4D Temporally Coherent Light-field Video

Armin Mustafa   Marco Volino   Jean-Yves Guillemaut   Adrian Hilton

CVSSP, University of Surrey, United Kingdom

{a.mustafa, m.volino, j.guillemaut, a.hilton} @surrey.ac.uk

Abstract

Light-field video has recently been used in virtual and augmented reality applications to increase realism and immersion. However, existing light-field methods are generally limited to static scenes due to the requirement to acquire a dense scene representation. The large amount of data and the absence of methods to infer temporal coherence pose major challenges in storage, compression and editing compared to conventional video. In this paper, we propose the first method to extract a spatio-temporally coherent light-field video representation. A novel method to obtain Epipolar Plane Images (EPIs) from a spare light-field camera array is proposed. EPIs are used to constrain scene flow estimation to obtain 4D temporally coherent representations of dynamic light-fields. Temporal coherence is achieved on a variety of light-field datasets. Evaluation of the proposed light-field scene flow against existing multi-view dense correspondence approaches demonstrates a significant improvement in accuracy of temporal coherence.

1. Introduction

Light-field cameras capture multiple densely spaced viewpoints using either lenticular arrays for a single sensor [13], a camera densely scanned over a static scene to capture a large number of views [11] or multiple densely spaced cameras [32]. This allows photo-realistic rendering of novel viewpoints, depth-of-field effects and high-accuracy dense reconstruction [11, 27]. Reconstruction from densely sampled light-fields commonly employs an Epipolar Plane Image (EPI) representation to estimate spatial correspondence between images [33].

Light-fields have primarily been captured for static scenes due to the requirement for a large number of viewpoints resulting in a high-volume of data. Recently motivated by applications in immersive content production for Virtual and Augmented Reality (VR/AR), arrays of video cameras could be employed to acquire light-fields of dynamic scenes. However, conventional video editing techniques are not suitable to edit such data as they fail to exploit the spatial and temporal redundancy encoded in light-fields due to multiple discrete views. Hence there is a need for effi-
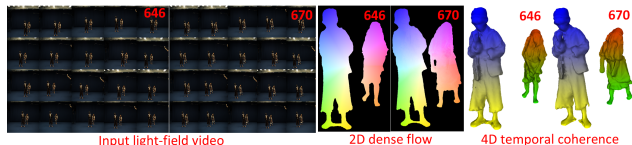

Figure 1. Proposed 4D temporally coherent dynamic light-fields

cient light-field video representations exploiting the spatio-temporal redundancy to enable compression and facilitate editing for live action VR. In this paper we propose the first method for spatio-temporally coherent dynamic light-field video, illustrated in Figure 1. This enables the use of light-field video for immersive VR content production. An EPI representation is used to estimate dense spatio-temporal correspondence from light-field video for temporal alignment.

Previous work on temporal alignment of complex dynamic objects has primarily focused on acquisition using multi-view cameras to reliably reconstruct the complete object surface at each frame using shape-from-silhouette and multiple view stereo [6, 23]. Robust techniques have been introduced for temporal alignment of the reconstructed non-rigid shape to obtain a 4D model based on tracking the complete surface shape or volume for complex motion [3, 4, 23, 8]. However, these approaches assume a reconstruction of the full non-rigid object surface at each time frame and do not easily extend to 4D alignment of partial surface reconstructions or depth maps. Recent work obtained reliable temporal alignment of partial surfaces for complex dynamic scenes [17, 26, 30, 16]. Dynamic Fusion [18] was introduced for 4D modelling from depth image sequences integrating temporal observations of non-rigid shape to resolve fine detail. However due to the high degree of spatial and temporal regularity and redundancy in the light-field video, existing methods for temporal alignment are not directly applicable. This paper exploits the spatio-temporal redundancy in light-field video for robust 4D spatio-temporally coherent representation of dynamic scenes. Sparse temporal correspondence tracks are obtained across the sequence for each view using temporal feature matching. These sparse temporal correspondence tracks are used to initialize dense scene flow estimation. A novel sparse-to-dense light-field
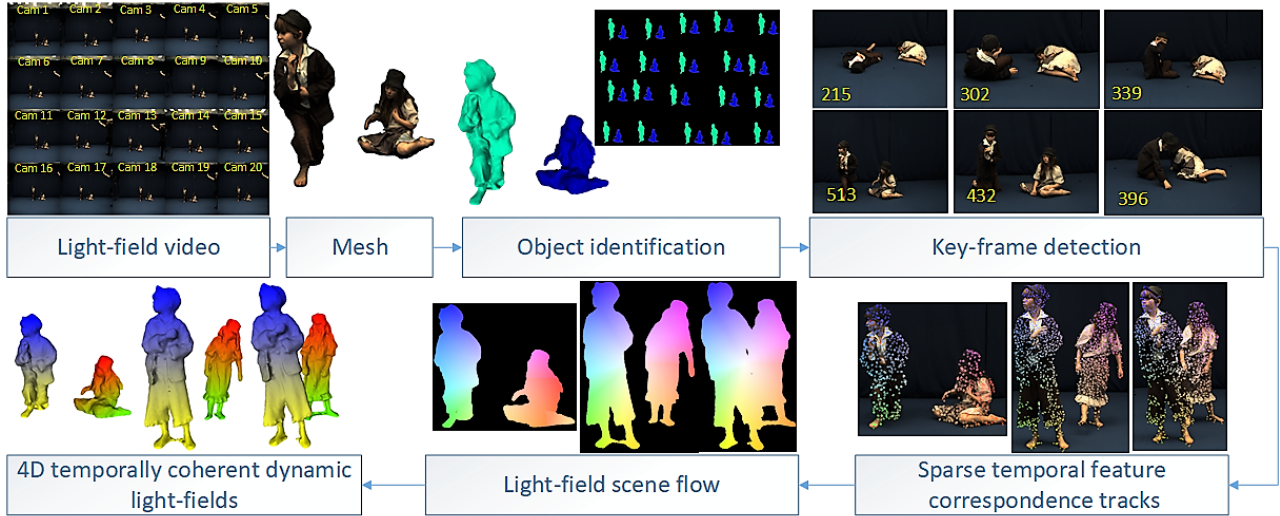
Figure 2. 4D Temporally Coherent Light-Field Video

scene flow is proposed exploiting EPIs to obtain a temporally coherent dense 4D representation of the scene, shown in Figure 2. The contributions of the proposed approach include:

- Temporally coherent 4D reconstruction of dynamic light-field video;
- EPI from sparse light-field video for spatio-temporal correspondence;
- Sparse-to-dense light-field scene flow exploiting EPI image information.

2. Related work

2.1. Light-fields

With the advent of virtual and augmented reality, light-field capture has been explored for live action immersive virtual reality experiences. Bolles et al introduced EPI [2] to represent light-field as a volume. This representation has been used effectively for depth estimation and segmentation [33, 28]. Methods have been proposed for depth-of-field effects [13] and image-based rendering [12, 7]. However existing light-field representations have been designed around capturing static objects. A recent method introduced oriented light-field windows to leverage the EPI information from light-fields and enable more robust and accurate pixel comparisons [21] to improve the performance for scene flow estimation for a pair of images. An efficient light-field representation for live action data should account for the high level of redundancy of light-field video and handle multiple resolutions, occlusions, and high variation in depth in complex scenes. Another challenging aspect relates to the development of creative tools dedicated to light-field editing [9]. There is currently a lack of an efficient representation to enable light-field video editing for content production. In this paper we propose a method to exploit the spatio-temporal redundancy in light-fields. EPI is used to obtain dense scene flow for 4D temporally coherent representation of dynamic light-field video to enable compression and efficient light-field video editing (colour grading, object removal, infilling etc.).

2.2. 4D reconstruction of dynamic scenes

For conventional single view depth sequences and multiple view reconstruction of dynamic scenes techniques have been introduced to align sequences using correspondence information between frames. Methods have been proposed to obtain sparse [15, 10, 35, 24] and dense [14, 34, 1] correspondence between consecutive frames for entire sequences. Existing sparse correspondence methods work independently on a frame-by-frame basis for a single view [24] or multiple views [10] and require a strong prior initialization [35]. Existing feature matching techniques either work in 2D [24] or 3D [14] or for sparse [10, 35] or dense [34] points. Other methods are limited to RGBD data [34] or stereo pairs [14] for dynamic scenes. Dense matching techniques include scene flow methods. Scene flow techniques [29, 1] typically estimate the pairwise surface or volume correspondence between reconstructions at successive frames but do not extend to 4D alignment or correspondence across complete sequences due to drift and failure for rapid and complex motion. In this paper we propose sparse-to-dense temporal alignment exploiting the high spatio-temporal redundancy in light-fields to robustly align light-field video captured with sparse camera arrays.

3. Methodology

The high spatial and temporal redundancy in light-field video makes it challenging to use in content production for AR/VR. The aim of this work is to obtain a 4D temporally coherent representation of dynamic light-field video exploiting spatio-temporal redundancy. This provides an efficient structured representation for light-field editing and use in immersive VR content production.

3.1. Overview

Given an independent per-frame surface or depth reconstruction from the light-field video captured with a sparse camera array the problem is to simultaneously estimate the temporal correspondence of the input light-field across all views for the entire sequence. This is achieved efficiently by estimating the temporal alignment of the reconstructed surface between each time frame and propagating this across all light-field camera views. A coarse-to-fine approach is introduced that initially estimates temporal correspondence based on sparse features and then estimates dense scene flow initialised from the sparse features. To ensure robust tracking, key-frames are identified as a reference to minimise drift in long-term tracking. An overview of the 4D temporally coherent reconstruction of light-field video is presented in Figure 2.

Light-field video: Light-field video of the dynamic scene is captured using a camera array. Multiple-view stereo is performed to obtain a per frame surface mesh reconstruction [20]. Throughout this work a sparse 5x4 array with a range of up to 50cm x 50cm is used to capture light-field video.

Object identification: The reconstructed point cloud is clustered in 3D [19] with each cluster representing a unique foreground object, shown in Figure 3.

Key-frame detection: Key-frames are detected for light-field video exploiting redundant spatial information across views to identify a set of unique reference frames for stable long-term tracking. Surface tracking is performed between key-frames to reduce the accumulation of errors in sequential tracking due to large non-rigid motion for long sequences ($\approx$ 700 frames).

Sparse temporal feature correspondence tracks: Reconstructed 3D points projected at all frames are matched frame-to-frame across the sequence to estimate sparse temporal feature tracks for each dynamic object for each light-field camera view. These sparse temporal correspondence tracks are used to initialize light-field scene flow to handle occlusions and improve robustness.

Light-field scene flow: We propose to estimate dense scene flow between images by exploiting EPI information from light-field video based on oriented light-field windows [21] initialised by the sparse feature correspondence per light-field view at each time instant. This exploits the spatio-temporal redundancy in light-fields to give 2D dense correspondences which are back-projected to the 3D mesh to obtain a 4D spatio-temporally coherent dynamic light-field video representation.
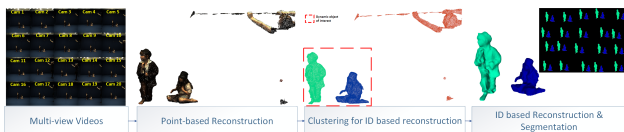

Figure 3. Object identification example on light-field frame

3.2. Key-frame detection

Aligning per-frame reconstruction for long sequences leads to drift due to accumulation of errors in alignment between frames and failure is observed due to large non-rigid motion. To tackle this problem we detect key-frames across the sequence. Key-frame detection exploits the spatial redundancy in light-field capture by fusing the appearance, distance and shape information across all views (N_v) in the sparse camera array. Temporal coherence is introduced between key-frames as explained in Section 3.3 and 3.4.

Appearance Metric ($M_{i,j}^c$)**:** This measures appearance similarity between frame i and j for each object region in light-field view c. It is the ratio of the number of temporal feature correspondences $Q_{i,j}^c$ to the total number of features in the object region at frame i, R_i^c and j, R_j^c: $M_{i,j}^c = \frac{2Q_{i,j}^c}{R_i^c + R_j^c}$

Distance Metric ($L_{i,j}^c$)**:** This metric measures the distance between frame i and j for each object in each view c, it is defined as: $L_{i,j}^c = \frac{j-i}{D_{max}^c}$ where $j > i$ and D_{max}^c is the maximum number of frames between key-frames for view c. This term ensures that the distance between two key-frames does not exceed D_{max}^c. This is set to 100.

Shape Metric ($I_{i,j}^c$)**:** Gives the shape overlap between pairs of frames for each object in the light-field video. It is defined as the ratio of the intersection of the aligned segmentation [5] $h_{i,j}^c$ to the union of the area $A_{i,j}^c$: $I_{i,j}^c = \frac{h_{i,j}^c}{A_{i,j}^c}$

Key-frame similarity metric: The metrics defined above are used to calculate the similarity between frames as follows: $D_{i,j} = 1 - \frac{1}{3N_v} \sum_{c=1}^{N_v} (M_{i,j}^c + I_{i,j}^c + L_{i,j}^c)$. All frames with similarity > 0.75 are selected as key-frames defined as $K_i = \{k^{i+1}, k^{i+2}, ..., k^{N_f^i}\}$ where $i = 1$ to N_k (number of key-frames in light-field video) and N_f^i is the number of frames between K_i and K_{i+1}.

3.3. Sparse temporal feature correspondence

Numerous approaches have been proposed to temporally align moving objects in 2D using either feature matching or optical flow. However these methods may fail in the case of occlusion, movement parallel to the view direction, large motions and visual ambiguity. To overcome these limitations we match sparse feature points from all light-field camera views at each time instant for each object. This is used to estimate the similarity between the object surface observed at different frames of the light-field video for key-frame detection and subsequently to initialize dense light-field scene flow between frames.

3D points corresponding to each dynamic object are projected at each frame to ensure spatial light-field coherence. The features for each light-field camera view are defined as: $F_i^c = \{f_1^c, f_2^c, ..., f_{R_i^c}^c\}$, where $c = 1$ to N_v and $N_v = 20$ in our case. R_i^c are the 3D points visible at each frame i. Nearest neighbour matching is used to establish matches
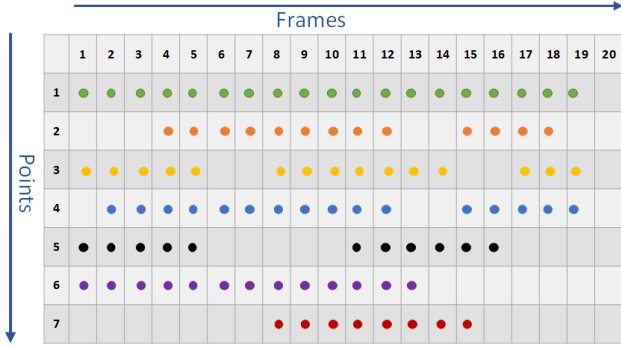
Figure 4. Example of sparse feature correspondence tracks between features. The ratio of the first to second nearest neighbour descriptor matching score is used to eliminate ambiguous matches ($ratio < 0.85$). This is followed by a symmetry test which employs the principle of forward and backward match consistency to remove erroneous inconsistent correspondences. To further refine the sparse matching and eliminate outliers we enforce local spatial coherence in the matching. For matches in an $m \times m$ ($m = 11$) neighbourhood of each feature we find the average Euclidean distance and constrain the match to be within a threshold ($\pm\eta < 2 \times$ Average Euclidean distance).

Sparse temporal correspondence tracks are obtained by performing exhaustive matching between all frames N_f^i for each key-frame for the entire sequence. Feature matching is performed between frames such that features at view c frame i, F_i^c are matched to features at view c to frames $j = \{i+1, ..., N_f^i\}$. This gives us correspondences for all the frames N_f^i with key-frame K_i. Point tracks are constructed from this correspondence information for key-frame K_i. The same process is repeated for key-points which are not part of point tracks at the corresponding key-frame for all frames $j = \{i+1, ..., N_f^i\}$. Any new point-tracks are added to the list of point tracks for key-frame K_i. The exhaustive matching between frames per key-frame handles reappearance and disappearance of points due to occlusion or object movement. An example of sparse feature correspondence tracks in shown in Figure 4. Frame 1 is a key-frame and each point track is represented by a unique colour with missing points showing partial feature correspondence tracks. Sparse correspondences are also obtained between all key-frames to initialise dense light-field flow.

3.4. Light-field scene flow

Dense temporal correspondence is estimated using a novel light-field scene flow method using pairwise dense correspondence which is estimated using the light-field oriented window matching. This combines information across all light-field views to achieve robust temporal correspondence. The sparse feature correspondences provides a robust initialisation of the proposed dense flow for large non-rigid shape deformation in the light-field video.

The high-level of inherent redundancy in the light-field images is used to improve the robustness and reliability of dense scene flow by using oriented light-field windows on the EPI for matching. This has been shown to improve the temporal correspondences for scene flow [21]. In this paper we propose a novel method to obtain EPIs from sparse light-field data to improve the quality of dense scene flow. Our approach uses oriented light-field windows [21] to represent surface appearance information for matching as illustrated in Figure 5.

3.4.1 Epipolar Plane Image (EPI)

Traditional light-field data is captured with a plenoptic camera or a dense camera array typically capturing 200-300 views [22]. An example dataset is illustrated in Figure 6 (a) captured with 289 images. Given a dense set of views the EPI provides a representation for estimating correspondence across views from the regular structure of the image i.e. slant lines correspond to the same surface point. The 2D slice of such representation has the same width as the captured image and height is given by the number of views in a camera array row (17 in Figure 6 (a)). With dense sampling the disparity between adjacent views is typically sub-pixel giving slant lines in the EPI. Approaches have been proposed to estimate depth and segmentation from this dense EPI representation [33, 28]. However for sparse camera sampling the disparity between views may be several pixels making it difficult to directly establish surface correspondence from the EPI, Figure 6 (b). In our case of sparse views with just 20 cameras, the height of this EPI reduces to 5 pixels which makes it challenging to utilize the regularity of the information from EPI obtained from light-fields.

In this paper we aim to use the EPIs to introduce spatio-temporal coherence in dynamic light-field video. We propose a novel method to create an EPI parametrized representation from sparse light-field capture. We assume that the calibration is known. All the images at each time instant are undistorted and rectified with respect to the reference camera. The depth information at each time instant is used to resample the light-field to create an EPI representation of sparse light-field data. The algorithm to obtain EPI from image, calibration and depth information is illustrated in Figure 5; the stages are as follows:

- The dense point-cloud P of each dynamic object is projected on the undistorted and rectified images. $B_{i,j}^c$ is the set of points in view c. For each row j of the projected points on the dynamic object a 2D EPI is obtained.
- The size of 2D EPI is set to $W \times H$, where W is the width of input images and $H = N_w \times \mu$ is estimated corresponding to the number of cameras in each row (N_w). μ is a constant introduced to increase
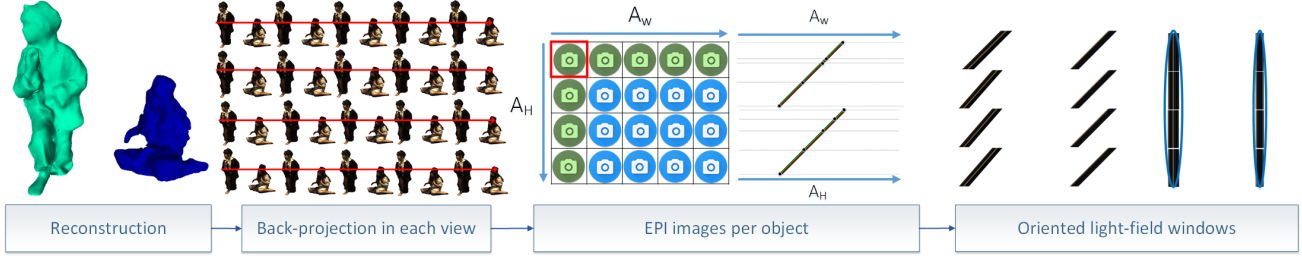
Figure 5. An example of EPI representation of sparse light-field (blue ellipse represents Gaussian weighted window)

the distance between views due to sparse camera sampling. It is set to 50 in our case to maximize performance. For each row of the camera array, H_o 2D EPIs are obtained for each dynamic object, where H_o is the height of the object. The corresponding projected points $B_{i,j}^c$ for each point $P_{i,j}$ are plotted on the 2D EPI with x coordinates the same as that of input images and y coordinates are estimated using the translation information from the calibration, defined as: $\frac{H \ X \ \text{Distance between consecutive cameras}}{\text{Maximum distance between pair of cameras}}$

- Given the set of image samples corresponding to a given surface point we fit a line in the EPI due to imperfect camera array. A scene point is represented by a line in the 2D EPI. A similar process is repeated for the entire point-cloud P to obtain multiple 2D EPIs. Given a 2 dimensional camera array there are 2 sets of EPIs obtained by resampling along epipolar lines in the vertical and horizontal directions

The EPIs obtained from sparse camera arrays are used to constrain dense light-field flow using oriented light-field windows.

3.4.2 Oriented light-field windows

Oriented light-field windows exploit the regularity information in the EPI to enable more robust and accurate pixel comparisons [21] over spatial windows which suffer from defocus blur and loss of precision. In this paper we propose to use these oriented light-field windows on the EPIs obtained from the sparse light-field data to estimate the light-field scene flow temporal correspondence.

Each scene point can be represented by an oriented window in the light-field ray space. The shear of the ray is related to the depth of the 3D point and the size of the window is defined by spatial and angular Gaussian weights. However the ray moves with object motion, hence there is a need to account for shear and translation. The oriented light-field window corresponding to a scene point for a light-field, L is computed as follows:

$$O_{d,x0,y0}(x,y,u,v) = (W * S_d * T_{x0,y0})[L] \qquad (1)$$

where, x_0, y_0 is the centre of the window, x, y represents the image plane and u, v represents the camera plane. S_d is the shear operation for depth d, defined as $S_d[L] = L_d(x,y,u,v) = L(x + u(1 - \frac{1}{d}), y + v(1 - \frac{1}{d}), u, v)$,
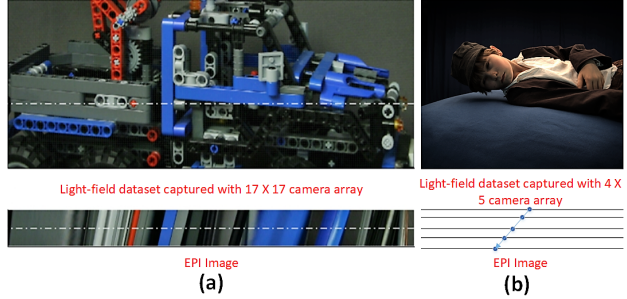
Figure 6. Illustration of EPI from (a) dense Stanford dataset [22] (EPI from [28]) and (b) sparse light-field capture

$T_{x0,y0}$ is the translation operator defined as, $T_{x0,y0}[L] = L_{x0,y0}(x,y,u,v) = L(x + x0, y + y0, u, v)$ and W is the Gaussian weighted windows defined as: $W[L] = L(x,y,u,v)N(x,y;\sigma_{xy}^2)A(u,v)$. $N()$ is the 2D Gaussian distribution windows and $A(u,v)$ is the corresponding row and column of the camera array (shown as green cameras in Figure 5, 3rd step) of the reference light-field view $((u_c, v_c)$ shown in red box), defined as:

$$A(u,v) = \begin{cases} 1, & \text{if } v = v_c \text{ or } u = u_c \\ 0, & otherwise \end{cases}$$

This formulation could be easily extended to dense camera array. We propose to use this oriented light-field window to robustly estimate the dense scene flow for light-field video.

3.4.3 Dense light-field scene flow

Oriented light-field windows are used to estimate flow between consecutive frames and between key-frames, Figure 7, using the sparse temporal feature correspondence as initialization for each view. Flow is estimated on the object region for each light-field view, as illustrated in Figure 8, to obtain dense temporal correspondence.

The flow is formulated as a translation of each pixel location $p = (x_p, y_p)$ in image I by $m_p = (\delta x_p, \delta y_p)$ in time. We formulate the computation of flow $\boldsymbol{M}$ per view for each dynamic object by minimization of the cost function:

$$E(\boldsymbol{M}) = \sum_{p \in I} \lambda_L E_L(p, m_p) + \lambda_C E_C(p, m_p) + \lambda_R E_R(p, m_p)$$

$$(2)$$

The cost consists of three terms: the light-field consistency E_L for the oriented light-field window alignment; the appearance term E_C for brightness coherency; and the regularization term E_R to avoid sudden peaks in flow and maintain the consistency. Colour and regularization terms are
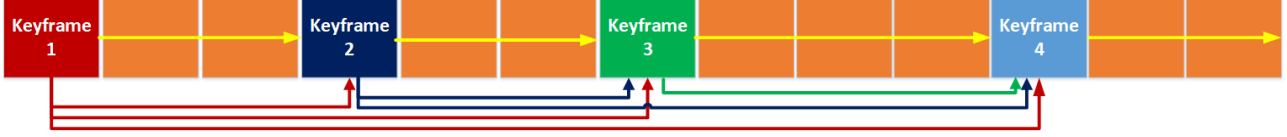
Figure 7. Illustration of dense correspondence between frames and key-frames for full sequence 4D alignment: arrows indicate sparse correspondence tracks and dense correspondence

common to optical flow problems [25] and the light-field consistency is introduced for the first time to improve dense flow for sparse light-field video.

Light-field Consistency: The 2D EPIs obtained from the sparse light-field views are used to define oriented light-field windows for each scene point. These windows encapsulate the observed multi-view light-field appearance of the corresponding surface point and can be matched over time to estimate the temporal correspondence, defined as:

$$E_L(p, m_p) = \left\| O_{d_t,p}(p, u, v, t) - O_{d_{t+1}, p+m_p}(p, u, v, t+1) \right\|^2$$

where $O_{d_t,m}()$ is the oriented light-field window at time t with depth d_t as defined in Equation 1.

Appearance Consistency: This adds the brightness consistency assumption to the cost function generalized for all N_v light-field cameras for both time steps. This term is obtained by integrating the sum of three penalizers over the reference image domain. $e_C^T()$ penalizes deviation from the brightness constancy assumption in time for same views; $e_C^V()$ penalizes deviation from the brightness constancy assumption between the reference view and each of the other views at time $t+1$. $e_C^S()$ forces the flow to be close to nearby sparse temporal correspondences.

$$E_C(p, m_p) = e_C^T(p, m_p) + e_C^V(p, m_p) + e_C^S(p, m_p)$$

$$e_C^T(p, m_p) = \sum_{i=1}^{N_v} \left\| (I_i(p, t) - I_i(p + m_p, t+1)) \right\|^2$$

$$e_C^V(p, m_p) = \sum_{i=2}^{N_v} \left\| (I_1(p, t) - I_i(p + m_p, t+1)) \right\|^2$$

$$e_C^S(p, m_p) = \sum_{i=1}^{N_v} e_C^S = \begin{cases} 0 & \text{if } p \in N \\ \infty & otherwise \end{cases}$$

where $I_i(p, t)$ is the intensity at point p and time t in camera i. This term denotes that the flow vector m is located within a window from a sparse constraint at p and it forces the flow to approximate the sparse 2D temporal correspondence tracks.

Regularization: This penalizes the absolute difference of the flow field to enforce motion smoothness and handle occlusions and areas with low confidence.

$$E_R(p, m_p) = \sum_{p,q \in N_p} \left\| \Delta m \right\|^2 \left(\lambda_R^L e_R^L(p, q, m_p, m_q) + \right.$$

$$\left. \lambda_R^C e_R^C(p, q, m_p, m_q) \right)$$

$$e_R^L(p, q, m_p, m_q) = \underset{q \in N_p}{mean} E_L(q, m_q) - \underset{q \in N_p}{min} E_L(q, m_q)$$

$$e_R^C(p, q, m_p, m_q) = \underset{q \in N_p}{mean} E_C(q, m_q) - \underset{q \in N_p}{min} E_C(q, m_q)$$
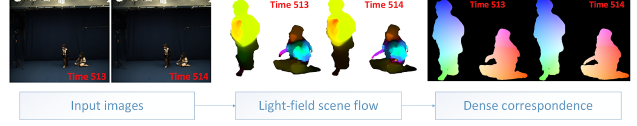

Figure 8. Light-field scene flow for two consecutive frames

where $\Delta m = m_p - m_q$ and we compute e_R^L and e_R^C as the minimum subtracted from the mean data energy within the search window N_p for each pixel p.

Occlusions: To detect occlusions, we compare the forward flow from t to $t+1$, and the backward flow from $t+1$ to t. Occluded pixels are robustly indicated by large differences in the forward/backward motion and excluded as outliers.

Optimization: For each pixel p, the energy is estimated as defined in eq. 2 on a window N_p, as in the SimpleFlow algorithm [25] to estimate the flow vector m. The optical flow is optimized over a multi-scale pyramid with warping between pyramid levels, resulting in a coarse-to-fine strategy that allows the estimation of large displacements by minimizing the equation defined as: $F(\boldsymbol{M}) = \arg\min E(\boldsymbol{M})$

3.4.4 4D temporally coherent light-field video

The estimated dense flow for each view is back projected to the 3D visible surface to establish dense 4D correspondence between frames (N_f^i) and between key-frames K_i as seen in Figure 7 to obtain 4D temporally coherent light-field video. Dense 4D correspondence is first obtained for the light-field view with maximum visibility of 3D points. To increase surface coverage correspondences are added in order of visibility of 3D points for different sparse light-field views. Dense temporal correspondence is propagated to new surface regions as they appear using the sparse feature correspondence tracks and respective dense light-field scene flow. Dense scene flow is estimated between key-frames for robust long-term surface tracking.

4. Results and Performance Evaluation

The proposed approach is tested on various light-field captures. The properties of the evaluation datasets are presented in Table 1. Algorithm parameters set empirically are constant for all results.

Sparse and dense correspondence are obtained on the sparse light-field dynamic data and the colour coded results are shown in Figure 9 for Walking dataset and in Figure 10 for Sitting and Waking up dataset using the method explained in Section 3.1. To illustrate the 2D dense alignment the silhouette of the dense mesh on key-frames is colour coded

Datasets	Sequence length	No. of views	Shot level	Resolution	Key-frames	Avg. sparse tracks
Walking	667	20	Far	2448 X 2048	15	1934
Sitting	694	20	Mid-level	2448 X 2048	13	1046
Wakingup	270	20	Close-up	2448 X 2048	7	2083
Running	140	20	Far	2448 X 2048	5	1278
Magician	353	20	Close-up	2448 X 2048	6	1312

Table 1. Properties of all datasets: Key-frames gives the number of key-frames detected for the dynamic sequence and Avg. sparse tracks gives the number of sparse temporal correspondence tracks averaged over the entire sequence for each object

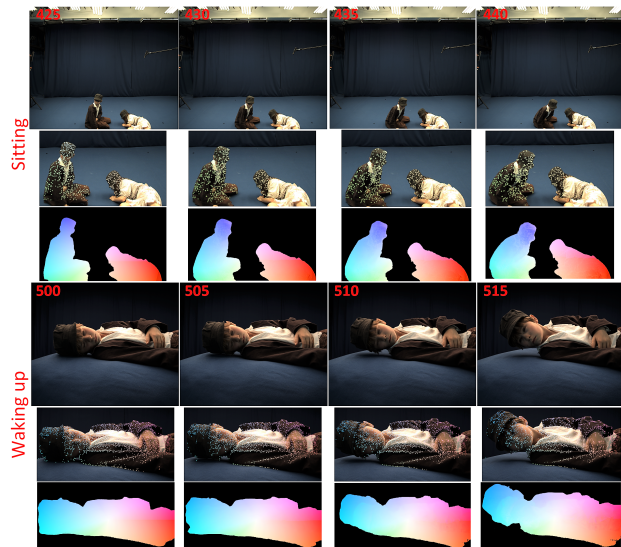Figure 9. Dense 2D and 4D correspondence for Walking sequence across key-frames

Figure 10. Sparse temporal correspondences and dense flow results on 2 light-field sequences: Sitting and Waking up

Figure 11. 4D temporal alignment between frames for Walking and Magician dataset

Datasets	Prop.	DFwLF	4DM	DF	SF
Walking	**0.45**	0.59	0.58	0.81	1.05
Sitting	**0.51**	0.73	0.71	1.13	1.83
Waking up	**0.39**	0.56	0.53	0.89	1.17
Running	**0.65**	0.87	0.92	1.23	1.95
Magician	**0.59**	0.82	0.83	1.05	1.67

Table 2. Silhouette overlap error for all the datasets. Prop. represents proposed approach, 4DM is 4DMatch, DF is Deepflow and SF is Simpleflow

and the colours are propagated between frames using dense scene flow explained in Section 3.4. Results of the proposed 4D temporal alignment, illustrated in Figure 11 shows that the colour of the points remains consistent between frames. The proposed approach is qualitatively shown to propagate the correspondences reliably over the entire light-field video for complex dynamic scenes with large non-rigid motion.

Qualitative evaluation: For comparative evaluation we use:(a) state-of-the-art dense flow algorithm Deepflow [31]; (b) dense flow without light-field consistency (DFwLF) in eq. 2; (c) a recent algorithm for alignment of partial surfaces (4DMatch) [17] and (d) Simple flow [25]. Qualitative results against DFwLF, 4DMatch, Deepflow and Simpleflow shown in Figure 12 indicate that the propagated colour map does not remain consistent across the sequence for large
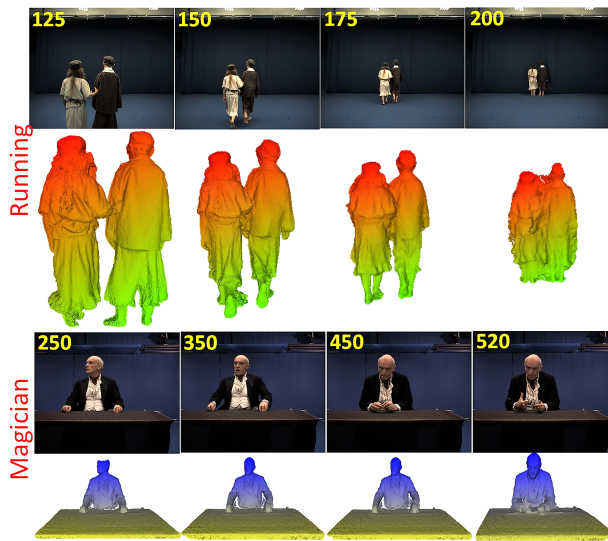
motion as compared to the proposed method (red regions indicate correspondence failure).

Quantitative evaluation: For quantitative evaluation we compare the silhouette overlap error (SOE). Dense correspondence over time is used to create propagated mask for each image. The propagated mask is overlapped
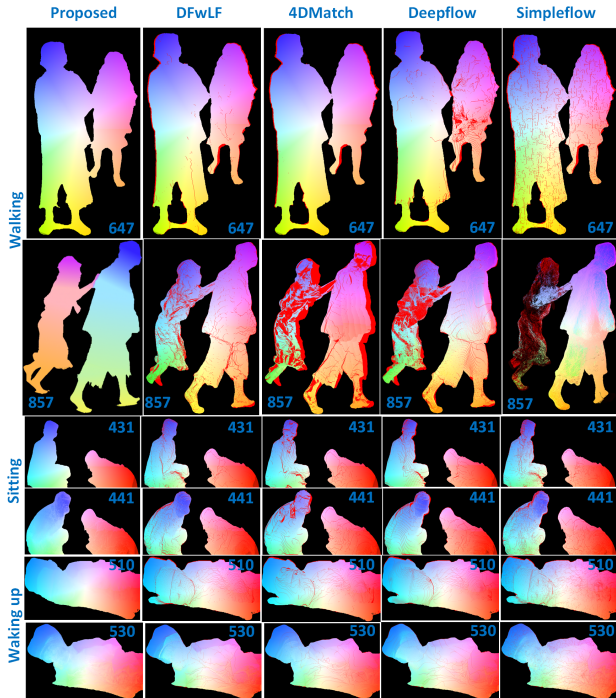
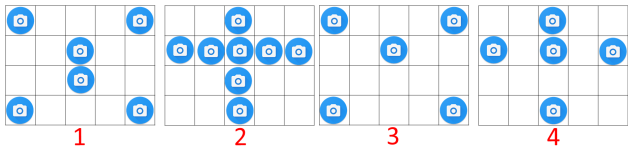Figure 12. Dense flow comparison results on different light-field sequences

Figure 13. Different light-field camera array configurations

	Frame-to-frame		Keyframe-to-frame	
Methods	**Mean**	**S.D.**	**Mean**	**S.D.**
Proposed	**3.78**	1.45	**4.34**	1.74
DFwLF	5.93	2.60	7.41	3.73
4DM	5.30	2.12	6.95	3.12
DF	6.28	3.77	15.79	6.36
SF	7.82	4.31	21.92	8.45

Table 3. Temporal coherence evaluation for Walking dataset against existing methods: S.D. is the standard deviation

with the silhouette of the projected surface reconstruction at each frame to evaluate the accuracy of the dense propagation. The error is defined as: $SOE = \frac{1}{MN} \sum_{i=1}^{N} \sum_{c=1}^{M} \frac{\text{Area of intersection}}{\text{Area of back-projected mask}}$. Evaluation against the different techniques is shown in Table 2 for all datasets. As observed the silhouette overlap error is lowest for the proposed approach showing relatively high accuracy.

We evaluate the temporal coherence across Walking sequence, by evaluating the variation in appearance for each scene point between frames and between key-frames and frames for state-of-the-art methods, defined as: $\sqrt{\frac{\Delta r^2 + \Delta g^2 + \Delta b^2}{3}}$, where Δ is the difference operator. Evaluation shown in Table 3 against state-of-the-art methods demonstrates the stability of long term temporal tracking

Config.	Walk	Sit	Wake	Run	Magician
1	95.15	92.58	89.64	91.03	90.82
2	91.33	86.73	86.98	88.78	89.90
3	90.76	85.15	86.05	87.21	86.21
4	87.82	82.40	85.16	86.56	80.91

Table 4. Completeness of dense 3D correspondence averaged over the entire sequence in % for different camera configurations

for proposed method. Evaluation of the proposed method against dense flow without light-field consistency (DFwLF) demonstrates the usefulness of information from the EPIs in the dense flow in Section 3.4.

Light-field camera array configuration: We evaluate the performance of the proposed 4D temporal alignment with different light-field camera configurations shown in Figure 13. The completeness of the 3D points at each time instant for all camera configurations as observed in Table 4 is defined as: $\frac{100}{MN} \sum_{i=1}^{N} \sum_{c=1}^{M} \frac{\text{Number of 3D points propagated in each configuration}}{\text{Number of 3D points propagated with full camera array}}$. The evaluation demonstrates a drop in 3D correspondence with the reduction in number of cameras in different camera configurations, specially for close-up shots (Wakeup and Magician). However configuration 1 and 3 with cameras in the corners provide a better coverage compared to 2 and 4.

Limitations: The proposed method fails for fast spinning objects; scenes with uniform appearance and highly crowded dynamic environments. This is due to the failure of sparse and dense correspondence due to high ambiguity.

5. Conclusion

This paper introduced the first algorithm to obtain a 4D temporally coherent representation of dynamic light-field video. A novel method to obtain EPIs from sparse light-field video for spatio-temporal correspondence is proposed. Sparse-to-dense light-field scene flow is introduced exploiting information from the EPIs. Dense correspondence is fused spatially for 4D temporally coherent light-field video. The proposed approach is evaluated on various light-field sequences of complex dynamic scenes with large non-rigid deformations to obtain a temporally consistent 4D representation and demonstrating accuracy of the resulting 4D alignment. 4D Light-field video provides a spatio-temporally coherent representation to support subsequent light-field video compression or editing to replicate the functionality of conventional video editing allowing the propagation of edits both spatially across views and temporally across frames.

References

[1] T. Basha, Y. Moses, and N. Kiryati. Multi-view scene flow estimation: A view centered variational approach. In *CVPR*,

pages 1506–1513, 2010. 2

[2] R. C. Bolles, H. Baker, and D. H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *IJCV*, 1(1):7–55, Mar 1987. 2

[3] C. Budd, P. Huang, M. Klaudiny, and A. Hilton. Global non-rigid alignment of surface sequences. *Int. J. Comput. Vision*, 102:256–270, 2013. 1

[4] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In *ECCV*, pages 326–339, 2010. 1

[5] G. D. Evangelidis and E. Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:1858–1865, 2008. 3

[6] J.-S. Franco and E. Boyer. Exact polyhedral visual hulls. In *Proc. BMVC*, pages 32.1–32.10, 2003. 1

[7] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 43–54, New York, NY, USA, 1996. ACM. 2

[8] C. Huang, C. Cagniart, E. Boyer, and S. Ilic. A bayesian approach to multi-view 4d modeling. *Int. J. Comput. Vision*, 116:115–135, 2016. 1

[9] A. Jarabo, B. Masia, A. Bousseau, F. Pellacini, and D. Gutierrez. How do people edit light fields? *ACM Transactions on Graphics (SIGGRAPH 2014)*, 33(4), 2014. 2

[10] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 2

[11] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph.*, 32(4):73:1–73:12, July 2013. 1

[12] M. Levoy and P. Hanrahan. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 31–42, New York, NY, USA, 1996. ACM. 2

[13] M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz. Light field microscopy. *ACM Trans. Graph.*, 25(3):924–934, July 2006. 1, 2

[14] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 2

[15] A. Mustafa and A. Hilton. Semantically coherent co-segmentation and reconstruction of dynamic scenes. In *CVPR*, 2017. 2

[16] A. Mustafa, H. Kim, J.-Y. Guillemaut, and A. Hilton. Temporally coherent 4d reconstruction of complex dynamic scenes. In *CVPR*, 2016. 1

[17] A. Mustafa, H. Kim, and A. Hilton. 4d match trees for non-rigid surface alignment. In *ECCV*, 2016. 1, 7

[18] R. Newcombe, D. Fox, and S. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015. 1

[19] R. B. Rusu. *Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments*. PhD the-sis, Computer Science department, Technische Universitaet Muenchen, Germany, 2009. 3

[20] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, pages 519–528, 2006. 3

[21] P. Srinivasan, M. Tao, R. Ng, and R. Ramamoorthi. Oriented light-field windows for scene flow. In *ICCV*, December 2015. 2, 3, 4, 5

[22] Stanford Graphics Laboratory. The (New) Stanford Light Field Archive, 2008. 4, 5

[23] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. In *ICCV*, pages 915–922, 2003. 1

[24] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, pages 438–451, 2010. 2

[25] M. W. Tao, J. Bai, P. Kohli, and S. Paris. Simpleflow: A non-iterative, sublinear optical flow algorithm. *Computer Graphics Forum (Eurographics 2012)*, 31(2), May 2012. 6, 7

[26] A. Tevs, A. Berner, M. Wand, I. Ihrke, M. Bokeloh, J. Kerber, and H.-P. Seidel. Animation cartography; intrinsic reconstruction of shape and motion. *ACM Trans. Graph.*, pages 12:1–12:15, Apr. 2012. 1

[27] V. Vaish, R. Szeliski, C. L. Zitnick, S. B. Kang, and M. Levoy. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *CVPR*, page 2331, 2006. 1

[28] S. Wanner and B. Goldluecke. Globally consistent depth labeling of 4D light fields. In *CVPR*, pages 41–48, June 2012. 2, 4, 5

[29] A. Wedel, T. Brox, T. Vaudrey, C. Rabe, U. Franke, and D. Cremers. Stereoscopic scene flow computation for 3D motion understanding. *IJCV*, 95:29–51, 2011. 2

[30] L. Wei, Q. Huang, D. Ceylan, e. Vouga, and H. Li. Dense human body correspondences using convolutional networks. *CoRR*, abs/1511.05904, 2015. 1

[31] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *ICCV*, pages 1385–1392, 2013. 7

[32] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, M. Adams, A.and Horowitz, and M. Levoy. High performance imaging using large camera arrays. *ACM Trans. Graph.*, 24(3):765–776, July 2005. 1

[33] K. Yücer, A. Sorkine-Hornung, O. Wang, and O. Sorkine-Hornung. Efficient 3D object segmentation from densely sampled light fields with applications to 3D reconstruction. *ACM Transaction in Graphics*, 35(3):22:1–22:15, March 2016. 1, 2, 4

[34] A. Zanfir and C. Sminchisescu. Large displacement 3d scene flow with occlusion reasoning. In *ICCV*, 2015. 2

[35] D. Zheng, E.and Ji, E. Dunn, and J.-M. Frahm. Sparse dynamic 3D reconstruction from unsynchronized videos. In *ICCV*, 2015. 2