4D Match Trees for Non-rigid Surface Alignment

Armin Mustafa(✉), Hansung Kim, and Adrian Hilton

CVSSP, University of Surrey, Guildford, UK
{a.mustafa,h.kim,a.hilton}@surrey.ac.uk

Abstract. This paper presents a method for dense 4D temporal alignment of partial reconstructions of non-rigid surfaces observed from single or multiple moving cameras of complex scenes. *4D Match Trees* are introduced for robust global alignment of non-rigid shape based on the similarity between images across sequences and views. Wide-timeframe sparse correspondence between arbitrary pairs of images is established using a segmentation-based feature detector (SFD) which is demonstrated to give improved matching of non-rigid shape. Sparse SFD correspondence allows the similarity between any pair of image frames to be estimated for moving cameras and multiple views. This enables the 4D Match Tree to be constructed which minimises the observed change in non-rigid shape for global alignment across all images. Dense 4D temporal correspondence across all frames is then estimated by traversing the 4D Match tree using optical flow initialised from the sparse feature matches. The approach is evaluated on single and multiple view images sequences for alignment of partial surface reconstructions of dynamic objects in complex indoor and outdoor scenes to obtain a temporally consistent 4D representation. Comparison to previous 2D and 3D scene flow demonstrates that 4D Match Trees achieve reduced errors due to drift and improved robustness to large non-rigid deformations.

Keywords: Non-sequential tracking · Surface alignment · Temporal coherence · Dynamic scene reconstruction · 4D modeling

1 Introduction

Recent advances in computer vision have demonstrated reconstruction of complex dynamic real-world scenes from multiple view video or single view depth acquisition. These approaches typically produce an independent 3D scene model at each time instant with partial and erroneous surface reconstruction for moving objects due to occlusion and inherent visual ambiguity [1–4]. For non-rigid objects, such as people with loose clothing or animals, producing a temporally coherent 4D representation from partial surface reconstructions remains a challenging problem.

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46448-0_13) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part I, LNCS 9905, pp. 213–229, 2016.
DOI: 10.1007/978-3-319-46448-0_13

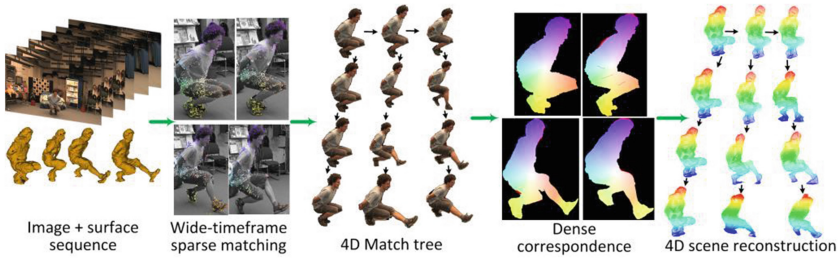

Fig. 1. 4D Match Tree framework for global alignment of partial surface reconstructions

In this paper we introduce a framework for global alignment of non-rigid shape observed in one or more views with a moving camera assuming that a partial surface reconstruction or depth image is available at each frame. The objective is to estimate the dense surface correspondence across all observations from single or multiple view acquisition. An overview of the approach is presented in Fig. 1. The input is the sequence of frames $\{F_i\}_{i=1}^{N}$ where N is the number of frames. Each frame F_i consists of a set of images from multiple viewpoints $\{V_c\}_{c=1}^{M}$, where M is the number of viewpoints for each time instant ($M \geq 1$). Robust sparse feature matching between arbitrary pairs of image observations of the non-rigid shape at different times is used to evaluate similarity. This allows a *4D Match Tree* to be constructed which represents the optimal alignment path for all observations across multiple sequences and views that minimises the total dissimilarity between frames or non-rigid shape deformation. 4D alignment is then achieved by traversing the 4D match tree using dense optical flow initialised from the sparse inter-frame non-rigid shape correspondence. This approach allows global alignment of partial surface reconstructions for complex dynamic scenes with multiple interacting people and loose clothing.

Previous work on 4D modelling of complex dynamic objects has primarily focused on acquisition under controlled conditions such as a multiple camera studio environment to reliably reconstruct the complete object surface at each frame using shape-from-silhouette and multiple view stereo [5–7]. 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 with impressive results for complex motion. 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.

The wide-spread availability of low-cost depth sensors has motivated the development of methods for temporal correspondence or alignment and 4D modelling from partial dynamic surface observations [8–11]. Scene flow techniques [12,13] 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. Existing feature matching techniques either work in 2D [14] or 3D [15] or for sparse [16,17] or dense [18] points. However these methods fail in the case of occlusion, large motions, background clutter, deformation, moving cameras and appearance of new parts of objects. Recent work has introduced approaches, such as DynamicFusion [8], for 4D modelling from depth image sequences integrating temporal observations of non-rigid shape to resolve fine detail. Approaches to 4D modelling from partial surface observations are currently limited to relatively simple isolated objects such as the human face or upper-body and do not handle large non-rigid deformations such as loose clothing.

In this paper we introduce the *4D Match Tree* for robust global alignment of partial reconstructions of complex dynamic scenes. This enables the estimation of temporal surface correspondence for non-rigid shape across all frames and views from moving cameras to obtain a temporally coherent 4D representation of the scene. Contributions of this work include:

- Robust global 4D alignment of partial reconstructions of non-rigid shape from single or multiple-view sequences with moving cameras
- Sparse matching between wide-timeframe image pairs of non-rigid shape using a segmentation-based feature descriptor
- 4D Match Trees to represent the optimal non-sequential alignment path which minimises change in the observed shape
- Dense 4D surface correspondence for large non-rigid shape deformations using optic-flow guided by sparse matching

1.1 Related Work

Temporal alignment for reconstructions of dynamic scenes is an area of extensive research in computer vision. Consistent mesh sequences finds application in performance capture, animation and motion analysis. A number of approaches for surface reconstruction [19,20] do not produce temporally coherent models for an entire sequence rather they align pairs of frames sequentially. Other methods proposed for 4D alignment of surface reconstructions assume that a complete mesh of the dynamic object is available for the entire sequence [21–25]. Partial surface tracking methods for single view [26] and RGBD data [8,27] perform sequential alignment of the reconstructions using frame-to-frame tracking. Sequential methods suffer from drift due to accumulation of errors in alignment between successive frames and failure is observed due to large non-rigid motion. Non-sequential approaches address these issues but existing methods require complete surface reconstruction [24,25]. In this paper we propose a non-sequential method to align partial surface reconstructions of dynamic objects for general dynamic outdoor and indoor scenes with large non-rigid motions across sequences and views.

Alignment across a sequence can be established using correspondence information between frames. Methods have been proposed to obtain sparse [14,16,17] and dense [13,15,18] correspondence between consecutive frames for entire sequence. Existing sparse correspondence methods work sequentially on a frame-to-frame basis for single view [14] or multi-view [16] and require a strong prior

initialization [17]. Existing dense matching or scene flow methods [12,13] require a strong prior which fails in the case of large motion and moving cameras. Other methods are limited to RGBD data [18] or narrow timeframe [15,28] for dynamic scenes. In this paper we aim to establish robust sparse wide-timeframe correspondence to construct 4D Match Trees. Dense matching is performed on the 4D Match Tree non-sequentially using the sparse matches as an initialization for optical flow to handle large non-rigid motion and deformation across the sequence.

2 Methodology

The aim of this work is to obtain 4D temporally coherent models from partial surface reconstructions of dynamic scenes. Our approach is motivated by previous non-sequential approaches to surface alignment [24,29,30] which have been shown to achieve robust 4D alignment of complete surface reconstructions over multiple sequences with large non-rigid deformations. These approaches use an intermediate tree structure to represent the unaligned data based on a measure of shape similarity. This defines an optimal alignment path which minimises the total shape deformation. In this paper we introduce the 4D Match Tree to represent the similarity between unaligned partial surface reconstructions. In contrast to previous work the similarity between any pair of frames is estimated from wide-timeframe sparse feature matching between the images of the non-rigid shape. Sparse correspondence gives a similarity measure which approximates the overlap and amount of non-rigid deformation between images of the partial surface reconstructions at different time instants. This enables robust non-sequential alignment and initialisation of dense 4D correspondence across all frames.

2.1 Overview

An overview of the 4D Match Tree framework is presented in Fig. 1. The input is a partial surface reconstruction or depth map of a general dynamic scenes at each frame together with single or multiple view images. Cameras may be static or moving and camera calibration is assumed to be known or estimated together with the scene reconstruction [3,20,31,32]. The first step is to estimate sparse wide-timeframe feature correspondence. Robust feature matching between frames is achieved using a robust segmentation-based feature detector (SFD) previously proposed for wide-baseline stereo correspondence [33]. The 4D Match Tree is constructed as the minimum spanning tree based on the surface overlap and non-rigid shape similarity between pairs of frames estimated from the sparse feature correspondence. This tree defines an optimal path for alignment across all frames which minimises the total dissimilarity or shape deformation. Traversal of the 4D Match Tree from the root to leaf nodes is performed to estimate dense 4D surface correspondence and obtain a temporally coherent representation. Dense surface correspondence is estimated by performing optical flow between

each image pair initialised by the sparse feature correspondence. The 2D optical flow correspondence is back-projected to the 3D partial surface reconstruction to obtain a 4D temporally coherent representation. The approach is evaluated on publicly available benchmark datasets for partial reconstructions of indoor and outdoor dynamic scenes from static and moving cameras: Dance1 [34]; Dance2, Cathedral, Odzemok, [35]; Magician and Juggler [36].

2.2 Robust Wide-Timeframe Sparse Correspondence

Sparse feature matching is performed between any pair of frames to obtain an initial estimate of the surface correspondence. This is used to estimate the similarity between observations of the non-rigid shape at different frames for construction of the 4D Match Tree and subsequently to initialize dense correspondence between adjacent pairs of frames on the tree branches. For partial reconstruction of non-rigid shape in general scenes we require feature matching which is robust to both large shape deformation, change in viewpoint, occlusion and errors in the reconstruction due to visual ambiguity. To overcome these challenges sparse feature matching is performed in the 2D domain between image pairs and projected onto the reconstructed 3D surface to obtain 3D matches. In the case of multiple view images consistency is enforced across views at each time frame.

Segmentation-Based Feature Detection: Several feature detection and matching approaches previously used in wide-baseline matching of rigid scenes have been evaluated for wide-timeframe matching between images of non-rigid shape. Figure 2 and Table 1 present results for SIFT [37], FAST [38] and SFD [33] feature detection. This comparison shows that segmentation-based feature detector (SFD) [33] gives a relatively high number of correct matches. SFD detects keypoints at the triple points between segmented regions which correspond to local maxima of the image gradient. Previous work showed that these keypoints are stable to change in viewpoint and give an increased number of accurate matches compared to other widely used feature detectors. Results indicate that SFD can successfully establish sparse correspondence for large non-rigid deformations as well as changes in viewpoint with improved coverage and number of features. SFD features are detected on the segmented dynamic object for each view c and the set of initial keypoints are defined as: $X^c = \{x^c_{F_0}, x^c_{F_1}, ..., x^c_{F_N}\}$. The SIFT descriptor [37] for each detected SFD keypoint is used for feature matching.

Table 1. Number of sparse wide-timeframe correspondences for all datasets.

No. of matches	Dance1	Dance2	Odzemok	Cathedral	Magician	Juggler
SFD	**416**	**1233**	**916**	**665**	**392**	**547**
SIFT	124	493	366	301	141	273
FAST	57	96	82	77	53	68

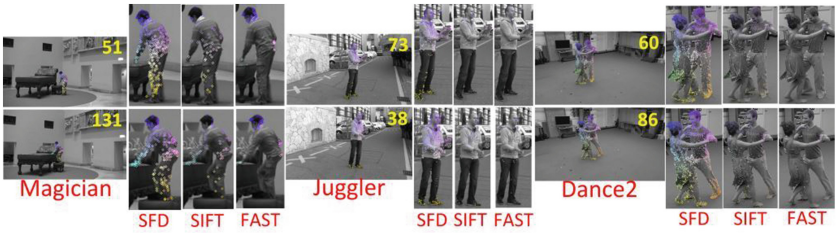

Fig. 2. Comparison of feature detectors for wide-timeframe matching on 3 datasets.

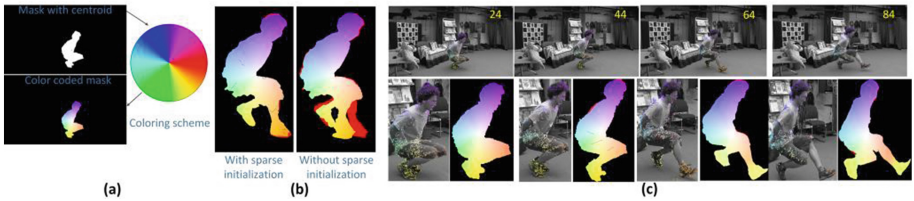

Fig. 3. Sparse feature matching and dense correspondence for the Odzemok dataset: (a)Color coding scheme, (b) Dense matching with and without the sparse match initialization and, (c) Sparse and dense correspondence example

Wide-Timeframe Matching: Once we have extracted keypoints and their descriptors from two or more images, the next step is to establish some preliminary feature matches between these images. As the time between the initial frame and the current frame can become arbitrarily large, robust matching technique are used to establish correspondences. A match $s^c_{F_i,F_j}$ is a feature correspondence $s^c_{F_i,F_j} = (x^c_{F_i}, x^c_{F_j})$, between $x^c_{F_i}$ and $x^c_{F_j}$ in view c at frames i and j respectively. Nearest neighbor matching is used to establish matches between keypoints $x^c_{F_i}$ from the i^{th} frame to candidate interest points $x^c_{F_j}$ in the j^{th} frame. The ratio of the first to second nearest neighbor descriptor matching score is used to eliminate ambiguous matches ($ratio < 0.85$). This is followed by a symmetry test which employs the principal of forward and backward match consistency to remove the erroneous correspondences. Two-way matching is performed and inconsistent correspondences are eliminated. To further refine the sparse matching and eliminate outliers we enforce local spatial coherence in the matching. For matches in an $m \times m$ ($m = 11$) neighborhood of each feature we find the average Euclidean distance and constrain the match to be within a threshold ($\pm\eta < 2 *$ Average Euclidean distance).

Multiple-View Consistency: In the case of multiple views ($M > 1$) consistency of matching across views is also enforced. Each match must satisfying the constraint: $\left\| s^{c,c}_{F_i,F_j} - (s^{c,k}_{F_j,F_j} + s^{k,k}_{F_i,F_j} + s^{c,k}_{F_i,F_i}) \right\| < \epsilon$ ($\epsilon = 0.25$). The multi-view consistency check ensures that correspondences between any two views remain

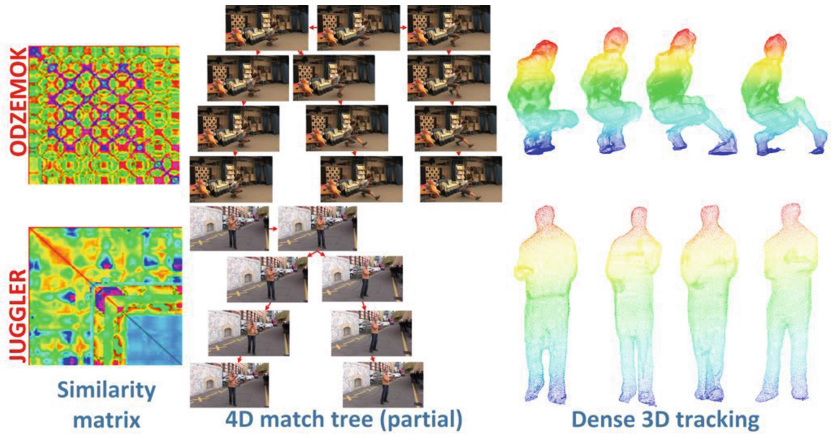

Fig. 4. The similarity matrix, partial 4D Match Tree and 4D alignment for Odzemok and Juggler datasets

consistent for successive frames and views. This gives a final set of sparse matches of the non-rigid shape between frames for the same view which is used to calculate the similarity metric for the non-sequential alignment of frames and initialise dense correspondence.

An example of sparse matching is shown in Fig. 3(c). For visualization features are color coded in one frame according to the colour map as illustrated in Fig. 3(a) and this color is propagated to feature matches at other frames.

2.3 4D Match Trees for Non-sequential Alignment

Our aim is to estimate dense correspondence for partial non-rigid surface reconstructions across complete sequences to obtain a temporally coherent 4D representation. Previous research has employed a tree structure to represent non-rigid shape of complete surfaces to achieve robust non-sequential alignment for sequences with large non-rigid deformations [24,29,30]. Inspired by the success of these approaches we propose the *4D Match Tree* as an intermediate representation for alignment of partial non-rigid surface reconstructions. An important difference of this approach is the use of an image-based metric to estimate the similarity in non-rigid shape between frames. Similarity between any pair of frames is estimated from the sparse wide-timeframe feature matching. The 4D Match Tree represents the optimal traversal path for global alignment of all frames as a minimum spanning tree according to the similarity metric.

The space of all possible pairwise transitions between frames of the sequence is represented by a dissimilarity matrix D of size $N \times N$ where both rows and columns correspond to individual frames. The elements $D(i,j) = d(F_i, F_j)$ are proportional to the cost of dissimilarity between frames i and j. The matrix is symmetrical ($d(F_i, F_j) = d(F_j, F_i)$) and has zero diagonal ($d(F_i, F_i) = 0$).

For each dynamic object in a scene a graph Ω of possible frame-to-frame matches is constructed with nodes for all frames F_i. $d(F_i, F_j)$ is the similarity metric between two nodes and is computed using information from sparse correspondences and intersection of silhouettes obtained from the back-projection of the surface reconstructions in each view.

Feature Match Metric: SFD keypoints detected for each view at each frame are matched between frames using all views. The feature match metric for non sequential alignment $M^c_{i,j}$ between frame i and j for each view c is defined as the inlier ratio $M^c_{i,j} = \frac{|s^c_{F_i,F_j}|}{R^c_{i,j}}$, where $R^c_{i,j}$ is the total number of preliminary feature matches between frames i and j for view c before constraining, and $|s^c_{F_i,F_j}|$ is the number of matches between view c of frame i and frame j obtained using the method explained in Sect. 2.2. $M^c_{i,j}$ is a measure of the overlap between partial surface reconstruction for view c at frames i and j. The visible surface overlap is a measure of their suitability for pairwise dense alignment.

Silhouette Match Metric: The partial surface reconstruction at each frame is back-projected in all views to obtain silhouettes of the dynamic object. Silhouettes between two frames for the same camera view c are aligned by an affine warp [39]. The aligned silhouette intersection area $h^c_{i,j}$ between frames i and j for view c is evaluated. A silhouette match metric $I^c_{i,j}$ is defined as: $I^c_{i,j} = \frac{h^c_{i,j}}{A^c_{i,j}}$, where $A^c_{i,}$ is the union of the area under the silhouette at frame i and j for view c. This gives a measure of the shape similarity between observations of the non-rigid shape between pairs of frames.

Similarity Metric: The two metrics $I^c_{i,j}$ and $M^c_{i,j}$ are combined to calculate the dissimilarity between frames used as graph edge-weights. The edge-weight $d(F_i, F_j)$ for Ω is defined as:

$$d(F_i, F_j) = \begin{cases} 0 & , \text{if } |s^c_{F_i,F_j}| < 0.006 * max(\mathrm{W}, \mathrm{H}) \\ \frac{1}{\sum_{c=1}^{M} M^c_{i,j} \times I^c_{i,j}} & , \text{otherwise} \end{cases} \quad (1)$$

where W and H are the width and height of the input image. Note small values of $d()$ indicates a high similarity in feature matches between frames. Figure 4 presents the dissimilarity matrix D between all pairs of frames for two sequences (red indicates similar frames, blue dissimilar). The matrix off diagonal red areas indicate frames with similar views of the non-rigid shape suitable for non-sequential alignment. A minimum spanning tree is constructed over this graph to obtain the 4D Match Tree.

4D Match Tree: A fully connected graph is constructed using the dissimilarity metric as edge-weights and the minimum spanning tree is evaluated [40,41]. Optimal paths through the sequence to every frame can be jointly optimised based on $d()$. The paths are represented by a traversal tree $T = (\mathbb{N}; E)$ with the nodes $\mathbb{N} = \{F_i\}_{i=1}^{N}$. The edges E are undirected and weighted by the dissimilarity

$e_{i,j} = d(F_i, F_j)$ for $e_{i,j} \in E$. The optimal tree T_o is defined as the minimum spanning tree (MST) which minimises the total cost of pairwise matching given by d:

$$T_o = \arg\min_{\forall T \in \Omega} \left(\sum_{\forall i,j \in T} d(F_i, F_j) \right) \tag{2}$$

This results in the 4D Match Tree T_o which minimises the total dissimilarity between frames due to non-rigid deformation and changes in surface visibility. Given T_o for a dynamic object we estimate the dense correspondence for the entire sequence to obtain a temporally coherent 4D surface. The tree root node M_{root} is defined as the node with minimum path length to all nodes in T_o. The minimum spanning tree can be efficiently evaluated using established algorithms with order $O(N \log N)$ complexity where N is the number of nodes in the graph Ω. The mesh at the root node is subsequently tracked to other frames by traversing through the branches of the tree T towards the leaves. Examples of partial 4D Match Trees for two datasets are shown in Fig. 4.

2.4 Dense Non-rigid Aligment

Given the 4D Match Tree global alignment is performed by traversing the tree to estimate dense correspondence between each pair of frames connected by an edge. Sparse SFD feature matches are used to initialise the pairwise dense correspondence which is estimated using optical flow [42]. The sparse feature correspondences provides a robust initialisation of the optical flow for large non-rigid shape deformation. The estimated dense correspondence is back projected to the 3D visible surface to establish dense 4D correspondence between frames. In the case of multiple views dense 4D correspondence is combined across views to obtain a consistent estimate and increase surface coverage. Dense temporal correspondence is propagated to new surface regions as they appear using the sparse feature matching and dense optical flow. An example of the propagated mask with and without sparse initialization for a single view is shown in Fig. 3(b). The large motion in the leg of the actor is correctly estimated with sparse match initialization but fails without (shown by the red region indicating no correspondence). Pairwise 4D dense surface correspondences are combined across the tree to obtain a temporally coherent 4D alignment across all frames. An example is shown for the Odzemok dataset in Fig. 3(c) with optical flow information for each frame. Figure 4 presents two examples of 4D aligned meshes resulting from the global alignment with the 4D match tree.

3 Results and Performance Evaluation

The proposed approach is tested on various datasets introduced in Sect. 2.1 and the properties of datasets are described in Table 2. Algorithm parameters set empirically are constant for all results.

Table 2. Properties of all datasets and their 4D Match Trees.

Datasets	Number of views	Sequence length	Resolution	Tree depth (frames)	Tree depth (%)
Dance1	8 static	200	780 × 582	65	33
Dance2	7 static, 1 moving	244	1920 × 1080	73	29
Odzemok	6 static, 2 moving	232	1920 × 1080	82	35
Cathedral	8 static	217	1920 × 1080	92	42
Magician	6 moving	400	960 × 544	127	32
Juggler	6 moving	400	960 × 544	104	26

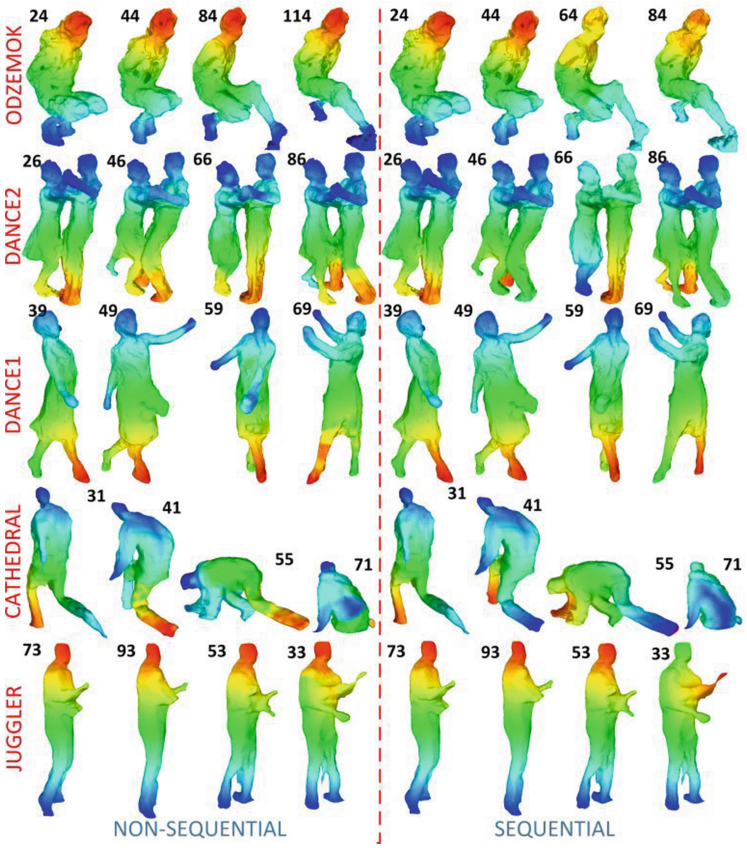

Fig. 5. Comparison of sequential and non-sequential alignment of all datasets.

3.1 Sequential vs. Non-sequential Alignment

4D Match Trees are constructed for all datasets using the method described in Sect. 2.3. The maximum length of branches in the 4D Match Tree for global alignment of each dataset is described in Table 2. The longest alignment path for all sequences is $< 50\,\%$ of the total sequence length leading to a significant reduction in the accumulation of errors due to drift in the sequential alignment process. Non-rigid alignment is performed over the branches of the tree to obtain temporally consistent 4D representation for all datasets. Comparison of 4D aligned surfaces obtained from the proposed non-sequential approach against sequential tracking without the 4D Match tree is shown in Fig. 5. Sequential tracking fails to estimate the correct 4D alignment (Odzemok-64, Dance2-66, Cathedral-55) whereas the non-sequential approach obtains consistent correspondence for all frames for sequences with large non-rigid deformations. To illustrate the surface alignment a color map is applied to the root mesh of the 4D Match tree and propagated to all frames based on the estimated dense correspondence. The color map is consistently aligned across all frames for large non-rigid motions of dynamic shapes in each dataset demonstrating qualitatively that the global alignment achieves reliable correspondence compared to sequential tracking.

3.2 Sparse Wide-Timeframe Correspondence

Sparse correspondences are obtained for the entire sequence using the traversal path in the 4D Match tree from the root node towards the leaves. Results of the sparse and dense 4D correspondence are shown in 6. Sparse matches obtained using SFD are evaluated against a state-of-the-art method for sparse correspondence Nebehay [43]. For fair comparison Nebehay is initialized with SFD keypoints instead of FAST (which produces a low number of matches). Qualitative results are shown in Fig. 7 and quantitative results are shown in Table 3. Matches obtained using the proposed approach are approx $50\,\%$ higher and consistent across frames compared to Nebehay [43] demonstrating the robustness of the proposed wide-timeframe matching using SFD keypoints.

Table 3. Quantitative evaluation for sparse and dense correspondence for all the datasets; Prop. represents proposed non-sequential approach and Matches depicts the number of sparse matches between frames averaged over the entire sequence.

	Silhouette overlap error								Matches	
Datasets	Seq	Prop.	Deepflow	SIFT	Nebehay	1 view	2 views	4 views	Prop.	Nebehay
Dance1	0.42	**0.35**	0.97	0.92	0.96	1.53	1.30	0.99	**416**	249
Dance2	0.83	**0.63**	1.36	1.43	1.38	2.13	1.78	1.47	**1233**	863
Odzemok	0.98	**0.89**	2.82	2.59	2.69	4.35	3.66	2.76	**916**	687
Cathedral	0.83	**0.69**	1.14	1.10	1.29	1.92	1.65	1.09	**665**	465
Magician	1.07	**0.86**	3.43	3.22	3.77	5.46	4.67	3.18	**392**	293
Juggler	0.78	**0.65**	1.24	1.19	1.31	2.12	1.76	1.44	**547**	437

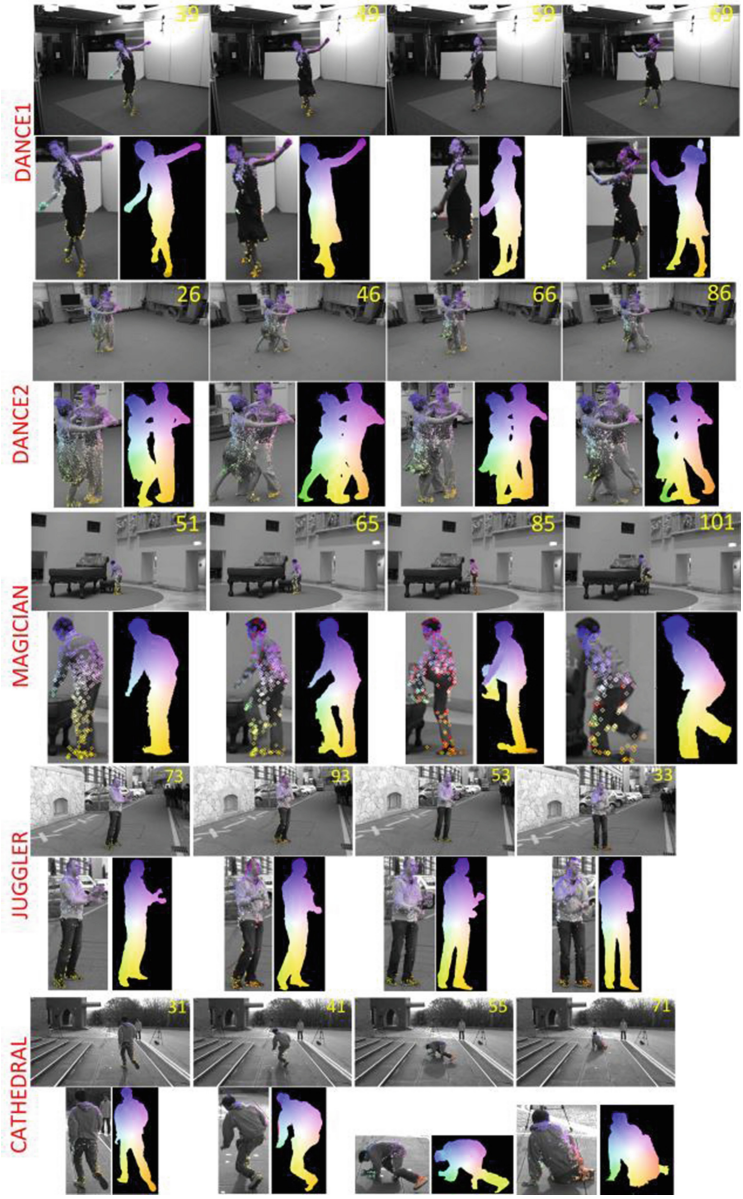

Fig. 6. Sparse and dense 2D tracking color coded for all datasets

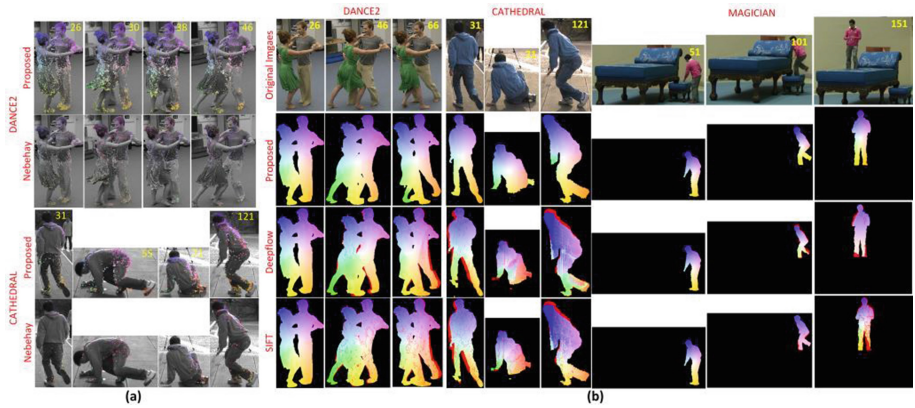

Fig. 7. Qualitative comparison: (a) Sparse tracking comparison for one indoor and one outdoor dataset and (b) Dense tracking comparison for two indoor and one outdoor datasets.

3.3 Dense 4D Correspondence

Dense correspondence are obtained on the 4D match tree and the color coded results are shown in Fig. 6 for all datasets. To illustrate the dense alignment the color coding scheme shown in Fig. 3 is applied to the silhouette of the dense mesh on the root node for each view and propagated using the 4D Match Tree. The proposed approach is qualitatively shown to propagate the correspondences reliably over the entire sequence for complex dynamic scenes.

For comparative evaluation of dense matching we use: (a) SIFT features with the proposed method in Sect. 2 to obtain dense correspondence; (b) Sparse correspondence obtained using Nebehay [43] with the proposed dense matching; and (c) state-of-the-art dense flow algorithm Deepflow [44] over the 4D Match Tree for each dataset. Qualitative results against SIFT and Deepflow are shown in Fig. 7. The propagated color map using deep flow and SIFT based alignment does not remain consistent across the sequence as compared to the proposed method (red regions indicate correspondence failure).

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 with the silhouette of the projected partial surface reconstruction at each frame to evaluate the accuracy of the dense propagation. The error is defined as:

$$SOE = \frac{1}{M * N} \sum_{i=1}^{N} \sum_{c=1}^{M} \frac{\text{Area of intersection}}{\text{Area of back-projected mask}}$$

Evaluation against sequential and non-sequential Deepflow, SIFT and Nebehay are shown in Table 3 for all datasets. As observed the silhouette overlap error is lowest for the proposed SFD based non-sequential approach showing relatively

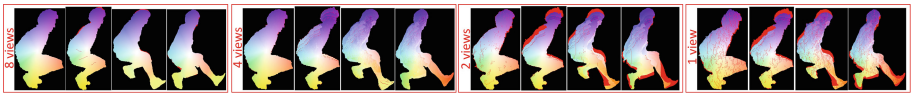

Fig. 8. Single and Multi-view alignment comparison results for Odzemok dataset

high accuracy. We also evaluate the completeness of the 3D points at each time instant as observed in Table 4:

$$completeness = \frac{100}{M * N} \sum_{i=1}^{N} \sum_{c=1}^{M} \frac{\text{Number of 3D points propagated}}{\text{Number of surface points visible from 'c' [45]}}$$

The proposed approach outperforms Deepflow, SIFT and Nebehay all of which result in errors as observed in Fig. 7 and Table 4.

Table 4. Evaluation of completeness of dense 3D correspondence averaged over the entire sequence in %.

Completeness(%)	Deepflow	SIFT	Nebehay	Sequential	Proposed (Non-sequential)			
				All views	1 view	2 views	4 views	All views
Dance1	81.56	83.28	82.55	91.52	60.78	71.65	81.30	**98.22**
Dance2	83.26	85.80	83.96	92.76	61.98	72.30	82.87	**99.36**
Odzemok	81.46	79.83	80.91	90.51	62.73	70.87	77.64	**98.19**
Cathedral	79.54	81.53	81.78	89.21	59.77	69.05	76.98	**97.40**
Magician	82.58	82.92	80.65	89.58	61.29	71.23	75.56	**97.53**
Juggler	79.09	80.11	81.33	91.89	59.54	68.40	78.81	**97.89**

3.4 Single vs Multi-view

The proposed 4D Match Tree global alignment method can be applied to single or multi-view image sequence with partial surface reconstruction. Dense correspondence for the Odzemok dataset using different numbers of views are compared in Fig. 8. Quantitative evaluation using *SOE* and *completeness* obtained from single, 2, 4 and all views for all datasets are presented in Tables 3 and 4 respectively. This shows that even with a single view the 4D Match Tree achieves 60 % completeness due to the restricted surface visibility. Completeness increases with the number of views to $> 97\,\%$ for all views which is significantly higher than other approaches.

4 Conclusions

A framework has been presented for dense 4D global alignment of partial surface reconstructions of complex dynamic scenes using 4D Match trees. 4D Match Trees represent the similarity in the observed non-rigid surface shape across the

sequence. This enables non-sequential alignment to obtain dense surface correspondence across all frames. Robust wide-timeframe correspondence between pairs of frames is estimated using a segmentation-based feature detector (SFD). This sparse correspondence is used to estimate the similarity in non-rigid shape and overlap between frames. Dense 4D temporal correspondence is estimated from the 4D Match tree across all frames using guided optical flow. This is shown to provide improved robustness to large non-rigid deformation compared to sequential and other state-of-the-art sparse and dense correspondence methods. The proposed approach is evaluated on single and multi-view sequences of complex dynamic scenes with large non-rigid deformations to obtain a temporally consistent 4D representation. Results demonstrate completeness and accuracy of the resulting global 4D alignment.

Limitations: The proposed method fails in case of objects with large deformations (high ambiguity), fast spinning (failure of optical flow), and uniform appearance or highly crowded dynamic environments where no reliable sparse matches can be obtained or surface reconstruction fails due to occlusion.

References

1. Zhang, G., Jia, J., Hua, W., Bao, H.: Robust bilayer segmentation and motion/depth estimation with a handheld camera. PAMI **33**, 603–617 (2011)
2. Jiang, H., Liu, H., Tan, P., Zhang, G., Bao, H.: 3D reconstruction of dynamic scenes with multiple handheld cameras. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 601–615. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33709-3_43
3. Taneja, A., Ballan, L., Pollefeys, M.: Modeling dynamic scenes recorded with freely moving cameras. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010. LNCS, vol. 6494, pp. 613–626. Springer, Heidelberg (2011). doi:10.1007/978-3-642-19318-7_48
4. Mustafa, A., Kim, H., Guillemaut, J., Hilton, A.: General dynamic scene reconstruction from wide-baseline views. In: ICCV (2015)
5. Kanade, T., Rander, P., Narayanan, P.J.: Virtualized reality: constructing virtual worlds from real scenes. IEEE MultiMedia **4**, 34–47 (1997)
6. Franco, J.S., Boyer, E.: Exact polyhedral visual hulls. In: Proceedings of BMVC, pp. 32:1–32:10 (2003)
7. Starck, J., Hilton, A.: Model-based multiple view reconstruction of people. In: ICCV, pp. 915–922 (2003)
8. Newcombe, R., Fox, D., Seitz, S.: DynamicFusion: reconstruction and tracking of non-rigid scenes in real-time. In: CVPR (2015)
9. Tevs, A., Berner, A., Wand, M., Ihrke, I., Bokeloh, M., Kerber, J., Seidel, H.P.: Animation cartography: intrinsic reconstruction of shape and motion. ACM Trans. Graph. **31**, 12:1–12:15 (2012)
10. Wei, L., Huang, Q., Ceylan, D., Vouga, E., Li, H.: Dense human body correspondences using convolutional networks (2015). CoRR abs/1511.05904
11. Malleson, C., Klaudiny, M., Guillemaut, J.Y., Hilton, A.: Structured representation of non-rigid surfaces from single view 3D point tracks. In: 3DV (2014)
12. Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U., Cremers, D.: Stereoscopic scene flow computation for 3d motion understanding. IJCV **95**, 29–51 (2011)

13. Basha, T., Moses, Y., Kiryati, N.: Multi-view scene flow estimation: a view centered variational approach. In: CVPR, pp. 1506–1513 (2010)
14. Sundaram, N., Brox, T., Keutzer, K.: Dense point trajectories by GPU-accelerated large displacement optical flow. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 438–451. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15549-9_32
15. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR (2015)
16. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: a massively multiview system for social motion capture. In: ICCV (2015)
17. Zheng, E., Ji, D., Dunn, E., Frahm, J.M.: Sparse dynamic 3D reconstruction from unsynchronized videos. In: ICCV (2015)
18. Zanfir, A., Sminchisescu, C.: Large displacement 3D scene flow with occlusion reasoning. In: ICCV (2015)
19. Lei, C., Chen, X.D., Yang, Y.H.: A new multiview spacetime-consistent depth recovery framework for free viewpoint video rendering. In: ICCV, pp. 1570–1577 (2009)
20. Mustafa, A., Kim, H., Guillemaut, J.Y., Hilton, A.: Temporally coherent 4D reconstruction of complex dynamic scenes. In: CVPR (2016)
21. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. ACM Trans. Graph. **27**, 97:1–97:9 (2008)
22. Tung, T., Nobuhara, S., Matsuyama, T.: Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In: ICCV, pp. 1709–1716 (2009)
23. Cagniart, C., Boyer, E., Ilic, S.: Probabilistic deformable surface tracking from multiple videos. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 326–339. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15561-1_24
24. Budd, C., Huang, P., Klaudiny, M., Hilton, A.: Global non-rigid alignment of surface sequences. Int. J. Comput. Vis. **102**, 256–270 (2013)
25. Huang, C., Cagniart, C., Boyer, E., Ilic, S.: A Bayesian approach to multi-view 4D modeling. Int. J. Comput. Vis. **116**, 115–135 (2016)
26. Russell, C., Yu, R., Agapito, L.: Video pop-up: monocular 3D reconstruction of dynamic scenes. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 583–598. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10584-0_38
27. Guo, K., Xu, F., Wang, Y., Liu, Y., Dai, Q.: Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In: ICCV (2015)
28. Bailer, C., Taetz, B., Stricker, D.: Flow fields: dense correspondence fields for highly accurate large displacement optical flow estimation. In: ICCV (2015)
29. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: CVPR (2012)
30. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Trans. Graph. **34**(4), 69:1–69:13 (2015)
31. Ji, D., Dunn, E., Frahm, J.-M.: 3D reconstruction of dynamic textures in crowd sourced data. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 143–158. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10590-1_10

32. Oswald, M.R., Stühmer, J., Cremers, D.: Generalized connectivity constraints for spatio-temporal 3D reconstruction. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 32–46. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_3
33. Mustafa, A., Kim, H., Imre, E., Hilton, A.: Segmentation based features for wide-baseline multi-view reconstruction. In: 3DV (2015)
34. 4D repository. In: Institut national de recherche en informatique et en automatique (INRIA) Rhone Alpes. http://4drepository.inrialpes.fr/
35. 4D and multiview video repository. In: Centre for Vision Speech and Signal Processing, University of Surrey, UK
36. Ballan, L., Brostow, G.J., Puwein, J., Pollefeys, M.: Unstructured video-based rendering: interactive exploration of casually captured videos. ACM Trans. Graph. **29**, 1–11 (2010)
37. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**, 91–110 (2004)
38. Rosten, E., Porter, R., Drummond, T.: Faster and better: a machine learning approach to corner detection. PAMI **32**, 105–119 (2010)
39. Evangelidis, G.D., Psarakis, E.Z.: Parametric image alignment using enhanced correlation coefficient maximization. IEEE Trans. Pattern Anal. Mach. Intell. **30**, 1858–1865 (2008)
40. Kruskal, J.B.: On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Am. Math. Soc. **7**, 48–50 (1956)
41. Prim, R.C.: Shortest connection networks and some generalizations. Bell Syst. Tech. J. **36**, 1389–1401 (1957)
42. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 363–370. Springer, Heidelberg (2003). doi:10.1007/3-540-45103-X_50
43. Nebehay, G., Pflugfelder, R.: Clustering of static-adaptive correspondences for deformable object tracking. In: CVPR (2015)
44. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: Deepflow: large displacement optical flow with deep matching. In: ICCV, pp. 1385–1392(2013)
45. Joo, H., Soo Park, H., Sheikh, Y.: Map visibility estimation for large-scale dynamic 3D reconstruction. In: CVPR (2014)