Space-Time Joint Multi-Layer Segmentation and Depth Estimation

Jean-Yves Guillemaut and Adrian Hilton
Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK
J.Guillemaut@surrey.ac.uk

Abstract

Video-based segmentation and reconstruction techniques are predominantly extensions of techniques developed for the image domain treating each frame independently. These approaches ignore the temporal information contained in input videos which can lead to incoherent results. We propose a framework for joint segmentation and reconstruction which explicitly enforces temporal consistency by formulating the problem as an energy minimisation generalised to groups of frames. The main idea is to use optical flow in combination with a confidence measure to impose robust temporal smoothness constraints. Optimisation is performed using recent advances in the field of graph-cuts combined with practical considerations to reduce runtime and memory consumption. Experimental results with real sequences containing rapid motion demonstrate that the method is able to improve spatio-temporal coherence both in terms of segmentation and reconstruction without introducing any degradation in regions where optical flow fails due to fast motion.

1. Introduction

Layered-depth estimation is the process of separating an image into different layers representing *e.g.* foreground and background together with recovering depth information at each pixel. We are concerned with the problem of recovering a layered-depth representation from multiple ($>$2) video streams captured by cameras in a general wide-baseline configuration. This problem is closely related to segmentation, which focuses on layer extraction, and reconstruction which is concerned with depth recovery. Traditionally in the image domain these problems have been formulated as labelling problems across the set of pixels $\mathcal{P}$ from the image lattice seeking a set of labels $\boldsymbol{l} = \{l_{\boldsymbol{p}}, \boldsymbol{p} \in \mathcal{P}\}$ minimising an energy function of the form [5, 17]

$$E(\boldsymbol{l}) = \sum_{\boldsymbol{p} \in \mathcal{P}} D_{\boldsymbol{p}}(l_{\boldsymbol{p}}) + \sum_{(\boldsymbol{p},\boldsymbol{q}) \in \mathcal{N}} V_{\boldsymbol{p},\boldsymbol{q}}(l_{\boldsymbol{p}}, l_{\boldsymbol{q}}). \qquad (1)$$

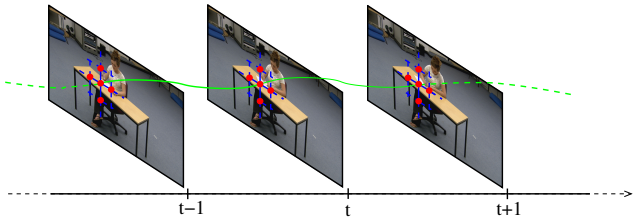


Figure 1. The space-time formulation combines constraints in the spatial domain (blue connections) and the temporal domain (green connections) defined by optical flow over adjacent frames.

The first term, called data term or unary term, measures the fit between the hypothesised labelling and the observations, while the second term, referred to as smoothness term or pairwise term, introduces regularisation constraints to encourage local smoothness of the solution across a spatial neighbourhood $\mathcal{N}$ connecting nearby pixels.

Transfer of these techniques from image to video domain poses several key challenges, primarily in terms of temporal consistency (independent processing of each frame would produce inconsistent results manifesting as flicker caused by noisy input data), secondly in terms of scalability (an increase in dimensionality significantly increases complexity particularly in the case of reconstruction which becomes 4D) and finally in terms of defining an adequate structure for solving the problem (the notion of proximity used to define neighbouring pixels in the spatial domain is less obvious in the temporal domain as connected pixels are usually no longer adjacent). Due to these shortcomings, many video-based techniques have been based on sequential application of image-based techniques to each frame of the video combined with a mechanism to propagate information between consecutive frames [7, 2, 23, 13, 20] or reduce noise via filtering [27, 3]. These approaches are usually prone to propagation of errors. We also argue that for optimality consistency should be enforced during labelling rather than as a post-process.

We propose a technique which explicitly enforces temporal consistency during layered depth estimation. The main idea is to generalise the traditional energy optimi-

sation approach to the temporal domain by replacing the spatial neighbourhood $\mathcal{N}$ by a spatio-temporal neighbourhood connecting nearby frames (see Fig. 1). Optical flow forms the basis for establishing temporal correspondences between consecutive frames. For robustness, the optical flow algorithm is coupled with an error detection framework to guarantee construction of a reliable temporal neighbourhood and avoid failure where optical flow is deemed unreliable. Extension of the energy is performed by generalising the smoothness term to enforce piece-wise smooth variations in the temporal domain as well as the spatial domain. To cope with fast moving surfaces without over smoothing them we utilise contextual information which measures the amplitude of the motion between temporal neighbours. Practical consideration are given to ensure scalability of the method in terms of run-time and memory consumption.

The paper is structured as follows. We start by giving an overview of related work in the field of temporally consistent segmentation and reconstruction. After stating the problem, we describe the proposed technique starting with the construction of a robust spatio-temporal neighbourhood, then describing the extension of each energy term to the temporal domain, and finally discussing practical considerations for scalability. Experiments with real footage containing fast motion are used to demonstrate the technique's ability to improve consistency and its resilience to fast motion. We finally conclude with a summary of contributions and recommendations for future work.

2. Related work

Temporally consistent segmentation One approach to video segmentation is to concatenate video frames along the temporal axis and then cast the problem as a volume segmentation problem [4, 19, 26]. In [4], a volumetric neighbourhood is defined by connecting adjacent pixels in both spatial and temporal domain producing a regular graph on which a binary graph-cut algorithm can be applied. In [19], temporal connections are extended to pixels lying within a predefined radius with elimination of unrelated connections based on colour distance. Optimisation is again performed using 3D graph-cuts.

Chuang *et al*. [7] used optical flow to propagate information from a set of manually defined trimaps, which partition each image into definite foreground, definite background and unknown regions, to the entire video sequence and then separately applied a Bayesian matting at each frame. During trimap propagation they utilised a measure of optical flow accuracy based on colour consistency to avoid propagation of errors. Manual editing is used to allow error correction. Bai *et al*. [2] used overlapping local classifiers which consist of a local colour model, a confidence value and a local shape model. They combined SIFT features suitable to perform a rigid alignment and optical flow suitable to capture additional non-rigid deformation to reliably propagate information between consecutive frames.

Temporally consistent reconstruction Temporal extension of reconstruction algorithms is more challenging than in the case of segmentation algorithms due to larger dimensionality and increased difficulty in correcting errors using key-frame editing techniques. Space-time stereo [30, 8] extends the traditional spatial window to the temporal domain performing matching using the spatio-temporal features and usually in combination with active lighting to artificially increase texture content. These techniques usually rely on static or quasi-static scenes as they do not perform motion estimation. In [27], a 4D bilateral filter is introduced to post-process displacement maps and impose smooth variations in both spatial and temporal domains. Recently, Richardt *et al*. [21] also used a spatio-temporal bilateral filtering approach to denoise and upsample the noisy depth input from a Kinect device.

Tao *et al*. [23] used colour-based segmentation to decompose a scene into piecewise planar patches following a constant velocity model. Temporal correspondences between patches are established based on colour segmentation or optical flow depending on texture content and used to iteratively predict plane parameters before final optimisation. In [3], Bleyer *et al*. used optical flow to establish temporal correspondences and then applied median filtering to smooth the depth map. In [28], Yang *et al*. generalise the bundle optimisation approach proposed in [29] in the case of a single moving camera seeing a static scene to multiple cameras observing a dynamic scene. Larsen *et al*. [18] proposed a belief propagation approach using spatio-temporal neighbouring linking each frame to its previous and next frames constructed using optical flow. To cope with optical flow errors they introduce a mechanism based on variance reduction to discard messages coming from outlying neighbours. In [31], Zhu *et al*. also use optical flow constraints to extend the traditional spatial MRF formulation to the temporal domain and demonstrate how this can be used to fuse stereo and time-of-flight data. Their algorithm was aimed at sensors arranged in a narrow-baseline configuration.

Vedula *et al*. [24] introduced the concept of scene flow which generalises optical flow to the 3D domain by establishing dense temporal correspondences between surface points. In [13], Gong used a view-dependent representation of the scene flow to iteratively predict disparity maps between consecutive frames. Liu and Philomin [20] followed a similar iterative approach using scene flow for disparity map prediction combined with a probabilistic framework to reduce propagation of errors between consecutive frames.

Contributions Our main contribution is the introduction of a temporally consistent approach for simultaneous seg-

mentation and reconstruction from multiple-view video. While a few methods have been previously proposed for joint segmentation and reconstruction [12, 32, 15, 14], these all performed frame-by-frame processing of the input data and therefore lacked temporal consistency. The proposed method is the first joint segmentation and reconstruction approach which explicitly enforces temporal consistency. The proposed approach also makes a number of improvements to existing temporally consistent reconstruction algorithms which do not simultaneously estimate segmentation. Firstly, it is able to operate under a wide-baseline camera configuration ($> 20^\circ$ between adjacent cameras). Secondly, contrary to the majority of techniques introduced earlier, it does not require iterative propagation of information, but instead directly operates on groups of adjacent frames, thus naturally enforcing temporal consistency. Finally the paper introduces several ideas to provide resilience to fast motion necessary to cope with errors in optical flow estimation and gives practical considerations to improve efficiency.

3. Problem statement

Let us consider a set of calibrated and synchronised cameras. The camera for which we would like to compute a layered-depth representation will be called reference camera, while other cameras will be referred to as auxiliary cameras. Let us denote by $\mathcal{P}_i$ the set of pixels in the reference image at frame i and by $\mathcal{P} = \bigcup_i \mathcal{P}_i$ the set of all pixels across a group of consecutive frames. Given a set of candidate layer labels $\mathcal{L} = \{l_i\}_{i=1}^{N_L}$ and depth labels $\mathcal{D} = \{d_i\}_{i=1}^{N_D}$, the objective is to identify an optimal labelling $(\boldsymbol{l}, \boldsymbol{d})$ across $\mathcal{P}$. Layer labels represent different layers in the scene such as foreground and background or larger number of layers corresponding to multiple foreground or background objects. Depth labels are obtained by discretising the 3D space using a regular 3D grid. To reduce the number of layer and depth labels, we assume that a coarse initial segmentation (obtained *e.g.* using background subtraction with threshold set to produce a conservative foreground estimate) and a coarse initial bounding volume (*e.g.* provided by applying visual hull reconstruction to the initial segmentation) are available; these are used to eliminate any labelling which does not belong to the initial foreground or the interior of the bounding volume.

4. Temporally consistent layered-depth estimation

4.1. General principle

The general idea to achieve temporal coherence is to simultaneously estimate a layered depth representation for multiple consecutive frames by introducing temporal constraints in the energy formulation. Solving this problem across the entire frame sequence would of course be completely impractical as it would require solving an extremely large optimisation problem which would be intractable due to prohibitively large memory and run-time requirements. To avoid this limitation each frame is processed by performing optimisation across a window of consecutive frames centred at the frame of interest. The window is translated across the entire sequence to process all frames in the sequence. This type of approach was previously used in [11] also in the context of space-time reconstruction. To emphasise the contribution of frames located in close proximity to the frame of interest, frames are weighted according to their distance from the central frame using Gaussian attenuation factors. It should be noted that although this approach iteratively slides a window across the entire sequence, each local group of frames is processed separately and as a result is not subject to propagation of errors.

4.2. Spatio-temporal neighbourhood definition

Spatial neighbourhood The spatial neighbourhood is defined in a conventional manner by considering pairs of spatially closed pixels in the image domain. Let us denote by $\mathcal{N}_s$ the set of pixel pairs $(\boldsymbol{p}, \boldsymbol{q})$ such that $\boldsymbol{p}$ and $\boldsymbol{q}$ belong to the same frame and are spatially connected. Spatially connected pixels are often adjacent pixels (most commonly systems of 4, 8 or 16 neighbours), however more complex types of neighbourhood can also be considered. In this paper, we assume a standard 4-connected spatial neighbourhood.

Temporal neighbourhood Definition of the temporal neighbourhood is less straightforward since adjacent images can contain fast motion resulting in temporal correspondences being non-contiguous and possibly spanning large displacements. Optical flow is used to compute a dense flow field between pairs of consecutive images. We use the variational approach proposed by Brox *et al.* in [6] because of its ability to preserve large displacements. Let us denote by $\boldsymbol{f}_{\boldsymbol{p}}^{ij} = (u_{\boldsymbol{p}}^{ij}, v_{\boldsymbol{p}}^{ij})$ the estimated displacement vector at pixel $\boldsymbol{p}$ from frame i to j. Although a temporal neighbourhood could be defined at this stage by considering all pixel correspondences between consecutive pairs of images, this is likely to produce a large number of false correspondences due to known limitations of optical flow algorithms (*e.g.* failure in untextured areas or areas undergoing a fast motion). To test the validity of each temporal correspondence and reject false matches, we perform cross checking by comparing the forward and backward flow for each pair of images and requiring them to be symmetrical up to a tolerance e_f (set to 3 pixels in the paper). The temporal neighbourhood is then defined as the set of all optical flow displacement between adjacent images which satisfy the forward backward consistency check, *i.e.* $\mathcal{N}_t = \{(\boldsymbol{p}, \boldsymbol{q}) | \boldsymbol{q} =$

Figure 2. From left to right: a pair of consecutive input images, its forward and backward flow and a visualisation of the forward backward error (for display purposes the error map has been normalised using histogram equalisation).

$\boldsymbol{p} + \boldsymbol{f}_{\boldsymbol{p}}^{ij}, \boldsymbol{p} \in \mathcal{P}_i, \boldsymbol{q} \in \mathcal{P}_j, |i-j| = 1, ||\boldsymbol{f}_{\boldsymbol{p}}^{ij} + \boldsymbol{f}_{\boldsymbol{q}}^{ji}|| \leq e_f\}$. Fig. 2 shows an example of two consecutive frames, the forward and backward flow and the forward backward error map. Note that the chosen optical flow algorithm can be arbitrary as long as it does not explicitly enforce forward backward symmetry.

4.3. Spatio-temporal energy formulation

Recovery of the layer labels l and depth labels d is cast as an energy optimisation across a local window of consecutive frames. The energy to be optimised is defined as

$$E(l,d) = E_{\text{col}}(l) + E_{\text{p}}(d) + E_{\text{con}}(l) + E_{\text{s}}(l,d). \quad (2)$$

The rest of this sub-section is focused on defining each terms with emphasis on the pairwise terms which define the novel temporal constraints. To make the generalisation to the temporal domain more evident, we adopt a similar formalism as in [15, 14] which defined a single frame layered depth estimation energy minimisation framework.

4.3.1 Unary terms

The first two terms in Eq. (2) are unary terms measuring the fit of the current labelling to the observations. These terms are identical to the temporally unconstrained formulation with the only difference that the set of pixels $\mathcal{P}$ spans multiple frames rather than a single frame. A brief summary of these terms is provided. For more details the reader is referred to [15, 14].

Colour term The colour term uses learnt colour models for each layer in order to encourage assignment of most likely layer labels. It is defined as

$$E_{\text{col}}(l) = \lambda_{\text{col}} \sum_{\boldsymbol{p} \in \mathcal{P}} -\log P(\boldsymbol{I}_{\boldsymbol{p}} | l_{\boldsymbol{p}}). \quad (3)$$

For each camera, colour models are learnt from a single manually drawn trimap partitioning the image into definite foreground, definite background and unknown regions. Each definite region is used to build a Gaussian mixture model for the background and foreground layers. In the case of the static background layer, the global model is complemented with a local colour model representing colour at each pixel as a Gaussian distribution. Local and global colour models are combined at each pixel using a linear combination similar to [22]. These colour models are used to compute the probability of a particular assignment $P(\boldsymbol{I}_{\boldsymbol{p}} | l_{\boldsymbol{p}})$.

Photo-consistency term The photo-consistency term measures the likelihood of a particular depth hypothesis based on a matching score and is defined as

$$E_{\text{p}}(d) = \lambda_{\text{p}} \sum_{\boldsymbol{p} \in \mathcal{P}} e_{\text{p}}(\boldsymbol{p}, d_{\boldsymbol{p}}) \quad (4)$$

with

$$e_{\text{p}}(\boldsymbol{p}, d_{\boldsymbol{p}}) = \begin{cases} S(\boldsymbol{p}, \boldsymbol{P}(\boldsymbol{p}, d_{\boldsymbol{p}})) & \text{if } d_{\boldsymbol{p}} \neq \mathcal{U}, \\ S_{\mathcal{U}} & \text{if } d_{\boldsymbol{p}} = \mathcal{U}. \end{cases} \quad (5)$$

In the previous equation, S is a matching score based on the normalised cross correlation (NCC) summed over all pairs of images defined by the reference camera and any of its neighbouring auxiliary cameras. An unknown depth label $\mathcal{U}$ is introduced in the formulation in order to cope with occlusions.

4.3.2 Pairwise terms

The last two terms in Eq. (2) are the pairwise terms responsible for regularising the optimisation problem by imposing local smoothness constraints both spatially and temporally.

Spatio-temporal contrast term The contrast term introduces a penalty based on a colour distance between neighbours to encourage layer discontinuities to follow high contrast regions such as edges. This term is decomposed into spatial and temporal components with respective contributions controlled by the parameters $\lambda_{\text{con}}^{\text{s}}$ and $\lambda_{\text{con}}^{\text{t}}$ as follows:

$$E_{\text{con}}(l) = \lambda_{\text{con}}^{\text{s}} \sum_{(\boldsymbol{p},\boldsymbol{q}) \in \mathcal{N}_s} e_{\text{con}}(\boldsymbol{p}, \boldsymbol{q}, l_{\boldsymbol{p}}, l_{\boldsymbol{q}}, \beta_s) + \lambda_{\text{con}}^{\text{t}} \sum_{(\boldsymbol{p},\boldsymbol{q}) \in \mathcal{N}_t} w_{\boldsymbol{pq}} u_{\boldsymbol{pq}} e_{\text{con}}(\boldsymbol{p}, \boldsymbol{q}, l_{\boldsymbol{p}}, l_{\boldsymbol{q}}, \beta_t) \quad (6)$$

with

$$e_{\text{con}}(\boldsymbol{p}, \boldsymbol{q}, l_{\boldsymbol{p}}, l_{\boldsymbol{q}}, \beta) = \begin{cases} 0 & \text{if } l_{\boldsymbol{p}} = l_{\boldsymbol{q}}, \\ \exp\left(-\beta ||I_{\boldsymbol{p}} - I_{\boldsymbol{q}}||^2\right) & \text{otherwise.} \end{cases} \quad (7)$$

The parameter β_s and β_t are set to $\beta = 1/(2\langle||I_{\boldsymbol{p}} - I_{\boldsymbol{q}}||^2\rangle)$ with the operator $\langle\rangle$ denoting the mean computed across the neighbourhoods $\mathcal{N}_s$ and $\mathcal{N}_t$ respectively. The temporal term adds robustness to errors in optical flow as it will favour discontinuities along the temporal axis to occur at high contrast regions where optical flow failures are most likely. The term $w_{\boldsymbol{pq}}$ defines the Gaussian attenuation weight according to the distance from the central frame in the window introduced in Section 4.1. The term $u_{\boldsymbol{pq}}$ uses contextual information to introduce robustness to large displacement where optical flow is likely to become less accurate and will be detailed in Section 4.3.3.

Spatio-temporal smoothness term The smoothness term encourages piecewise smooth variations in depth. As previously, this term is decomposed into spatial and temporal components with relative contributions controlled by the parameters $\lambda_{\text{s}}^{\text{s}}$ and $\lambda_{\text{s}}^{\text{t}}$ as follows:

$$E_{\text{s}}(l,d) = \lambda_{\text{s}}^{\text{s}} \sum_{(\boldsymbol{p},\boldsymbol{q})\in\mathcal{N}_s} e_{\text{s}}(l_{\boldsymbol{p}}, d_{\boldsymbol{p}}, l_{\boldsymbol{q}}, d_{\boldsymbol{q}}, d_{\text{max}}^{\text{s}}) + \lambda_{\text{s}}^{\text{t}} \sum_{(\boldsymbol{p},\boldsymbol{q})\in\mathcal{N}_t} w_{\boldsymbol{pq}} u_{\boldsymbol{pq}} e_{\text{s}}(l_{\boldsymbol{p}}, d_{\boldsymbol{p}}, l_{\boldsymbol{q}}, d_{\boldsymbol{q}}, d_{\text{max}}^{\text{t}}) \quad (8)$$

with

$$e_{\text{s}}(l_{\boldsymbol{p}}, d_{\boldsymbol{p}}, l_{\boldsymbol{q}}, d_{\boldsymbol{q}}, d_{\text{max}}) = \begin{cases} \min(|d_{\boldsymbol{p}} - d_{\boldsymbol{q}}|, d_{\text{max}}) & \text{if } l_{\boldsymbol{p}} = l_{\boldsymbol{q}} \text{ and } d_{\boldsymbol{p}}, d_{\boldsymbol{q}} \neq \mathcal{U}, \\ 0 & \text{if } l_{\boldsymbol{p}} = l_{\boldsymbol{q}} \text{ and } d_{\boldsymbol{p}}, d_{\boldsymbol{q}} = \mathcal{U}, \\ d_{\text{max}} & \text{otherwise.} \end{cases} \quad (9)$$

A truncated linear distance with truncation thresholds $d_{\text{max}}^{\text{s}}$ and $d_{\text{max}}^{\text{t}}$ in spatial and temporal domains is used so as to avoid overpenalising large discontinuities.

4.3.3 Use of contextual information

In the temporal domain, direct application of the pair-wise penalty terms may result in oversmoothing particularly at points undergoing a rapid motion. In order to avoid this shortcoming, contextual information provided by the amplitude of the displacement vector $f(\boldsymbol{p},\boldsymbol{q})$ is used to define the weighting factor $u_{\boldsymbol{pq}} = \exp(-\gamma[f(\boldsymbol{p},\boldsymbol{q})]^2)$. Perceptually, it should be noted that temporal incoherences are most noticeable in static regions. Incidentally, these are also the regions where attenuation using contextual information will be minimal and the effect of temporal smoothing maximal.

4.4. Efficient energy minimisation using graph-cuts

Optimisation of the energy function defined in Eq. (2) is NP-hard. However an approximate solution with strong optimality properties can be obtained using the alpha-expansion algorithm based on graph-cuts [5]. Application of this algorithm requires the energy to be regular [17]. After observing that the spatial and temporal terms have a similar structure differing only in terms of coefficients independent of the labels, proof of the regularity follows immediately from the known regularity of the spatial component of the energy (see [14] for a proof in that case). As previously observed, a key challenge in extending layered depth estimation to the temporal domain is scalability. A naive implementation would see a dramatic increase in the number of labels which would make optimisation prohibitively slow and require a prohibitively large amount of memory. We propose a number of practical considerations to make this problem tractable. Firstly data terms are pre-computed for the first window of frames and stored in a list; then as the window is sliding along the sequence, data for the frame no longer required is discarded and replaced with data for the new frame. This guarantees that data terms are computed only once for each frame. Second, to reduce run-time during graph-cut optimisation, an optimised graph-cut implementation re-using the flow between cycles of the alpha-expansion algorithm is used [1]. Finally, memory consumption and run-time during graph-cut optimisation are further reduced by improving the initialisation and reducing the number of labels. For that, an initial solution is computed using the fast temporally unconstrained algorithm and used to reduce the number of hypotheses to a narrow band centred around the temporally unconstrained reconstruction. In practice, only labels located within 5 cm from the initial reconstruction are considered.

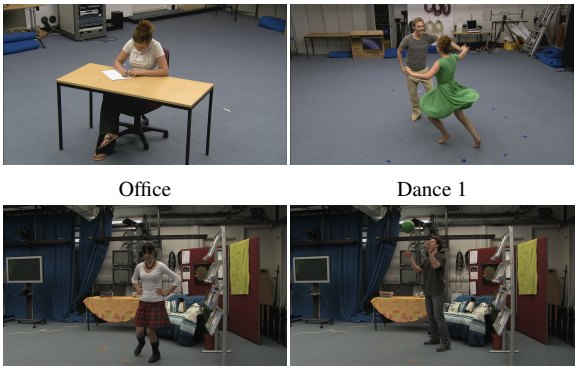


Figure 3. Example image for each test sequence.

5. Experimental results

Experiments are performed using four indoor datasets exhibiting various degrees of background complexity and foreground motion (see Fig. 3 for example images). The first two datasets were captured using eight Thomson Viper studio cameras, while seven Canon XHG1 camcorders were used for the other two datasets. In all cases, cameras are

	Run-time (s)	Memory requirement (GB)
Office	29/98	0.9/6.6
Dance 1	19/55	1.0/6.2
Dance 2	18/69	1.0/6.3
Ball	15/55	0.9/5.7

Table 1. Average run-time and memory consumption for last iteration (optimisation performed on a narrow band). Each cell give values for the frame-independent joint segmentation and reconstruction [15, 14] followed by proposed method.

static and configured in an arc around the foreground with at least 20° baseline separation between adjacent cameras. All cameras capture HD data and are fully calibrated and synchronised using genlock. All experiments where performed with a temporal window of size 7 (*i.e.* three neighbours on either side of the frame of interest) and $\lambda^{\mathrm{t}}_{\mathrm{con}} = \lambda^{\mathrm{s}}_{\mathrm{con}}$, $\lambda^{\mathrm{t}}_{\mathrm{s}} = 0.1$, $d^{\mathrm{t}}_{\mathrm{max}} = 100\times$[depth sampling step] and $\gamma = 1$. Other parameters were set as specified in [14] for each dataset. Run-times are provided in Table 1.

Segmentation results To assess the performance of the proposed algorithm in terms of segmentation capabilities, a comparison is carried out against the following algorithms:

- Background cut [22]: This is a frame-independent segmentation algorithm which generalises the chroma-keying and difference-keying approaches by combining global colour models and local colour models with contrast information in an MRF optimisation framework solved using graph-cuts. As in [22], the global foreground and background colour models are represented as GMMs learnt from a small number (one or two for the entire dataset) of manually annotated key-frames, while the local background colour models are represented as normal distributions at each pixel.

- Frame-independent joint segmentation and reconstruction [15, 14]: Results were generated for this algorithm by setting the window size to one in the proposed algorithm, effectively disabling the temporal constraints.

- Temporal filtering of frame-independent joint segmentation and reconstruction output. At each pixel in the reference image, we define a 1D temporal window of dimension 7 (same dimension as in the proposed method) centred at the pixel of interest and connecting pixels in neighbouring frames using the set of optical flow measurements. A pixel $\boldsymbol{p}$ is assigned the label which receives maximum support over the temporal neighbourhood $\mathcal{N}_t(\boldsymbol{p})$, that is:

$$l_{\mathrm{F}}(\boldsymbol{p}) = \arg\max_{l^*} \sum_{\boldsymbol{q}\in\mathcal{N}_t(\boldsymbol{p})} w_{\boldsymbol{p},\boldsymbol{q}} T(l_{\boldsymbol{q}} = l^*), \qquad (10)$$

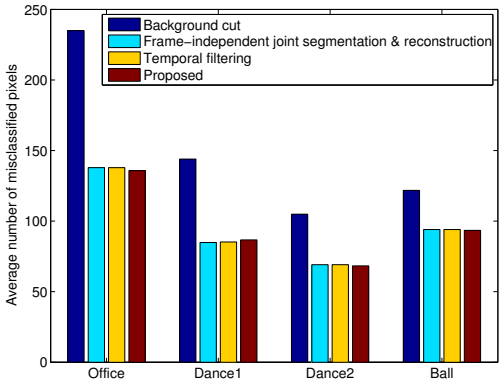


Figure 5. Evaluation of the accuracy of different segmentation techniques.

where the the function T returns 1 if its argument is true and 0 otherwise and $w_{\boldsymbol{p},\boldsymbol{q}}$ are the temporal attenuation weights used in the proposed method.

Segmentation results can be seen in Fig. 4 in the case of a single frame and in the supplementary video for full sequences. Segmentation accuracy is quantitatively assessed by computing the number of misclassified pixels using ground truth data obtained from manual annotation of a minimum of 30 frames for each dataset (see Fig. 5). The results demonstrate that the proposed method performs better than other segmentation techniques in static areas and similarly in other areas. While the improvement may seem modest in terms of reduction of the number of misclassified pixels, qualitatively flicker is usually attenuated with the temporally constrained algorithm (see supplementary video). This improvement cannot be measured by the quantitative evaluation method which only considers spatial accuracy.

Reconstruction results The proposed method produces a set of view-dependent depth maps, which can be combined into a global mesh representation using Poisson surface reconstruction [16]. These results are compared against the following techniques:

- Furukawa and Ponce [10]: This is one of the most accurate image-based reconstruction techniques according to the Middlebury evaluation benchmark. For best results and to allow comparison with other techniques, the algorithm was initialised using the segmentation output from the following technique.

- Frame-independent joint segmentation and reconstruction [15, 14].

- Temporal bilateral filtering of frame-independent joint segmentation and reconstruction output. Using the temporal window introduced in the previous section

Ground truth | Background cut | Frame-independent joint segmentation & reconstruction | Temporal filtering | Proposed

Figure 4. Segmentation results for one frame (results for full sequences can be seen in the supplementary video). For visualisation purposes, misclassified pixels are highlighted in red.

for temporal filtering we compute a filtered depth at each pixel $\boldsymbol{p}$ according to the formula:

$$d_{\mathrm{F}}(\boldsymbol{p}) = \frac{1}{W_{\boldsymbol{p}}} \sum_{\boldsymbol{q} \in \mathcal{N}_t(\boldsymbol{p})} w_{\boldsymbol{p},\boldsymbol{q}} G_d(|d_{\boldsymbol{p}} - d_{\boldsymbol{q}}|) d_{\boldsymbol{q}}, \quad (11)$$

where G_d is the Gaussian function for depth range values and the $W_{\boldsymbol{p}}$ is the normalisation weight such that

$$W_{\boldsymbol{p}} = \sum_{\boldsymbol{q} \in \mathcal{N}_t(\boldsymbol{p})} w_{\boldsymbol{p},\boldsymbol{q}} G_d(|d_{\boldsymbol{p}} - d_{\boldsymbol{q}}|). \quad (12)$$

The standard deviation for G_d was set to $\sigma_d = 2\,\mathrm{cm}$ allowing optimum smoothing of the static parts of the scene while minimising artefacts in dynamic scene parts. Only depth measurements for foreground pixels are considered when computing the filtered depth value; this avoids artefacts at object boundaries. As in the proposed technique, the obtained depths maps are then merged into a single representation using Poisson surface reconstruction [16].

Results in the case of a single frame are shown in Fig. 6 and Fig. 7 while results for full sequences can be seen in the supplementary video. Due to the difficulty to obtain valid ground truth reconstruction data for dynamic scenes, it was not possible to perform a quantitative evaluation of reconstruction accuracy. However the improvement resulting from the incorporation of temporal constraints can be seen in the video. It is apparent that even the best reconstruction techniques such as [10] or [15, 14] can produce disturbing temporal artefacts when applied independently to each frame of a sequence. In contrast, the proposed approach increases temporal coherence. These improvements are most noticeable on static or slow moving object such as the table top. Additionally, it can be observed that the temporal constraints preserve reconstruction quality in areas undergoing a fast motion and where optical flow fails. Temporal constraints appear also beneficial for increasing spatial consistency in poorly textured areas such as the person's legs in Fig. 6 and Fig. 7. Bilateral filtering of the frame-dependent joint segmentation and reconstruction output is also able to improve temporal consistency in quasi-static areas where noise is small but often fails in areas where noise in larger such as the person's legs.

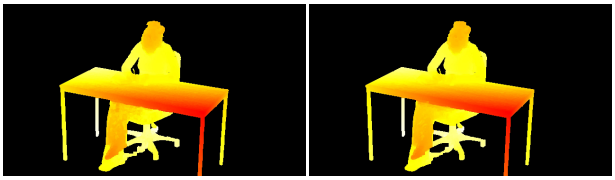
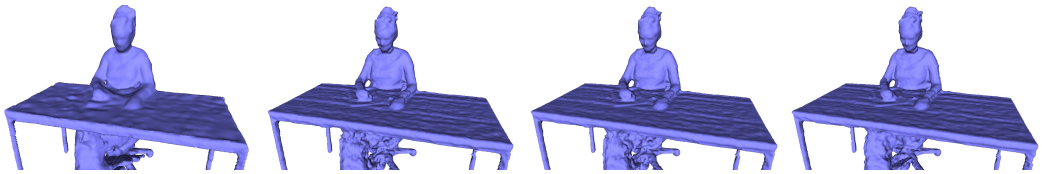


Figure 6. Example of depth map recovered for the office scene using frame-independent joint segmentation and reconstruction (left) and the proposed approach (right). Results with full sequences are available in the supplementary video.

6. Conclusions and future work

We introduced the first temporally consistent joint segmentation and reconstruction algorithm based on spatio-temporal energy optimisation. Generalisation to the spatio-temporal domain is non-trivial and requires several modifications to provide robustness to errors in optical flow and fast motion and scalability to allow use of sufficiently large temporal windows. Results demonstrate that the method is able to reduce noise on surfaces where optical flow correspondences can be reliably inferred without resulting in a reduction in quality in less constrained areas such as those undergoing a fast motion. To cope with fast motion and prevent any deterioration in these areas we introduce contextual information based on the amplitude of the local displacement to weight the influence of the temporal smoothing effect. This appears to work well in practice and correlates well with the fact that perception of inconsistencies is usually higher on static or slow moving objects. An interesting avenue for future work would be the extension of the framework to more general motion models. Possible approaches to achieve this would be to consider higher order cliques. Further work could also be carried out to reduce run-time and increase the temporal window size. This could be achieved through use of GPU acceleration [25] or scalable graph-cut optimisation algorithms [9].

Acknowledgements This work was supported by the EU FP7 project SCENE and the TSB project SyMMM.



Furukawa & Ponce | Frame-independent joint segmentation & reconstruction | Temporal bilateral filtering | Proposed

Figure 7. Reconstruction results with different techniques (results with full sequences available in supplementary video).

References

[1] K. Alahari, P. Kohli, and P. Torr. Reduce, reuse & recycle: Efficiently solving multi-label MRFs. In *CVPR*, 2008. 5

[2] X. Bai, J. Wang, D. Simons, and G. Sapiro. Video snapcut: robust video object cutout using localized classifiers. *SIGGRAPH*, 28(3), 2009. 1, 2

[3] M. Bleyer and M. Gelautz. Temporally consistent disparity maps from uncalibrated stereo videos. In *International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 383 –387, 2009. 1, 2

[4] Y. Boykov and G. Funka-Lea. Graph cuts and efficient N-D image segmentation. *IJCV*, 70(2):109–131, 2006. 2

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001. 1, 5

[6] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, volume 3024, pages 25–36, 2004. 3

[7] Y.-Y. Chuang, A. Agarwala, B. Curless, D. Salesin, and R. Szeliski. Video matting of complex scenes. *SIGGRAPH*, 21:243–248, July 2002. 1, 2

[8] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: a unifying framework for depth from triangulation. *PAMI*, 27(2):296–302, 2005. 2

[9] A. Delong and Y. Boykov. A scalable graph-cut algorithm for N-D grids. In *CVPR*, 2008. 7

[10] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 32(8):1362–1376, 2010. 6, 7

[11] B. Goldlücke, I. Ihrke, C. Linz, and M. Magnor. Weighted minimal hypersurface reconstruction. *PAMI*, 29(7):1194–1208–688, 2007. 3

[12] B. Goldlücke and M. Magnor. Joint 3D-reconstruction and background separation in multiple views using graph cuts. In *CVPR*, volume 1, pages 683–688, 2003. 3

[13] M. Gong. Enforcing temporal consistency in real-time stereo estimation. In *ECCV*, pages 564–577, 2006. 1, 2

[14] J.-Y. Guillemaut and A. Hilton. Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. *IJCV*, 93(1):73–100, 2011. 3, 4, 5, 6, 7

[15] J.-Y. Guillemaut, J. Kilner, and A. Hilton. Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complex dynamic scenes. In *ICCV*, pages 809–816, 2009. 3, 4, 6, 7

[16] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Symp. on Geometry Processing*, pages 61–70, 2006. 6, 7

[17] V. Kolmogorov and R. Zabih. What energy function can be minimized via graph cuts? *PAMI*, 26:147–159, 2004. 1, 5

[18] E. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *ICCV*, 2007. 2

[19] Y. Li, J. Sun, and H.-Y. Shum. Video object cut and paste. *SIGGRAPH*, 24(3):595–600, 2005. 2

[20] F. Liu and V. Philomin. Disparity estimation in stereo sequences using scene flow. In *BMVC*, 2009. 1, 2

[21] C. Richardt, C. Stoll, N. Dodgson, H.-S. Seidel, and C. Theobalt. Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos. *Computer Graphics Forum (Proc. Eurographics)*, 31(2), 2012. 2

[22] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum. Background cut. In *ECCV*, volume 3954, pages 628–641, 2006. 4, 6

[23] H. Tao, H. Sawhney, and R. Kumar. Dynamic depth recovery from multiple synchronized video streams. In *CVPR*, volume 1, pages 118–124, 2001. 1, 2

[24] S. Vedula, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *PAMI*, 27(3):475–480, 2005. 2

[25] V. Vineet and P. Narayanan. Solving multi-label MRFs using incremental alpha-expansion move on the GPUs. *ACCV*, 2009. 7

[26] J. Wang, P. Bhat, R. Colburn, M. Agrawala, and M. Cohen. Interactive video cutout. *SIGGRAPH*, 24:585–594, July 2005. 2

[27] M. Waschbüsch, S. Würmlin, and M. Gross. 3d video billboard clouds. *Proceedings of Eurographics, Computer Graphics Forum*, 26(3):561–569, 2007. 1, 2

[28] M. Yang, X. Cao, and D. Dai. Multiview video depth estimation with spatial-temporal consistency. *BMVC*, 2010. 2

[29] G. Zhang, J. Jia, T.-T. Wong, and H. Bao. Consistent depth maps recovery from a video sequence. *PAMI*, 31(6):974 – 988, 2009. 2

[30] L. Zhang, B. Curless, and S. Seitz. Spacetime stereo: shape recovery for dynamic scenes. In *CVPR*, volume 2, pages 367–74, 2003. 2

[31] J. Zhu, L. Wang, J. Gao, and R. Yang. Spatial-temporal fusion for high accuracy depth maps using dynamic MRFs. *PAMI*, 32(5):899–909, 2010. 2

[32] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *SIGGRAPH*, pages 600–608, 2004. 3