Probabilistic and Voting Approaches to Cue Integration for Figure-Ground Segmentation

Eric Hayman and Jan-Olof Eklundh

Computational Vision and Active Perception Laboratory (CVAP)
Dept. of Numerical Analysis and Computer Science
KTH, SE-100 44 Stockholm, Sweden
{hayman, joe}@nada.kth.se
http://www.nada.kth.se/~hayman/ECCV2002*

Abstract. This paper describes techniques for fusing the output of multiple cues to robustly and accurately segment foreground objects from the background in image sequences. Two different methods for cue integration are presented and tested. The first is a probabilistic approach which at each pixel computes the likelihood of observations over all cues before assigning pixels to foreground or background layers using Bayes Rule. The second method allows each cue to make a decision independent of the other cues before fusing their outputs with a weighted sum. A further important contribution of our work concerns demonstrating how models for some cues can be learnt and subsequently adapted online. In particular, regions of coherent motion are used to train distributions for colour and for a simple texture descriptor. An additional aspect of our framework is in providing mechanisms for suppressing cues when they are believed to be unreliable, for instance during training or when they disagree with the general consensus. Results on extended video sequences are presented.

1 Introduction

A key capability for active observers is to segment foreground objects out from the background. This enables higher level processes such as object recognition, object grasping and vehicle navigation to take place. It is also a topic of significant interest in video coding where the MPEG-4 format requires different objects to be represented separately for efficient transmission of data and for interactive editing of sequences.

To achieve this goal robustly, systems can make use of multiple cues, the cues chosen in this work are *motion, colour, contrast and prediction*, although cues based on stereo or more advanced models of texture could readily be incorporated. The use of multiple cues not only provides more measurements when estimating a state, but, crucially, different cues may be complementary in that one may succeed when another fails. For example segmentation based on motion or stereo tends to suffer in untextured regions, but it is precisely in such patches that colour analysis is at its most effective. A second potential cause of failure is outlying measurements, which plague many computer vision algorithms.

Overall system performance may be improved in two main ways, either by enhancing each individual cue, or by improving the scheme for integrating the information from

* Visit the website to see the movies accompanying this paper.

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2352, pp. 469–486, 2002.
© Springer-Verlag Berlin Heidelberg 2002

the different modalities. The contributions of this paper concern the latter, we do not attempt to improve on existing techniques for the individual components.

We present and compare two alternative integration techniques. The first is a probabilistic approach where the likelihood of observing the data given a model of each layer is computed before Bayes rule is applied to classify the pixels. The second scheme is a voting method, the key difference being that each cue makes an independent decision regarding layer membership before these decisions are combined using a weighted sum. The advantage of voting in data fusion is that measurements drawn from very different spaces can easily be combined [6]. With probabilistic methods more care must be taken in designing the model of each so that the different cues combine in the desired manner. However, therein also lies their strength since it forces the user to be explicit in terms of designing the model and specifying what parameters are used and what assumptions are made. As in [2,21,23] the final output from both algorithms is a measure of belief of membership to a layer and is a continuous number between 0 and 1. The sum of these over all layers must be equal to 1 at all pixels. Layers are allowed to compete for pixels rather than assigning a pixel to the first layer which can describe the observations.

A contaminated mixture of Gaussians (described in Section 3) is used to model both the foreground and background layers in each modality. The use of mixture models permits a single layer to encode an *entire* object consisting of a number of distinct, homogeneous, regions. Any additional cue which can also be modelled in this manner can easily be inserted into the framework.

A key contribution of our work is that our system not only integrates information over different modalities, but also over *time* by learning models for cues based on the output of the overall algorithm in previous frames. During initialization, the motion cue provides an estimate of the segmentation, and this classification of pixels is then used to train models for other modalities via the Expectation-Maximization (EM) algorithm [8, 4]. This builds on the notion that figure-ground separation is simpler if there is a solution from previous time-steps, and it ensures temporal stability, which cannot be guaranteed by repeatedly applying traditional single-frame segmentation methods based on colour, texture and proximity. Thus the approach adopted in this paper strongly resembles the tracking paradigm where new images are searched for evidence supporting a particular hypothesis. However, our purely data driven approach is in contrast to most work with multiple cues in the tracking community which relies on prior models of at least some cues in terms of colour distributions or shape or appearance models, although [25] deals with learning non-parametric distributions for skin colour and head orientation for face-tracking. Our work is motivated largely by applications in autonomous mobile robots which may encounter a large variety of objects and where assistance by a human in bootstrapping is at best inconvenient and at worst impossible. This scenario further constrains our approach to being sequential and feasible for real-time implementation. The models are adapted over time, initially so that more data is used to train the models, and subsequently to handle changes such as illumination variations. In terms of Clark and Yuille's classification scheme for data fusion algorithms [7], our techniques are recurrent strongly-coupled fusion since the outputs of modalities affect other modalities via the training mechanism.

Both schemes incorporate reliability measures. In the Bayesian approach these are modelled via hyper-priors on the model parameters, whereas in the voting algorithm this capability is provided by the weights in the final integration stage. The reliabili-

ties are used to suppress cues while their models are being learnt, and to incorporate self-evaluation outputs from the different modalities. For instance, if little independent motion is detected, motion computation, and subsequently motion segmentation, could be unreliable. A further use of reliabilities is based on history; if a cue has been unreliable in recent frames, then it should be given less weight than the more reliable modalities also in the current frame [26]. In the future we also envisage using reliabilities to express belief concerning pixels being occluded in motion and stereo algorithms.

Previously a number of different techniques have been used to assign pixels to layers according to some coherent property. Particularly well studied is motion segmentation both in 2D [11,5,27,2], and in 2.5 and 3D [10,17] whereas [23] used depth information. With regard to combing cues, in independent work Khan and Shah [12] proposed a framework with many similarities to ours. They use optic flow, colour and a spatial cue (similar to our prediction cue) and train colour distributions. Each region is only assigned a single colour. Thus the foreground and background risk being split into several regions rather than being assigned to a single foreground or background layer. They weight cues depending on reliability measures for optic flow. Other work closely related to ours is that of Tao *et al.* [21] who computed posterior estimates of pixel assignment as well as model parameters for real-time vehicle tracking and segmentation from airborne cameras using motion, appearance and shape information. Altunbasak *et al.* [1] iteratively enhanced the output of a motion segmentation algorithm with results from a colour or intensity based segmentation step. After motion segmentation, entire regions from the colour or intensity segmentation are assigned to layers, and then the motion parameters are recomputed. In [22] colour segmentation was used in a similar manner to improve stereo depth maps. Nordlund and Eklundh [15] combine motion and stereo by detecting and tracking peaks in histograms in a combined motion-stereo space. Global techniques such as normalized cuts [19] can incorporate any kind of information available by setting the edge weights in a graph, but as yet this and related algorithms are slow. For single images Malik *et al.* [14] provide an illuminating discussion of the integration of texture and contour information whereas Belongie *et al.* [3] considers fusion of texture and colour cues. A number of these techniques are based on the EM algorithm, or a generalization of it, to provide mathematically rigorous algorithms [2,3,22,23].

The segmentation problem is closely related to tracking, and our work is also inspired by literature in that field. Whereas in tracking the goal is usually to recover the image of a fixed point (x, y) on the tracked object, in segmentation a dense map of membership to each layer is required. Impressive real-time tracking has been demonstrated by Triesch and von der Malsburg [26], Kragic [13] and Spengler and Schiele [20] who recover saliency maps from each independent cue and combine these using a weighted sum. The coordinates of the fixation point are then given by the peak in the combined saliency map. The weights are continuously updated according to the general consensus. The voting technique presented in the current paper is based on this integration scheme. In figure-ground segmentation the problem may be posed in much the same manner, except that in the final stage the weighted sum of saliency maps is itself the output, or alternatively a binary mask derived from it by a simple thresholding technique. Furthermore, while in tracking it is frequently just the foreground object which is modelled, segmentation is most naturally posed as finding which of two or more layers best explains the measurements at each pixel. In this paper we demonstrate how reweighting cues in voting also applies to segmentation, and we present an adaptation suitable for the

Bayesian integration scheme. For probabilistic tracking with multiple cues, Toyama and Horvitz [24] and Sherrah and Gong [18] use Bayesian networks to incorporate reliability measures. The former work uses discrete variables whereas the latter generalizes the technique to continuous variables while maintaining computational tractability by using Gaussian mixture models. Also of note is the co-inference approach of Wu and Huang [28] where shape and colour distributions are adapted online with impressive results.

The rest of the paper is organized as follows. The next two sections discuss the voting and probabilistic fusion schemes and the choice of probability density functions before Section 4 presents the inclusion of reliability measures. Section 5 describes the particular cues used in this work while Section 6 outlines how their models are automatically initialized and subsequently updated. Experimental results are presented in Section 7. Conclusions are drawn and avenues for future research discussed in Section 8.

2 The Integration Schemes

2.1 Weighted Voting

The key feature of the voting method is that each cue makes a decision regarding layer membership independent of all other cues before the results are combined. Let the likelihood of the observations $Z_{i,k}$ from cue k at pixel i given the model $M_{j,k}$ of layer j be denoted by $p_{i,k}(Z_{i,k} \mid M_{j,k})$. Then for each cue, the posterior probability of layer membership is found using Bayes' Rule

$$p_{i,k}(j \mid Z_{i,k}) = \frac{p_{i,k}(Z_{i,k} \mid M_{j,k})\, p(j)}{\sum_j p_{i,k}(Z_{i,k} \mid M_{j,k})\, p(j)} \tag{1}$$

where $p(j)$ is the prior probability of layer j. These priors may be used to express belief concerning the size of the foreground relative to the background. The cues are combined to obtain an overall value for degree of membership to a given layer using a weighted sum of the responses from each cue

$$\text{score}_i(j) = \sum_k w_k\, p_{i,k}(j \mid Z_{i,k})$$

where these scores are normalized to sum to one over all layers. The weights are used to express the confidence in each cue. This data fusion scheme is also known as a *linear opinion pool*.

An important property of voting is that the influence of a single cue is bounded. Considering equation (1) it matters little if the ratio $p_{i,k}(Z_{i,k} \mid M_{j,k})/p_{i,k}(Z_{i,k} \mid M_{j',k})$ is 10 or infinite for all layers $j' \neq j$ since the posterior degree of membership from that cue alone subsequently varies only between 0.9 and 1. Regardless how strongly the evidence from a single cue supports a given layer, this layer assignment may still be defeated if two or more other cues vote in favour of an alternative hypothesis, even if they have significantly lower ratios $p_{i,k}(Z_{i,k} \mid M_{j,k})/p_{i,k}(Z_{i,k} \mid M_{j',k})$. This can be a desirable property since it provides some robustness to outliers.

It is worthwhile commenting on how the final result is influenced by uncertain cues, that is cues with equal $p_{i,k}(j \mid Z_{i,k})$ for all layers j at a certain pixel. This is best illustrated

with an example. Assume that there are two cues, k and k', and two layers j and j' and that the measurements yield $p_{i,k}(j \mid Z_{i,k}) = 0.8, p_{i,k}(j' \mid Z_{i,k}) = 0.2, p_{i,k'}(j \mid Z_{i,k'}) = 0.5$, and $p_{i,k'}(j' \mid Z_{i,k'}) = 0.5$, implying that cue k' is completely uncommitted to either layer. Combining the cues assuming equal weights gives $\mathrm{score}_i(j) = 0.65$ and $\mathrm{score}_i(j') = 0.35$. The uncertain cue k' "blurs" the final response, although a binary mask derived as $\arg\max(j)$ remains unchanged. Similarly one may verify that when all cues provide the same output, e.g. $p_{i,k}(j \mid Z_{i,k}) = 0.8$, $p_{i,k}(j' \mid Z_{i,k}) = 0.2$, $p_{i,k'}(j \mid Z_{i,k'}) = 0.8$, and $p_{i,k'}(j' \mid Z_{i,k'}) = 0.2$, the final scores also take these values, $\mathrm{score}_i(j) = 0.8$ and $\mathrm{score}_i(j') = 0.2$.

2.2 A Bayesian Approach to Segmentation

From a probabilistic point of view, the voting algorithm described above is invalid, as will be highlighted below while presenting our second approach to cue integration.

When deriving the probabilistic integration scheme we make the assumption that the observations from the cues are independent, and so the total likelihood of the observations given the combined model $M_j = \{M_{j,1} ... M_{j,k}\}$ over all cues k for layer j at pixel i is

$$p_i(Z_i \mid M_j) = \prod_k p_{i,k}(Z_{i,k} \mid M_{j,k}).$$

A posterior estimate of layer membership is then given by Bayes' Rule as

$$p_i(j \mid Z_i) = \frac{\prod_k p_{i,k}(Z_{i,k} \mid M_{j,k})\, p(j)}{\sum_j \prod_k p_{i,k}(Z_{i,k} \mid M_{j,k})\, p(j)}. \tag{2}$$

This is also known as an *independent opinion pool*.

The important difference between this and the voting scheme is that the observations from different cues are now combined *before* layer membership is evaluated as opposed to computing layer membership for each cue and only then combining the results.

Upon consideration of equation (2) it is easily seen that completely uncertain cues have no effect on the posterior estimate, in contrast with voting which blurred the layer assignment scores. Repeated measurements reinforce each other: with uniform priors and $p_{i,k}(Z_{i,k} \mid M_{j,k}) = 0.8$, $p_{i,k}(Z_{i,k} \mid M_{j',k}) = 0.2$, $p_{i,k'}(Z_{i,k'} \mid M_{j,k'}) = 0.8$, and $p_{i,k'}(Z_{i,k'} \mid M_{j',k'}) = 0.2$, the posterior layer assignments are $p_i(j \mid Z_i) \approx 0.941$ and $p_i(j' \mid Z_i) \approx 0.059$.

If a measurement coincides exactly with a sharp peak in one cue's probability density function (pdf), the likelihood ratio $p_{i,k}(Z_{i,k} \mid M_{j,k}) / p_{i,k}(Z_{i,k} \mid M_{j',k})$ can be very large, and equation (2) shows that this cue can completely overpower the remaining cues, unless they too have measurements coinciding with equally sharp peaks. Sharply peaked distributions can arise when a cue is trained from a compact cluster of data. Hence different cues can have very different looking pdf's, and it is by no means trivial to ensure that the resulting interaction *between* the cues is the desired one. Voting avoids this problem by to a great extent only permitting interplay between competing layer models, and not so much between the models for different cues. The flip side of the coin is that it is difficult to ascertain precisely what that interaction between cues is.

3 Choosing Probability Distribution Functions

In our work we assume that the likelihood probability density functions may be described by a contaminated mixture of H Gaussians

$$p_{i,k}(Z_{i,k} \mid M_{j,k}) = \frac{p_0}{A} + (1 - p_0) \sum_{h=1}^{H} \alpha_h \mathcal{N}(\mu_h, \square_h) \ ,$$

$$\mathcal{N}(\mu_h, \square_h) = \frac{1}{(2\pi)^{d/2} |\square_h|^{1/2}} e^{-\frac{1}{2}(Z_{i,k}-\mu_h)^\top \boldsymbol{\Sigma}_h^{-1}(Z_{i,k}-\mu_h)} \tag{3}$$

where d is the dimensionality of the measurement space and each Gaussian is described by its mean μ and covariance matrix $\square$. The Gaussians are weighted by factors α_h where $\sum_h \alpha_h = 1$. $|\cdot|$ denotes the matrix determinant. The mixture of Gaussians is augmented by a uniform distribution which integrates to p_0 over the hyper-volume A of the domain of the d-dimensional measurements $Z_{i,k}$. Note that the probability density function integrates to unity as required.

The contamination with a uniform distribution is crucial. Without it, if no layer fits a measurement well such that $p_{i,k}(Z_{i,k} \mid M_{j,k})$ is very small for all j, the ratio of likelihoods for that cue for layers j and j', $p_{i,k}(Z_{i,k} \mid M_{j,k})/p_{i,k}(Z_{i,k} \mid M_{j',k})$, can still be several orders of magnitude. Since it is unreasonable that the posterior probabilities should be biased so much by an unreliable measurement (an outlier) we require probability density functions with heavy tails, and adding the uniform distribution has indeed the desired effect. Thus p_0 represents the expected proportion of data poorly described by the Gaussians. In our experiments p_0 was set to 0.1.

Contaminations were also used in [23,21].

4 Incorporating Reliabilities

One of the main advantages of voting is that the weights provide a simple way of expressing confidence in the output of each cue. Below we review how the weights may be continuously updated before discussing how such features may be included in probabilistic cue integration.

4.1 Self-Organized Adaptation in Weighted Voting

In their tracking paper, Triesch and von der Malsburg [26] proposed a scheme *Democratic integration* for adapting the weights online by investigating which cues were in agreement with the final result. A score q_k is computed for each cue based on the rms values, $\bar{E}$, of the difference between the membership maps computed by that cue alone and the membership maps obtained by fusing the outputs from all cues [1]. In our implementation the scores are computed as $q_k = e^{-a\bar{E}}$, with $a = 2$, since this gives highest scores for zero error and tails off suitably towards the worst rms error of 1. These q_k are

[1] This is one of the error measures suggested by Triesch and von der Malsburg, but their smaller state space permits them to use a simpler scheme in their tracking implementation.

then normalized to sum to one over all k, $\sum_k q_k = 1$. Assuming that also the weights w_k used in the integration sum to one, the weights are adapted according to

$$\tau \dot{w}_k = q_k - w_k \tag{4}$$

where τ is a parameter which determines the rate of change of the weights. The normalization of w_k and q_k assures that the updated weights also sum to one.

4.2 Incorporating Reliabilities into the Probabilistic Fusion Scheme

A rigorous scheme for expressing belief in the validity of cues is through hyper-priors on the model parameters. We choose to apply hyper-priors to the means μ in the Gaussian mixture models in equation (3). Dropping the subscript h, let one such component be $\mathcal{N}(\mu, \square)$. The mean μ of this distribution is itself assumed to be a Gaussian random variable centred on its current estimate μ^* and with covariance matrix $\square^*$, in summary $p(\mu|\mu^*, \square^*) = \mathcal{N}(\mu^*, \square^*)$. The parameter μ can be eliminated by marginalization

$$p(Z|\mu^*, \square, \square^*) = \int p(Z|\mu^*, \square, \square^*, \mu)\ p(\mu|\mu^*, \square, \square^*)\ d\mu$$

using the property that a convolution of two Gaussians $\mathcal{N}(\mu_1, \square_1)$ and $\mathcal{N}(\mu_2, \square_2)$ is itself a Gaussian $\mathcal{N}(\mu_1 + \mu_2, \square_1 + \square_2)$. Computationally the hyper-priors are thus applied simply by replacing each Gaussian $\mathcal{N}(\mu, \square)$ in equation (3) by a flatter distribution $\mathcal{N}(\mu, \square + \square^*)$.

What remains to be established in the model is what $\square^*$ should be; assuming that we have a measure of reliability R varying from 0 to 1 for each cue, we require a mapping $\square^*(R)$ such that $\square^*$ grows as R tends to 0, and such that $\square^*$ is small when R is close to 1. A difficulty in defining this mapping is that for those cues which are trained online we lack precise advance knowledge of the pdf's covariance, $\square^*(R)$ must be suitable for both relatively flat and sharp distributions. Although a number of functions $\square^*(R)$ could be suitable, our current implementation uses

$$\square^*(R) = a(1-R)\square + b\left(\frac{1}{R} - 1\right)\mathtt{I} \tag{5}$$

where $\mathtt{I}$ is the identity matrix and $\square$ is the covariance of the original Gaussian, as before. The first term is linear in $\square$, expressing the belief that the flatter a peak is, the less certain is its location. The second term is independent of $\square$ and ensures that also extremely sharp distributions are blurred. Suitable values for the constants a and b are chosen in advance for each cue and are the same in all sequences in our experiments. The values are also set to express prior belief in the reliability of each cue. For instance, we believe colour to be more trustworthy than the simple texture descriptor, and we can ensure that this is reflected in experiments by setting a and b appropriately. The values of these parameters for each cue are given in Section 5. In fact we used $a = 5$ for all cues and only varied b.

In the preceding discussion we neglected the uniform component in the contaminated Gaussian mixture models. However, convolving a uniform distribution with a Gaussian gives a good enough approximation to a uniform distribution that the the pdf

$p(Z|\mu^*, \square, \square^*)$ is still a contaminated Gaussian. This is described in more detail in the long version of this paper [9]. There we also consider applying hyper-priors not to the means but to the *covariances* in the likelihood pdf's. With colour, for instance it is not unreasonable to assume that the uncertainty does indeed lie in the *location* of the peak in hue-saturation space, but with other cues we in fact know the mean exactly (e.g. as 0 or 1) and it would be more appropriate to assign hyper-priors to the covariance matrices in the Gaussians. The pdf after marginalization takes a much less simple form, but can often be well approximated by a contaminated Gaussian.

The long version of this paper also contains a comparison between hyper-priors and *logarithmic opinion pooling* as adopted by Khan and Shah [12] by investigating how this procedure alters the pdf's in the contaminated Gaussians. We have implemented logarithmic opinion pooling and found that it works well and has the added benefit of not requiring further parameters. The results will not be reported in this paper.

4.3 Self-Organized Weighting of Cues in the Probabilistic Technique

A simple adaptation of Triesch and von der Malsburg's self-organized reweighting scheme has proved effective for the Bayesian integration approach. The rms error $\bar{E}$ is computed for each cue k and converted to a score q_k between 0 and 1 as in Section 4.1. New values for the reliabilities R_k are then given by equation (4) with the R_k replacing the weights w_k, but now the scores and reliabilities are *not* normalized to sum to one since this would cause a different amount of blurring of the Gaussian mixture models depending on how many cues are present[2].

When computing the scores we use the contaminated Gaussian mixture models *without* the blurring caused by applying the hyper-priors. This places an extra computational burden on the system as all likelihoods must be computed twice, but it cannot be avoided since computing a score for a completely blurred model is meaningless.

5 Implementation Details: The Individual Cues

Analogous to Clark and Yuille's classification scheme for weak and strong coupling of cues in data fusion [7], we distinguish between cues which operate independently of all other cues, and those which are coupled to others through the training scheme.

5.1 Motion

The only independent cue in our system is motion. Currently we use Black and Anandan's algorithm (and code) [5] to fit planar affine motion models between temporal pairs of images. In this direct and robust approach the dominant motion is first computed and pixels which satisfy this model are removed. The process may be repeated for further layers by fitting affine motion models to the remaining data. While Black and Anandan stop here and report binary segmentation masks, we instead allow layers to compete for the pixels. This is done by defining a likelihood function for each layer j as

[2] Even so, there is still some reliance on the number of cues since the error $\bar{E}$ will in general increase with the number of cues.

$$p_{i,k}(Z_{i,k} \mid M_{j,k}) = \frac{p_0}{A} + (1 - p_0)\mathcal{N}(I_i^t - I_{i,j}^{t-1}, \sigma_m^2),$$

where $I_i^t - I_{i,j}^{t-1}$ is the greyscale difference between pixels i from the current image and those from the image from the previous time-step aligned with the current image according to the recovered motion model. σ_m is a non-adaptive parameter. The dominant motion is assumed to belong to the background layer.

For the background layer we also perform a temporal integration following Irani *et al.* [11] by comparing the current image with an internal representation of the layer based on the previous images and recovered motion models. This scheme reduces the effect of random noise in the greyscale images, and blurs out independently moving objects. We found that this scheme was not always appropriate for foreground layers since our sequences do not contain rigid motion, in which case it is somewhat arbitrary what motion model is computed. As a result the appearance model becomes blurred and harms rather than improves the motion computation in subsequent frames.

In addition to letting the general consensus dictate the reliability R of motion we multiply R by a number r between 0 and 1 which is evaluated by counting the fraction f of pixels which were used to compute motion models for the foreground layers [18]. r is given by $\min(1.0, f/\beta)$ where β is a constant of 0.05 in our experiments.

Having normalized the greyvalues to lie between 0 and 1, the parameters in equation (5) for applying hyper-priors are $a = 5, b = 0.1$ for this cue.

5.2 Colour

Our colour segmentation cue is based on the work of Raja *et al.* [16] who model colour distributions with adaptive Gaussian mixture models in hue-saturation space. Although that paper was mainly motivated by tracking, they do also present some segmentation results. The key difference in our system is that models are initialized online rather than offline, as will be discussed in Section 6.

Pixels lacking colour, that is their saturation falls below a threshold, are assigned zero reliability. Similarly, saturation of the CCD causes inaccurate estimation of colour components, and so such pixels are also removed.

The constants in equation (5) are $a = 5, b = 0.001$.

5.3 Image Contrast

As a very simple texture cue, we compute the standard deviation of greyscale values within 7×7 patches at two scales and form a feature vector by concatenating the contrast response over the scales. A contaminated mixture of two Gaussians is initialized and subsequently adapted online. This cue was also used in [26,13], though only at a single scale, and the model was given in advance.

Responses from larger filter banks could be used to replace or complement the contrast measure. However, as feature vectors grow in size the computational cost increases rapidly unless filter responses can be considered independent of each other.

The constants in equation (5) are $a = 5, b = 0.005$ are for this cue.

5.4 Prediction

Akin to what has been done in tracking [26,13], the segmentation masks from the previous frame are used to predict the layer assignment in the current frame. The background mask is warped according to the affine motion model recovered in the previous frame. We do not warp the foreground layer since we have less faith in the foreground motion model due to non-rigid movement. This mask is then blurred by a 9×9 Gaussian mask with $\sigma = 4$ giving a value Z between 0 and 1 at each pixel. The likelihood of a pixel belong to layer j is modelled with a contaminated normal distribution on the prediction value Z

$$p_{i,k}(Z_{i,k} \mid M_{j,k}) = \frac{p_0}{A} + (1 - p_0)\mathcal{N}(Z - 1, \sigma_p^2)$$

where $A = 1$ and the parameter $\sigma_p = 1/\pi$ is chosen such that the the Gaussian tails off suitably towards 0 as Z goes to 0. We acknowledge that the assumption that the observations from this cue is independent of the others is somewhat bold. The spatial cue of Khan and Shah [12] is, in fact, very similar to this[3].

For this cue we used $a = 5, b = 0.5$ in equation (5).

6 Model Initialization and Adaptation

The two cues which require learning and subsequent adaptation are colour and contrast.

When starting the algorithm we first run the motion segmentation algorithm for five frames so that the appearance model for the background layer settles. Then, the layer classification from motion segmentation is used to train the colour model using the Expectation Maximization (EM) algorithm [8,4]. Prediction is introduced after an additional ten frames, and does not not require training. After a further ten frames a contrast model is trained also using EM. The reason for the wait is that the contrast measure risks converging to a poor solution if the training data straddles object boundaries, as is often the case with a segmentation obtained purely by motion.

The colour and contrast mixture models are trained only using data which are classified as foreground or background with a degree of certainty over a threshold of 0.6. The update equations for EM for Gaussian mixture models can be found in [4,9] and are of closed form for Gaussian mixture models. We have also tried using all data weighted by the current segmentation mask. In effect we then have two sets of missing variables, the first concerns layer membership while the second set describes the assignment to the different components in each Gaussian mixture model. Although this is appealing from a probabilistic viewpoint, we found that EM frequently converges more slowly or fails to converge at all if too many pixels are uncertain during training. Therefore we do not use this weighting in the experiments reported in this paper. Similarly we prefer not to include the uniform component in the pdf when training mixture models because too much data tends to get assigned to the uniform distribution. Furthermore, the contamination was really introduced to provide robustness to outlying measurements, well-spread data should instead be assigned to a Gaussian with high covariance.

[3] Interestingly, the manner in which they pose the cue as the probability of measuring the *spatial location* given the previous mask makes it possible to justify the independence of observations from this cue and those from colour and optic flow.

Due to EM's well-documented tendency to converge to local minima, it is important to provide the algorithm with a good starting point. This we achieve with k-means clustering which in turn is initialized by placing cluster centres as follows: with colour, initial estimates are evenly spaced on the 0.6 saturation circle in hue-saturation space whereas the contrast distributions are initialized at regular intervals along the diagonal direction $(1, 1)$ of its measurement space. The diagonal is chosen because most data lies close to this line; the contrast measures at two scales are highly correlated.

Currently we make no attempt to automatically determine the best number of components in the mixture models, although standard techniques such as cross-validation or minimum description length could be used. However, we note that having too many components does not cause problems since components not supported by the data shrink to zero α_h and do not have any further effect. For each layer we currently use a mixture of four Gaussians for colour distributions and two for the contrast cue.

The approach taken in this paper for updating the colour and contrast distributions follows that of Raja *et al.* [16]. This procedure is performed separately for colour and contrast. The current cue model, image data and layer assignments are used to compute a new estimate of the model by effectively performing an iteration of the EM algorithm. The degree of membership (i.e. the hidden variables) for each component is computed in an E step, and these values are then used to recompute the model parameters. The updated model is computed as a weighted sum of this and previous models over a fixed time window. The relevant equations may be found in [16,9]. Again, we have extended this adaptation scheme to incorporate the uniform distribution, but do not use it here since we find that better results are obtained when neglecting this component.

During training and adaptation we place a lower bound on the eigenvalues of the covariance matrices to avoid numerical instabilities caused by components shrinking to zero covariance.

7 Experimental Results

As yet we only attempt to find a single foreground layer in addition to the background. The complete algorithm runs at approximately 1/3 Hz on a 1.5 GHz Pentium 4 for 164×132 images. Of this 2 seconds is used for motion computation while the remaining cues and the fusion process require 0.5 seconds in total. A further half second is needed for self-organized reweighting of cues in the Bayesian method since all likelihoods must be recomputed. The one-off cost of training initial colour and contrast models took 0.5-1 and 0.2-0.5 seconds respectively. The "tennis" sequence ran at 1/13 Hz due to its larger 348×276 images. There is significant scope for speeding up this code.

Videos of all results can be viewed from the webpage
`http://www.nada.kth.se/~hayman/ECCV2002`.

Experiment 1: The "tennis" sequence. Both the probabilistic and voting integration techniques were tested on 850 frames of the "tennis" sequence in which a player hits some shots while the camera pans. The original sequence and the segmented foreground object are shown in Figure 1. Here the Bayesian integration technique was used. Reliabilities were incorporated via hyper-priors, as described in Section 4.2 and the method

of Section 4.3 was used to recompute cue reliabilities in the probabilistic algorithm according to general consensus. Every 100th frame is shown, and we do not show results from the first frames when the models are being trained.

Figure 2 shows the contributions of the different cues. Motion does not provide information in the large textureless regions of the image. Colour successfully distinguishes between the skin of the tennis player and the green background and red playing surface. The colour cue provides no information, unfortunately, for the player's white T-shirt since these pixels are rejected for colour analysis due to their lack of saturation. However, the contrast cue does a surprisingly good job of latching onto the T-shirt and also her shoes, and as a result the final segmentation is very satisfactory. The pixels on the court which are incorrectly assigned to the foreground straddle the red surface and white line, and the resulting colour is indistinguishable from flesh colour. Results are good despite the motion of the foreground object being far from rigid.

Experiment 2: The "mother and daughter" sequence. A second experiment illustrates the self-organization of cues with both Bayesian and voting integration schemes on 240 frames of the "mother and daughter" MPEG-4 test sequence. Figure 3 shows the original sequence and segmentation results. During initialization the mother's head and hand move. At first the segmentation is poor while only motion is used, but results improve as models for the remaining cues are trained and their associated reliabilities increase. The mother is successfully segmented out from the background throughout the sequence. Also the face of the daughter is included since it shares the same colour characteristics as her mother's face. The segmentation of the daughter's hair and clothing is more erratic, largely because it is unclear whether these regions were used for the initial training of colour and contrast models. After a while the hair is assigned to the background, apart from one small region on the top of her head which bears a strong resemblance both in colour and texture to the mother's hair. Rather than showing the final masks, Figure 3 illustrates the layer assignment where bright regions indicate pixels which are assigned to the foreground with a high degree of certainty. Results are good with both integration techniques, but note that voting gives a graded solution while the output from the Bayesian method is almost binary, as expected. The sharp results from the Bayesian scheme are more reliable for generating thresholded masks whereas the voting scheme in some sense gives more meaningful results by clearly indicating regions which are assigned with less certainty due to disagreement of the cues.

Figures 4a and b show plots of the evolution of the reliabilities and 4c and d the outputs from each cue for a particular frame. Especially the Bayesian scheme successfully suppresses the contrast cue which is in disagreement with the general consensus.

The erratic motion in this sequence caused the motion cue to frequently give poor results, but this did not have much effect on the final segmentation since the cue's associated reliability was low. In an unreported experiment we confirmed that failing to lower the reliability of the motion cue gave a dissatisfactory overall segmentation.

Experiment 3: The "silent" sequence. A final demonstration of our self-organized probabilistic system is given on the 450 frame "silent" MPEG-4 test sequence in Figure 5. The motion is very non-rigid, and an additional challenge is that the background is more cluttered, though static. Even so, the colour and contrast models converge correctly, although a few areas in the background are misclassified since they share the same colour

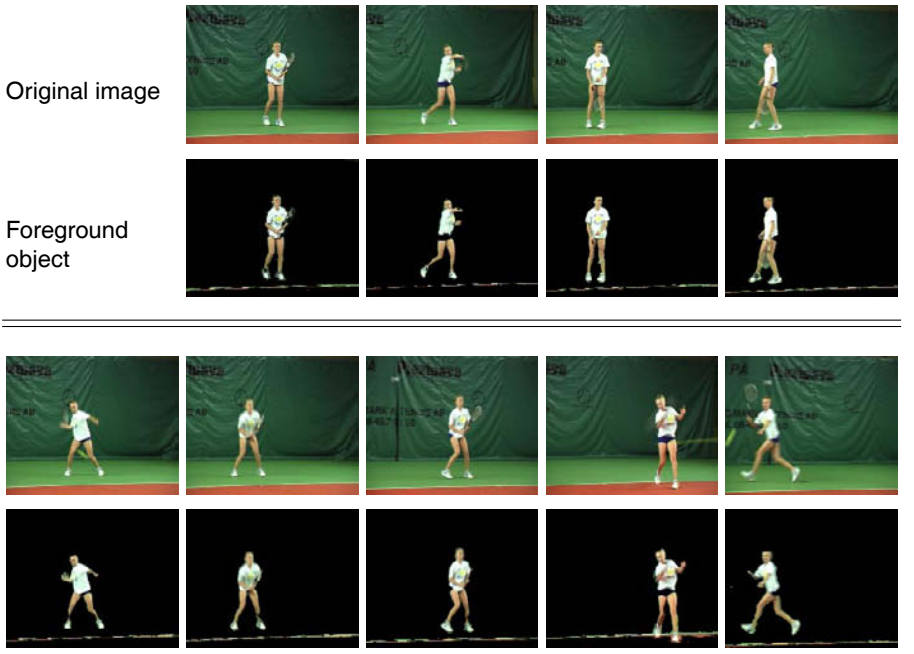
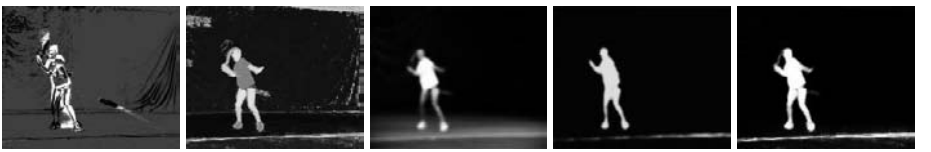


Fig. 1. Every 100th frame of the original "tennis" sequence and segmentation results from the Bayesian integration method. A threshold mask from the foreground layer is applied to the original images.



(a) Motion (b) Colour (c) Contrast (d) Prediction (e) Combined

Fig. 2. Results from the different cues at a frame in the "tennis" sequence. The bright regions correspond to pixels which are classified as foreground with high certainty. The different cues complement each other. For instance, the colour cue is completely uncertain in the player's white T-shirt, but the contrast cue correctly assigns this region to the foreground.

characteristics as the lady's top. The self-organized voting scheme also performed well, but space limitations prevent us from recording those results here.

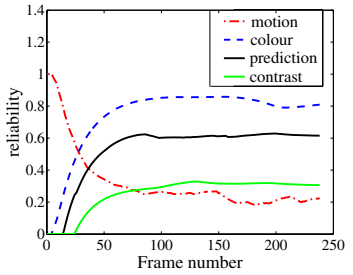
8 Conclusions

In this paper we have demonstrated how the use of multiple cues improves segmentation results on video sequences. Even using fairly basic techniques for each individual cue, the *combined* segmentation masks are very accurate, and the output is robust to one

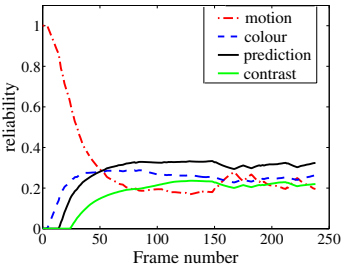
Fig. 3. Segmentation results from the "mother and daughter" MPEG-4 test sequence. The top rows show the original sequence whereas the middle and bottom rows show the belief of membership to the foreground for the Bayesian and voting integration schemes respectively. Bright areas indicate a higher degree of certainty. After the first few frames results improve as additional cues are trained and their reliabilities increase. Both integration schemes used self-organized weighting of cues according to general consensus.

or more of the modalities failing or being completely uncertain. Voting and Bayesian integration techniques were implemented and evaluated. Both performed well.

An important aspect of this work in comparison with much previous research in tracking is that colour and contrast distributions were learnt online rather than offline, thus no human interaction is required. During initialization, regions which are classified from the motion cue are used to train Gaussian mixture models for the remaining cues. These models are then used for segmentation in subsequent frames. Thus the approach is entirely data driven.



(a) The evolution of the cue reliabilities for Bayesian cue integration



(b) Cue reliabilities in the voting method

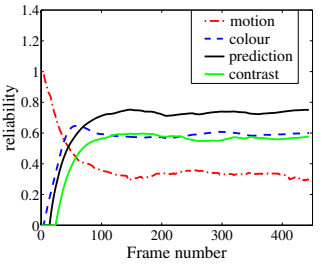
(c) Cue combination by Bayesian integration

(d) Cue combination by voting

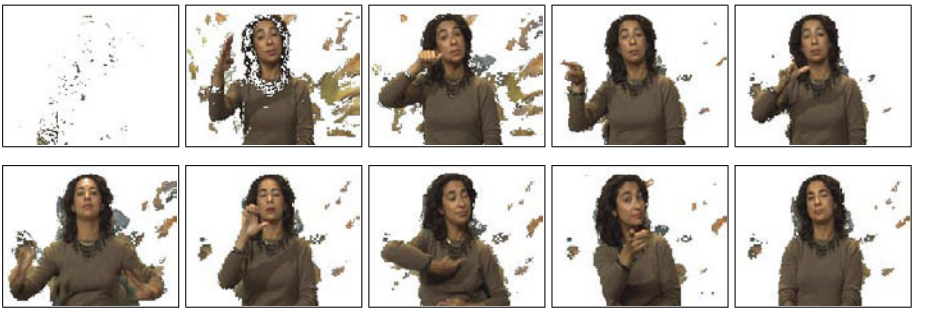
Fig. 4. Results from the Bayesian and voting schemes for cue integration for the "mother and daughter" MPEG-4 test sequence: (a) and (b) show that the contrast cue is assigned low reliability, particularly so in the probabilistic method. The lack of influence of this cue on the final result is demonstrated in (c) and (d) for the final frame of the sequence. The mother and child's faces are still accepted as foreground. The uncertain cue is completely suppressed in the Bayesian technique and blurs the response in the voting method.

In our future work we intend to incorporate stereo into the framework. Having more cues will allow us to better investigate the outlier rejection properties of voting. Furthermore, since objects close to the camera are likely to be foreground, stereo provides an additional mechanism for the initialization of other cues. Although our work has demonstrated that automatic bootstrapping of cues is certainly possible, we find that it is difficult to recover if the initialization is poor. Therefore *reinitialization* needs to be addressed, also because it is desirable for additional Gaussians to be included in the mixture models if a new colour or texture becomes visible. Reinitialization could

(a) An original still image



(b) The evolution of the cue reliabilities



(c) The foreground person segmented out from the background. As before, results are poor in the first few frames before models are trained for cues other than motion.

Fig. 5. Segmentation results from the "silent" MPEG-4 test sequence.

also prevent colour and texture models from getting trapped; a property of bottom-up adaptive systems like ours is that if one cue dominates due to a sharply peaked pdf, other cues adapt also to that data, which is fine if the strong cue is correct, but dangerous otherwise since parts of the foreground, say, may become permanently attached to the background. Currently, motion is the only cue which can prevent such problems, but interestingly this cue was assigned low reliability in the experiments presented here due to non-rigid motion.

Additional extensions concern incorporating a more detailed texture model and permitting additional foreground layers to be introduced while the system is running.

Acknowledgments. The authors would like to express their gratitude to Michael Black for the use of his code for parametric motion computation [5]. EH was funded by a Marie Curie Fellowship of the European Community programme "Improving Human Research Potential" (Contract no. HPMFCT-2000-00650).

References

1. Y. Altunbasak, P.E. Eren, and A.M. Tekalp. Region-based parametric motion segmentation using color information. *Graphical Models and Image Processing*, 60(1):13–23, Jan 1998.
2. S. Ayer and H. Sawhney. Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding. In *Proc. Int. Conf. on Computer Vision*, pages 777–784, 1995.

3. S. Belongie, C. Carson, H. Greenspan, and J. Malik. Color- and texture-based image segmentation using the Expectation-Maximization algorithm and its application to content-based image retrieval. In *Proc. Int. Conf. on Computer Vision*, pages 675–682, 1998.
4. J. Bilmes. A gentle tutorial on the EM algorithm and application to gaussian mixtures and Baum-Welch. Technical Report TR-97-021, International Computer Science Institute, Berkeley, CA, April 1997.
5. M.J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow-fields. *CVIU*, 63(1):75–104, January 1996.
6. C. Bräutigam, J.-O. Eklundh, and H.I. Christensen. A model-free voting approach for integrating multiple cues. In *Proc. European Conf. on Computer Vision*, 1998.
7. J. J. Clark and A. L. Yuille. *Data Fusion for Sensory Information Processing Systems*. Kluwer Academic Press, 1990.
8. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.*, 39 B:1–38, 1977.
9. E. Hayman and J. O. Eklundh. Figure-ground segmentation of image sequences from multiple cues, 2002. The long version of this conference paper is available at `http://www.nada.kth.se/~hayman/ECCV2002`.
10. M. Irani and P. Anandan. A unified approach to moving object detection in 2D and 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6), 1998.
11. M. Irani, B. Rousso, and S. Peleg. Computing Occluding and Transparent Motions. *International Journal of Computer Vision*, 12(1):5–16, 1994.
12. S. Khan and M. Shah. Object based segmentation of video using color, motion and spatial information. In *Proc. Computer Vision and Pattern Recognition*, pages II:746–751, 2001.
13. D. Kragić. *Visual Servoing for Manipulation: Robustness and Integration Issues*. PhD thesis, Royal Institute of Technology (KTH), Stockholm, Sweden, 2001.
14. J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *Proc. Int. Conf. on Computer Vision*, 1999.
15. P. Nordlund and J.-O. Eklundh. Towards a seeing agent. In *First International Workshop on Cooperative Distributed Vision, Kyoto, Japan*, pages 93–123, 1997.
16. Y. Raja, S.J. McKenna, and S. Gong. Colour model selection and adaptation in dynamic scenes. In *Proc. European Conf. on Computer Vision*, pages 460–474, 1998.
17. H.S. Sawhney, Y. Guo, and R. Kumar. Independent motion detection in 3D scenes. *IEEE Trans. on Patt. Analysis and Machine Intelligence*, 22(10):1191–1199, October 2000.
18. J. Sherrah and S. Gong. Continuous global evidence-based bayesian modality fusion for simultaneous tracking of multiple objects. In *Proc. Int. Conf. on Computer Vision*, pages II: 42–49, 2001.
19. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Patt. Analysis and Machine Intelligence*, 22(8), Aug 2000.
20. M. Spengler and B. Schiele. Towards robust multi-cue integration for visual tracking. In *Computer Vision Systems, July 2001, Vancouver,BC*, 2001.
21. H. Tao, H.S. Sawhney, and R. Kumar. Dynamic layer representation with applications to tracking. In *Proc. Computer Vision and Pattern Recognition*, pages II:134–141, 2000.
22. H. Tao, H.S. Sawhney, and R. Kumar. A global matching framework for stereo computation. In *Proc. Int. Conf. on Computer Vision*, pages I: 532–539, 2001.
23. P.H.S. Torr, R. Szeliski, and P. Anandan. An integrated Bayesian approach to layer extraction from image sequences. *IEEE Trans. on Patt. Analysis and Machine Intelligence*, 23(3):297–303, March 2001.
24. K. Toyama and E. Horvitz. Bayesian modality fusion: Probabilistic integration of multiple vision cues for head tracking. In *Proc. Asian Conference on Computer Vision*, 2000.
25. K. Toyama and Y. Wu. Bootstrap initialization of nonparametric texture models for tracking. In *Proc. 6th European Conf. on Computer Vision, Dublin*, 2000.

26. J. Triesch and C. von der Malsburg. Self-organized integration of adaptive visual cues for face tracking. In *Proc Int, Conf. on Automatic Face and Gesture Recognition, Grenoble, France*, 2000.
27. J.Y.A. Wang and E.H. Adelson. Spatio-temporal segmentation of video data. In *SPIE: Image and Video Processing II, San Jose*, Feb 1994.
28. Y. Wu and T.S. Huang. A co-inference approach to robust visual tracking. In *Proc. Int. Conf. on Computer Vision*, pages II: 26–33, 2001.