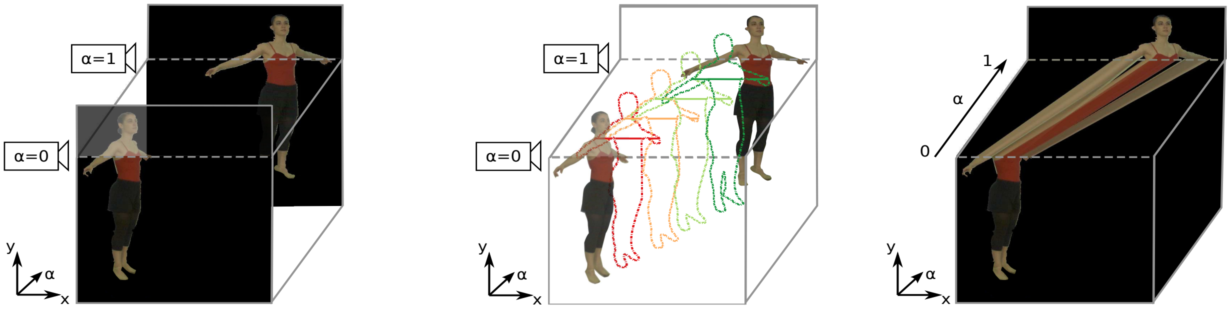
Wide-baseline foreground object interpolation using silhouette shape prior

Cédric Verleysen, Thomas Maugey, Pascal Frossard and Christophe De Vleeschouwer.



(a) Left and right reference views are positioned, after epipolar rectification, as a function of their normalized distance α to the left view.

(b) Reference views and intermediate silhouette priors.

(c) The intermediate views continuum is equivalent to a vertical stack of Epipolar Plane Images.

Fig. 1: (a) Two images of an object, captured by a wide-baseline stereo camera. (b) To drive the interpolation of intermediate views, our method generates a sequence of plausible object silhouettes between those in the reference views using a Gaussian Process latent variable model [1] that has been learned from previous observations of the same object by the same cameras. (c) The sequence of silhouette priors is exploited to disambiguate the reconstruction of Epipolar Plane Images, which are used to construct intermediate views [2].

***Abstract*—We consider the synthesis of intermediate views of an object captured by two widely spaced and calibrated cameras. This problem is challenging because foreshortening effects and occlusions induce significant differences between the reference images when the cameras are far apart. That makes the association or disappearance/appearance of their pixels difficult to estimate. Our main contribution lies in disambiguating this ill-posed problem by making the interpolated views consistent with a plausible transformation of the object silhouette between the reference views. This plausible transformation is derived from an object-specific prior that consists of a nonlinear shape manifold learned from multiple previous observations of this object by the two reference cameras. The prior is used to estimate the evolution of the epipolar silhouette segments between the reference views. This information directly supports the definition of epipolar silhouette segments in the intermediate views, as well as the synthesis of textures in those segments. It permits to reconstruct the Epipolar Plane Images (EPIs) and the continuum of views associated with the Epipolar Plane Image Volume, obtained by aggregating the EPIs. Experiments on synthetic and natural images show that our method preserves the object topology in intermediate views and deals effectively with the self-occluded regions and the severe foreshortening effect associated with wide-baseline camera configurations.**

***Index Terms*—Epipolar plane image, light-field, shape prior.**

I. Introduction and Overview

Virtual view synthesis aims at rendering images of a real scene from viewpoints different from those recorded by a camera. Our paper restricts the synthesis problem to the interpolation of images observed by a virtual camera positioned at an arbitrary point along the wide-baseline connecting two reference cameras. Formally, the two reference images are denoted $\mathbf{I}_l$ and $\mathbf{I}_r$, and referred to as left and right images in the rest of the paper. Any intermediate synthesized image is denoted $\mathbf{I}_\alpha$, $0 \leq \alpha \leq 1$, with α defining the distance to the left image, normalized by the baseline length (Figure 1). From an application perspective, the smooth transition in synthetic views between two reference viewpoints is especially desired in the field of video production [3], where it prevents the visual discomfort caused by the loss of visual cues and landmarks during abrupt camera switching.

In previous works, the image interpolation problem has been addressed successfully when the distance between the cameras is small compared to their distance to the scene. State-of-the-art methods generally decompose the scene into background and foreground objects and reconstruct them independently [4] [5]. The existing solutions however remain largely unsatisfactory in wide-baseline setups. In particular, the scene's proximity to the cameras causes many projective discrepancies between the two views (*e.g.*, occlusions, foreshortening effects), which hamper the computation of dense correspondences and damage the interpolated views [6]. Our work is original in that it specifically addresses the case of objects that are close to the wide-baseline cameras. It considers that the background can be reconstructed using state-of-the-art methods, and after segmentation of a foreground object, e.g. using [7], it provides a generic solution to reconstruct intermediate views of this object, including when it is close to the cameras. That ability to handle proximity makes our solution unique in the related literature.

Part of this work has been supported by the Belgian NSF. Cédric Verleysen and Christophe De Vleeschouwer are with ICTEAM, UCL, Belgium. Thomas Maugey is with INRIA, France. Pascal Frossard is with the LTS4 laboratory, EPFL, Switzerland.

As depicted in Figure 1, our paper adopts the usual Epipolar Plane Image (EPI) formalism [2] to reconstruct the images of the foreground object between the two widely spaced reference cameras. This formalism builds on epipolar rectification [8], which applies homographies to the views, so that all horizontal lines with the same ordinate in the transformed images correspond to the same epipolar plane. Hence, after epipolar rectification, the 3D stack obtained by gathering, as a function of α, the images captured along the baseline by a dense array of cameras is named Epipolar Plane Image Volume [9] [10] (EPIV) because any transverse cross-section of this 3D volume corresponds to an EPI. As depicted in Figure 1(c), each EPI describes how the pixels of one epipolar line in a reference view move to the other reference. The light field theory [2] states that these transitions are always linear and their slopes are inversely proportional to the scene's depth. Despite its richness, the EPIV estimation has been limited so far to very narrow-baseline setups [10].

We propose in this paper to close this gap and to reconstruct the EPIV of an arbitrary foreground object in the wide-baseline scenario. After epipolar rectification, the silhouette of the foreground object in each reference view is described by a sequence of background/foreground segments along each epipolar line. Our contribution primarily aims at matching those segments between the two reference images, and then using this information to guide the EPIs reconstruction. As illustrated in Figure 1(b), our new method builds on intermediate object silhouette priors to guide the EPIs reconstruction. To generate this prior, an object-specific 2D-shape manifold is learned from a set of silhouettes that represent the various appearances of the object shape, when projected on a planar view. The flow chart of our algorithm is presented in Figure 2. The steps involved in the EPIs reconstruction are further illustrated in Figure 3. We note that the set of silhouettes used to learn the object-specific manifold is composed of silhouettes of the same object moving in front of the same stereo pair, as shown by the black components in Figure 2. The training set is typically collected before the instant at which the object image has to be interpolated. At the interpolation time, the learned object-specific 2D-shape manifold is exploited to guide the deformation of the pair of object silhouettes observed by the reference views, as shown in Figure 3. Specifically, after projecting the reference silhouettes in the low-dimensional manifold, a sequence of shapes that likely represents the deformation of the object silhouette between the stereo pair is generated by sampling the geodesic path computed between the two projections of the reference silhouettes on the manifold (Figure 3, left side). The resulting sequence of plausible silhouettes is then used to guide the transformation (*i.e.*, the scaling/translation/vanishing) of epipolar foreground segments between the two reference views (Figure 3, right side).

Interestingly, the priors are used not only to disambiguate the association of pixels between reference views, but also to determine how the occluded parts vanish/appear while moving from one reference view to the other. This gives a unique ability to reconstruct visually pleasant and topologically consistent images in the presence of significant self-occlusions or severe foreshortenings inherent in wide-baseline stereo setups.

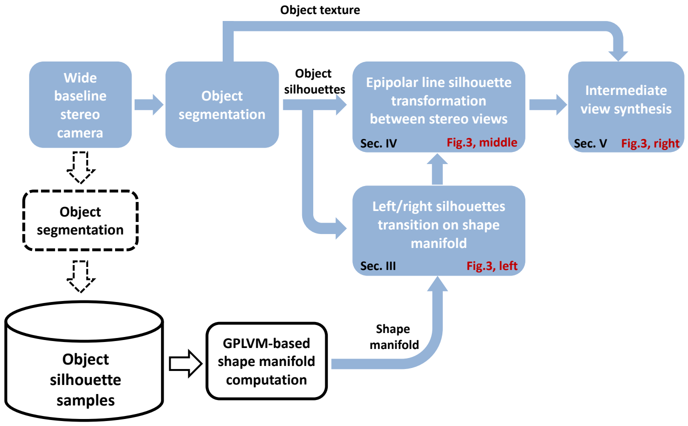

Object texture
Wide baseline stereo camera
Object segmentation
Object silhouettes
Epipolar line silhouette transformation between stereo views
Sec. IV
Fig.3, middle
Intermediate view synthesis
Sec. V
Fig.3, right
Object segmentation
Left/right silhouettes transition on shape manifold
Sec. III
Fig.3, left
Shape manifold
Object silhouette samples
GPLVM-based shape manifold computation

Fig. 2: Flow chart of our algorithm. Black components are only activated in an initialization stage, to capture the object-specific prior. Black solid lines depict the blocks involved in the generation of the object-specific GPLVM shape manifold, from a set of silhouettes that correspond to various and representative object poses. The dashed lines indicate that the representative silhouettes are collected from previous observations of the object, assumed to be moving in front of the stereo setup. Blocks and arrows in blue implement the interpolation of intermediate images, and are run each time an intermediate view has to be interpolated. The behavior of the three blocks following the object segmentation is described in Sections III, IV, and V, and is illustrated in Figure 3.

The rest of our paper is organized as follows. Section II surveys the recent advances in virtual view reconstruction, and identifies the limitations of earlier methods in our envisioned wide-baseline stereo acquisition setup. Section III explains how to capture and embed a prior about the plausible silhouettes of the object in a low-dimensional silhouette manifold, which can be exploited to guide the reconstruction of the EPIs between two reference images, as detailed in Section IV. The view synthesis process is described in Section V. Section VI then validates our framework by generating topologically valid intermediate views on both real and synthetic images captured by two cameras with very different viewpoints. The advantages induced by shape priors are further demonstrated by comparing our method with a set of conventional and state-of-the-art approaches.

II. Related work and challenges

View synthesis methods are generally categorized into two groups: model-based rendering and image-based rendering.

In model-based rendering, a 3D shape model of the observed scene is reconstructed explicitly from multiview images. Appropriate texture is then mapped on the model, which is projected onto any arbitrary viewpoint. Methods such as projective grid space [11] [12], visual-hull [13] [14] [15] [16], 3D model adjustment [17], and shape-from-video [18] belong to this category. These methods have the advantage of synthesizing intermediate views representing the actual 3D scene. How-

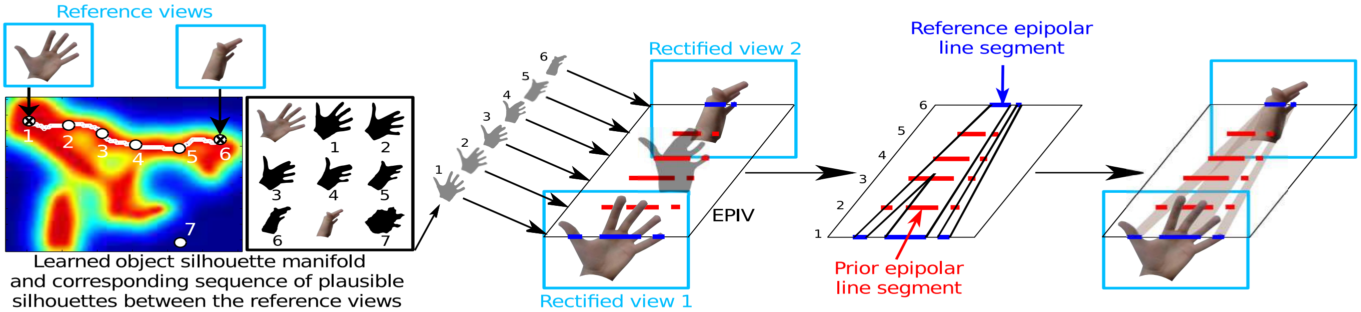


Fig. 3: The reconstruction of Epipolar Plane Images (EPIs) is driven by incorporating prior knowledge about the deformations of the object silhouette. This knowledge is learned from previous observations of the object by (one of) the two reference views. It is captured by a low-dimensional Gaussian Process latent space, from which intermediate 2D prior silhouettes are extracted in-between the projected reference ones (in the left part of the figure, each point of the 2D latent space represents a silhouette, and warmer colors indicate a higher 'plausibility'). These intermediate 2D priors are then appropriately placed in the EPIV to be converted into a set of 1D priors that disambiguate the matching of epipolar silhouette segments between the reference epipolar lines. Eventually, this matching drives the reconstruction of the EPI (right part of the figure).

ever, the quality of the virtual view is highly dependent on the accuracy of the estimated 3D model [19]. To obtain an accurate 3D model, model-based rendering methods therefore rely on a dense coverage of the scene, which requires a large number of precisely calibrated video cameras [20]. The trade-off between the accuracy of the reconstruction and the number of cameras is often relaxed when the distance between the object and the reference cameras is large compared with the baseline distance separating these cameras [21]. In this particular case, a simple (set of) planar model(s) (called *billboards*) enables one to generate realistic intermediate views of the object. However, when the distance to the scene decreases, planar *proxies* become insufficient to approximate the 3D structure of the object [22]. Therefore, to synthesize intermediate views in-between wide-baseline reference cameras, state-of-the-art methods generally decompose the scene into its background and its foreground objects and reconstruct them independently [4]. The still background 3D geometry is typically acquired based on state-of-the-art active 3D acquisition systems [23] [24], or estimated based on piecewise-planar 3D geometry approximations [25] [26] [22] [27]. In contrast, the capture of dynamic foreground geometry is impractical, and current estimation solutions are restricted to specific contexts, where the foreground is still relatively far from the cameras [18] [5] [22]. Hence, the interpolation of dynamic foreground objects situated relatively close to the pair of cameras remains a largely unsolved question [28] [6].

In contrast, image-based rendering (IBR) methods [29] create the virtual views directly in the image color space without explicit reconstruction of a 3D piecewise-smooth surface. Such methods are further classified into arbitrary-view and baseline interpolation approaches. On the one hand, arbitrary-view IBR approaches determine the pixel color values of each virtual view in a way that is geometrically and/or photometrically consistent with $N \geq 2$ reference views. These methods focus on optimizing multiple depth maps (either the ones of the virtual views [5] or those of the reference views [22]) and/or the virtual image's color [30]. However, the dense estimation of a depth map is possible only when all the 3D points corresponding to a pixel in the reconstructed view are observed with at least two reference views. This requires a sufficiently dense coverage of the scene with many cameras. On the other hand, baseline interpolation approaches determine region correspondences or pixel correspondences (disparity) between two reference views and generate the intermediate views by interpolation [31] or morphing [32]. They rely on dense correspondences between the views, generally for small-baseline configurations. This trend culminates in light-field reconstruction approaches [2], which require tens or hundreds of narrow-baseline[1] cameras/lenses [33] to determine a continuous (sub-pixel) correspondence between the references.

So far, image-based rendering techniques have thus been restricted to dense acquisition setups, where many images of the same 3D scene are captured by cameras that are close to each other compared with their distance to the 3D scene. To the best of our knowledge, no image-based rendering method has been able to provide effective synthesis with a wide-baseline setup composed of only two reference cameras.

The main source of failure in wide-baseline stereo rendering lies in the strong geometrical deformations (including occlusions) induced by the projections. More specifically:

- *The foreshortening effect* causes a distance or an object to appear shorter/wider than it is because it is angled toward the viewer. As a consequence, a given 3D object will be represented by a different number of pixels in different views. This implies that finding correspondences with fixed-template matching methods fails [34]. The same holds when a pixel correspondence is optimized by graph-cut [35], belief propagation [36], or dynamic programming [37] approaches, which generally enforce the pixel uniqueness constraint, *i.e.*, a pixel in an image corresponds to at most one pixel in another image.
- *The self-occlusion effect* occurs when part of an object hides another region of the same object. It drastically

[1] The reference views are separated by a few microns (microlens arrays) to a few centimeters in narrow-baseline setups.

limits the correspondence-based interpolation methods [32] in a wide-baseline configuration.

- The ambiguous correspondences induced by the large difference in viewpoints results in *sparse disparity/depth maps*, leading to large holes in the reconstructed intermediate view. Multiple methods exist to fill in these holes [38] [39] [40], but they are based either on globally non-valid hypotheses (*e.g.*, holes should contain patterns that are visible in the non-occluded parts) or on computationally expensive (post-)processing techniques [41].

Our work explicitly addresses these issues by computing correspondences between image segments, and by constraining those correspondences to be consistent with a plausible deformation of the projected object silhouette between the reference views. Therefore, our approach is able to properly deal with occlusion of segments or their shrinkage/elongation.

III. Object silhouette priors

This section derives a sequence of object silhouettes that defines *a priori* a plausible transition between the wide-baseline stereo images. We rely on a set of object silhouettes that have been observed before the interpolation time by (one of) the reference cameras, and assume that those silhouettes of the object moving in front of the cameras are representative of the silhouettes observed when moving the cameras along the baseline. We then describe each silhouette based on Elliptic Fourier Descriptors (Section III-A), and map the resulting set of high-dimensional features to a lower dimensional latent space (Section III-B), in which a smooth and topologically consistent sequence of intermediate silhouettes is interpolated (Section III-C) and registered (Section III-D) between the silhouettes from the actual cameras.

A. High-dimensional silhouette description

We use Elliptic Fourier Descriptors (EFD) [42] to describe each silhouette. Therefore, the silhouette contour is first represented as a parametric curve $(x(t), y(t))$ in a 2D coordinate system with arbitrary origin and horizontal direction aligned with rectified epipolar lines. The origin of the contour is selected arbitrarily to be the point with the smallest x-coordinate among the ones having the same y-coordinate as the silhouette's center of mass. Given those definitions, as detailed in [42], Elliptic Fourier Descriptors then represent the shape of a silhouette as a sum of N elliptic harmonics[2]. We have:

$$x(t) = a_0 + \sum_{n=1}^{N} \Big(a_n \ \cos \frac{2\pi n t}{T} + b_n \sin \frac{2\pi n t}{T} \Big),$$

where T is the perimeter of the contour. The second coordinate of the shape contour, $y(t)$, is defined analogously as a function of the coefficients c_0, c_n, and d_n. To make the silhouette features translation invariant, we ignore a_0 and c_0 and describe the shape of an object silhouette as a high dimensional feature vector composed of N sets of harmonic coefficients (a_n, b_n, c_n, d_n). Note that scale invariance naturally results from the normalization of the curvilinear coordinate t by T.

B. Low-dimensional silhouette manifold

To capture the structure of the high-dimensional object silhouette space, we collect M silhouette instances that have been previously observed by (one of) the reference cameras. We then map the M instances of high-dimensional EFD vectors to a low-dimensional latent space that captures the high-dimensional data set distribution. Here we follow the pioneering approach of [43, 44], which uses a nonlinear dimensionality reduction technique called *Gaussian Process Latent Variable model* (GPLVM) [1].

Let $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_M]^T$ denote the M high-dimensional silhouette instances. Each silhouette instance $\mathbf{w}_i$ is defined as a vector of dimension $D = 4 \cdot N$ that collects the N elliptic harmonics of the silhouette, each harmonic being defined by 4 parameters. The corresponding set of low dimensional latent variables computed by GPLVM is denoted by $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_M]^T$, where the variable $\mathbf{v}_i$ is the latent point of dimensionality d corresponding to the silhouette $\mathbf{w}_i$, with $d << D$ ($d = 2$ in our case). To associate a latent point to each silhouette in $\mathbf{W}$, GPLVM considers that each dimension of the silhouette data samples results from a Gaussian process. Given a definition of the covariance kernel in the latent space, it determines how the latent points $\mathbf{V}$ associated with the data samples $\mathbf{W}$ should be distributed in order to maximize the likelihood $P(\mathbf{W}|\mathbf{V}, \boldsymbol{\theta})$ of the observed high-dimensional samples, with $\boldsymbol{\theta}$ denoting the kernel parameters. As in [45], this likelihood is defined as the product of D separate Gaussian processes sharing the same covariance/kernel function. It assigns a precision to each point in the latent space, where the precision is defined as the inverse of the (shared) variance of the D posterior Gaussian distributions resulting from the projections of the latent point in the D-dimensional data space [45]. A small precision corresponds to a large variance and thus to an ambiguous mapping. By construction of the latent space model, which aims at maximizing the likelihood of the data set $\mathbf{W}$, high precision latent points correspond to shapes that are frequent in $\mathbf{W}$. We visualize the precision by coloring the latent space pixels. Warmer, *i.e.*, red, (resp. colder, *i.e.*, blue) pixels correspond to higher (smaller) precision and are more (less) likely to generate a valid shape, *i.e.*, a shape that is similar to the training samples (see for example Figure 3, where GPLVM has learned a latent space from $M = 150$ hand silhouettes).

In its native formulation, GPLVM ensures that points that are close in the latent space remain close in the data space. Hence, points that are far apart in the data space are far apart in the latent space. However, it does not guarantee that points that are close in the data space will also be close in the latent space. To push GPLVM to preserve local distances, we follow [46] and impose *back-constraints* in the computation of the latent variables by forcing the form of the mapping from the data space to the latent space. As a result, the learned latent space becomes more adapted to our interpolation purpose, since it guarantees that the transition between two close points in the latent space maps to a smooth and topologically coherent silhouette transition in the high-dimensional space.

[2] N is set to 50 in most of our experiments.

C. Interpolating intermediate silhouettes on the manifold

Here we derive a sequence of plausible 2D silhouettes between the reference views. We first project the two reference silhouettes onto the latent space (points 1 and 6 in the left part of Figure 3), using the mapping function learned in Section III-B. We then use the Dijkstra's shortest path algorithm [47] to interpolate a plausible transition between these low-dimensional reference silhouettes (white path in Figure 3), and obtain the corresponding high-dimensional silhouette prior by back-projecting some ($P+1$ uniformly sampled) points of this path to the high-dimensional space. The black silhouettes in the left part of Figure 3 depict the silhouettes obtained by back-projection of the points represented in white in the latent space (in this example, $P=5$).

More precisely, the shortest path connecting the two reference silhouettes in the latent space is computed in a graph defined as follows. Each node in the graph corresponds to a point in a discretized latent space. Edges connect neighboring nodes only, with the cost c_{ij} between two neighboring nodes i and j being defined to increase when the precision (see Section III-B) of node j decreases, namely $c_{ij} = -\log(\text{precision}_j + \epsilon)$, where ϵ avoids numerical instabilities. By computing the shortest path on this graph, we obtain a smooth sequence of silhouettes that are well supported by the training set, and could thus be observed with a reasonable likelihood when moving the viewpoint from one reference view to the other (reconstruction scenario), or when the 3D object moves with respect to a single viewpoint (learning scenario). In general, there is no strict guarantee that the intermediate silhouettes derived from the latent space correspond to the actual ones. In practice, however, the similarities between the actual and GPLVM interpolated silhouettes have been confirmed in all our experiments.

D. Registering the silhouette priors with the reference ones

Due to the translation and scale invariance of the shape features (see Section III.A), the intermediate shapes extracted from the manifold define the object silhouettes up to a scaling and translation. To exploit them during the EPI reconstruction, we have to register them in the EPIV. This is performed by:

1) translating the intermediate shapes in such a way that their centers of mass coincide with the linear interpolation of the centers of mass of the two reference silhouettes,
2) scaling the translated prior shapes based on the linear interpolation of the height of the object between the two reference silhouettes.

Figure 1(b) shows examples of registered prior shapes, sampled along the manifold shortest path, according to their normalized distance α to the left reference camera.

IV. Transformation of epipolar line silhouettes

In a given view and along an epipolar line, the object silhouette defines a sequence of consecutive background and foreground line segments, named epipolar line silhouette (ELS). Examples of reference and prior ELSs are illustrated by the blue and red lines in Figure 4(a). This section explains how the sequence of ELS priors helps in estimating how the actual ELS segments evolve between the reference views, thereby supporting the EPI reconstruction (see Figure 3). For this purpose, we represent the ELS segments transformation between the views by the displacement and potential fusion of the segments' borders, simply named ELS borders. After some terminology definition in Section IV-A, Section IV-B derives a method to match ELS borders between the reference views, in a way that is consistent with the available ELS priors. Section IV-C then explains how these priors can be used to control the vanishing/occlusion of unmatched segments.

A. Terminology and notations

Along a given epipolar line, the ELS consists in a sequence of $\mathcal{S}$ adjacent background and foreground segments. It is denoted by $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{\mathcal{S}}]$, with each component $\mathbf{s}_i$ defining the spatial support and the type (foreground vs. background) of the i^{th} segment. A corresponding sequence of ELS borders is denoted $\mathbf{B} = [\mathbf{b}_0, \mathbf{b}_1, \cdots, \mathbf{b}_{\mathcal{S}}]$, with $\mathbf{b}_{k-1}$ and $\mathbf{b}_k$ denoting the coordinates of the beginning and the end of the segment $\mathbf{s}_k$, respectively. The modality $m(\mathbf{b}_k)$ of the border $\mathbf{b}_k$ defines the type of segment ended by $\mathbf{b}_k$.

As depicted in Figure 4, in a given EPI, we use $\mathbf{S}^L$ and $\mathbf{S}^R$ to denote the two sequences of ELS segments observed in the left and right reference views, respectively. $\mathbf{B}^L$ and $\mathbf{B}^R$ denote the two corresponding sequences of ELS borders. We also introduce $\{\mathbf{S}^{\alpha_p}\}_{p\in[0;P]}$ and their corresponding ELS borders $\{\mathbf{B}^{\alpha_p}\}_{p\in[0;P]}$ to refer to the $(P+1)$ ELSs derived from the prior silhouettes (see red lines in Figure 4). After registration of the prior silhouettes with the reference views (Section III-D), the extraction of ELSs from each prior silhouette is straightforward, except regarding the beginning of the first segment and the end of the last one. In our work, those two points are defined based on a linear (with respect to α) interpolation of their corresponding points in the left and right reference views. For the sake of clarity, in the rest of the paper, the reference (blue) and prior (red) ELS segments will be represented on a front view, as shown in Figure 4(b).

B. Matching epipolar line silhouette borders

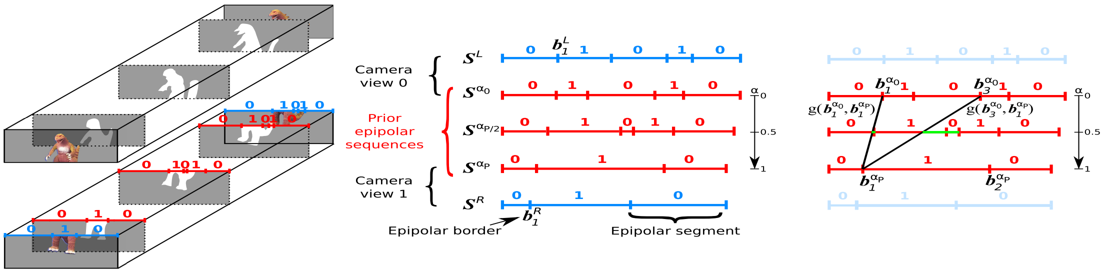
To match the left and right reference ELS borders, i.e. $\mathbf{B}^L$ and $\mathbf{B}^R$, we aim at designing an algorithm that:

- promotes the associations that are consistent with the prior $\{\mathbf{B}^{\alpha_p}\}_{p\in[0;P]}$, describing a plausible deformation of the ELS between the reference views;
- tolerates unmatched borders to reflect the potential vanishing/appearance of ELS segment between the views.

Mathematically, within a given EPI, the left/right ELS border matching is formulated as a constrained minimization problem. For all $0 \le i \le \mathcal{S}^L$ and $0 \le j \le \mathcal{S}^R$, let δ_{ij} indicate whether $\mathbf{b}_i^L$ and $\mathbf{b}_j^R$ are matched ($\delta_{ij}=1$) or not ($\delta_{ij}=0$), and define

$$\delta_i^L = \max_j \delta_{ij}, \qquad \text{and} \qquad \delta_j^R = \max_i \delta_{ij}, \tag{1}$$

to indicate whether $\mathbf{b}_i^L$ and $\mathbf{b}_i^R$, respectively, are matched or not. The optimal ELS border matching decisions $\boldsymbol{\delta}^*$ are then



(a) (b) (c)

Fig. 4: (a) Prior information about the plausible deformations of the object silhouette is used to determine the cost of matching the left epipolar border $\mathbf{b}_k^L$ to the right epipolar border $\mathbf{b}_n^R$. (b) For clarity, the actual and prior epipolar line silhouettes (ELSs) are presented in a frontal picture. (c) The cost $g(\mathbf{b}_k^{\alpha_0}, \mathbf{b}_n^{\alpha_P})$ of matching the k^{th} and n^{th} border of the left and right priors, respectively, is defined by summing, over the intermediate views, the shortest distance measured between one of the intermediate prior borders and the linear interpolation of $\mathbf{b}_k^{\alpha_0}$ and $\mathbf{b}_n^{\alpha_P}$, as depicted in green in (c).

defined as:

$$\begin{aligned}\boldsymbol{\delta}^* = \underset{\{\delta_{ij}\}}{\operatorname{argmin}} \Big(&\sum_{i,j} \delta_{ij} \cdot d(\mathbf{b}_i^L, \mathbf{b}_j^R) \\ &+ \sum_i (1 - \delta_i^L) \cdot u(\mathbf{b}_i^L) + \sum_j (1 - \delta_j^R) \cdot u(\mathbf{b}_j^R) \Big), \\ \text{subject to} \quad &\sum_i \delta_{ij} \leq 1 \quad \text{, and} \quad \sum_j \delta_{ij} \leq 1, \end{aligned} \tag{2}$$

with $u(.)$ denoting the penalty cost induced by unmatched borders, and $d(\mathbf{b}_i^L, \mathbf{b}_j^R)$ measuring how different the $\mathbf{b}_i^L$ and $\mathbf{b}_j^R$ borders are, given the priors. We observe that, if we admit that the relative left/right order of borders has to be preserved during the matching, this minimization problem becomes equivalent to a sequence alignment problem. This problem can be solved efficiently based on dynamic programming, using the Needleman-Wunsch (NW) algorithm [48][49]. Formally, the ordering constraint writes

$$\forall i, i', j, j' \mid \delta_{ij} \cdot \delta_{i'j'} = 1, \; (i - i') \cdot (j - j') \geq 0,$$

and makes the minimization problem of Equation (2) solvable with the standard NW algorithm.

We now define the costs used in Equation (2). The cost $u(\mathbf{b}_i^v)$, induced by skipping the i^{th} border in view v, is defined to be equal to $\max\left(l(\mathbf{s}_k^v), l(\mathbf{s}_{k+1}^v)\right)$, with $\mathbf{s}_k^v$ denoting the k^{th} ELS segment in view v, and $l(\mathbf{s}_k^v)$ measuring the length of $\mathbf{s}_k^v$. This definition ensures that borders delimiting long ELS segments have less chance of being unmatched. This reflects the fact that long segments are less likely to vanish than shorter ones. The cost $d(\mathbf{b}_i^L, \mathbf{b}_j^R)$ of matching the i^{th} border of $\mathbf{B}^L$ with the j^{th} border of $\mathbf{B}^R$ is defined differently depending on whether the two borders have the same modality or not. When the borders have different modalities, the cost is set to ∞, to forbid their matching, as it is desirable to prevent the association of a foreground segment with a background one. In contrast, when the borders have the same modalities, the distance cost $d(\mathbf{b}_i^L, \mathbf{b}_j^R)$ is defined to promote the associations that are supported by the prior $\{\mathbf{B}^{\alpha_p}\}_{p \in [0;P]}$. To account for the deformation of the reference silhouette when it is projected to the latent space, we further decompose $d(\mathbf{b}_i^L, \mathbf{b}_j^R)$ into two parts. The first part considers each reference view independently and measures, in each view, the distance between the actual and prior borders. It is denoted $h(\mathbf{b}, \mathbf{b}')$ for two (actual and prior) borders $\mathbf{b}$ and $\mathbf{b}'$ belonging to the same epipolar line. The second part, denoted $g(\mathbf{b}_k^{\alpha_0}, \mathbf{b}_n^{\alpha_P})$, estimates how well the association of two prior borders in the left and right viewpoints (corresponding to α_0 and α_P, respectively), is supported by the intermediate prior borders ($0 < \alpha_p < 1$). Eventually, the cost $d(\mathbf{b}_i^L, \mathbf{b}_j^R)$ of matching the i^{th} border of $\mathbf{B}^L$ with the j^{th} border of $\mathbf{B}^R$ is defined as:

$$d(\mathbf{b}_i^L, \mathbf{b}_j^R) = \min_{k,n} \left(h(\mathbf{b}_i^L, \mathbf{b}_k^{\alpha_0}) + g(\mathbf{b}_k^{\alpha_0}, \mathbf{b}_n^{\alpha_P}) + h(\mathbf{b}_n^{\alpha_P}, \mathbf{b}_j^R) \right), \tag{3}$$

which corresponds to the cost of the shortest path computed in the directed graph connecting the left and right reference borders through two intermediate layers, corresponding to left and right priors. The cost $d(\mathbf{b}_i^L, \mathbf{b}_j^R)$ is thus straightforward to compute using the Dijkstra's algorithm [47], as long as $h(,)$ and $g(,)$ are known. We discuss these distances below.

In the left view, the distance $h(\mathbf{b}_i^L, \mathbf{b}_k^{\alpha_0})$ between the i^{th} left reference border, and the k^{th} border of $\mathbf{B}^{\alpha_0}$ is defined to be the normalized Hamming distance between the reference and prior ELS, when $\mathbf{b}_i^L$ and $\mathbf{b}_k^{\alpha_0}$ are aligned[3]. Because $\mathbf{S}^{\alpha_0}$ and $\mathbf{S}^L$ are seen from the same camera viewpoint, the foreshortening effect does not influence the lengths of their epipolar segments, making the Hamming distance relevant due to this one-to-one pixel correspondence. In the right view, the distance $h(\mathbf{b}_n^{\alpha_P}, \mathbf{b}_j^R)$ between the n^{th} prior border in $\mathbf{B}^{\alpha_P}$ and the j^{th} reference border in $\mathbf{B}^R$, is defined analogously.

Then, in order to define the cost $g(,)$ of associating two borders between the first and last priors, we assume that the transition between the two borders should ideally follow a linear displacement between the reference views. It means that the displacement should be proportional to the variable α defining the normalized distance to the left view. With rectified cameras, the displacement proportionality assumption is strictly verified when the silhouette borders correspond to the same physical 3D point [10], independently of the viewpoint. In other cases, since the actual 3D point supporting the silhouette border generally does not move a lot when

[3] After alignment, the two ELSs generally do not fully overlap. Hence, the Hamming distance is computed on the part of the epipolar line where the two strings of ELS pixels overlap, and the normalization consists in dividing the distance by the length of this overlap.

the viewpoint changes, the assumption is also approximately valid. Under this assumption, we define $g(,)$ to reflect the discrepancy between the actual and the ideal linear displacement. In practice, for $\mathbf{b}_k^{\alpha_0}$ and $\mathbf{b}_n^{\alpha_P}$, this discrepancy, denoted $g(\mathbf{b}_k^{\alpha_0}, \mathbf{b}_n^{\alpha_P})$, is measured as the sum over the intermediate views ($p \in \{1, \cdots, P-1\}$) of the distances (normalized by the width of the image) between the linear interpolation of $\mathbf{b}_k^{\alpha_0}$ and $\mathbf{b}_n^{\alpha_P}$ in the intermediate view α_p and the closest prior borders having the same modality in $\mathbf{B}^{\alpha_p}$. This distance is illustrated with a green color code and for one intermediate view in Figure 4(c). A detailed formal definition of $g(,)$ is provided in [50].

C. Appearing/vanishing trajectories

We now study how unmatched borders move while going from one reference view to the other. In practice, since the displacement of a border along the EPI is assumed to be proportional to the normalized distance α [10], we only have to evaluate the corresponding coefficient of proportionality, called *border displacement slope* in the rest of the paper. In our new model, the unmatched border displacement slopes are estimated as a function of the ELS priors, in three steps.

First, the prior borders that correspond to occluded segments are identified as the ones that are sufficiently far from the linear interpolation of matched borders (*i.e.*, farther than 5% of the image width in our experiments). We name them *occluded prior borders* (dark red borders in Figure 5(a)), and group them in two distinct sets according to their modality. Each set is processed independently. Given the simplicity of the criterion adopted to identify occluded prior borders, each set might include false positives.

Hence, the second step of our approach considers a RANSAC-based procedure [51] to fit linear vanishing trajectories to each set of occluded prior borders. At each RANSAC iteration, two occluded prior borders are randomly selected to define a linear border displacement model. The occluded prior borders that are close enough (in our experiments, 5% of the width of the image) to the trajectory model are considered to be inliers. The linear model that maximizes the amount of inliers is kept, and its inliers are removed from the corresponding set of occluded prior borders, before subsequent applications of RANSAC. This simple greedy algorithm works well in practice, due to the relatively small number of outliers in the set of occluded prior borders. The outcome of the RANSAC-based fitting procedure is illustrated in Figure 5(b).

Finally, in a third step, the displacement slopes associated with those trajectories are assigned to the unmatched borders in the reference views (Figure 5(c)). The cost of assigning a trajectory to a reference border is simply defined to be the $\mathcal{L}_1$ distance between the border and the position defined by the trajectory in the reference view. The assignment problem is then solved using the Hungarian algorithm [52], so as to assign one and only one trajectory to each unmatched border while minimizing the sum of assignment costs.

V. View synthesis

This section describes how the intermediate view epipolar lines are synthesized based on the matching of reference silhouette epipolar borders computed in Section IV. For occluded epipolar segments, we propose to propagate the texture from the reference view in which the segment is visible, to the vanishing point, using a linear morphing strategy (Section V-A). For pairs of matched silhouette segments, two strategies are envisioned and compared in our experiments. The first one consists in a simple linear view morphing (Section V-B). The second one is based on a pixel-wise pixel association that is computed under strict ordering constraint assumption (Section V-C).

A. Texture occlusion

Let y denote the EPI index, and (x_v, α_v) denote the point in the EPI where the occluded segment vanishes, such as illustrated at the intersection of the two bold black lines in Figure 5(c). The linear morphing reconstructs the segment in view α as:

$$\mathbf{I}_\alpha\left(\frac{\alpha}{\alpha_v}\cdot(x_v - x_L) + x_L, y\right) = \mathbf{I}_L(x_L, y)$$

if the vanishing epipolar line segment belongs to the left reference image (occlusion), or as:

$$\mathbf{I}_\alpha\left(\frac{(\alpha-1)\cdot(x_v - x_R)}{\alpha_v - 1} + x_R, y\right) = \mathbf{I}_R(x_R, y)$$

if the vanishing segment belong to the right reference image (disocclusion). In these equations, x_L and x_R denote the position along the epipolar line in the left and right views, respectively. Their possible values are defined to cover the support of the occluded segment.

B. View morphing

The morphing-based combination of textures relies on epipolar rectification to synthesize the intermediate textures by linear interpolation of the reference textures. Between two matched segments, the texture is defined as [32]:

$$\mathbf{I}_\alpha(x_\alpha, y) = (1-\alpha)\cdot\mathbf{I}_L(x_L, y) + \alpha\cdot\mathbf{I}_R(x_R, y) \qquad (4)$$

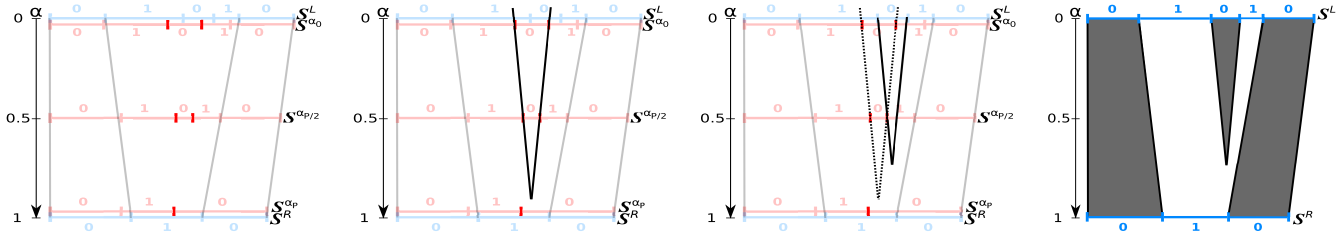
with y denoting the EPI index, and x_L denoting the ordinate of a pixel in the left view. x_α and x_R are computed as follows:

$$\begin{aligned} x_\alpha &= (1-\alpha)\cdot x_L + \alpha\cdot x_R & (5)\\ x_R &= \frac{l(\mathbf{s}_j^R)}{l(\mathbf{s}_i^L)}\cdot\left(x_L - p(\mathbf{b}_i^L)\right) + p(\mathbf{b}_j^R) & (6)\end{aligned}$$

with $\mathbf{s}_i^L$ denoting the epipolar line segment including x_L and $\mathbf{s}_j^R$ denoting the epipolar line segment matched to $\mathbf{s}_i^L$. Then $p(\mathbf{b}_i^L)$ is the position of the left border of the epipolar line segment including x_L and $p(\mathbf{b}_j^R)$ defines the position of the corresponding matched border.

C. Ordering preservation

The ordering constraint assumes that the left-right relation between the projection of 3D points belonging to the same epipolar plane is preserved when changing the observer point of view. It is not strictly valid, especially in a wide-baseline stereo setup, but it is usually valid within pairs of matched



(a) (b) (c) (d)

Fig. 5: As detailed in Section IV.C, the vanishing trajectories are estimated by (a) identifying the so-called occluded prior borders (in dark red, with light gray lines depicting matched borders trajectories), (b) fitting linear trajectories to these prior borders, and (c) associating the slopes of these trajectories (dotted lines) with the occluded reference borders (solid lines). Finally, the EPI of the object silhouette is reconstructed by (d) adding the vanishing trajectories to the set of trajectories describing the transitions between the associated reference borders.

segments. Under this assumption, the problem of associating the pixels of a pair of matched epipolar segments can be formulated based on the Needleman-Wunsch algorithm and solved using a dynamic programming approach, as detailed in [37]. Once the pixels have been associated, the interpolation of an intermediate segment is straightforward, and reads:

$$\mathbf{I}_\alpha(x_\alpha, y) = (1-\alpha)\cdot\mathbf{I}_L(x_L, y) + \alpha\cdot\mathbf{I}_R(x_R, y) \tag{7}$$

with x_L and x_R denoting the x coordinates of a pair of pixels associated between two matched epipolar segments, and y denoting the epipolar line index. The reconstructed pixel abscissa x_α is defined by (5).

VI. Results

In this section, we demonstrate the performance of our approach on well-known datasets, namely the synthetic *Kung-Fu Girl* sequence [53] and the real *Ballet* sequence [40]. Other validations on the *Dino* dataset [54] are presented in [50]. Although these multiview datasets contain numerous images acquired by multiple (small-baseline) cameras, we consider only a pair of widely separated cameras from these sets to learn our shape priors model and to reconstruct the intermediate views. In practice, only a small number of training samples (typically forty to sixty) is required. They all correspond to previous observations of the object moving in front of a single reference view. To evaluate our method, for each dataset, we present five intermediate views sampled uniformly in-between the left and right reference views. The videos showing the continuous transition from the left to the right cameras, as well as the source code of our implementation, are provided at http://infoscience.epfl.ch/record/200492.

A. The Kung-Fu Girl dataset

For the *Kung-Fu Girl* dataset, we have selected two wide-baseline cameras separated by an angular difference of 45°. They correspond to the cameras $b05$ and $b07$ in the dataset, and the 120^{th} frames of those two cameras are considered as the reference ones. The view captured by the left (or right) camera is shown on the left (respectively right) of the first row in Figure 6. The image shown in-between corresponds to the ground-truth captured by a camera situated approximatively at the middle ($\alpha \simeq 0.5$) in-between these two reference views.

The second row in Figure 6 represents the intermediate views generated based on a conventional visual-hull reconstruction [14][31]. The numerous artifacts in the reconstructed intermediate views reveal the limitations of such a straightforward model-based approach in our wide-baseline stereo.

The third row in Figure 6 represents the intermediate views generated based on the dense (pixel) correspondence between corresponding epipolar lines obtained by dynamic programming [37] [55], with matching cost set to the l_2-norm of the pixels' colors and the skipping penalty $u(.)$ arbitrarily set to 0.5. Two kinds of artifact can be observed in these reconstructed views. First, they are topologically incoherent. This can be observed between the legs of the Kung-Fu girl, near her neck, and on her left hand, where some parts separate from her body. Second, holes appear in the reconstructed intermediate views. This artifact, caused by the foreshortening effect, is generally avoided by imposing a smooth disparity/depth-map [29], at the cost of increased complexity.

To impose smoothness along the epipolar lines explicitly, as shown in the 4th row of Figure 6, we use epipolar line segments as matching elements. The method extends that of [37] by considering epipolar line segments (and not pixels) to be basic image elements. It corresponds to the approach we introduced in Section IV-B, but without prior silhouettes knowledge. Hence, each epipolar border is matched by NW [55], considering that $d(\mathbf{b}_i^L, \mathbf{b}_j^R) = h(\mathbf{b}_i^L, \mathbf{b}_j^R)$ (instead of Equation 3). We observe in Figure 6 that the reconstructed intermediate views are smoother but still exhibit some topologically incoherent transitions (see head and legs).

To guide the reconstruction of the EPIs in such a way that they provide topologically coherent intermediate views, a latent space has been learned from sixty silhouettes of the Kung-Fu girl captured by each of the two wide-baseline cameras (*i.e.*, b05 and b07), at 30 time instants uniformly sampled between the first and the hundredth frames. These silhouettes have been described using seventy elliptic harmonics. The fifth row in Figure 6 illustrates this learned latent space.

The advantage of considering these priors is illustrated in the last row in Figure 6, where intermediate views have been generated by the method proposed in this paper, using linear

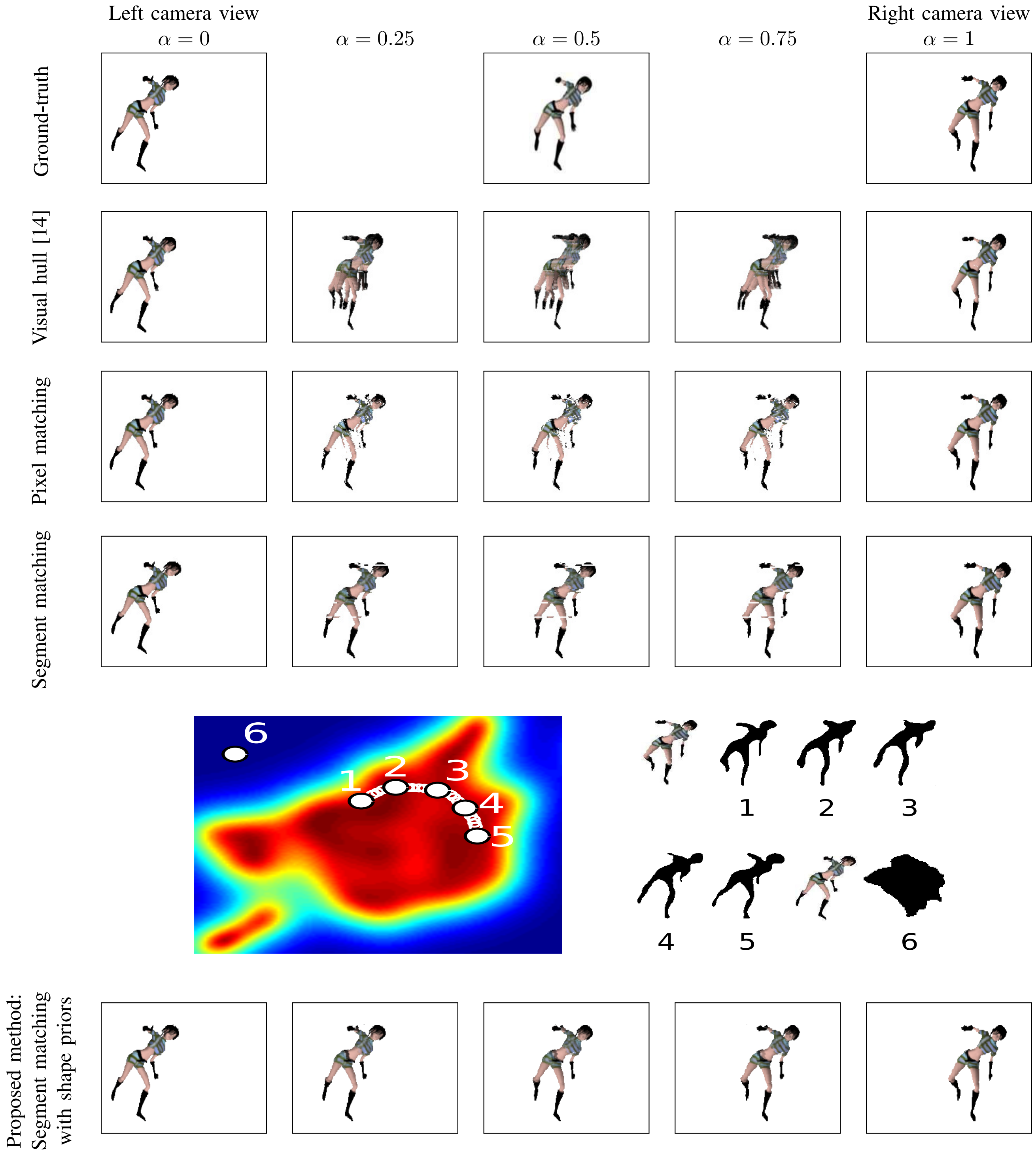


Fig. 6: Instead of projecting an estimated 3D model [14] (second row) or determining a dense (pixel) match (third row), epipolar line segments are used as basic matching elements (fourth row). In the last row, our method guides the epipolar segments' matching so that the shapes of the intermediate silhouettes are topologically consistent with the plausible deformations of the object silhouette, as learned and described by a low-dimensional latent space (fifth row).

view morphing for texture synthesis (Section V-B). Only five intermediate priors have been used to reconstruct the EPIs, i.e. $P = 4$. The intermediate views reconstructed by our method offer a topologically coherent transition of the Kung-Fu girl from the left to the right reference view.

B. The Ballet sequence

The second sequence, called *Ballet* [40], has been captured using eight cameras placed along a 1D arc, spanning about 30° end-to-end. While two neighboring cameras in this array constitute a small-baseline stereo pair, the outer cameras represent a wide-baseline configuration, with strong self-occlusions and foreshortening effects between the two external viewpoints (especially on the dancer's arms). For this sequence, we visually compare the images interpolated with different methods, but also provide quantitative values derived from the comparison between real and interpolated images. The covering percentage measures the fraction of the pixels in the real foreground image that is covered by the interpolated image. The PSNR is computed on this covered area only. Both values are averaged on the intermediate views that are available in the dataset. Those quantitative metrics should however be treated with caution due to the fact that the images that are available in the Ballet dataset (and that have been used as a ground truth) have been captured by cameras that are distributed along an arc of a circle, and not along the baseline. Hence, for each camera view in the Ballet dataset, we have computed the PSNR between the image captured by the real camera, and the image interpolated at the position along the baseline that is the closest to the real camera position.

In Figure 7, we compare the images reconstructed at intermediate viewpoints by five methods, using only the two extreme wide-baseline reference views (in contrast to the use of the small-baseline multiview pairs presented in [40]). The images presented in Figure 7 correspond to frame 30. In the first row, the intermediate views are generated by view morphing, based on multiple depth maps as proposed in [40][4]. Since the depth map estimated from the extreme wide-baseline views is very poor, we provide the images reconstructed from the textures in the two extreme views, based on the depth maps computed with neighboring cameras (small-baseline configuration). Even with this additional information, the small depth inaccuracies (equivalently weak pixel correspondences) lead to merging non-corresponding textures, *i.e.*, ghosting artifacts. The second row in Figure 7 illustrates the intermediate views obtained by a state-of-the-art (narrow-baseline) stereo method [56] top-ranked in February 2015 in the well-known Middlebury Stereo Evaluation [57] [58] [55]. By combining a cost-filtering approach, especially adapted to manage the occlusions, with a global (fully connected Markov Random Field) optimization, this method achieves impressive results on small-baseline stereo setups. However, as expected, the strong geometrical and photometric changes, as well as the foreshortening effects make this algorithm pretty vulnerable in our wide-baseline stereo setup. In the third row, we use only the external views and test wide-baseline stereo matching by applying the Needleman-Wunsch algorithm [55] to pixels, as done in [37]. Despite the high PSNR, it produces visually subpar results. We observe that strong foreshortening effects produce holes in the reconstructed intermediate views. By applying [55] to the segment representation, we find dense correspondences, but topological inconsistencies subsist (see fourth row in Figure 7). The last row in Figure 7 illustrates the result obtained by our complete method using silhouette priors and morphing-based texture interpolation. The latent space has been learned on the silhouettes captured by the same wide-baseline stereo pair (camera 0 and camera 7) between the 50^{th} and 90^{th} frames. These silhouettes have been described with fifty harmonics of Elliptic Fourier Descriptors, and six intermediate priors are used to drive the epipolar segment matching (P=6). We observe that our method prevents severe artifacts (holes or topologically inconsistent matching) in the interpolated image. We however observe that deriving the inner texture from Equation (4) might result in blurry ghosting artifacts, caused by wrong pixel associations between the matched segments. This explains the relatively small PSNR value observed in the last row of Figure 7, compared to the third row (using [37] on complete epipolar lines). To convince the reader that a more accurate inner texture interpolation would mitigate this issue, Figure 8 compares the images reconstructed using the morphing-based combination of inner texture (as described in Section V-B) with the ones obtained by the approach presented in Section V-C, which basically uses [37] between associated segments rather than complete lines. We observe that, indeed, finer association of inner pixels significantly improves the sharpness of edges in the reconstructed images. It also increases the PSNR much beyond the values obtained in Figure 7. Overall, the PSNR and covering values presented in the second line of Figure 8 reveal that our method outperforms the previous art. The gains in PSNR, compared to the methods in Figure 7, range from 10 to 6 dB. Moreover, the artifacts appearing in the third row of Figure 7, which uses pixel-wise association [37] over entire epipolar lines (instead than within the segments matched by our method), confirm that our proposed matching of silhouette epipolar segments provides a worthwhile pre-processing step, preventing the failure of conventional pixel-based matching techniques. Other inner texture interpolation methods than [37] could obviously be envisioned to follow this pre-processing step, *e.g.*, based on the floating textures concept [59].

C. Discussion

Our method appears to generate topologically coherent intermediate views, due to the additional silhouette prior obtained from the latent space. It also deals efficiently with the foreshortening effect, as for example observed from the chest of the dancer, which is slanted in the left view but almost fronto-planar in the right one. Finally, to the best of the authors' knowledge, this paper is the first one to interpolate occluded parts in intermediate views explicitly, as shown by the Kung-Fu girl's left shoulder and the space between the dancer's legs. Next to these very encouraging results, two

[4] The pixel correspondences are obtained by projection of the pixels of one reference view at the depth indicated by the depth map, and back-projection of these $3D$ points in the other reference view.

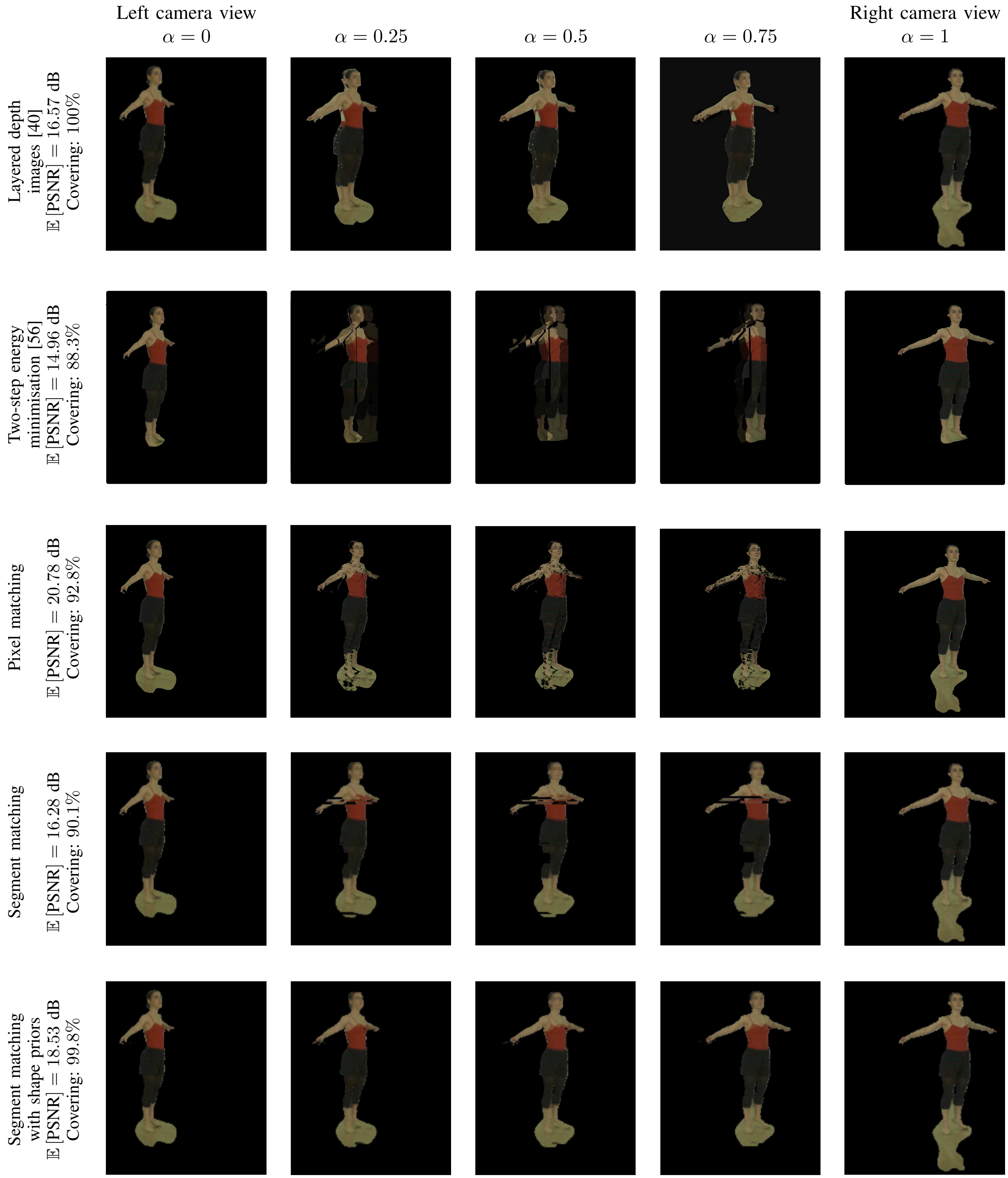

Left camera view
$\alpha = 0$
$\alpha = 0.25$
$\alpha = 0.5$
$\alpha = 0.75$
Right camera view
$\alpha = 1$
Layered depth images [40]
$\mathbb{E}\,[\mathrm{PSNR}] = 16.57$ dB
Covering: 100%
Two-step energy minimisation [56]
$\mathbb{E}\,[\mathrm{PSNR}] = 14.96$ dB
Covering: 88.3%
Pixel matching
$\mathbb{E}\,[\mathrm{PSNR}] = 20.78$ dB
Covering: 92.8%
Segment matching
$\mathbb{E}\,[\mathrm{PSNR}] = 16.28$ dB
Covering: 90.1%
Segment matching with shape priors
$\mathbb{E}\,[\mathrm{PSNR}] = 18.53$ dB
Covering: 99.8%

Fig. 7: Comparison between the interpolated intermediate views generated based on matching of layered representation [40] (first row), which exploits intermediate depth maps in addition to external views; a state-of-the-art narrow-baseline method [56] that is top-ranked in the Middlebury Stereo Evaluation [57] (second row); dynamic programming on pixels [37] (3rd row); dynamic programming on our proposed epipolar line segment representation (4th row); and our method (5th row).

limitations of our approach, however, deserve to be mentioned. The first one can be observed on the fingers of the dancer's right hand in Figure 7, which separate from her right hand when the viewpoint changes from the left to the right reference view, showing a topologically incoherent transition. This is due to the limited accuracy of the priors, determined from a low-dimensional space representing the approximations of training shapes as a set of N smooth harmonics (ellipses). When high frequency details, such as the dancer's fingers, are not represented by the priors, their matching cannot be properly driven, and their transition may become topologically incoherent. The second limitation is related to the fact that, despite providing a smooth and plausible transition between the views, there is no guarantee that the shortest geodesic path between the two reference views corresponds to the true evolution of the object silhouette across viewpoints. In particular, the proposed method could encounter difficulties for a combination of object and acquisition setup that produces the same rectified silhouette in both reference views (same point on the latent space), while inducing different silhouettes in the intermediate views. This case is reasonably unlikely for complex objects. However, a possible path to address those hypothetical challenging cases could be to extend the EFD features with features capturing the deformation of the silhouette in a temporal neighborhood. Finally, in presence of multiple objects, attention should be paid to the separation of the foreground silhouette into distinct object silhouettes, and to the possible interaction between them. This question is tightly connected to the application scenario at hand, and is beyond the scope of our paper. Also regarding the practical deployment of our method, we note that the processing time of our algorithm (Matlab implementation, code available on line at http://infoscience.epfl.ch/record/200492, Intel I5 CPU 2.4GHz and 8Gb of RAM) shows encouraging performance: on average 4.2s to describe a 768×1024 image in epipolar line segments, 0.06s to match all the epipolar lines independently, and 0.16s to render an intermediate view. Since the epipolar lines are processed independently, real-time implementation is within reach, *e.g.*, based on GPU parallelization.

VII. Conclusions

This paper interpolates intermediate views between two cameras in wide-baseline configurations. Our method relies on learned information about the silhouettes of an object in the intermediate views to guarantee consistency between the synthesized silhouettes and the ones present in the two reference viewpoints. We learn this prior information from previous observations of the object moving in front of the two wide-baseline cameras. The learned prior is then exploited to determine the 1D transformation of epipolar line segments when moving from one view to the other. Although the coherence is imposed independently on each epipolar line, the fact that the 1D constraints are derived from 2D priors promotes consistency across epipolar lines. We then demonstrate that this new framework not only has the advantage of generating consistent and smooth virtual transitions of correspondences between the viewpoints, but it can also handle the vanishing of occluded informations. Finally, we have shown that our method outperforms state-of-the-art view interpolation methods by generating topologically coherent intermediate views of an object, despite occlusions and severe foreshortening effects that are inherent in wide-baseline camera settings.

References

[1] N. D. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," NIPS, vol. 2, p. 5, 2004.

[2] M. Levoy and P. Hanrahan, "Light field rendering," in Conf. on Computer Graphics and Interactive Techniques. ACM, 1996, pp. 31–42.

[3] J. Owens, Television sports production. Taylor & Francis, 2012.

[4] A. Smolic, "3D video and free viewpoint video: from capture to display," Journal of Pattern recognition, vol. 44, no. 9, pp. 1958–1968, 2011.

[5] P. Goorts, C. Ancuti, M. Dumont, S. Rogmans, and P. Bekaert, "Real-time video-based view interpolation of soccer events using depth-selective plane sweeping," in International Conf. on Computer Vision Theory and Applications. IEEE, 2013.

[6] C. D. Castillo and D. W. Jacobs, "Wide-baseline stereo for face recognition with large pose variation," in Conference on Computer Vision and Pattern Recognition. IEEE, 2011, pp. 537–544.

[7] L. Sun, Q. De Neyer, and C. De Vleeschouwer, "Multimode spatiotemporal background modeling for complex scenes," in European Signal Processing Conference. IEEE, 2012, pp. 165–169.

[8] R. Hartley and A. Zisserman, Multiple view geometry in computer vision. Cambridge University Press, 2003.

[9] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," International Journal of Computer Vision, vol. 1, no. 1, pp. 7–55, 1987.

[10] A. Criminisi, S. B. Kang, R. Swaminathan, R. Szeliski, and P. Anandan, "Extracting layers and analyzing their specular properties using epipolar-plane-image analysis," Computer Vision and Image Understanding, vol. 97, no. 1, pp. 51–85, 2005.

[11] H. Saito and T. Kanade, "Shape reconstruction in projective grid space from large number of images," in Conference on Computer Vision and Pattern Recognition, vol. 2. IEEE, 1999, pp. 49–54.

[12] S. Yaguchi and H. Saito, "Arbitrary viewpoint video synthesis from multiple uncalibrated cameras," Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 34, no. 1, pp. 430–439, 2004.

[13] A. Laurentini, "The visual hull concept for silhouette-based image understanding," Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 2, pp. 150–162, 1994.

[14] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, "Image-based visual hulls," in Conf. on Computer Graphics and Interactive Techniques. ACM, 2000, pp. 369–374.

[15] T. Matsuyama and T. Takai, "Generation, visualization, and editing of 3D video," in Proc. of the First International Symposium on 3D Data Processing Visualization and Transmission. IEEE, 2002, pp. 234–245.

[16] G. Slabaugh, R. Schafer, and M. Hans, "Image-based photo hulls," in Proc. of the First International Symposium on 3D Data Processing Visualization and Transmission. IEEE, 2002, pp. 704–862.

[17] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel, "Free-viewpoint video of human actors," Transactions on Graphics, vol. 22, no. 3, pp. 569–577, 2003.

[18] M. Germann, T. Popa, R. Keiser, R. Ziegler, and M. Gross, "Novel-view synthesis of outdoor sport events using an adaptive view-dependent geometry," in Computer Graphics Forum, vol. 31, 2012, pp. 325–333.

[19] S. M. Seitz, "Image-based transformation of viewpoint and scene appearance," Ph.D. dissertation, University of Wisconsin-Madison, 1997.

[20] A. Smolic, K. Mueller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, "3D video and free viewpoint video-technologies, applications and mpeg standards," in International Conference on Multimedia and Expo. IEEE, 2006, pp. 2161–2164.

[21] B. Goldlucke and M. Magnor, "Real-time microfacet billboarding for free-viewpoint video rendering," in International Conference on Image Processing, vol. 3. IEEE, 2003, pp. 713–717.

[22] L. Ballan, G. J. Brostow, J. Puwein, and M. Pollefeys, "Unstructured video-based rendering: interactive exploration of casually captured videos," Transactions on Graphics, vol. 29, p. 87, 2010.

[23] M. Reynolds, J. Dobos, L. Peel, T. Weyrich, and G. J. Brostow, "Capturing time-of-flight data with confidence," in Conference on Computer Vision and Pattern Recognition. IEEE, 2011, pp. 945–952.

[24] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison et al., "KinectFusion: real-time 3D reconstruction and interaction using a moving depth

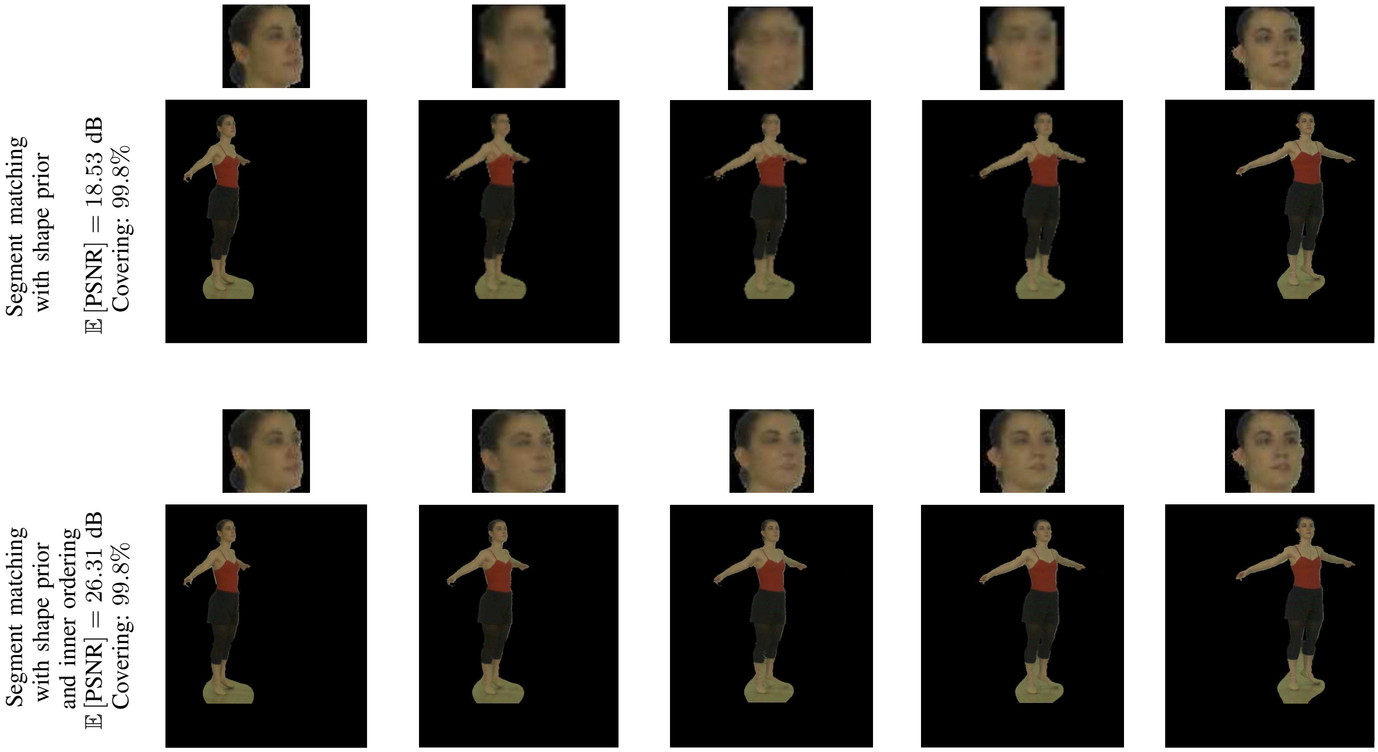


Fig. 8: Comparison between the interpolated intermediate views generated based on our proposed epipolar silhouette segment matching, when using view morphing (top, Section V-B) or dynamic programming on pixels [37] (bottom, Section V-C) to reconstruct the inner texture of each segment. Zoomed-in versions of the face area are provided for easier comparison.

camera," in Proc. of the 24th annual symposium on User interface software and technology. ACM, 2011, pp. 559–568.

[25] C. Verleysen and C. De Vleeschouwer, "Piecewise-planar 3D approximation from wide-baseline stereo," in Conference on Computer Vision and Pattern Recognition. IEEE, 2016, pp. 3327–3336.

[26] A. Bodis-Szomoru, H. Riemenschneider, and L. V. Gool, "Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels," in Conference on Computer Vision and Pattern Recognition. IEEE, 2014, pp. 469–476.

[27] I. Kitahara and Y. Ohta, "Scalable representation for 3D video display in a large-scale space," in Conf. on Virtual Reality. IEEE, 2003, pp. 45–52.

[28] S. B. Kang, Y. Li, X. Tong, and H.-Y. Shum, "Image-based rendering," Journal of Computer Graphics and Vision, vol. 2, no. 3, 2006.

[29] H.-Y. Shum, S.-C. Chan, and S. B. Kang, Image-based rendering. Springer Science and Business Media, 2008.

[30] A. W. Fitzgibbon, Y. Wexler, A. Zisserman et al., "Image-based rendering using image-based priors," International Journal of Computer Vision, vol. 63, no. 2, pp. 141–151, 2005.

[31] N. Inamoto and H. Saito, "Virtual viewpoint replay for a soccer match by view interpolation from multiple cameras," Transactions on Multimedia, vol. 9, no. 6, pp. 1155–1166, 2007.

[32] S. M. Seitz and C. R. Dyer, "View morphing," in Conf. on Computer Graphics and Interactive Techniques. ACM, 1996, pp. 21–30.

[33] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Computer Science Technical Report, vol. 2, no. 11, 2005.

[34] O. Veksler, "Fast variable window for stereo correspondence using integral images," in Conference on Computer Vision and Pattern Recognition, vol. 1. IEEE, 2003, pp. I–556.

[35] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 11, pp. 1222–1239, 2001.

[36] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," International Journal of Computer Vision, vol. 70, no. 1, pp. 41–54, 2006.

[37] I. J. Cox, S. L. Hingorani, S. B. Rao, and B. M. Maggs, "A maximum likelihood stereo algorithm," Computer Vision and Image Understanding, vol. 63, no. 3, pp. 542–567, 1996.

[38] S. Mattoccia, "Accurate dense stereo by constraining local consistency on superpixels," in Int. Conf. on Pattern Recognition. IEEE, 2010.

[39] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in Int. Conf. on Pattern Recognition. IEEE, 2006, pp. 15–18.

[40] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in Transactions on Graphics, vol. 23, no. 3. ACM, 2004, pp. 600–608.

[41] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz, "Patchmatch belief propagation for correspondence field estimation," International Journal of Computer Vision, pp. 1–12, 2012.

[42] F. P. Kuhl and C. R. Giardina, "Elliptic Fourier features of a closed contour," Computer Graphics and Image Processing, vol. 18, no. 3, pp. 236–258, 1982.

[43] V. A. Prisacariu and I. Reid, "Nonlinear shape manifolds as shape priors in level set segmentation and tracking," in Conference on Computer Vision and Pattern Recognition. IEEE, 2011.

[44] V. Prisacariu and I. Reid, "Shared shape spaces," in ICCV, 2011.

[45] N. D. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," Journal of Machine Learning Research, vol. 6, pp. 1783–1816, 2005.

[46] N. D. Lawrence and J. Quiñonero-Candela, "Local distance preservation in the GP-LVM through back constraints," in International Conference on Machine Learning. ACM, 2006, pp. 513–520.

[47] E. W. Dijkstra, "A note on two problems in connexion with graphs," Numerische mathematik, vol. 1, no. 1, pp. 269–271, 1959.

[48] V. I. Levenstein, "Binary codes capable of correcting deletions, insertions, and reversals," Soviet Physics Doklady, vol. 10, no. 8, pp. 707–710, 1966.

[49] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," Journal of molecular biology, vol. 48, no. 3, pp. 443–453, 1970.

[50] C. Verleysen, "3D estimation and view synthesis in wide-baseline stereo," Ph.D. dissertation, Université catholique de Louvain, 2015. [Online]. Available: https://dial.uclouvain.be/pr/boreal/en/object/boreal%3A167678/datastreams

[51] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis," Communications of the ACM, vol. 24, no. 6, pp. 381–395, 1981.

[52] H. W. Kuhn, “The Hungarian method for the assignment problem,” Naval research logistics quarterly, vol. 2, no. 1-2, pp. 83–97, 1955.

[53] M.-P.-I. Informatik, “A synthetic test sequence for multi-view reconstruction and rendering research,” 2005. [Online]. Available: http://www.mpi-inf.mpg.de/departments/irg3/kungfu/

[54] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, “A comparison and evaluation of multi-view stereo reconstruction algorithms,” in Conference on Computer Vision and Pattern Recognition, vol. 1. IEEE, 2006, pp. 519–528.

[55] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” International Journal of Computer Vision, vol. 47, no. 1-3, pp. 7–42, 2002.

[56] M. Mozerov and J. van de Weijer, “Accurate stereo matching by two-step energy minimization,” Transactions on Image Processing, vol. 24, no. 3, pp. 1153–1163, 2015.

[57] D. Scharstein, H. Hirschmüller, Y. Kitajima et al., “High-resolution stereo datasets with subpixel-accurate ground truth,” in German Conference on Pattern Recognition. Springer, 2014, pp. 31–42.

[58] H. Hirschmuller and D. Scharstein, “Evaluation of cost functions for stereo matching,” in Conference on Computer Vision and Pattern Recognition. IEEE, 2007, pp. 1–8.

[59] M. Eisemann, B. De Decker, M. Magnor, P. Bekaert, E. De Aguiar, N. Ahmed, C. Theobalt, and A. Sellent, “Floating textures,” in Computer Graphics Forum, vol. 27, no. 2. Wiley, 2008, pp. 409–418.