IDIAP RESEARCH REPORT

ROBUST COLOUR SEGMENTATION USING MAP ADAPTATION

Mark Barnard [*] Jean-Marc Odobez [*]

IDIAP–RR 03-77

DECEMBER 2003

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[*] Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP) P.O. Box 592, CH-1920 Martigny, Switzerland. barnard@idiap.ch

IDIAP Research Report 03-77

Robust Colour Segmentation using MAP Adaptation

Mark Barnard Jean-Marc Odobez

December 2003

Abstract. A vital task in sports video annotation is to detect and segment areas of the playfield. This is an important first step in player or ball tracking and detecting the location of the play on the playfield. In this paper we present a technique using statistical models, Gaussian mixture models (GMMs) and Maximum a Posteriori (MAP) adaptation. This involves first creating a generic model of the playfield colour and then using unsupervised MAP adaptation to adapt this model to the colour of the playfield in each game. This technique provides a robust and accurate segmentation of the playfield. In order to test the robustness of the method we tested it on a number of different sports that have grass playfields, rugby, soccer and field hockey.

1 Introduction

Colour is an important feature in recognition of patterns within images. Here we address the specific problem of playfield segmentation in sports. In our case we will focus on sports where the playfield is grass. The technique we present here could, however, be used for any type of playing area, such as a basketball or tennis court.

a)  b)  c) 

Figure 1: Images from soccer and rugby.

While the colour of grass is generally green, this colour can vary depending on the individual playfield, the presence of shadows or the viewing angle. There is also a lot of noise present in the images from sports videos. This noise can come from green players shirts, green in the crowd or advertising signs (Figure 1b). The amount of this noise also depends on the sport. In soccer due to the nature of the game there are many wide shots where the majority of the image is the playfield (Figure 1a). In rugby however there are more medium and close shots with many players gathered in the same area of the playfield (Figure 1c). The player close up in Figure 1b illustrates a case where green is the dominant colour but there is no grass in the image. So any unsupervised segmentation technique must be general enough to account for the variations in playfield colour, while being focused enough to filter noise in the image.

1.1 Colour Representation

It is important for a colour representation to be invariant to different viewing angles and to different illumination conditions. The standard *Red, Green, Blue* (RGB) colour space represents not only the colour but also the brightness and so is not invariant to illumination conditions.

One common transformation from RGB space is to the *Hue, Saturation, Value* (HSV) space. The HSV colour representation is based on the achromatic axis of the RGB space, where $R = G = B$. The values for H, V and S are given by

$$H = \arctan(\sqrt{3}(G - B), 2R - G - B),$$

$$V = \frac{(R + G + B)}{3}, S = 1 - \frac{\min(R, G, B)}{V}.$$

In HSV space, as we approach the achromatic axis and the saturation approaches zero, the value of the hue becomes unstable. So acromatic pixels, or near achromatic pixels, with no colour, will have hue value with no meaning [1]. These achromatic pixels are a source of noise in the hue space, being given hue values that do not correspond to their actual colour.

In order to represent a colour in a way that is invariant to intensity and is also less effected by acromatic pixels we use cromatic r and g as features [2]. These are derived by normalising R and G

with respect to the intensity so

$$r = \frac{R}{R+G+B} \text{ and } g = \frac{G}{R+G+B}.$$

We are also able the shorten the feature vector by one as b is linearly dependent on r and g.

1.2 Previous Work

Previous approaches to playfield segmentation can be divided into those that require human intervention during testing and those that are unsupervised. One method needing intervention is to build a model for the colour and then set the threshold for recognition during the test [3]. We believe an automatic solution to this problem would prove more useful in any video annotation system.

Previous work on unsupervised methods of playfield extraction has included using statistics of RGB values in each frame [4]. This method assumes that in each frame the playfield is the largest area in the image, however this if often not the case (Fig 1b). One unsupervised method which makes no prior assumption on the nature of the data is simply to take a predefined range of the hue, for example $\frac{1}{3}\pi \leq H \leq \frac{5}{6}\pi$, as grass [5]. Another proposed method collects statistics on the hue for the first five minutes of the video [6]. This method relies on the assumption that the most dominant single colour in the first five minutes will be the colour of the playfield. This is, however, not always the case: may be the actual play might not start until five or ten minutes into the broadcast or there may be a long break in the play when the playfield is not visible. While these techniques often work well in ideal conditions, for example in soccer data as shown in Figure 1(a), they are deemed to fail in noisy conditions where grass is not the dominant colour, such as rugby games with players in green shirts, where the playfield has turned to mud or grass in shadow.

1.3 Our Approach

The method we propose here is unsupervised and is also robust to large amounts of noise in the image. This method involves modeling colour using a *Gaussian Mixture Model* (GMM) and then adapting this model using *Maximum a Posteriori* (MAP) adaptation [7]. A GMM is trained on grass images from a variety of different soccer games. This creates a general model of playfield grass. We then train a noise model from images containing no grass. The features used in the training of these models are the chromatic r and g features. A validation set is used to the hyperparameters of the algorithm. The models are then tested on games of rugby, field hockey and soccer not used in the training set. During the testing, the grass model is adapted, using MAP adaptation, to the particular colour of the playfield in each of the test games.

2 Gaussian Mixture Models (GMM)

In order to model the colour of grass in $x = (r, g)$ space we use a 2 dimensional GMM. In a GMM the likelihood of the data x is given by

$$p(x) = \sum_{i=1}^{N} w_i \cdot \mathcal{N}(x_i; \mu_i, \Sigma_i) \tag{1}$$

where N is the number mixtures in the model, $W = \{w_i\}$ is the set of mixture weights, $\mu = \{\mu_i\}$ are the set of means and $\Sigma = \{\Sigma_i\}$ is the set of convariance matrices of the Gaussian mixture. Hence a GMM is fully parameterised by $\theta = \{W, \mu, \Sigma\}$. In our case, Σ_i are diagonal matrices.

Given a training data set X of observations, θ is estimated using the *Maximum Likelihood* (ML) principle, that is the likelihood of X is maximised with respect to θ: So we select the parameters $\hat{\theta}$ such that,

$$\hat{\theta} = \arg\max_{\theta} p(X|\theta). \tag{2}$$

The normal method of training GMMs is to use the *Expectation Maximization* (EM) algorithm [8].

3 Maximum a Posteriori Adaptation (MAP)

The ML principle can be applied when we are given labeled data, for example grass data only. This is suitable for offline learning. In online learning, as all data may not correspond to the right label, prior knowledge is necessary to constrain the space of solutions for θ. This can be achieved using MAP adaptation, where prior knowledge is given in the form of a prior distribution over θ, $p(\theta)$. Using the MAP principle we select $\hat{\theta}$ such that it maximizes the *a posteriori* likelihood,

$$\hat{\theta} = \arg\max_{\theta} P(\theta|X) = \arg\max_{\theta} p(X|\theta) \cdot p(\theta). \tag{3}$$

The contributions of the data likelihood, $p(X|\theta)$, and the prior distribution, $p(\theta)$, can be balanced by introducing a weighting factor, α, in equation 3. So, in pratice we maximise $p(X|\theta)^{(1-\alpha)} \cdot p(\theta)^{(\alpha)}$.

A common use for MAP adaptation is in speech and face verification [9], in this case a general world model of speakers or faces is trained. This model is then adapted, using MAP, to the particular speaker or face. In our case, we train a general model of grass and then use MAP to adapt this general model to the specific colour of the grass in each particular game.

When using MAP adaptation, different parameters can be chosen to be adapted [10]. In our case we adapt all the parameters. We adapt the weights because we may have different green colours within the playfield area, in which case we want to model these different colours with different mixtures. When only one colour is present the weight of the other mixtures will be adapted to zero. We also adapt the variances in order to move from a broader generic playfield model to a model focusing on the particular playfield in the current data.

The parameters of a mixture i are adapted using the following set of update equations [7] [9]

$$\hat{w}_i = \alpha \cdot w_i^{pr} + (1-\alpha) \cdot w_i^{ml}, \tag{4}$$

$$\hat{\mu}_i = \alpha \cdot \mu_i^{pr} + (1-\alpha) \cdot \mu_i^{ml}, \tag{5}$$

$$\begin{aligned} \hat{\Sigma}_i &= \alpha \cdot (\Sigma_i^{pr} + (\hat{\mu}_i - \mu_i^{pr})(\hat{\mu}_i - \mu_i^{pr})^T) \\ &+ (1-\alpha) \cdot (\Sigma_i^{ml} + (\hat{\mu}_i - \mu_i^{ml})(\hat{\mu}_i - \mu_i^{ml})^T), \end{aligned} \tag{6}$$

where α is a weighting factor on the prior parameters, w_i^{pr}, μ_i^{pr} and Σ_i^{pr} are the prior weight, mean and variance. The parameters estimated by ML, w_i^{ml},μ_i^{ml} and Σ_i^{ml}, are given by the following equations [11]

$$w_i^{ml} = \frac{1}{M} \sum_{i=1}^{M} p(i|x_i, \theta), \tag{7}$$

$$\mu_i^{ml} = \frac{\sum_{i=1}^{M} x_i p(i|x_i, \theta)}{\sum_{i=1}^{M} p(i|x_i, \theta)}, \tag{8}$$

$$\Sigma_i^{ml} = \frac{\sum_{i=1}^{M} p(i|x_i, \theta)(x_i - \mu_i^{ml})(x_i - \mu_i^{ml})^T}{\sum_{i=1}^{M} p(i|x_i, \theta)}, \tag{9}$$

where M is the number of data examples.

4 Algorithm description

In our approach to this problem we train two GMMs. One is a general model of playfield colour and the other is a noise model. Both of these models are trained offline. In recognition we use a likelihood ratio of the playfield model and the noise model in preference to applying a fixed threshold to the playfield likelihood.

Starting from the offline trained model ($\theta_0^{grass} = \theta_{offline}^{grass}$), the process consists of iterating between two steps: (1) selecting data X_k to be used for adaptation (2) updating θ_k from θ_{k-1} using X_k.

4.1 Selecting Adaptation Data

The data to be used for adaptation X_k are gathered from pixel values extracted from images sampled every second. The image pixels used for adaptation are selected in a two step process.

1. Colour Selection. We recognise grass pixels by thresholding the ratio of the likelihood of a pixel feature with respect to the current grass model θ_k and its likelihood with respect to the noise model, so a pixel is labeled as grass if

$$\frac{p(C_p|\theta_k^{grass})}{p(C_p|\theta^{non-grass})} > d, \tag{10}$$

where C_p is the (r, g) feature for pixel p and d is a threshold determined on a validation set.

2. Prior and Morphological filtering. For each frame, a binary image is created using equation (10). As there is a higher prior probabilty of noise in the top half of the image, for example the stadium, advertisments, the crowd or trees, we consider only the bottom half of the image for potential adaptation pixels. This binary image is then morphologically filtered using the open operation which is a combination of erosion and then dilation. In our case we use a 5×5 square structuring element for this operation. This filtering eliminates small areas of noise and leaves only larger areas of pixels for adaptation.

4.2 Updating Model Parameters

The parameters are updated every 30 seconds using data X_k and applying equations (4) to (9). During adaptation we also control the weighting α on the prior model by making α dependent on the number of pixels that have been selected for adaptation according to

$$\alpha_k = 1 - \frac{N_p}{c \cdot f \cdot s} \tag{11}$$

where N_p is the number of pixel selected for adaptation over the past 30 seconds, f is the number of frames, in our case typically 30, s is the number of pixels in each frame and c is a tuning parameter which was set on the validation set. So it can be seen that if the number of pixels found for adaptation is very small then a larger weight is given to prior parameters. In this way the rate of adaptation depends on the data. This is important if the playfield is not dominant during the time we collect the adaptation data. If we have a long interuption in a game, resulting in many crowd shots or player close ups, then our model will give more weight to the prior model, avoiding adaptation to noisy data.

The output from the algorithm is a binary image produced by the likelihood ratio of grass to noise and morphological filtered with a 3x3 structuring element. These are the image shown in the results section in Figure 2.

5 Experiments and Results

In our experiments, we divided the data into three sets: a training set of data from 6 different games of soccer, a validation set of data from 4 different games of soccer and a testing set of data from 8 games of rugby, 2 games of soccer and one game of field hockey.

The offline playfield colour model was trained on 163 grass images taken from the training set and a noise model was trained using 87 images containing no grass. The values of c, the adaptation rate tuning parameter, and d, the recognition threshold, were then adjusted for optimal performance using the validation set. The optimal sampling rate, one frame per second, and adaptation interval, every 30 seconds, were also selected using the validation set.

The performance of the models was tested on a variety of different games with grass playfields to show the generic nature of this method. We tested our algorithm against simple hue segmentation [5] and also against the technique of finding the mean and standard deviation of the dominant hue for the first five minutes of the game [6].

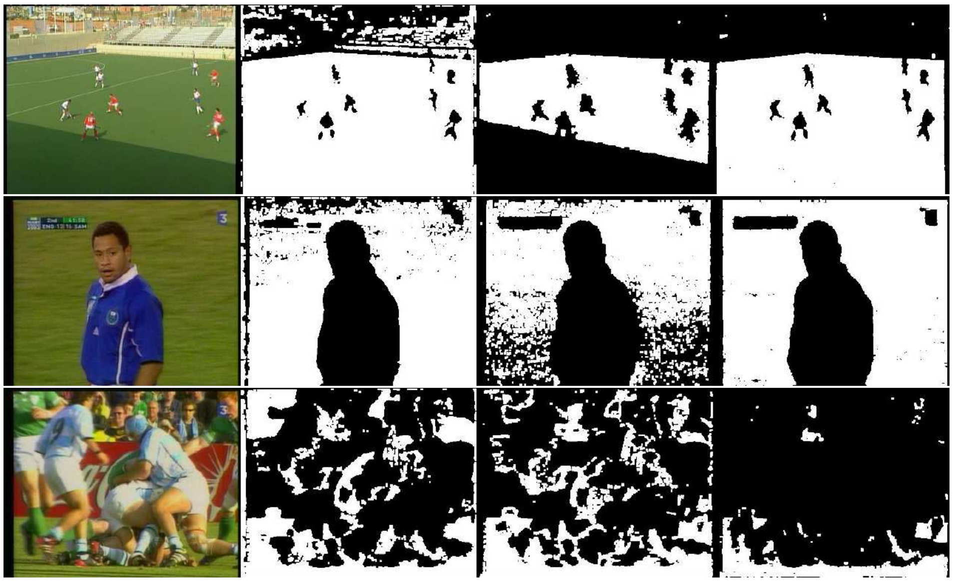

Figure 2: Results for the techniques tested. 1st column: original image; 2nd column: hue thresholding [5]. 3rd column: dominant hue [6]. Fourth column: our technique

Game	Simple hue thresholding [5]	Statistical hue thresholding [6]	Adapted model
Ire vs Arg	0.812	0.865	**0.932**
Eng vs SA	0.740	0.881	**0.927**
Eng vs Sam	0.858	0.918	**0.963**
Field hockey	0.828	0.840	**0.952**
Soccer	0.835	0.922	**0.953**

Table 1: Results for each game showing the average recognition rate.

In order to have a quantitative measure of the performance of our algorithm, we extracted 20 images from each of three games of rugby, one game of soccer and one game of field hockey. These images were taken at exactly 30 second intervals from the start of each recording and then the play field was segmented by hand. As a performance measure, we simply used the recognition rate of the pixel label. The results of all these tests can be seen in Table 1. In these empirical tests, the method we propose performs better than either of the other methods.

Results for some individual frames are shown in Figure 2. In the first image it can be seen that the simple thesholding of the hue, while recognising all the grass also produces a lot of noise from the background. The statistical thresholding technique has a tendency to underestimate the noise variance, especially when different shades of green are present (here part of the play field in shadow is not recognised). In our method it can be seen that using more than one Gaussian mixture allows for better recognition of both sections of the field with very little noise in the background. Similarly, for the other images, it can been seen that the algorithm we propose produces a more accurate and robust result.

6 Conclusions

In this paper we present an automatic adaptation method to segment the playfield in sports videos. In our algorithm, starting from a broad general model of the playfield colour learned off-line, a MAP adaptation step is iteratively applied to drive the general model to a specific one associated with the available video stream. It is very important when using adaptation to control both the data that you adapt with and the rate of adaptation. We control the data for adaptation by using the prior model for recognition and morphological filtering to eliminate potential noise. The rate of adaptation is dynamically controlled depending on the amount of adaptation data that has been collected. The results clearly show that our proposed algorithm gives a more accurate and robust segmentation of the play field than current unsupervised techniques. The models used in our approach are also generic to any sport played on grass, as can be seen in the test result for models trained on soccer but tested on rugby and field hockey. While we have not done any experiments with playfields of other colours, such as basketball or tennis, clearly this technique could be implemented for these sports.

The output of the playfield segmentation is currently being used to generate features for event recognition with rugby data. The binary images are also being used as a starting point for player and ball tracking.

References

[1] W. Skarbek and A. Koschan, "Colour image segmentation - a survey," Tech. Rep., Technical University Berlin, 1994.

[2] Theo Gevers and Arnold W. M. Smeulders, "Color based object recognition," in *ICIAP (1)*, 1997, pp. 319–326.

[3] Ahmet Ekin, A. Murat Tekalp, and Rajiv Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, to appear.

[4] S. Choi, Y. Seo, H. Kim, and K.S. Hong, "Where are the ball and players? soccer game analysis with color-based tracking and image mosaick," in *ICIAP*, Florence, Italy, Sept 1997.

[5] Okihisa Utsumi, Koichi Miura, Ichiro Ide, Shuichi Sakai, and Hidehiko Tanaka, "An object detection method for describing soccer games from video," in *ICME2002*, Aug 2002, pp. 45–48.

[6] Peng Xu, Lexing Xie, Shih-Fu Chang, Ajay Divakaran, Anthony Vetro, and Huifang Sun, "Algorithms and system for segmentation and structure analysis in soccer video," in *Proc. ICME*, Tokyo, Japan, Aug 22-25 2001.

[7] J. L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech Audio Processing*, vol. 2, pp. 291–298, April 1994.

[8] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood for incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B*, vol. 39, pp. 1–38, 1977.

[9] Mariéthoz, J. and Bengio, S., "A comparative study of adaptation methods for speaker verification," in *ICSLP*, Denver, USA, September 2002.

[10] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, 2000.

[11] J Bilmes, "A gentle tutorial of the EM algirthm and its application to parameter estimation for gaussian mixture and hidden markov models," Tech. Rep. TR-97021, ICSI), U.C. Berkeley, 1998.