Wide Baseline Multi-View Video Matting using a Hybrid Markov Random Field

Tinghuai Wang, John Collomosse and Adrian Hilton
Centre for Vision Speech and Signal Processing (CVSSP)
University of Surrey
Guildford, United Kingdom.
Email: {tw00051 | j.collomosse | a.hilton}@surrey.ac.uk

***Abstract*—We describe a novel framework for segmenting a time- and view-coherent foreground matte sequence from synchronised multiple view video. We construct a Markov Random Field (MRF) comprising links between superpixels corresponded across views, and links between superpixels and their constituent pixels. Texture, colour and disparity cues are incorporated to model foreground appearance. We solve using a multi-resolution iterative approach enabling an eight view high definition (HD) frame to be processed in less than a minute. Furthermore we incorporate a temporal diffusion process introducing a prior on the MRF using information propagated from previous frames, and a facility for optional user correction. The result is a set of temporally coherent mattes solved for simultaneously across views for each frame, exploiting similarities across views and time.**

I. Introduction

The creative industries are turning to multi-camera systems for the acquisition of 3D assets in movie and game production. In particular, 4D performance capture (4DPC) [1] is maturing as a technology for reconstruction of moving 3D models from synchronised multiple view video. 4DPC is underpinned by the extraction of polyhedral [2] or voxel [3] visual hulls, which require segmentation of the actor's silhouette from background within each captured view – the process of *matting*. Unfortunately multi-view matting has received little attention and remains unsolved in the general case. This often necessitates use of distinctively coloured (chroma-key) backgrounds [2], [3], [4] so limiting practical scenarios for 4DPC.

This paper contributes a novel framework for multiple view video matting, encouraging both temporal coherence and spatial coherence between views. We define a Markov Random Field (MRF) across pixels in all views and super-pixel correspondences identified between pairs of views. The MRF is solved once per frame to extract the binary mattes for all views simultaneously. Our solution of all views in a single step contrasts with existing methods that either propagate labels between pairs of independently solved views [5], or build foreground appearance models across all views but then independently solve the segmentation for each view [6]. Additionally, we contribute a propagation strategy in which soft labelling constraints are carried forward in time to influence the MRFs defined over subsequent frames, so enhancing the coherence of the resulting matte sequence. Our experimental setup, typical of a 4DPC scenario, comprises 8 cameras surrounding the actor resulting in views spaced at 45 degree intervals. We instantiate our framework using colour, texture and disparity cues, and calibrated cameras to determine super-pixel correspondences between views under the epipolar constraint. However alternative cues and correspondence strategies might be incorporated and our method does not in principle require calibrated cameras.

A. Related Work

Video segmentation is a fundamental problem that has attracted considerable research effort, the majority focusing on monocular sequences. Monocular approaches perform either spatio-temporal (3D) analysis, or frame-by-frame processing with temporal information passing (2D+t). Methods in the former category either perform unsupervised clustering on pixels of a space-time video stack [7], [8], or develop hierarchical graph [9], [10] or surface [11] representations over space-time. However such approaches often under-segment fast moving objects and become computationally infeasible for videos of even moderate size, suggesting extensions to multi-view video to be intractable. The 2D+t category segments 2D frames independently and then associates [12] or tracks [13] labels over time to identify and prune sporadic regions. Minimal user-interaction is often used to correct for tracker drift [13], [14]. Our multi-view work adapts the probabilistic strategy of Wang *et al* [15] who perform graph-cut optimization on a per-frame basis, using label priors propagated forward over time on monocular video. Both our problem domain and the structure of our MRF differ significantly however due to the incorporation of multiple views and super-pixel correspondences.

Prior multi-view segmentation algorithms typically draw upon geometric constraints implicit within 4DPC, simultaneously deriving mattes whilst performing visual hull estimation. Zeng *et al* [16] iteratively carve an over-estimate of the hull to obtain a consistent set of mattes. Graph-cut [17] and convex optimization [18] have also been explored for simultaneous estimation. Weaker geometric information is incorporated by Sarim *et al* who propagate trimaps between views by matching along the epipolar line [5]. Recently Djelouah *et al* [6] proposed a geometry-free approach, building foreground and background appearance models simultaneously across views using expectation maximisation. However to perform the segmentation itself a set of independent graph-cuts (via GrabCut [19]) are used to extract a matte from each view in isolation using those models. We adopt a complementary approach, extracting mattes using a single novel graph-cut over an MRF spanning all views, using models built independently per view and propagated over time.

II. Multi-view Segmentation Framework

A 4DPC wide-baseline camera configuration typically comprises 8 calibrated high definition (HD) cameras in a circular formation about the subject. Consequently large variations in view angle occur and subject resolution is often limited by the desire to maximise capture volume. Pixel-level stereo correspondence is unreliable on such wide-baseline footage. However larger-scale similarities can be exploited by using superpixel matching to acquire the correspondence across views. Furthermore each superpixel provides a texture or colour homogeneous domain within which richer visual features can be exploited without being affected by clutter from the neighbouring superpixels. In our MRF representation, nodes are formed by both pixels and superpixels from multiple view images, and edges are the inter-view correspondences among superpixels, intra-view links among superpixels and underlying pixels, and intra-view links among each pixel lattice. This hybrid representation enables a robust multiple view segmentation which incorporates both the intra- and inter-view appearance and geometry consistency.

We propose a multiple-layer MRF $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ composed of pixels and superpixels. We formulate multiple view video segmentation as a labelling problem of assigning each node $i \in \mathcal{V}$ ($\mathcal{V} = \{\mathcal{V}_p \cup \mathcal{V}_s\}$) in frame I_t a value x from the binary label set $\mathcal{L} = \{0, 1\}$, where $\mathcal{V}_p$ and $\mathcal{V}_s$ represent pixel and superpixel nodes respectively. We seek to infer a consistent labelling $\mathrm{x} = (x_i)_{i\in\mathcal{V}_p} \cup (x_j)_{j\in\mathcal{V}_s}$ of these nodes that better separates the foreground ($l = 1$) and background ($l = 0$) area, driven by both the appearance model and camera geometry properties on the current frame and labelling priors propagated forwarded from past frames. We formulate the problem as the minimization of:

$$E(\mathrm{x}) = \sum_{i\in\mathcal{V}_p} \psi_i^p(x_i) \quad + \sum_{i\in\mathcal{V}_s} \psi_i^s(x_i) + \qquad (1)$$
$$\theta_\lambda \cdot (\sum_{i\in\mathcal{V}_p, j\in\mathcal{N}_i} \psi_{ij}^p(x_i, x_j) \quad + \sum_{i\in\mathcal{V}_s, j\in\mathcal{N}_i^*} \psi_{ij}^s(x_i, x_j)).$$

where $\mathcal{N}_v$ and $\mathcal{N}_v^*$ denote the set of neighbours $\{u \in \mathcal{V} | (u, v) \in \mathcal{E}\}$ of pixel or superpixel v respectively, and θ_λ is a parameter. The first two terms ψ_i^p and ψ_i^s are unary potentials encoding the likelihood of pixels and superpixels given the appearance models or labelling priors. The following two terms ψ_{ij}^p and ψ_{ij}^s represent pairwise terms of pixels and superpixels respectively which encourage coherent inter- and intra-view labelling.

A. Graph Construction

We first describe how the hybrid graph is constructed. Our algorithm starts by segmenting all views of the current frame to produce a set of superpixel maps (∼2500 per image) using SLIC [20]. SLIC clusters pixels in a combined five-dimensional RGB colour and image coordinate space to efficiently generate compact, near uniform superpixels. Pixels and all the generated superpixels form the nodes in the graph as shown in Fig. 1. Four types of edges exist:

1) pixel-to-pixel edges in the 4-connected neighbourhood within each image,
2) edges between the superpixels and their composit pixels within each image,
3) superpixel-to-superpixel edges, with adjacency defined on 4-connected basis, within each image's superpixel map,
4) superpixel-to-superpixel edges between neighbouring views.

The first three types of edge are defined using connectivity within each image independently, whilst the fourth type is defined by the superpixel matching between adjacent views, as detailed later in Sec. II-C. Our hybrid graph exploits both the pixel/superpixel level observations and the inter-view information using sparse superpixel matching across the multi-view video frame.

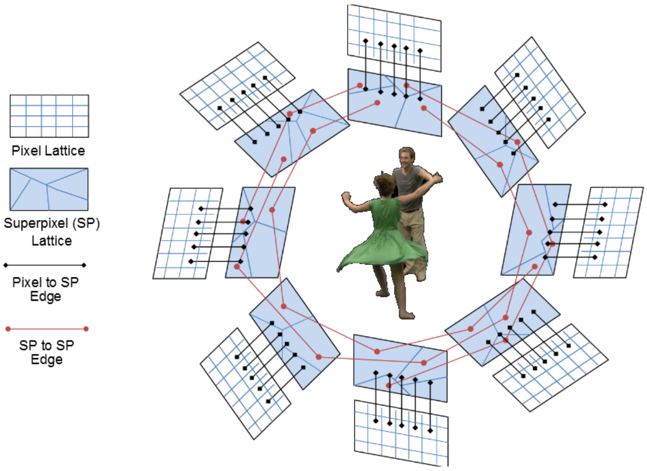

Fig. 1. Illustrating MRF connectivity between pixels and super-pixels.

B. Definition of Energy Potentials

We now describe how the the energy potentials are defined respectively in eq. 2 which consists of both unary and pairwise potentials of pixels and superpixels.

1) Pixel Unary Potentials: The pixel unary term ψ_i^p exploits the fact that different appearance homogeneous regions tend to follow different appearance models. This encourages assignment of pixels to the label following the most similar appearance model (we write the parameters of such models Θ). The unary term is defined as the negative log-likelihood of a label being assigned to pixel i. It can be computed from the appearance model for each label. To provide more discriminative power for accurate segmentation, the unary term incorporates colour and texture features as well as *prior* labelling probabilities. The pixel unary term is defined:

$$\psi_i^p(x_i) = \theta_{col}^p \psi_{col}^p(x_i) + \theta_{tex}^p \psi_{tex}^p(x_i) + \theta_{lab}^p \psi_{lab}^p(x_i). \qquad (2)$$

where θ_{col}^p, θ_{tex}^p and θ_{lab}^p are weights of colour potential $\psi_{col}^p(x_i)$, texture potential $\psi_{tex}^p(x_i)$ and *prior* labelling potential $\psi_{lab}^p(x_i)$ respectively. We define colour potential:

$$\psi_{col}^p(x_i) \quad = \quad \frac{1}{|W_i|} \sum_{j\in W_i} w(i,j) \cdot \qquad (3)$$
$$-\log P_g(I_t(j)|x_i; \Theta_{col}).$$

where the colour models (Θ_{col}) are represented by a mixture of Gaussians (GMM) in RGB colour space, W_i is a squared window (15×15 in the implementation) centred at pixel i, $|W_i|$ is the cardinality of W_i, and $w(\cdot,\cdot)$ assigns a support weight to each pixel within a support window which is defined as $w(i,j) = \exp(-\frac{||I_i - I_j||}{\sigma})$. σ represents a parameter (30 in the implementation), and $||I_i - I_j||$ denotes the L_1 distance between I_i and I_j in RGB colour space. The unary potentials are averaged over a small local window which acts as a local edge-preserving smoothness constraint.

Colour potential alone is not very discriminative and we incorporate texture potential to achieve more accurate segmentation. To this end, we adopt textons [21] which have been proven effective in categorizing general objects.

For extracting texton histograms, we use a filter bank comprising 36 bar and edge filters, 1 Laplacian of Gaussian (LoG) and 1 Gaussian filter. The 36 bar and edge filters (6 orientations and 3 scales for each) are applied to the luminance channel only, producing 36 filter responses. The Gaussian filter is applied to each CIELab channel, thus producing 3 filter responses. The LoG is also applied to the L channel only, thus producing 1 filter response. We quantize filter responses to 400 textons by running K-means clustering and each pixel in I_t is assigned to the nearest cluster centre to generate the texton map T_t. We define the texture potential as:

$$\begin{aligned} \psi^p_{tex}(x_i) &= \frac{1}{|W_i|}\sum_{j\in W_i} w(i,j)\cdot -\log P_g(T_t(j)|x_i;\Theta_{tex}). \\ P_g(T_t(i)|x_i = l_n;\Theta_{tex}) &= \mathcal{H}^n(T_t(i)). \end{aligned} \tag{4}$$

The texture model Θ_{tex} of the n^{th} label l_n is represented by a discrete probability model given the normalized texton histogram $\mathcal{H}^n$ learned from the texton map in the first frame. Similar to the Gaussian weighting in the colour potential, we aggregate texture potentials over a small window centred at i.

The labelling *prior* potential exploits the fact that pixels with a higher probability propagated from particular labelled region tend to have consistent label assignment. Unlike other interactive or automatic segmentation algorithms which use the labelling prior as a hard constraint, we incorporate labelling prior as a soft constraint which is inferred from a probabilistic motion estimation framework which inherently takes into account the motion estimation errors. Details of the process by which labels are propagated over time in our framework are given in Sec. II-D.

2) Superpixel Unary Potentials: Superpixel unary potentials exploit a higher-level tendency that a group of pixels follow different appearance models. Superpixels provide an unique domain where rich visual features and machine learning techniques can be utilised. Superpixel segmentation also imposes a soft constraint to guide the label boundaries towards high contrast image edges. We extract three features from each superpixel: a histogram of Textons; a histogram of SIFT visual words; and a histogram of CIELab colours.

We extract Texton histogram H^t_i from each superpixel i in a similar manner as in Sec. II-B1. We quantise SIFT features into a codebook containing 200 visual words using k-means, and extract a histogram H^b_i from each superpixel. We convert the pixel colours into the perceptually uniform CIE Lab colour space, where we extract a 23×3 bin histogram H^c_i from each superpixel. We concatenate the histograms from three features to form a histogram H_i of 669 bins from superpixel i. In this case, learning GMMs with such a high dimensionality is infeasible. Rather, we train a linear SVM classifier over the extracted feature.

The pixel unary term is defined as

$$\psi^s_i = -\log P(H_i|x_i). \tag{5}$$

where $P(H_i|x_i)$ is obtained from the trained SVM classifier.

3) Pixel Pairwise Potentials: The pixel pairwise terms encourage coherence in region labelling and discontinuities to occur at high contrast locations, which is computed using RGB colour distance, disparity map between adjacent views using α-expansion stereo matching algorithm [22], and the appearance disparity between pixel and superpixel:

$$\begin{aligned} \psi^p_{ij}(x_i,x_j) &= \psi^p_{col}(x_i,x_j) + \theta^p_{disp}\psi^p_{disp}(x_i,x_j) \\ &\quad +\theta^p_{sp}\psi^p_{sp}(x_i,x_j). \end{aligned} \tag{6}$$

where θ^s_{disp} and θ^p_{sp} are weights of the disparity map compatibility $\psi^p_{disp}(x_i,x_j)$ and pixel-to-superpixel pairwise $\psi^p_{sp}(x_i,x_j)$ respectively.

We define the pixel colour pairwise term as

$$\psi^p_{col}(x_i,x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ e^{-\beta_{col}||I_t(i)-I_t(j)||^2} & \text{if } x_i \neq x_j, \end{cases} \tag{7}$$

where β_{col} is chosen to be contrast adaptive [19]:

$$\beta_{col} = \frac{1}{2}\langle ||I_t(i) - I_t(j)||^2\rangle^{-1}. \tag{8}$$

where $\langle\cdot\rangle$ denotes expectation over an image sample.

Similarly the pixel disparity map compatibility as

$$\psi^p_{disp}(x_i,x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ e^{-\beta_{disp}||D_t(i)-D_t(j)||^2} & \text{if } x_i \neq x_j. \end{cases} \tag{9}$$

where β_{disp} is a parameter (10 in the implementation).

The introduction of the appearance disparity between pixel and superpixel is based on the premise that pixel labelling helps to overcome errors propagated from the superpixel, the superpixel level enforces spatial grouping, which is defined as

$$\psi^p_{sp}(x_i,x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ e^{-\beta_{p,sp}} & \text{if } x_i \neq x_j. \end{cases} \tag{10}$$

where $\mathcal{H}^t_j$ and $\mathcal{H}^c_j$ denote the histograms of Textons and Lab colours of superpixel j respectively, $T_t(i)$ and $C_t(i)$ are the Texton label and Lab colour value at pixel i, and $\beta_{p,sp}$ is a contrast adaptive parameter:

$$\beta_{p,sp} = \frac{1}{2}\langle(1 - \frac{1}{2}(\mathcal{H}^t_j(T_t(i)) + \mathcal{H}^c_j(C_t(i)))\rangle^{-1}. \tag{11}$$

4) Superpixel Pairwise Potentials: The superpixel pairwise explores both the cross-view consistency and intra-view smoothness of superpixel labelling, defined as

$$\psi^s(x_i,x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ e^{-\beta_{sp}} & \text{if } x_i \neq x_j. \end{cases} \tag{12}$$

where $D_{KL}(H_i||H_j)$ denotes the KL-divergence between full feature histogram H_i of superpixel i and H_j of superpixel j, and β_{sp} is a contrast adaptive parameter:

$$\beta_{sp} = \frac{1}{2}\langle\min(D_{KL}(H_i||H_j), D_{KL}(H_j||H_i))\rangle^{-1}. \tag{13}$$

C. Superpixel Matching

With wide-baseline camera set-ups and natural backgrounds, stereo matching is no longer reliable to find dense pixel-level correspondence. We overcome this problem by exploiting both the superpixel-level appearance measure and camera geometry to identify potential correspondence of regions between each pair of adjacent views.

For each superpixel S_i^{k-1} in view V_{k-1}, we compute the distance between S_i^{k-1} and a set of relevant superpixels $\mathcal{S}_i^k$ in view V_k. This relevance is determined by the camera geometry. Specifically, $\mathcal{S}_i^k$ consists of all the superpixels on the epipolar line in view V_k corresponding to the centroid of superpixel S_i^{k-1} in view V_{k-1}. The distance between superpixels is defined as

$$\underset{j \in \mathcal{S}_i^k}{\operatorname{argmax}} \quad e^{\beta_{fgp}(P_{fg}^i - P_{fg}^j)^2} \min(D_{KL}(H_i || H_j), \\ D_{KL}(H_j || H_i)). \tag{14}$$

where β_{fgp} is a parameter (2 in the implementation) and P_{fg} denotes the foreground probability of superpixel which are defined as

$$P_{fg} = \frac{P(H_i | x_i = 1)}{P(H_i | x_i = 1) + P(H_i | x_i = 0)}. \tag{15}$$

D. Bootstrapping and Motion Propagation

We bootstrap the first frame of segmentation by manually supplying a relatively accurate mask for the desired object using image-based object selection tools, such as GrabCut [19] for all views at $t = 0$. To encourage spatio-temporal coherence between consecutive frames, we adopt a probabilistic motion diffusion strategy similar to Wang *et al.* [15]. The purpose of the motion diffusion model is to propagate forward the labels of past frames — forming a distribution of labelling priors for segmentation of the current frame. In contrast to the motion diffusion model proposed in [15], which diffuses the propagated subset of pixels to close vicinity and assumes the predicted position as the centre of an isotropic Gaussian distribution, we adopt an oriented anisotropic Gaussian distribution to better capture the inherent errors in motion estimation along the motion vector.

Specifically, let Ω_n ($n \in \{0, 1\}$) be a region of interest in frame I_{t-1} labelled as x_n ($x_n \in \{0, 1\}$). We form a subset of pixels $\mathcal{O}_n \subset \Omega_n$ by sampling from a morphologically dilated skeleton of region Ω_n for propagation, to account for the impact from imprecise motion estimation close to boundary. Let J_k^{t-1} ($k = 1, 2, 3, \cdot, |\mathcal{O}_n|$) be pixels in $\mathcal{O}_n$ with position z_k^{t-1}. For each pixel $J_k^{t-1} \in \mathcal{O}_n$, its position $\mu_k^t = (\mu_x, \mu_y)$ in frame I_t is computed based on the motion vector from optical flow. We treat the predicted position as the centre of a oriented anisotropic Gaussian distribution,

$$p(z^t | z_k^{t-1}; \mu_k^t, \sigma_u, \sigma_v, \theta) = \\ \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[\frac{((z_x^t - \mu_x)\cos\theta + (z_y^t - \mu_y)\sin\theta)^2}{2\sigma_u^2} + \\ \frac{(-(z_x^t - \mu_x)\sin\theta + (z_y^t - \mu_y)\cos\theta)^2}{2\sigma_v^2} \right]. \tag{16}$$

where $z^t = (z_x^t, z_y^t)$ is a position in frame I_t. σ_u and σ_v are the variances in the direction of motion vector (axis u, with orientation θ) and the axis v which is orthogonal to u respectively. We use the local motion coherence to encode the motion estimation error. This accommodates the per-pixel local motion estimation errors σ_u along the direction of motion vectors as in [15]. σ_v is set to be $\sigma_u/3$. Based on this novel definition of oriented anisotropic per-pixel motion distribution, the likelihood $p(z^t | l_i)$ of the pixel at z^t being assigned with label x_i propagated from previous frame in the sequence, following [15]. We define the *prior* labelling potential as

$$\psi_{lab}^p(x_i) \quad = \quad -\log p(z_i | x_i). \tag{17}$$

which is encoded directly in the unary term of our energy function (eq. 2), which comprises a sum of appearance and labelling potentials (described in Sec. II-B1, eq. 2).

E. Multi-resolution Solution of MRF

To substantially reduce the computational cost of segmentation across all views, we adopt the multilevel banded graph cut [23] for iteratively segmenting each frame into foreground and background regions. In practice, we adopted a three-level banded graph cut with a downscaling factor of 2, to achieve a trade-off between accuracy and complexity, which is ~ 8 times as fast as the original graph cut algorithm and consumes ~ 4 times less memory.

III. Results and Discussion

We evaluate our algorithm using two publicly available indoor multi-view video sequences; KARATE and SALSA shot using 8 synchronised HD Canon cameras spaced at ~ 45 degree intervals in a circular arrangement on 1.5 metre tripods surrounding the performance [24]. Sequences are shot against general clutter in the studio, causing background appearance to vary across views. Both sequences contain a variety of complex and fast moving actions by one (KARATE) or two (SALSA) people. In both cases 60 frames of the sequence were processed, then compared with a ground truth mask defined manually at 5 frame intervals across all of the 8 camera views. The ground truth data was not made available to any of the algorithms.

We qualitatively (Fig. 2) and quantitatively (Figs. 3-4) compare the performance of our proposed approach with seven alternative approaches commonly used for matte generation: two forms of RGB chroma-keying, motion-differencing, and the algorithms GrabCut [19], Geodesic segmentation [25], and Geodesic Graph-cut [26]. Additionally we compare against Sarim *et al.*'s multi-view segmentation method that propagates tri-maps between views using patch matching along the epipolar constraint, but independently solves views following propagation. Our method not only propagates information between views but solves all views simultaneously using our novel inter-view MRF approach.

Fig. 2 provides representative samples of multi-view frames and their output. To generate the mattes using our multi-view method, user segmentation is performed in the first frame and very minor corrections e.g. a small scribble in a view is required approximately every 5-10 frames due to information propagation between views and over time.

The first RGB chroma-keying method we compare to *(RGB ratio)* uses a thresholded ratio of blue (b) to red (r) and green (g) i.e. $\frac{b}{r+g+b} > T_{RGBratio}$ to produce the matte on a per-pixel basis. This is commonly used in 4DPC matting [4]. The second chroma-keying method *(RGB eigen)* computes a per-pixel background likelihood via Mahalanobis distance from

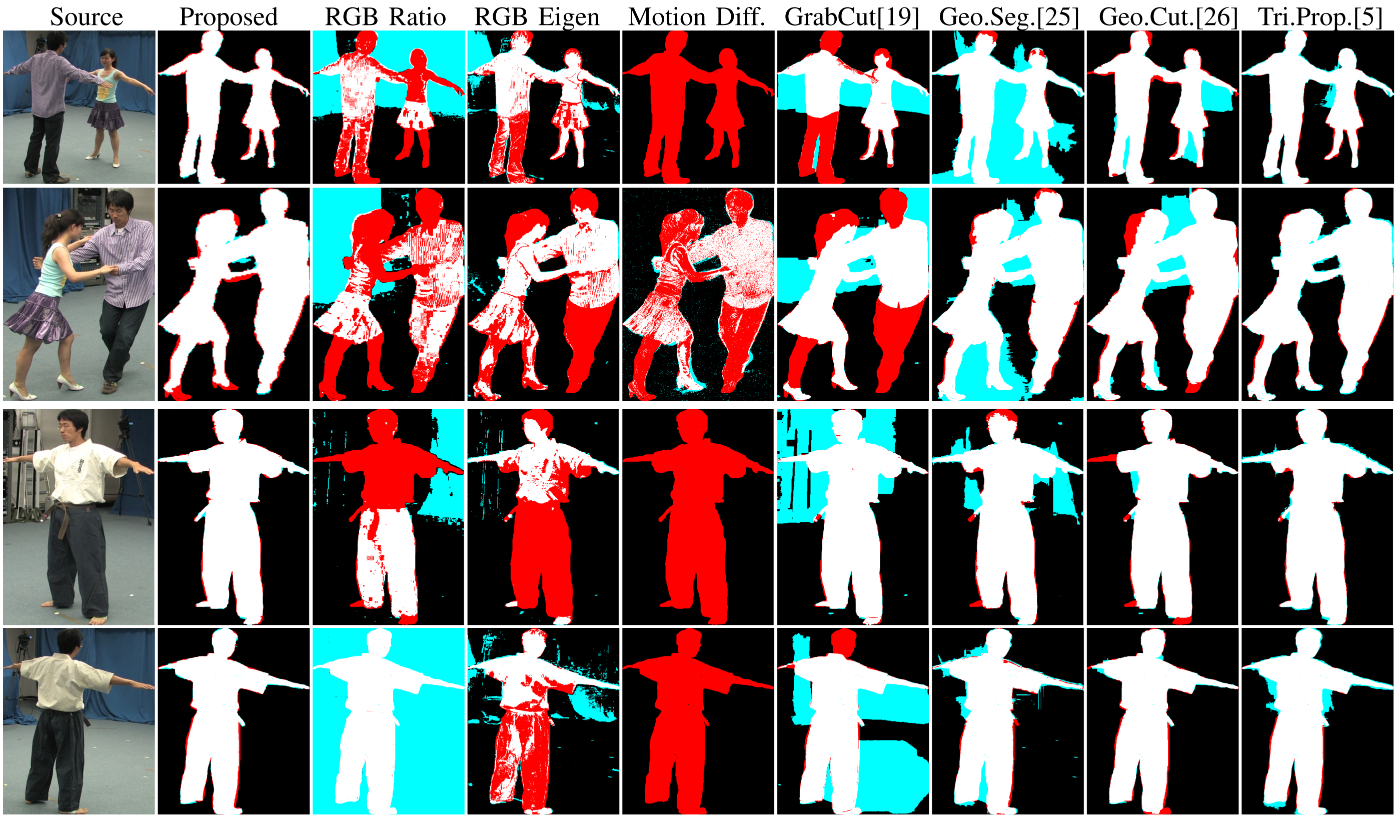

Fig. 2. Multi-view segmentation results for SALSA (upper) and KARATE (lower). Red: under-detected vs. groundtruth. Blue: over-detected vs. groundtruth.

an Eigenmodel built from the mean (μ) and covariance (C) of RGB colour samples of the background (obtained manually in the first frame). Each pixel's colour c is thresholded via $(c-\mu)C^{-1}(c-\mu)^T < T_{RGBeigen}$. Inter-frame motion subtraction (Motion Differencing) is commonly used to create mattes by differencing adjacent frames and thresholding (T_{modiff}). This is performed independently for each camera view in our comparison. To ensure fair comparison we exhaustively search the range of the thresholds ($T_{RGBratio}, T_{RGBeigen}, T_{modiff}$) for each of these methods to find the optimum level with respect to our objective evaluation measure (eq.18).

Methods GrabCut [19], Geodesic segmentation [25] and Geodesic Graph-cut [26] are commonly used in image processing to interactively obtain object mattes from a set of scribbles generated manually over the image. We apply these independently to each view and frame, generating scribbles for each ground truthed frame. The tri-map propagation (TriProp) approach of Sarim *et al.* was also compared against, and was initialised manually at $t = 0$ as per [5] using only a single tri-map key as input.

In all cases we obtain a score $E(M, G) \in [0, 1]$ representing the accuracy of the mask generated by an algorithm M with respect to the manually defined ground truth mask G:

$$E(M, G) = \frac{M \cap G}{M \cup G}. \tag{18}$$

On SALSA, the accuracy of the proposed algorithm was 87.9%, versus 8.5% for RGB difference, 37.4% for Eigenmodel RGB, 36.3% for motion difference, 59.5% for GrabCut, 40.6% for geodesic segmentation, 57.6% for geodesic graph-cut, and 86.1% for tri-map propagation. The error-bars within Fig. 3 illustrate greater robustness of our approach across views with respect to interactive methods [19], [25], [26] which despite frame- and view-specific guidance fail to produce consistent results in all views. The inter-view matching and temporal propagation of information regularise the mattes under our proposed approach. The presence of two differently dressed actors in this sequence does not reduce robustness versus KARATE, where with tri-map propagation the region between performers is frequently classified false-positive.

Accuracy on KARATE was 88.1% for the proposed algorithm, versus 39.2% for RGB difference, 62.8% for Eigenmodel RGB, 48.1% for motion difference, 60.2% for GrabCut, 63.4% for geodesic segmentation, 77.3% for geodesic graph-cut and 83.5% for tri-map propagation. Although the methods compared against perform significantly better on this footage, our approach exhibits significantly greater inter-view robustness (error-bars of Fig. 4) and overall performance improvement. Note that the closest rival interactive method (geodesic graph-cut) requires *ab initio* per-view and per-frame user interaction whereas our approach requires partial and minor correction (only a few strokes) every 5-10 frames. In this sequence tri-map propagation is out-performed by our proposed approach by $\sim$ 4.6% per frame yet exibits comparable robustness across views (c.f. error-bars).

A limitation of our approach is the relatively high computational complexity. The basic chroma and difference keying approaches compared against are near-instantaneous, and the optimized implementations of the interactive approaches [19], [25], [26] take 8-10s per HD (1920×1080) frame to compute (i.e. over 8 views). Our approach takes approximately 40-60s per multi-view frame to solve the MRF. Nevertheless this is

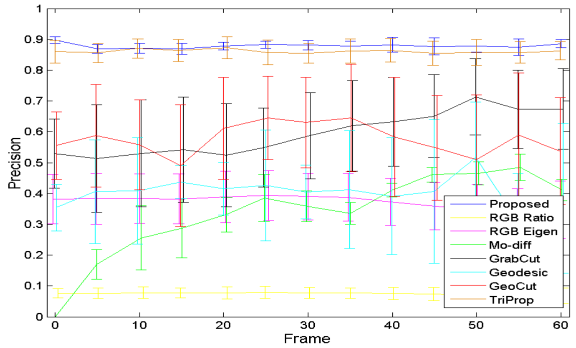

Fig. 3. SALSA comparison (errorbars computed across views).

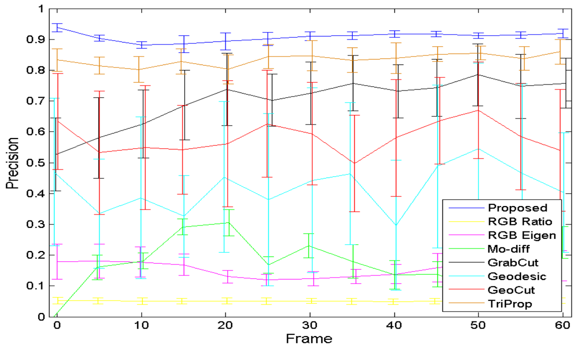

Fig. 4. KARATE comparison (errorbars computed across views).

practical for processing 4DPC as an offline process.

IV. Conclusion

We have demonstrated a novel multi-view video matting method suitable for incorporation into a 4DPC pipeline. The key contributions of this method are 1) the propagation of appearance and spatial information across views using superpixel matching and a novel MRF that solves the mattes simultaneously across all views, 2) the temporal propagation of appearance and spatial information forward in time. Both contributions enhance inter-view and temporal coherence, and so improve accuracy of the matte. Although the second contribution broadly follows the pipeline of Wang *et al.* [15] for matte propagation in monocular video, such an approach has not previously been applied to multi-view video — a domain in which removing the requirement for per-frame manual initialisation of matting is particularly attractive due to the overhead of marking up multiple views. In addition, stable application of [15] required extension of that method to use anisotropic rather than isotropic motion diffusion. We have quantitatively and qualitatively compared against seven alternative approaches commonly used in video matte generation (including monocular and multi-view techniques) on two public datasets[1], showing significant gain in accuracy (Fig. 3-4). Future work will explore alternative appearance models for colour texture to further enhance robustness in cases of illumination change and common foreground-background appearance.

[1] Ground truth mark-up and results are available for research purposes at http://cvssp.org/cvssp3d.

References

[1] C. Budd, P. Huang, M. Klaudinay, and A. Hilton, “Global non-rigid alignment of surface sequences,” *IJCV*, vol. 102, no. 1-3, pp. 256–270, 2012.

[2] J.-S. Franco and E. Boyer, “Exact polyhedral visual hulls,” in *BMVC*, vol. 1, 2003, pp. 329–338.

[3] A. Laurentini, “The visual hull concept for silhouette-based image understanding,” *IEEE PAMI*, vol. 16, no. 2, pp. 150–162, Feb. 1994.

[4] J. Starck and A. Hilton, “Surface capture for performance-based animation,” *IEEE Computer Graphics and Applications*, vol. 27, no. 3, pp. 21–31, 2007.

[5] M. Sarim, A. Hilton, and J.-Y. Guillemaut, “Temporal trimap propagation for video matting using inferential statistics,” in *ICIP*, 2011, pp. 1745–1748.

[6] A. Djelouah, J.-S. Franco, E. Boyer, F. Leclerc, and P. Pérez, “N-Tuple Color Segmentation for Multi-View Silhouette Extraction,” in *ECCV*. Firenze, Italy: University of Florence, Oct. 2012.

[7] J. Wang, B. Thiesson, Y. Xu, and M. Cohen, “Image and video segmentation by anisotropic kernel mean shift,” in *ECCV*. Springer Berlin Heidelberg, 2004, pp. 238–249.

[8] S. Paris, “Edge-preserving smoothing and mean-shift segmentation of video streams,” in *ECCV*, 2008, pp. 460–473.

[9] M. Grundmann, V. Kwatra, M. Han, and I. A. Essa, “Efficient hierarchical graph-based video segmentation,” in *CVPR*, 2010, pp. 2141–2148.

[10] I. Budvytis, V. Badrinarayanan, and R. Cipolla, “MoT - mixture of trees probabilistic graphical model for video segmentation,” in *BMVC*, 2012, pp. 72.1–72.11.

[11] J. P. Collomosse, D. Rowntree, and P. M. Hall, “Stroke surfaces: Temporally coherent artistic animations from video,” *IEEE TVCG*, vol. 11, no. 5, pp. 540–549, 2005.

[12] F. Moscheni, S. K. Bhattacharjee, and M. Kunt, “Spatiotemporal segmentation based on region merging,” *IEEE PAMI*, vol. 20, no. 9, pp. 897–915, 1998.

[13] W. Brendel and S. Todorovic, “Video object segmentation by tracking regions,” in *ICCV*, 2009, pp. 833–840.

[14] X. Bai, J. Wang, D. Simons, and G. Sapiro, “Video snapcut: robust video object cutout using localized classifiers,” in *ACM SIGGRAPH*, vol. 28, no. 3, 2009, p. 70.

[15] T. Wang and J. P. Collomosse, “Probabilistic motion diffusion of labeling priors for coherent video segmentation,” *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 389–400, 2012.

[16] G. Zeng and L. Quan, “Silhouette extraction from multiple images of an unknown background,” in *ACCV*, 2004.

[17] D. Snow, P. Viola, and R. Zabih, “Exact voxel occupancy with graph cuts,” in *CVPR*. IEEE, 2000.

[18] D. Cremers and K. Kolev, “Multiview stereo and silhouette consistency via convex functionals over convex domains,” *IEEE PAMI*, vol. 33, no. 6, pp. 1161–1174, 2011.

[19] C. Rother, V. Kolmogorov, and A. Blake, “”grabcut”: interactive foreground extraction using iterated graph cuts,” *ACM SIGGRAPH*, vol. 23, no. 3, pp. 309–314, 2004.

[20] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, “SLIC superpixels compared to state-of-the-art superpixel methods,” *IEEE PAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[21] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi, “Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context,” *IJCV*, vol. 81, no. 1, pp. 2–23, 2009.

[22] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. F. Tappen, and C. Rother, “A comparative study of energy minimization methods for markov random fields with smoothness-based priors,” *IEEE PAMI*, vol. 30, no. 6, pp. 1068–1080, 2008.

[23] H. Lombaert, Y. Sun, L. Grady, and C. Xu, “A multilevel banded graph cuts method for fast image segmentation,” in *ICCV*, 2005, pp. 259–265.

[24] M. Sarim, “Wide baseline image matting,” Ph.D. dissertation, CVSSP, University of Surrey, Sep. 2010.

[25] X. Bai and G. Sapiro, “A geodesic framework for fast interactive image and video segmentation and matting,” in *ICCV*, 2007, pp. 1–8.

[26] B. L. Price, B. S. Morse, and S. Cohen, “Geodesic graph cut for interactive image segmentation,” in *CVPR*, 2010, pp. 3161–3168.