Shadow Classification and Evaluation for Soccer Player Detection

J. Renno, J. Orwell, D.Thirde, G.A. Jones
Digital Imaging Research Centre, Kingston University
{j.renno, j.orwell, d.thirde, 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. Unfortunately, many of the existing techniques are sensitive to shadows, falsely classifying them as foreground. In this work an unsupervised learning procedure that determines the RGB colour distributions of the *foreground* and *shadow* classes of feature data is proposed. A novel skelatonisation and spatial filtering process is developed for identifying components in the foreground segmentation that are *most-likely* to belong to each class of feature. A pixel classification mechanism is obtained at by approximating both classes of feature data by N Gaussian parametric models. To assess our technique's performance and reliability, a comparison is made with other published works.

1 Introduction

Detection of moving objects is essential for automatic monitoring of human activity. A common method for extracting the moving or *foreground* regions from a video sequence is known as *background subtraction* [6, 8, 4]. 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 usually classified as foreground or background by using a *statistical* or *deterministic* approach 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 automatic identification and removal of these shadows to improve player tracking performance.

Several authors have proposed techniques for identifying shadows in outdoor environments. Cucchiara *et al* [1] detect shadows by: 1) transforming the feature's dimensions into the HSV (Hue, Saturation and Value) colour space and 2) classify the feature as shadow if: a significant difference in brightness is observed with little variation in colour.

BMVC 2004 doi:10.5244/C.18.86

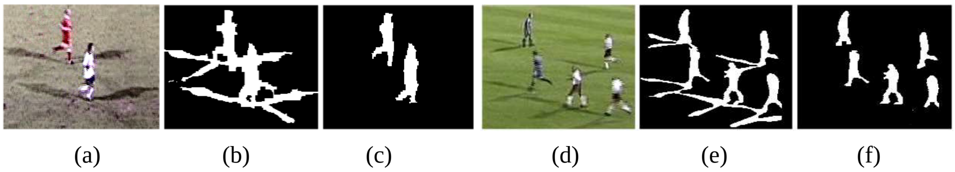


Figure 1: (a) & (d) illustrate the shadows caused by flood-lights; (b) & (e) illustrate the segmentation of the foreground with no shadow suppression, (c) & (f) represent the desired segmentation

A similar approach is adopted by McKenna *et al* [6] for shadow detection. The feature's membership to the models is used to determine the presence of shadows. Horprasert *et al* [4] develop their own computational colour model for feature analysis. Their feature space is used to classify a feature into one of four possible classes. A feature is classified as a shadow in much the same way as that proposed by Cucchiara *et al*[1]; the relative colour and brightness changes are analysed to determine when a shadow overcasts a surface. The novelty of this colour space is adopted in the works of Bowden *et al* [5] to classify shadows in a similar manner. Each of these techniques use information relating to a shadows appearance to derive the classification criteria. We propose a technique that exploits the geometric properties of people shadows to develop an unsupervised method for feature sampling. This method *learns* the appearance of both the shadow and foreground features, providing a more reliable mechanism for feature classification.

2 Pixel Classification

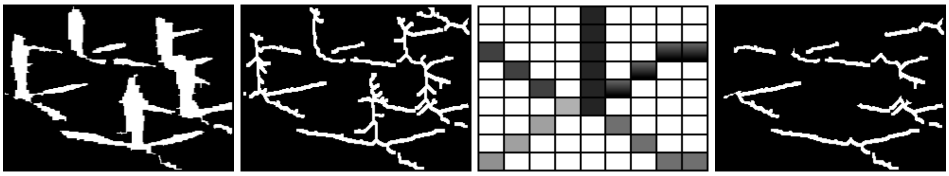
The aim of the proposed algorithm is the successful classification and removal of shadows resulting from floodlights present in the football stadium. The algorithm works by exploiting information relating to the shadow's *shape, size, orientation, luminosity, originating position* and *appearance model* [7] to determine the colour distributions of both the foreground and shadow classes. In this section we discuss the unsupervised process of obtaining the foreground and shadow feature data, the parametric modelling of the feature data, and the pixel classification mechanism. For reliable foreground detection, a background model is required that can adapt to any changes in the environment. Examples of common changes include the amount of direct sunlight, wind-blown trees and periodically rotating advertising boards. Given this requirement, an adaptive background subtraction technique based upon the work of Stauffer and Grimson [8] is used.

2.1 Shadow Feature Sampling

The first step towards identifying shadows is the analysis of the foreground segmentation recovered from background modelling. Whether any of the changed pixels in the foreground feature space were the result of shadows is unknown. The following discusses how information about a shadow's shape, orientation, location and size can be used to identify the *most-probable* shadow components within the segmentation mask.

2.1.1 Skeletonisation

We propose to exploit information relating to the spatial distribution of the foreground features, to develop a process for identifying the shadow features. In a football stadium, the light sources positioned in the stadium corners result in four shadows that propagate from the base of each player. Each shadow has a similar shape and size to that of the attached player and is orientated towards each of the opposing light sources. Given this information it is desirable to label a foreground player's feature components using information relating to its shape. An object's skeleton provides an intuitive, compact representation of a shape that can be used to determine an objects sub-parts and their connectivity [3]. Using a skelatonisation algorithm similar to that proposed by Zhang and Suen [9], skaletonisation of the foreground is performed to obtain the medial axis of the foreground players and their shadows - *see* Figure 2(b). The sub-parts of the medial axis provide a more reliable information source even during periods of player occlusion. An example of a typical skelatonisation result can be seen in Figure 2(c). From the skelatonisation result, it can be seen that shadows exhibit lengthy skeletal sub-parts with a non-vertical orientation in the image plane (Figure 2(d)). In the remainder of this paper a skeleton is considered to be comprised of nodes that form inter-connected branches - see Figure 2. Occasionally the term leaf will be used to describe a skeleton's surface branches.



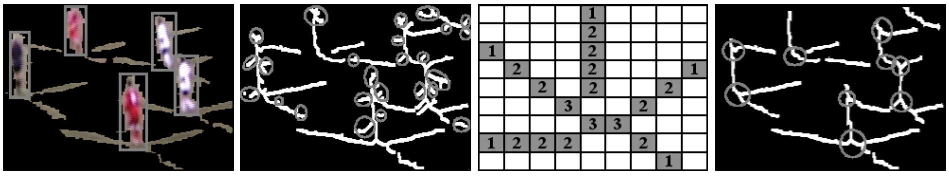
(a) Segmentation Mask (b) Skeleton Mask (c) Node Labelling (d) Shadow branches

Figure 2: Examples of (a) Segmentation mask, (b) Skelatonisation, (c) Node groupings into branches and (d) Desired shadow branches.

2.1.2 Skeletal Medial Axis Analysis

A skeleton's medial axis is comprised of numerous branches, each representing groupings of foreground and/or shadow nodes (Figure 2(c)). To determine a set of shadow sample locations in the current video frame, we attempt to identify the skeleton branches that result from shadows. By assuming that people stand vertically in the image-plane, spatial filtering and appearance models [7] are used to identify the shadow braches. This process is implemented in the following three steps:-

Medial Axis Simplification - To reduce the complexity of the skeletons we attempt to remove the leaves below a certain size (S). Due to the fact that a person's image-plane size depends on their location w.r.t the camera, the value of S should be related to the leaf's position. In the works of Renno *et al* [7] a technique was developed that automatically computes a person's image-plane bounding-box model. We apply this model here to set the value of S to the model's *width* attribute, since it is reasonable to assume that any leaf belonging to shadow will extend outside of a person's *real* bounding-box model - see Figure 3(a).



(a) People Bounding-boxes (b) Unwanted leafs (c) Node Connectivity (d) Trimmed Skeletons

Figure 3: Examples of (a) Bounding-box models overlaid onto foreground objects, (b) Illustration of the unwanted leafs on the skeleton mask, (c) labelling of a node to its neighbour count and (d) Trimmed skeletons with and the areas of class ambiguity.

To determine the skeleton leafs that constitute noise (Figure 3(b)), a connectivity based node labelling scheme is applied that assigns a label to each node; this label being equivalent to the number of its neighbouring nodes - *see* Figure 3(c). Leaves are spawned from the nodes that have a label value of one; its nodes are determined as those connected nodes that lie in the path to a node with a label value greater than two or another spawn node. A candidate leaf is deleted if:-

$$\left\|\mathbf{X_1}^{(\mathbf{i})} - \mathbf{X_2}^{(\mathbf{i})}\right\| < S^{(i)} \tag{1}$$

where $X_1{}^i$ and $X_2{}^i$ are the image coordinates for each of the i'th leaf end points, and S is

$$\mathbf{S}_{(i)} = \frac{\Omega^{\varphi}\left(j_{(i)} - j_h\right)}{2} \tag{2}$$

where j_i is the average vertical image coordinate of the ith branch. j_h and Ω and φ are the *image horizon*, bounding-box *width model* and the *person* class respectively [7]. This procedure ensures that most of the unwanted skeleton leafs are removed - Figure 3(d).

Spatial Filtering - A spatial filtering process is employed to determine the class memberships for each node of the skeletons. This is achieved by accentuating the appearance of the skeletons vertical branches through the application of two spatial filters: 1) 5x5 kernel representing five 1D Gaussians and 2) 5x5 kernel representing an average operator. Each of the filter processes is followed by a thinning process: a logical AND of the resulting filtered skeleton and the original skeleton mask. The Gaussian filter accentuates the vertical branches; the smoothing operator softens noise regions resulting from *not-quite*-vertical branches and branch intersections as well as increasing the similarity between non-vertical skeleton branches. The resulting skeleton node intensities are put into a histogram and the intensity threshold (τ) identified as: the intensity bin in the histogram with the highest frequency. The spatial filtering process can be seen in Figure 4.

Shadow Node classification - Applying τ as the class threshold for the filtered skeletons and selecting the non-zero nodes that are below τ, yields the shadow nodes of the skeletons (Figure 2(d)). Each node represents a possible shadow feature sample location in the current video-frame. There remains one problem with the sample nodes: intersection points where branches of differing classes meet represent the origins of regions where unknown boundaries between foreground and shadow exist. Due to the fact that node class values are undeterminable in these regions, circular exclusion areas around each of

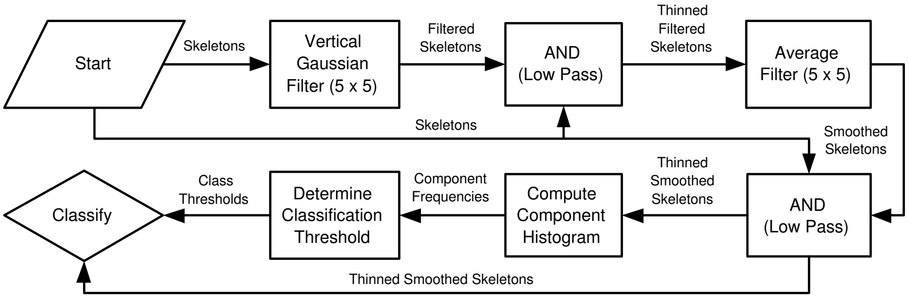

Start
Skeletons
Vertical Gaussian Filter (5 x 5)
Filtered Skeletons
AND (Low Pass)
Thinned Filtered Skeletons
Average Filter (5 x 5)
Skeletons
Smoothed Skeletons
Thinned Smoothed Skeletons
AND (Low Pass)
Compute Component Histogram
Component Frequencies
Determine Classification Threshold
Class Thresholds
Classify
Thinned Smoothed Skeletons

Figure 4: The spatial filtering process used to determine the foreground and shadow components

the intersection locations (Figure 3(d)) are defined. To reflect scale in the image-plane, the diameters of the exclusion regions are defined as $S_{(i)}$. Any shadow nodes that reside in any of the exclusion zones are removed to leave the candidate shadow sample points $\mathbf{X} = \{\mathbf{x_1}, ..., \mathbf{x_n}\}$. The sample points in $\mathbf{X}$ represent the *most-probable* locations of the shadow features in the current video-frame.

2.1.3 Shadow Sampling

Sampling of the shadow feature data is achieved by storing pixels from the current video-frame from the locations identified in $\mathbf{X}$. For each new video-frame, sampling is repeated until a set of N_α shadow training feature vectors is obtained. The set of shadow features is denoted by α, where $\alpha = \{\mathbf{f}_1, ..., \mathbf{f}_{N_\alpha}\}$ and $\mathbf{f}$ is the feature vector representative of a pixel's $\{R, G, B\}$ values.

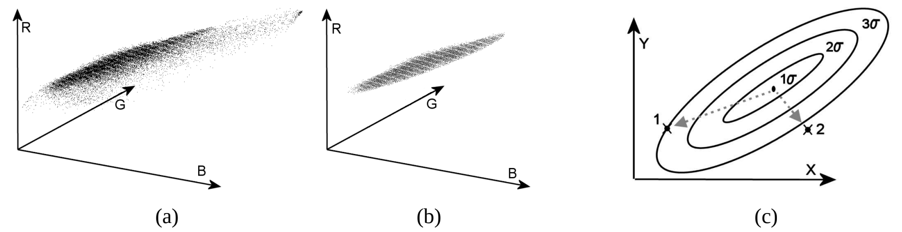
2.2 Foreground Feature Sampling

To determine the features in the segmentation mask that constitute real foreground, it is desirable to remove any feature resulting from a shadow. Empirical analysis reveals that the majority of the shadow features form a single cluster in the feature space (Figure 5a), surrounded by a sparse field of outlying feature vectors; these are assumed to be the result of erroneous labelling during shadow sampling - *see* Section 2.1. The *Gaussian* is a good statistical model for clustered data and easily parameterised using the mean (μ) and covariance (Σ) of the feature data. Using the shadow feature data, a *tri-variate Gaussian* model is developed that to obtain the probability density function :-

$$p(\mathbf{f}|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{3}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{f}-\mu)^T\Sigma^{-1}(\mathbf{f}-\mu)\right) \tag{3}$$

To determine a feature's membership to the shadow feature model requires a decision boundary between the Gaussian model and the feature space. In this work the shadow Gaussian model is approximated to a hyper-ellipsoid, enabling a decision boundary to be defined in units of the Gaussian's standard-deviation (σ) - *see* Figure 5c. The *Mahanalobis distance* metric is used to determine the normalised distance between a feature and the Gaussian mean vector. Assuming that 95% of the shadow feature data is accurate, the classification distance threshold (D) is set to: the number of standard-deviations from a distribution's mean value that encompasses 95% of the feature data ($D \approx 2.7$); rejecting the remaining 5% as noise.

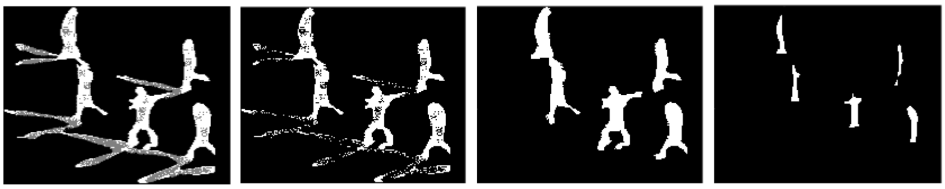
$$\left\{(\mathbf{f}-\mu)^T\Sigma^{-1}(\mathbf{f}-\mu)\right\}^{\frac{1}{2}} < D \tag{4}$$



(a) (b) (c)

Figure 5: (a) Distribution of the sampled shadow features, (b) Hyper-Ellipsoid approximation of distribution of (a) and (c) Demonstration of the mahanalobis distance showing that point 1 is closest to the distribution's mean value, regardless of the euclidean distances of each feature

Classification according to (4) is performed for those features identified in the segmentation mask - *see* Figure 6(a). Those features classified as shadow are removed from the segmentation mask *see* Figure 6(b) and the remaining points grouped into regions of connected components. Regions of a size lower than that of the system noise floor are removed. The remaining regions undergo a morphological filtering process that consists of one iteration of closing and three iterations of eroding. Closing fills small holes in the regions resulting from mis-classification by (4) and eroding ensures that the majority of non-foreground components are removed - *see* Figure 6(c,d). The remaining region components of the segmentation mask $\mathbf{Y}$, where $\mathbf{Y} = \{\mathbf{y_1}, ..., \mathbf{y_n}\}$ represent the locations of the foreground features in the current video-frame - *see* Figure 6d.



(a) (b) (c) (d)

Figure 6: Segmentation mask processing - (a) Shadow classification, (b) Shadow suppression, (c) Component grouping and region analysis followed by closing and (d) Erosion of the regions resulting in the remaining foreground samples

Sampling of the foreground features is performed in the same manner as for that for shadows - *see* Section 2.1.3. A feature vector for each of the points specified in $\mathbf{Y}$ is sampled from the current video-frame to develop the foreground feature training set β, where $\beta = \{\mathbf{f_1}, ..., \mathbf{f_{N_\beta}}\}$. As with shadows, the sampling process is repeated for each new video-frame until a set of N_β features is obtained.

2.3 Feature Modelling and Classification using Gaussian Mixtures

Tri-variate Gaussian mixtures models are used to approximate the shadow (α) and foreground (β) feature data. The Gaussian mixture models for both shadow (θ_α) and foreground (θ_β) are determined using the technique proposed in Figueiredo *et al* [2]. Their work demonstrates an unsupervised *expectation maximisation* algorithm that hypothesises the optimal: 1) number of component densities within each mixture required to encapsulate the feature data (K), 2) the parameters for each of the Gaussian distribution functions $\phi_i = \{\mu_i, \Sigma_i\}$ and 3) the prior probabilities (i.e. mixing parameters) of the component Gaussian models ω. The probability of observing the feature $\mathbf{f}$ given the mixture model θ is

$$P(\mathbf{f}|\theta_i) \ = \ \sum_{i=1}^{K} \omega_i \times \eta\left(\mathbf{f}, \phi_i\right), \tag{5}$$

where η is the probability density function for a Gaussian parameterised by ϕ - *see* equation 3. Using *Bayes theorem*, the *conditional* and *prior* probabilities are combined to compute the *posterior* probability, that is, the probability that the Gaussian density function θ_i is responsible for generating data point $\mathbf{f}$.

$$P(\theta_i|\mathbf{f}) = \ \frac{p(\mathbf{f}|\theta_i)P(\theta_i)}{P(\mathbf{f})} \ \equiv \ \frac{p(\mathbf{f}|\theta_i)P(\theta_i)}{\sum_{j=1}^{J} p(\mathbf{f}|\theta_j)P(\theta_j)} \tag{6}$$

The classification of each feature identified in the original segmentation mask (*see* - Figure 2a) is determined by the *maximum a posteriori probability* rule. The posterior probabilities $P(\theta_\alpha|\mathbf{f})$ and $P(\theta_\beta|\mathbf{f})$ for each class of Gaussian mixture are computed. Feature classification is achieved by labelling the feature as the class (α, β) of the Gaussian mixture that achieved the largest posterior probability value - *see* equation 7

$$\theta_i = \arg \max_{i \in (\alpha, \beta)} \{P(\theta_i|\mathbf{f})\} \tag{7}$$

In this paper the class prior probabilites $P(\theta_\alpha)$ and $P(\theta_\beta)$ are assumed to be equal (=0.5).

3 Shadow Classifiers

In this section we propose the use of three previously published shadow classification algorithms [6, 1, 4] for comparative and evaluative purposes. Each of the classifier variants is to be applied to the segmentation mask developed during our background subtraction process [8].

3.1 Classifier A : HSV colour space conversion

This classifier uses the same criteria for classifying shadows originally proposed by Cucchiara et al [1]. Their shadow classification is applied to the ungrouped features detected in our segmentation mask. Each flagged feature in the segmentation mask as well as our background model is transformed into the HSV colour space.

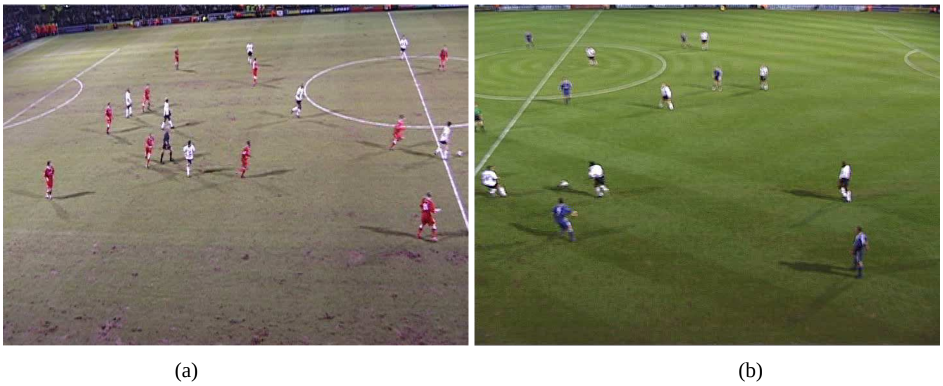

(a)
(b)

Figure 7: Video Datasets of (a) Middlesborough and (b) Newcastle - at Fulham (Camera 2)

3.2 Classifier B : Normalised RGB and Gradient models

This classifier is based upon the technique for classifying shadows proposed by McKenna *et al* [6]. The implementation of this technique remains relatively unchanged except for their background subtraction process. This is replaced by the technique used in this paper to determine the features constituting changed features. Classification of pixels is determined using the published criteria for those pixels flagged in our segmentation mask.

3.3 Classifier C : Computational Colour Model

This classifier is based upon a technique for feature classification proposed by Horprasert *et al* [4]. Our implementation of this technique is adapted to only use the colour model for shadow classification purposes.

4 Evaluation

An objective evaluation methodology requires the availability of ground-truth data. This enables quantitative performance measures and therefore direct comparisons between the proposed methods. In addition any differences between real and ground-truth segmentations provide interesting insights into algorithm performance. This section briefly describes the ground-truth data format and the quantitative evaluation metrics that are used. The four described shadow classification methods were applied to the two flood-lit football sequences (shown in Figure 7), each of 1000 frames. The performances of each classifier were evaluated using the above metrics.

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 8.

Two evaluation metrics are proposed to measure classifier performance by directly comparing the segmentation result to the ground truth data. The first metric determines how much of the shadow is removed by:-

$$\mathbf{DR} = \left(1 - \left(N^{fp}/N_0^{fp}\right)\right) \times 100 \tag{8}$$

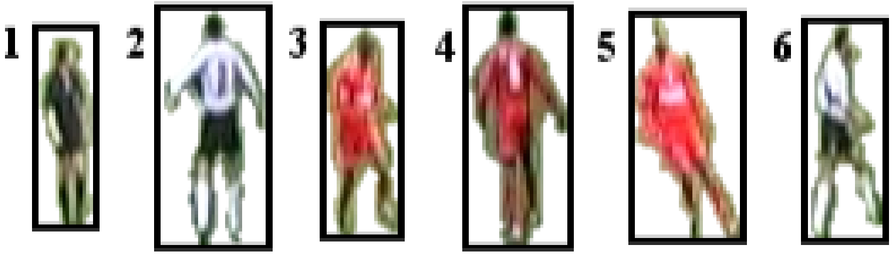


Figure 8: Target postures with the ground-truth overlayed.

where N_0^{fp} and and N^{fp} are the number of false-positives (shadow and background noise) before and after shadow classification respectively. The second metric computes the accuracy of the segmentation after the identified shadows have been removed by using the signal to noise *SNR* ratio:-

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

where N^{tp} and N^{fp} are the number of true and false positives. True positives relate to the number of correctly classified foreground features with respect to the ground-truth.

Numerical results are tabulated in Table 4. An example of the results are also shown qualitatively in Figure 9(c-f). In terms of the shadow detection rate performance, the proposed method is effective, correctly detecting 93% of input shadows (on average). The other methods vary from between 55 to 84%. The SNR represents the overall accuracy of the classifiers by measuring their impact upon both foreground and shadow segmentations. A high SNR of 10dB is achieved by the proposed classifier indicating a high percentage of foreground to shadow in the signal. Classifier (b) performs poorly achieving a SNR of 4.3 indicating poor segmentation of shadow and foreground.

Algorithm	Datasets			
	Middlesborough: Camera 2		Newcastle: Camera 2	
	DR(%)	**SNR(dB)**	**DR(%)**	**SNR(dB)**
Detection [8]	**0**	**-2.2**	**0**	**-2.6**
Cucchiara *et al* [1]	**74**	**10.3**	**75**	**10.0**
Horprasert *et al*[4]	**84**	**10.5**	**79**	**11.1**
McKenna & Jabri [6]	**57**	**4.3**	**55**	**4.6**
Proposed Classifier - Sect 2	**93**	**10.6**	**91**	**9.8**

Table 1: Shadow classifier performance

5 Conclusions

In this paper a novel shadow classification method was presented and compared to three published methods for shadow classification, in the specific domain of floodlit football games. The evaluation of the classifiers comprised the amount of shadow removed, and the SNR of foreground to any remaining shadow and background noise. The proposed method is shown to be highly effective at separating objects from their shadows. It is

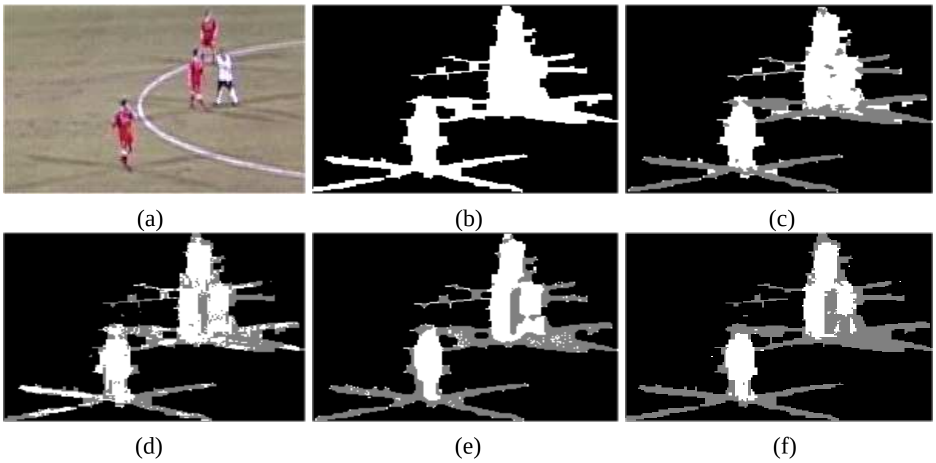


Figure 9: empirical results of shadow classification:- (a) Original Image, (b) Segmentation Mask, (c-e) Classifier's A-C respectively - *see* Section 3 and (f) Proposed Classifier.

completely unsupervised, and works well across a variety of sensor types, camera locations and match conditions without the need for ad hoc refinement of parameters. It is sufficiently efficient to be included in processes running at real-time frame rates. In future we expect to generalise the method to work in other scenarios, i.e. outside the football stadium.

References

[1] R. Cucchiara, C. C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, October 2003.

[2] M. Figueiredo and A.K. Jain. "Unsupervised learning of finite mixture models". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, March 2002.

[3] P. Gollina and W.E.L. Grimson. "Fixed Topology Skeletons". In *IEEE Trans. Computer Vision and Pattern Recognition, Vol 1*, pages 1010–1017, Hilton Head Island, USA, June 13-15 2000.

[4] T. Horprasert, D. Harwood, and L.S. Davies. A robust background subtraction and shadow detection. In *Asian Conference on Computer Vision*, Taipei, Taiwan, January 8-11 2000.

[5] P. KaewTraKulPong and R. Bowden. "An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection". In *2nd European Workshop on Advanced Video-Based Surveillance Systems*, pages 149–158, Kingston upon Thames, UK, Sept 4 2001.

[6] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80(1):42–56, October 2000.

[7] J-P.R. Renno, J. Orwell, and G.A. Jones. Learning surveillance tracking models for the self-calibrated ground plane. In *British Machine Vision Conference*, Cardiff, UK, September 2002.

[8] 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, Vol 2*, pages 246–252, Fort Collins, Colorado, June 23-25 1999.

[9] T.Y. Zhang and C.Y. Suen. "A Fast Parallel Algorithm for Thinning Digital Patterns". *Commun. ACM*, 27(3):236–240, March 1984.