Bayesian Regularization of the Video Matting Problem
Nicholas Apostolo and Andrew W. Fitzgibbon
Department of Engineering Science, University of Oxford
{nema,awf}@robots.ox.ac.uk

Objective
To regularize the inverse problem of Video Matting in a Bayesian framework using priors on the distribution of alpha values and the spatio-temporal consistency of image sequences.

Video Matting
Video matting is a classic image processing problem involving the extraction of a foreground object from an arbitrary background in a sequence of images. It is most prevalent in the film industry for special effects shots that require the superposition of an actor onto a new background.

The Strategy: MAP Estimation
The Bayesian formulation of the video matting problem is one of finding the maximum a posteriori (MAP) estimate of the foreground image $F$ and the alpha-matte $\alpha$ given $C$ and $B$:

$$\{F, \alpha\} = \arg \max P(F, \alpha | C, B)$$

Using Bayes rule, the posterior can be expressed as a combination of priors on $F$, $\alpha$, $C$, and $B$, and a conditional probability in $C$ and $B$:

$$P(F, \alpha | C, B) = \frac{P(C | B, F, \alpha)P(F | \alpha)P(\alpha)}{P(C | B)}$$

The MAP estimation problem can be converted into an energy minimisation by taking the negative log-likelihood of the posterior:

$$E(F, \alpha) = - \log (P(C | B, F, \alpha)) - \log (P(F)) - \log (P(\alpha)) + \log (P(C)) + \log (P(B))$$

When taking the minimum of this wrt $F$ and $\alpha$, the last two terms are constant and drop out.

The Problem: The Energy Equation
Video matting is inherently under-constrained – 7 unknowns (red) in 3 equations per colour pixel. Background extraction removes $B_x$, $B_y$ and $B_t$ to reduce the problem to one of 4 unknowns in 3 equations:

$$C(x) = \alpha(x) F(x) + (1 - \alpha(x)) B(x)$$

Given an sequence of images $C$, solve for $\alpha$, $F$ and $B$.

Two main approaches are typically used:
- Colour-based approaches exploit statistical information about the foreground and background pixel distributions to extract the foreground object from a single frame (Smith and Blinn, 1996; Chuang et al., 2001; Ruzon and Tomasi, 2000).
- Motion-based techniques exploit the temporal domain and use motion based tracking in combination with image statistics to extract the foreground object (Wexler et al., 2002; Chuang et al., 2002).

The main problem with these systems is that they often require some level of human assistance to disambiguate the problem.

The Strategy: The Energy Equation - Reconstruction Error and Constraints
Assuming Gaussian, white, additive image noise, the conditional density in $C$ and $B$ and the corresponding energy function are:

$$P(C | B, \alpha, F(x), \alpha(x)) = \eta \exp \left( - \frac{1}{2\sigma} \sum_{x} (C(x) - \alpha(x) F(x) - (1 - \alpha(x)) B(x))^2 \right)$$

$$E_{\alpha}(F(x), \alpha(x)) = ||C(x) - \alpha(x) F(x) - (1 - \alpha(x)) B(x)||^2$$

where the constant $\eta$ has been dropped and the fraction, $1/2\sigma$, is submerged by the weights $\phi$ and $\psi$ in $E(F, \alpha)$.

Two constraints are imposed on the values of $\alpha$ and $F$: $0 \leq \alpha(x) \leq 1$ and $0 \leq F(x) \leq 1$.

$$E(\alpha, F) = \sum_{p} \left[ \frac{1}{2 \sigma} (C(p) - \alpha(p) F(p) - (1 - \alpha(p)) B(p))^2 \right]$$

$$+ \phi \left( (1 - \eta) \log (\alpha(p)) + (1 - \tau) \log (1 - \alpha(p)) \right)$$

$$+ \psi \sum_{|j| \leq 1} \left[ W(p) \left( \alpha(p + x) + \alpha(p - x) \right) \right]$$

where $W(p) = \exp (-d(p,x) / \mu)$ at pixel $x$.

Alpha Distribution Prior
A Beta function is used to model the distribution of alpha values as primarily 0’s or 1’s and has the density and energy functions:

$$P(\alpha) = \frac{\alpha(1 - \alpha)^{\beta - 1}}{\Gamma(\beta)}$$

$$\log (P(\alpha)) = \log \left( \frac{\alpha(1 - \alpha)^{\beta - 1}}{\Gamma(\beta)} \right)$$

$$E_{\alpha}(\alpha) = \sum_{p} \left[ W(p) (\alpha(p + x) + \alpha(p - x) \right]$$

where $\beta$ is the prior weight high

Spatio-temporal Consistency Prior
The spatio-temporal consistency prior smooths alpha along, but not across, edges in space and time:

$$E_{\alpha}(\alpha) = \sum_{p} \left[ W(p) (\alpha(p + x) + \alpha(p - x) \right]$$

where $W(p)$ is the gradient vector of the image, $\|W(p)\|^2$ at pixel $x$. 

References