N_G} \sum_{\forall g \in G_{(t)}} \min_i \|\mathbf{g} - \mathbf{r}_i\|$$

where $G = \{g_1, \ldots, g_{(N_G)}\}$ and $r_i \in R = \{r_1, \ldots, r_{(N_R)}\}$. The sets represent the ground-truth and tracker positions that consist of N_G ground truth boxes and N_R tracked targets. The tracker used to evaluate the error is a single view Kalman tracker of the object bounding box and centroid [7].

5. RESULTS

In this section the the four performance metrics defined in Section 4 are used to evaluate each of the four second stage classifiers discussed in Section 3.2. For reference, the first stage output (*i.e.*, foreground detection with no shadow suppression) is also evaluated.

The video data used for this evaluation is a single sequence of 1000 frames captured during a floodlit U.K. Premiership game on the 12th January 2002. Ground truth data of all players was recorded by hand. In Fig. 4, four graphs are plotted: one for each evaluation metric. For the True Positive (TP) and False Positive (FP) graphs, the results are plotted w.r.t to the stage 1 detection parameter : the sensitivity. From inspection of these graphs, that parameter is then set to 3.4, and used for the SNR and TErr metrics (shown in the lower part of Fig. 4).

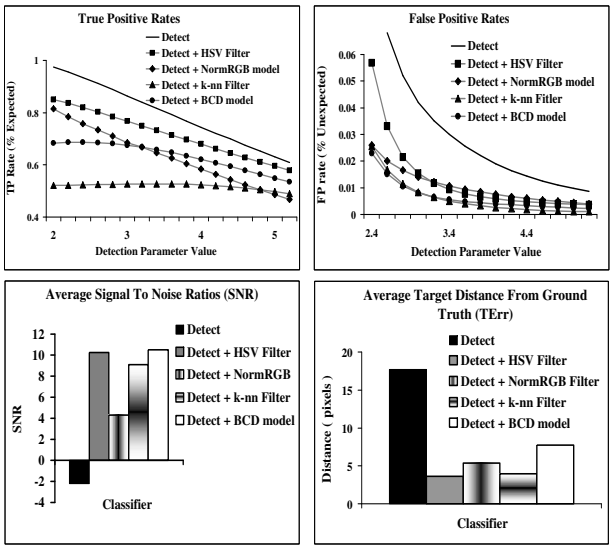

Fig. 4. Evaluation results :- **top-left**:Detection rate, **top-right**:False Positive rate, **bottom-left**:Signal to noise ratio and **bottom-right**:Tracking error.

Looking first at the TP and FP results, it is clear that each of the second stage shadow classifiers reduce the number of true (i.e., player) and false (i.e. shadow and other noise) pixels classified as foreground, compared with the output of the first stage detection. It is also clear that the detection sensitivity parameter trades-off performance between the True and False positive performance. One interesting difference between classifiers is the TP rate. Because the ground truth bounding box area is not the true representation of the ideal segmentation for a person target, many background pixels located within the ground truth bounding box may be incorrectly classified as foreground. However the graph shows that the *k-nn* and *BCD* classifiers produce a near uniform response to the changes in the sensitivity, indicating a robustness to the inaccuracies of the ground-truth. The FP graph shows that each of the second stage classifiers significantly reduce the number of falsely detected foreground pixels.

The SNR is the relationship between the average FP and TP rates over the entire video sequence. The SNR shows that all of the second stage classifiers achieve a significantly greater SNR, compared to that achieved by the first stage detection. This improved SNR is caused by the significant reduction in the FP rate, compared to that of the TP rate.

The TErr represents the average distances between the tracked targets and the nearest ground truth bounding box base position. The graph of the TErr shows that each of the second stage classifiers, out-perform the tracking performance of that achieved by the first stage detection. This final graph illustrates the better performing classifiers: the *k-nn* and the *HSV colour model* classifiers.

6. CONCLUSIONS

In this paper four shadow classification methods were demonstrated. Each was evaluated against a complex foreground segmentation algorithm using four evaluation metrics. It was shown that each of the second stage classifiers produced significant improvements in foreground segmentation and tracking performance. Future work includes classification of shadows caused by natural sunlight.

7. REFERENCES

[1] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Computer Vision and Image Understanding*, vol. 80, no. 1, pp. 42–56, October 2000.

[2] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking.," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Fort Collins, Colorado, June 23-25 1999, vol. 2, pp. 246–252.

[3] T. Horprasert, D. Harwood, and L.S. Davies, "A robust background subtraction and shadow detection," in *Asian Conference on Computer Vision (ACCV'2000)*, Taipei, Taiwan, January 8-11 2000.

[4] R. Cucchiara, C. C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *PAMI*, vol. 25, no. 10, pp. 1337–1342, October 2003.

[5] S. Nene and S. Nayar, "A simple algorithm for nearest neighbor search in high dimensions," *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI'97*, vol. 19, 1997.

[6] T. Ellis, "Performance metrics and methods for tracking in surveillance," in *Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS2002)*, Copenhagen, Denmark, June 2002, pp. 26–31.

[7] M. Xu and T. Ellis, "Partial observation vs. blind track-ing through occlusion," in *British Machine Vision Conference 2002 , BMVC'02*, Kobe,Japan, 2002, pp. 777–786.