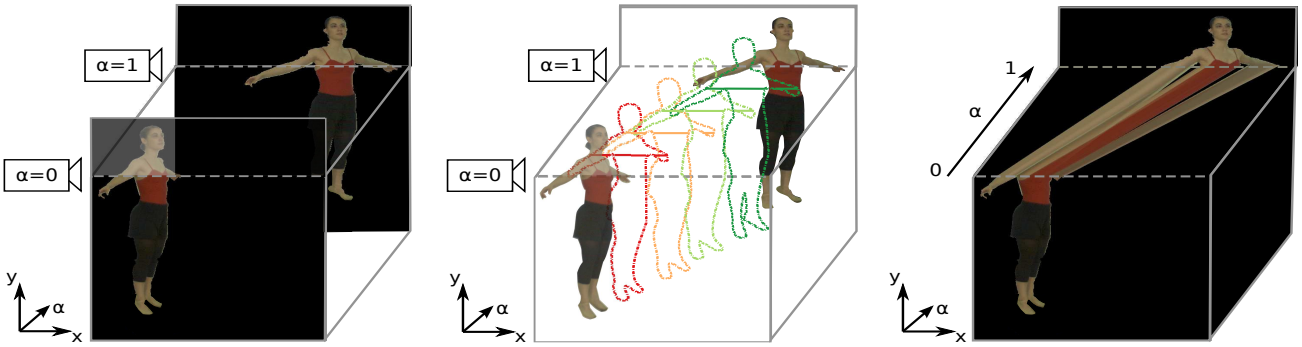
Wide-baseline object interpolation using shape prior regularization of epipolar plane images

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



(a) Reference views

(b) Reference views and silhouette priors.

(c) Intermediate views continuum associated with the Epipolar Plane volume.

Fig. 1: Intermediate views are generated from only two reference views (a). For that purpose, our method derives intermediade silhouette shapes between the reference silhouettes (b) from a Gaussian Process latent variable model, typically learnt from a set of images previously captured by the same two reference cameras. It then exploits the resulting shape priors to match epipolar silhouette segments between the views, thereby reconstructing Epipolar Plane Images (EPIs) and the continuum of views associated with the Epipolar Plane Image Volume, obtained by aggregating the EPIs (c).

***Abstract*—We consider the synthesis of intermediate views of an object captured by two widely spaced calibrated cameras. Based only on those two views, our paper reconstructs the object Epipolar Plane Image Volume [1] (EPIV), which describes the object when moving the synthetic viewpoint in-between the two reference cameras. This problem is challenging when the cameras are far apart because occlusions and the foreshortening effect induce significant differences between the two reference views. Our main contribution consists in disambiguating this ill-posed problem by making interpolated views consistent with an object shape prior. This prior is learnt from the object silhouettes segmented in images captured by the two reference views. It consists of a nonlinear shape manifold representing the plausible silhouettes of the object described by Elliptic Fourier Descriptors. Dynamic programming is then used to associate the epipolar silhouette segments between the two reference views while minimizing the distance between the linearly interpolated silhouettes and the prior. Eventually, conventional (non-)linear interpolation methods are considered to synthesise the texture in intermediate epipolar silhouette segments. Experiments on synthetic and natural images show that the proposed method preserves the object topology in intermediate views while dealing effectively with the self-occluded regions and the severe foreshortening effect associated with wide-baseline camera configurations.**

***Index Terms*—Wide-baseline, epipolar plane image, light field, view interpolation, free viewpoint, shape priors.**

I. Introduction

Virtual view synthesis aims at rendering images of a real scene from different viewpoints than the ones acquired by the cameras. Our paper restricts the synthesis problem to the interpolation of images of a dynamic object observed by a virtual camera positioned at an arbitrary point along the wide baseline connecting two reference cameras. The graceful transition in synthetic views between two reference viewpoints is a feature greatly in demand, especially in the field of video production [2]. For example, in the rendering of cultural and sport events, conventional acquisition systems switch abruptly between the cameras, making the viewer uncomfortable. By generating a graceful transition between the reference viewpoints, view interpolation gives the ability to understand how the rendered viewpoint changes, *i.e.*, the feeling of being "inside the scene".

To synthesise intermediate views in-between reference cameras, state-of-the-art methods generally decompose the scene into its background and its dynamic foreground objects and reconstruct them independently [3]. The still background 3D geometry is typically acquired based on state-of-the-art active 3D acquisition systems [4] [5], or estimated based on piecewise planar 3D geometry approximations [6] [7] [8] [9]. Hence, the interpolation of dynamic foreground objects situated relatively close to the pair of cameras is the most complex question among both [10] [11]. The fundamental issues encountered to reconstruct a foreground object lie in (1) the availability of only two reference views and (2) the object's proximity to the cameras compared with the distance between those cameras. The first factor prevents dense 3D estimation, while the second causes many projective discrepancies between the two views (occlusions, foreshortening effects), which again hamper the computation of dense correspondences and lead to holes in the interpolated views [11].

C. Verleysen and C. De Vleeschouwer are with the ICTEAM institute, Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium.
T. Maugey is with Inria, Rennes, France.
P. Frossard is with the LTS4 laboratory, EPFL, Lausanne, Switzerland.

This paper assumes that the object silhouette can be extracted from the reference views[1], as generally assumed by state-of-the-art foreground synthesis methods [13] [14] [8]. Based only on two very different views captured by a pair of wide-baseline cameras, such as the ones shown in Figure 1(a), our scheme reconstructs intermediate views of the object along the baseline in the form of an Epipolar Plane Image Volume [1] (see Figure 1(c)). The specifity of our method lies in the fact that it regularizes the Epipolar Plane Images (EPIs) reconstruction based on a sequence of plausible intermediate object silhouettes (see Figure 1(b)). This sequence is derived from a low-dimensional manifold that is learnt from the previous observations of the dynamic object moving in front of the wide-baseline stereo pair. Given the sequence of silhouette priors, dynamic programming is considered to associate the epipolar silhouette segments between the two reference views while minimizing the distance between the linearly interpolated intermediate silhouettes and their respective prior. Silhouette segment textures are then synthesised using conventional methods. Interestingly, the priors are used not only to disambiguate the association, but also to determine how occluded parts vanish/appear while moving from one reference view to the other. This gives our work the unique ability to reconstruct visually pleasant and topologically consistent images in presence of the significant self-occlusions and the severe foreshortenings inherent to wide baselines.

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 introduces our proposed Epipolar Plane Images interpolation formalism. Section IV 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 constrain the reconstruction of the EPIs between two reference images, as detailed in Section V. The view synthesis process is described in Section VI. Section VII 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 multi-view images. Adequate texture is then mapped on the mode, and projected onto any arbitrary viewpoint. Methods such as projective grid space [15] [16], visual-hull [17] [18] [19] [20], 3D model adjustment [21], and shape from video [13] belong to this category. These methods have the advantage of synthesising intermediate views representing the actual $3D$ scene. However, the quality of the virtual view is highly dependent on the accuracy of the estimated 3D model [22]. 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 [23]. 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 great compared with the baseline distance separating these cameras [24]. 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 of the object [8]. This makes model-based rendering inappropriate to render close (dynamic) scenes between wide-baseline cameras.

In contrast, image-based rendering (IBR) methods [25] create the virtual view directly in the image colour 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 colour 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 optimising multiple depth maps (either the ones of the virtual views [14] or those of the reference views [8]) and/or the virtual image's colour [26]. 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 only two reference views and generate the intermediate views by interpolation [27] or morphing [28]. They are restricted to the reconstruction of images on the baseline between a pair of reference cameras, generally for small-baseline configurations, and rely on dense correspondence between the views. This trend culminates in light-field reconstruction approaches [1], which require tens or hundreds of narrow-baseline[2] cameras/lenses [29] to determine a continuous (sub-pixel) correspondence between the reference views. 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 totally different number of pixels in different views. This implies that finding correspondences

[1]In this paper, the foreground is extracted by thresholding the $\mathcal{L}_2$ color distance with a Gaussian mixture model of the background [12].

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

with fixed-template matching methods fails [30]. The same holds when a pixel correspondence is optimised by graph-cut [31], belief propagation [32], or dynamic programming [33] 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 limits the correspondence-based interpolation methods [28] in a wide-baseline configuration.
- The lack of non-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 [34] [35] [36], but they are based either on globally non-valid hypothesis (*e.g.*, holes should contain patterns that are visible in the non-occluded parts) or on computationally expensive (post-)processing [37].

Our work addresses these issues explicitly by computing correspondences between image segments (from which dense correspondences can be inferred), and by constraining those correspondences to be consistent with a plausible deformation of the projected object silhouette between the reference views (thereby dealing with occlusion of segments, or their shrinkage/elongation due to the foreshortening effect).

III. Wide-baseline interpolation algorithm

This paper adopts an Epipolar Plane Image (EPI) formalism to reconstruct the image of a foreground object between two widely spaced cameras. The images between the reference viewpoints are described by the object Epipolar Plane Image Volume [38] [39] (EPIV), which is obtained by arranging in a 3D stack the images captured by a dense array of cameras that are distributed uniformly along the baseline with their image plane coplanar and vertically aligned. This implies epipolar rectification [40] of the images, which associates each horizontal line in one image with rows with the same ordinate in the other images, as illustrated in Figure 2. As a consequence, two corresponding pixels belong to the same horizontal plane in the EPIV and any transverse cross-section of this 3D cube, *i.e.*, each EPI (see Figure 1(c)) describes how the pixels of one epipolar line in a view move to the other view. The light field theory [1] states that these transitions are always linear and their slopes are inversely proportional to the scene's depth. The EPVI is rich and includes depth information. However, its estimation has been limited so far to very narrow-baseline setups. We adopt a new object-based approach to reconstruct the EPIV, as illustrated in Figure 2. After epipolar rectification of the reference views, the shape of the foreground object is described by a sequence of background/foreground segment along each epipolar line in each view. Our contribution primarily aims at matching those segments between the two reference images, and using this information to guide EPI synthesis. Therefore, we :

1) learn a low-dimensional silhouette manifold to describe prior plausible transformations of the object silhouette when changing the viewpoint along the baseline (see Figure 2, left side). In absence of intermediate cameras, the training set is composed of silhouettes of the object moving in front of the stereo pair.
2) locate, in this low-dimensional space, the silhouettes observed at the interpolation time in the reference views, and interpolate on the manifold a sequence of silhouettes that likely represents the deformation of the object silhouette between the reference views.
3) use this sequence of plausible silhouettes to define how the object silhouette epipolar foreground segments are transformed (*i.e.*, through scaling/translation/vanishing) between the two reference views (Figure 2, right side), and reconstruct the intermediate silhouette segments by mapping texture on the foreground epipolar segments.

The different components of our novel view interpolation algorithm are described in detail in the next sections.

IV. Object silhouette priors

This section derives a sequence of object silhouettes that defines *a priori* a plausible transition from the left reference silhouette to the right one. This sequence is used in Section V to regularize the reconstruction of the EPIV.

The main challenge of our approach lies in the definition of a low-dimensional space that ensures that the interpolation step results in a smooth and topologically coherent sequence of silhouette priors. As detailed below, our work *learns* this space from a set of silhouette samples. These samples should ideally describe how the object silhouette evolves in intermediary views as the viewpoint is moved along the baseline. They should thus be derived from intermediate view observations. However, since such observations are not available in many practical cases (*i.e.*, because only the images captured by the stereo pair are available), all our experiments instead learn the manifold from a set of silhouettes that have been previously observed by one of the reference views. Thereby, we implicitly assume that the silhouettes of the object moving in front of a reference view are representative of the silhouettes observed when a camera is moved between the two reference views.

To learn and exploit the low-dimensional manifold from this set of silhouette samples, we follow the pioneering approach of [41, 42]: we first describe the silhouette based on Elliptic Fourier Descriptors (Section IV-A), map those high-dimensional features to a lower dimensional latent space (Section IV-B), and builds on the resulting manifold to interpolate and register a sequence of plausible silhouettes between the reference views (Section IV-C and IV-D, respectively).

A. High-dimensional silhouette description

We use Elliptic Fourier Descriptors (EFD) [43] to associate a high-dimensional feature vector with each silhouette. Therefore, the silhouette contour is first represented as a sequence of 2D coordinates $(x(t), y(t))$ in a 2D referential with arbitrary origin but the X direction of which is aligned with rectified epipolar lines. The origin of the contour is selected arbitrarilyto be the point with the smallest x-coordinate among the ones having the same y-coordinate as the silhouette's centre of mass. Given those definitions, Elliptic Fourier Descriptors

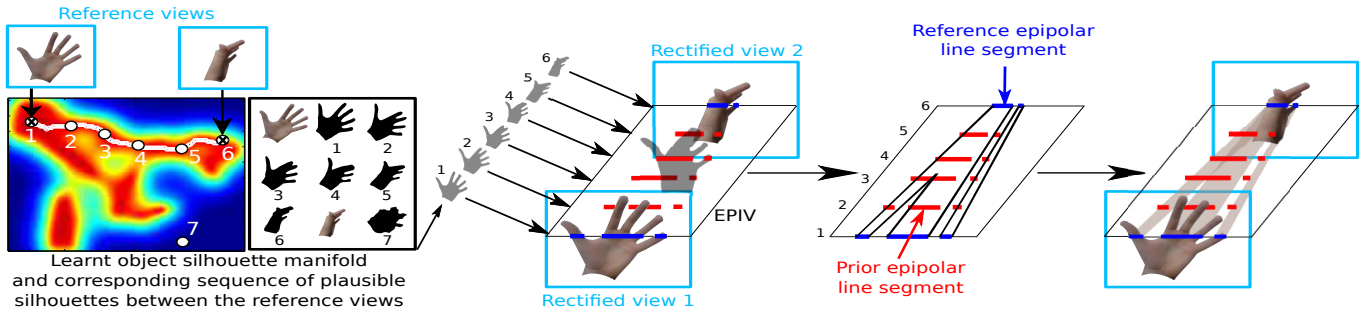


Fig. 2: The reconstruction of Epipolar Plane Images (EPIs) is regularized by incorporating prior knowledge about the deformations of the object silhouette. This knowledge is learnt from previous observations of the object by (one of) the two reference views and is captured by a a low-dimensional Gaussian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

then represent the shape of a silhouette as a sum of N elliptic harmonics, based on:

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

where T is the perimeter of the contour and:

$$\begin{aligned} a_0 &= \frac{1}{T} \sum_{p=1}^{K} \left(\frac{\Delta x_p}{2\Delta t_p} (t_p^2 - t_{p-1}^2) + \xi_p \ (t_p - t_{p-1}) \right) \\ a_n &= \frac{T}{2n^2\pi^2} \sum_{p=1}^{K} \left(\frac{\Delta x_p}{\Delta t_p} \left(\cos \frac{2\pi n t_p}{T} - \cos \frac{2\pi n t_{p-1}}{T} \right) \right) \\ b_n &= \frac{T}{2n^2\pi^2} \sum_{p=1}^{K} \left(\frac{\Delta x_p}{\Delta t_p} \left(\sin \frac{2\pi n t_p}{T} - \sin \frac{2\pi n t_{p-1}}{T} \right) \right) \end{aligned}$$

where

$$\xi_p = \sum_{j=1}^{p-1} \Delta x_j - \frac{\Delta x_p}{\Delta t_p} \sum_{j=1}^{p-1} \Delta t_j$$

with K being the number of sampling points in the contour, t_p the curvilinear coordinates on the shape, x_p the abscissa projection of t_p, $\Delta x_p = x_p - x_{p-1}$, and $\Delta t_p = \sqrt{(\Delta x_p)^2 + (\Delta y_p)^2}$. The second coordinate of the shape contour, $y(t)$, is defined completely analogously in terms of coefficients c_0, c_n, and d_n by replacing Δx_p with Δy_p. 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 the contour perimeter T.

B. Low-dimensional silhouette manifold

We then map M instances of high-dimensional EFD feature vectors to a low-dimensional latent space that represents the different plausible silhouettes. We use a nonlinear dimensionality reduction technique called *Gaussian Process Latent Variable model* (GPLVM) [44]. This technique is used because GPVLM enables one to work in low dimension while still capturing most of the shape variability.

More specificallyl, GPLVM considers that each dimension of the data samples results from a Gaussian process and represents those samples in a lower dimensional embedded space known as the latent space. Hence, GPLVM describes a data set $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_M]^T$ composed of M original data points (*e.g.*, M reference silhouettes represented with EFD) collected in a D dimensional space ($D = 4 \cdot N$ here), with a lower dimensional set of latent variables $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_M]^T$, where each variable is a latent point of dimensionality d, with $d << D$ ($d = 2$ in our case).

Given a definition of the covariance kernel in the latent space, GPLVM determines how the latent points $\mathbf{X}$ associated with the training data samples $\mathbf{Y}$ should be distributed so as to maximise the likelihood $P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})$ of the observed training samples, with $\boldsymbol{\theta}$ denoting the kernel parameters. As in [45], this likelihood is defined as the product of D separate Gaussian processes having the same shared covariance/kernel function. Sharing the covariance function leads to an a posteriori shared level of uncertainty (=variance) in each process. This permits to associate a precision with 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. A small precision corresponds to a large variance and thus to an ambiguous mapping. By contruction of the latent space model, which aims at maximising the likelihood of training data samples, high precision latent points correspond to shapes that are well represented in the training set. In our paper, we visualize the precision by varying the colour of the latent space pixels. Warmer, *e.g.*, red, (colder, *e.g.*, 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 Figure 2).

In its native formulation, GPLVM ensures that points that are close in the latent space remain close in the data space. Equivalently, it keeps the points that are far apart in the data

space 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 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 [46]. As a result, the learnt 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.

As an example, in the left part of Figure 2, GPLVM has learnt a 2-dimensional latent space from a set of $M = 150$ shapes of silhouettes captured on video sequence representing hands' gestures and described by thirty-five elliptic harmonics. The colour map in Figure 2 represents the precision of this learnt latent space, where the regions with the warmest colours are more likely to represent the shape of a hand.

C. Interpolating intermediate silhouettes on the manifold

To derive a sequence of plausible 2D silhouettes between the reference views, we first project the left and right reference silhouettes on the latent space (points 1 and 6 in the left part of Figure 2), based on the smooth mapping function introduced to force the preservation of local distance from data to latent space [46]. We then use a shortest path algorithm to interpolate a plausible transition between these low-dimensional reference silhouettes, and obtain the corresponding high-dimensional silhouette prior by projecting some (uniformly sampled) points of this path from the latent space to the image spaces. The black silhouettes in the left part of Figure 2 illustrate the silhouettes obtained by back-projection (from the latent space to the shape space) of the points represented in white in the latent space (learnt object manifold).

More precisely, the path connecting the two reference silhouettes in the latent space is obtained by computing, in a discretized latent space, the shortest geodesic path between the two reference latent points (white path in Figure 2). This path is computed using the Dijkstra's algorithm [47]. Each node in the graph corresponds to a point in the discretized latent space, and edges connect neighbouring nodes only, with the cost c_{ij} between two neighboyring nodes i and j being defined to be inversely proportional to the precision of node j ($c_{ij} = -\log(\text{precision}_j + \epsilon)$, where ϵ avoids numerical instabilities). By GPLVM construction, a short sequence of high-precision neighbours in the latent space corresponds to a sequence of silhouettes that is (i) highly likely, in the sense that the silhouettes correspond to high precision latent points, and (ii) topologically consistent, in the sense that consecutive silhouettes differ from each other only slightly. In so doing, we obtain a smooth sequence of silhouettes that likely corresponds to the sequence observed when moving the viewpoint from one reference view to the other (reconstruction scenario), or when the 3D objet 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 and are highly likely in most realistic practical cases, where the silhouettes observed from two distinct viewpoints are different[3].

D. Registering the silhouette priors with the reference ones

The set of prior foreground silhouettes derived hereabove represents a smooth and topologically consistent interpolation between the projections of the two reference silhouettes on the latent space. However, due to the translation and shape invariance of the shape features (see Section IV.A), these priors describe the 2D shapes of the silhouettes up to a scaling and translation. To exploit them during the EPIV reconstruction, we have thus to register them in the EPIV approximately. This alignment is performed in three consecutive steps by:

1) Translating the oriented prior shapes in such a way that their centres of mass coincide with the linear interpolation of the centres of mass of the two reference silhouettes,
2) Scaling the translated and orientated prior shapes based on the linear interpolation of the height of the object between the two reference silhouettes.

Figure 1(b) and Figure 3(a) show examples of registered prior shapes, sampled along the manifold shortest path in proportion to their distance to the reference cameras. The next section explains how to exploit this prior knowledge to match the epipolar silhouette segments betwen the reference views.

V. Transformations of epipolar line segments

This section explains how the sequence of $2D$ silhouette priors help in estimating how the epipolar line silhouette segments evolve between the reference views, thereby supporting EPI reconstruction, as illustrated in the right side of Figure 2.

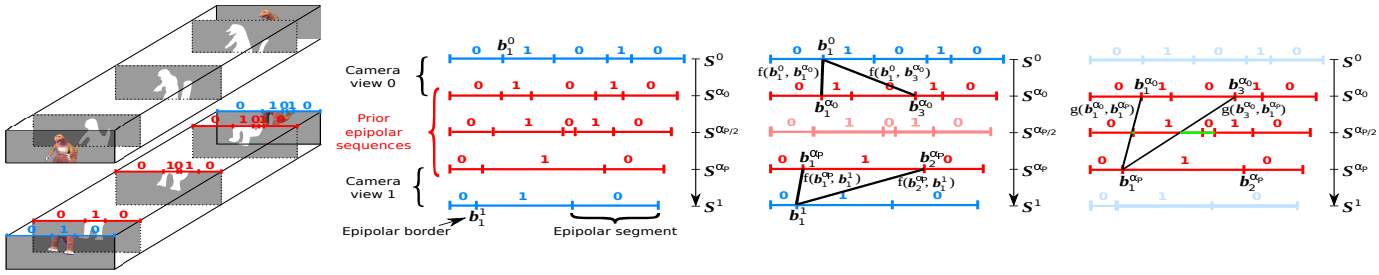
Due to the epipolar rectification of the reference images, the set of possible geometric transformations of an epipolar line silhouette segment is restricted to the combination of an horizontal translation, a 1D scaling, and a potential split-up or merge with other epipolar line silhouette segments. In the following text, without loss of generality, we represent those combined transformations based on the displacement and potential fusion of the segments' borders.

We first introduce some notations.

Let $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{\mathcal{S}}]$ denote a sequence of consecutive foreground and background epipolar line segments defined along a rectified epipolar line as illustrated on one of the blue or red lines of Figure 3(a). For more clarity, in the figures these reference (blue) and prior (red) epipolar line segments will be represented as a front view, as shown in Figure 3(b). The number of segments constituting the rectified epipolar line is denoted by $\mathcal{S} = |\mathbf{S}|$. Each segment $\mathbf{s}_k \in \mathbf{S}$ (with $k \in \{1, 2, \cdots, \mathcal{S}\}$) is characterized by a binary value, denoted $v(\mathbf{s}_k)$, depending on whether it corresponds to foreground (1) or background (0) information, and by its normalized length $l(\mathbf{s}_k)$ relative to the length of the entire sequence $\mathbf{S}$.

We associate a sequence of epipolar borders $\mathbf{B} = [\mathbf{b}_0, \mathbf{b}_1, \cdots, \mathbf{b}_{\mathcal{S}}]$ with each sequence $\mathbf{S}$, where $\mathbf{b}_{k-1}$ and

[3]When the same silhouette is observed in distinct camera viewpoints, multiple sequences of intermediate silhouettes may connect two identical pairs of reference silhouettes, because these identical pairs are observed from distinct viewpoints. In those cases, the GPLVM shortest path captures the shortest latent-space sequence only, and fails in capturing the others.



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

Fig. 3: (a) Prior information about the plausible deformations of the object silhouette is used to determine the cost of matching the left epipolar border $\mathbf{b}_1^0$ to the right epipolar border $\mathbf{b}_1^1$ (b). This cost is defined by minimising the sum of (c) the cost f of moving from a reference border to a prior one and (d) the discrepancy g with the prior (see the text for details)).

$\mathbf{b}_k$ represent the beginning and the end, respectively, of the segment $\mathbf{s}_k$ ($\forall k \in \{1, \cdots, \mathcal{S}\}$). The position of a border is then defined as $p(\mathbf{b}_k) = \sum_{x=0}^{k} \delta(x > 0) \cdot l(\mathbf{s}_x)$, with $k \in \{0, 1, \cdots, \mathcal{S}\}$ and $\delta(.)$ being the Kronecker function. The modality $m(\mathbf{b}_k)$ of the border $\mathbf{b}_k$ defines its type of transition (foreground to background or background to foreground), *i.e.*, $m(\mathbf{b}_k) = v(\mathbf{s}_k)$ if $k \in \{1, 2, \cdots, \mathcal{S}\}$, and $m(\mathbf{b}_0) = 0$.

To determine how the 2D object silhouette, which is represented in each EPI by a set of epipolar borders, evolves when changing the viewpoint, we proceed in two steps:

1) We identify and match the reference epipolar borders (blue borders in Figure 3(a)) that have a corresponding border in the other reference view. This is done by introducing an original cost-function to drive the matching process in a way that is consistent with the available silhouette priors (see Section V-A).
2) We then approximate the vanishing trajectories of all the unmatched borders in a way that is consistent with the prior information (see Section V-B).

These two steps are described in detail below.

A. Matching epipolar borders

For a given EPI, let $\mathbf{B}^0$ and $\mathbf{B}^1$ denote the two sequences of reference epipolar borders that delimit the epipolar segments of the left and right reference silhouettes, respectively. Thus, as illustrated on Figure 3(b), $\mathbf{b}_i^0$ refers to the i^{th} epipolar border in the first reference view (starting at index 0). Similarly, $\mathbf{b}_j^1$ is the j^{th} epipolar border in the second reference view.

We match pairs of borders with the standard algorithm of Needleman and Wunsch [48], but have to adapt its underlying cost functions to account for our problem specificities.

The Needleman and Wunsch (NW) algorithm, which was initially introduced by Levenstein[49] to measure the minimal number of edits (*i.e.*, insertions, deletions or substitutions) between two strings, has been used extensively to compare sequences of characters [48]. Given an alphabet of characters $\mathcal{C}$ and a measure of dissimilarity $d(.,.)$ between any pair of characters in $\mathcal{C}$, the NW algorithm aligns two sequences of characters in a way that (1) preserves the order of the characters within each sequence [50], (2) matches the most similar characters together by minimising the sum of dissimilarities between them, and (3) tolerates unmatched characters at the cost of some skipping penalty $u(.)$. Its optimisation scheme, which determines the associations and unmatched characters based on the matching cost $d(.,.)$ and skipping cost $u(.)$, is described in the supplementary material (http://infoscience.epfl.ch/record/200492).

We now define the borders' matching and skipping costs ($d(.,.)$ and $u(.)$, respectively), so as to take advantage of the available intermediate prior silhouettes. In particular, we want to ensure that:

- long segments are less likely to vanish than shorter ones. In other words, borders that delimit long reference epipolar segments have less chance of being unmatched. Therefore, the skipping cost $u(\mathbf{b}_k)$ of the reference border $\mathbf{b}_k$ is defined to be equal to $\max\left(l(\mathbf{s}_k), l(\mathbf{s}_{k+1})\right)$;
- reference borders are unmatched by pairs of consecutive borders, so that their skipping can be interpreted as a vanishing/appearing segment. Since, by definition, a border separates two segments having different foreground/background values, the modes of consecutive borders are different. Skipping borders by pairs is thus equivalent to constraining each border to match borders having the same modality only. Hence, the distance between two borders with different modalities in two camera views should be set to ∞.
- the matching of reference borders between the two reference views shall be consistent with the prior that is available about the plausible deformation of the silhouette between the two views. The rest of this section explains how this is achieved through proper definition of the distance metric $d(.,.)$ between borders of the same modality.

Let us recall that the silhouette priors are represented by a sequence of $P+1$ foreground images, in which the p^{th} image, with $p \in [0; P]$, describes *a priori* the silhouette of the object as observed at a relative intermediate position $\alpha_p = \frac{p}{P}$ between the left and the right reference views. Those $P+1$ silhouette priors thus represent *a priori* a linear sampling of the continuous smooth transformation of the silhouette from the left to the right reference views. As illustrated in Figure 3(a) and (b), they provide, for a given epipolar line, a set of intermediate sequences of segments $\{\mathbf{S}^{\alpha_0}, \cdots, \mathbf{S}^{\alpha_p}, \cdots, \mathbf{S}^{\alpha_P}\}$ and their associated sequences of borders $\{\mathbf{B}^{\alpha_0}, \cdots, \mathbf{B}^{\alpha_p}, \cdots, \mathbf{B}^{\alpha_P}\}$. We define the cost of matching a border in $\mathbf{B}^0$ with a border in $\mathbf{B}^1$ by measuring

how this matching is in line with the prior sequences $\mathbf{B}^{\alpha_p}$ (with $p \in [0; P]$).

To account for the fact that the alignment of the prior silhouettes in the EPIV is prone to a translation error (as discussed in Section IV-D), we decompose the cost of matching the i^{th} border of $\mathbf{B}^0$ with the j^{th} border of $\mathbf{B}^1$, *i.e.*, $d(\mathbf{b}_i^0, \mathbf{b}_j^1)$ into two metrics. The first metric measures the quality of the alignment, in each reference view, between the prior and the reference borders. It is defined to be independent of a global and rigid translation of the prior. The second metric estimates how well the association of two prior borders that are extracted from the left and right viewpoints (corresponding to α_0 and α_P, respectively), is supported by the intermediate prior borders ($0 < \alpha_p < 1$).

Specifically, the **first metric**, illustrated in Figure 3(c), quantifies the likelihood of matching each reference epipolar border of $\mathbf{B}^0$ (respectively $\mathbf{B}^1$) with each of the prior borders of $\mathbf{B}^{\alpha_0}$ (respectively $\mathbf{B}^{\alpha_P}$) observed from a reference viewpoint. To define the associativeness $f(\mathbf{b}_i^0, \mathbf{b}_k^{\alpha_0})$ between the i^{th} reference border of $\mathbf{B}^0$, *i.e.*, $\mathbf{b}_i^0$, and the k^{th} border of $\mathbf{B}^{\alpha_0}$, *i.e.*, $\mathbf{b}_k^{\alpha_0}$, we rely on the fact that two borders are likely to be in correspondence when they have similar neighbourhoods. Because $\mathbf{S}^{\alpha_0}$ and $\mathbf{S}^0$ are seen from the same camera viewpoint, the foreshortening effect does not influence the lengths of their epipolar segments. This cost can be measured by the complementary of the normalised Hamming correlation (detailed in the supplementary material), *i.e.*, the number of positions at which the reference and prior sequences have identical values when they are aligned on the borders of interest. We highlight the fact that this metric is invariant to a rigid translation and is thus adapted to consider the translation error-prone prior. The metric $f(\mathbf{b}_l^{\alpha_P}, \mathbf{b}_j^1)$ to match the l^{th} prior border in $\mathbf{B}^{\alpha_P}$ with the j^{th} reference border in $\mathbf{B}^1$, observed in the other reference view, is defined analogously.

The **second metric** evaluates the cost of associating a border of the first prior $\mathbf{B}^{\alpha_0}$ with a border of the last prior $\mathbf{B}^{\alpha_P}$, as illustrated in Figure 3(d). We assume a linear displacement between two corresponding borders, meaning that the displacement is proportional to α. With rectified cameras, the linearity is strictly verified when the silhouette borders correspond to the same physical 3D point [39], independently of the viewpoint. In other cases, since the actual 3D point supporting the silhouette border generally does not move a lot when the viewpoint changes, the linearity assumption is also reasonably valid. Hence, we evaluate the discrepancy between a linear displacement and the actual transformations given by the priors $\mathbf{B}^{\alpha_p}$ (with $p \in \{1, \cdots, P-1\}$). Formally, we define the prior deformation cost $g(\mathbf{b}_k^{\alpha_0}, \mathbf{b}_l^{\alpha_P})$ of matching the k^{th} border of $\mathbf{B}^{\alpha_0}$ with the l^{th} border of $\mathbf{B}^{\alpha_P}$ to be the sum of the $\mathcal{L}_1$ interpolation residues, *i.e.*, the distance between the linear interpolation of $\mathbf{b}_k^{\alpha_0}$ and $\mathbf{b}_l^{\alpha_P}$ in the intermediate views α_p and the closest prior borders having the same modality in $\mathbf{B}^{\alpha_p}$ (with $p \in \{1, \cdots, P-1\}$. This is illustrated with green colour codes in Figure 3(d). The formal derivation of the prior deformation cost $g(\mathbf{b}_k^{\alpha_0}, \mathbf{b}_l^{\alpha_P})$ is given in supplementary material.

Finally, the cost $d(\mathbf{b}_i^0, \mathbf{b}_j^1)$ of matching the i^{th} border of $\mathbf{B}^0$ with the j^{th} border of $\mathbf{B}^1$ is defined as:

$$d(\mathbf{b}_i^0, \mathbf{b}_j^1) = \min_{k,l} \left(f(\mathbf{b}_i^0, \mathbf{b}_k^{\alpha_0}) + g(\mathbf{b}_k^{\alpha_0}, \mathbf{b}_l^{\alpha_P}) + f(\mathbf{b}_l^{\alpha_P}, \mathbf{b}_j^1) \right) \tag{1}$$

where the minimum is determined by the Dijkstra's algorithm [47]. By definition, a small $d(\mathbf{b}_i^0, \mathbf{b}_j^1)$ reflects the existence of a prior border that moves smoothly while going from one extreme prior view to the other (*i.e.*, small $g(\mathbf{b}_k^{\alpha_0}, \mathbf{b}_l^{\alpha_P})$), and good coherence between the prior and the actual reference borders in each reference view (*i.e.*, small $f(\mathbf{b}_i^0, \mathbf{b}_k^{\alpha_0})$ and $f(\mathbf{b}_l^{\alpha_P}, \mathbf{b}_j^1)$ values). Thereby, a small $d(\mathbf{b}_i^0, \mathbf{b}_j^1)$ promotes the matching of the borders $\mathbf{b}_i^0$ and $\mathbf{b}_j^1$.

Using $d(\mathbf{b}_i^0, \mathbf{b}_j^1)$ and $u(\mathbf{b}_i^0)$, the NW algorithm determines the optimal border associations and identifies (pairs of) unmatched borders.

B. Appearing/vanishing trajectories

We now present an original method to handle vanishing trajectories of unmatched borders. This is equivalent to analysing how occluded parts vanish or appear when changing the viewpoint, which is rather unique in the related literature. An occluded epipolar segment is defined by two consecutive occluded epipolar borders, and we know from border matching in which reference view the occluded epipolar segment is visible. In the following description, we consider the vanishing when moving from this view to the other and assume that the learnt latent space embeds an instance of vanishment of this occluded part. As illustrated in Figure 4, our method estimates from the prior the speed at which each occluded segment shrinks (vanishes) when changing the viewpoint. Since the borders' displacements along the EPI are assumed to be proportional to α [39], we have only to evaluate the two constant border displacement speeds and propagate this prior information to the occluded reference segments. This is done as follows:

1) identifying the prior borders that correspond to a segment that is subject to occlusion. We name them occluded prior borders (dark red borders in Figure 4(a)).
2) fitting linear trajectories to these prior borders (Figure 4(b)).
3) assigning the slopes (vanishing speeds) of those linear trajectories to the corresponding occluded reference borders (Figure 4(c)).

We present each of these steps in detail in the following paragraphs.

1) Identification of prior borders defining the occluded prior segments: Obviously, only the prior borders that do not support one of the associations/matchings of reference borders computed by the algorithm presented in Section V-A should be considered to explain the vanishing of occluded segments. Hence, we first select as *occluded prior borders* the prior borders that are sufficiently far from the linear trajectories followed between the pairs of associated reference borders or, more specifically, between their corresponding priors $\mathbf{B}^{\alpha_0}$ and $\mathbf{B}^{\alpha_P}$ at the reference viewpoints. In our experiments, we used a simple heuristic threshold, set to 5% of the image width, to decide whether a prior border was sufficiently far from the linearly interpolated trajectories. This may, however, lead to

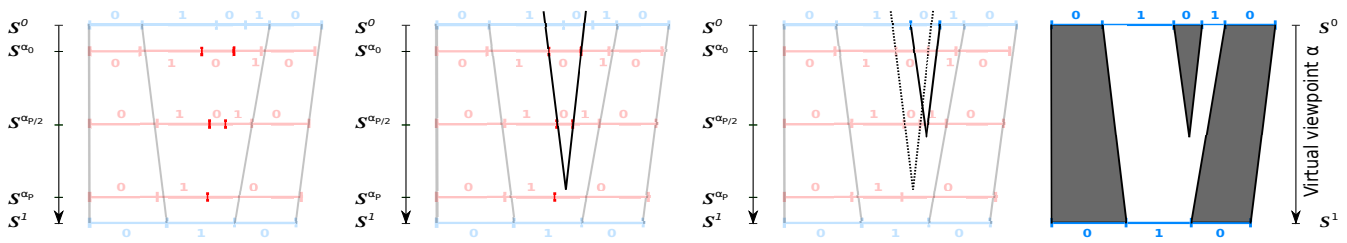


Fig. 4: The vanishing trajectories are estimated by (a) identifying the occluded prior borders (in dark red), (b) fitting linear trajectories to these prior borders, and (c) associating the slopes of these trajectories (dotted lines) with the occluded reference borders (solid lines). Finally, the EPI 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.

many false positives. Hence, the following section proposes a robust way to estimate the vanishing/appearing paths from this initial set of occluded prior borders.

2) Robust fitting of linear trajectories: This section shows how to determine the linear trajectories of the l occluded reference borders from an imperfect set of prior occluded borders. Specifically, the set of l occluded reference borders can be divided into l_0 occluded reference borders representing a transition from foreground to background (*i.e.*, having a mode value of 0) and l_1 borders representing a transition from background to foreground, such that $l_0 + l_1 = l$. Hence, we propose to divide the set of prior occluded borders into two sets, based on their modes. Then, l_0 linear trajectories (respectively l_1 linear trajectories) are estimated on the subset of occluded prior borders having a mode of 0 (respectively 1). This is done by applying the RANSAC algorithm [51] l_0 times (respectively l_1 times) sequentially, *i.e.*, by estimating a linear trajectory on the subset of occluded prior borders of mode 0 (respectively 1), removing the prior borders that are inlier to this estimated model, estimating a new linear trajectory on this new subset, and so on.

At each RANSAC iteration, two borders are randomly selected from the set of occluded prior borders and the linear trajectory passing through these borders is estimated. All the prior borders located in a small and conservative $\mathcal{L}_1$ distance (*e.g.*, 5% of the width of the image) are considered inliers to the trajectory model. This simple greedy algorithm works well in practice, due to the relatively small number of outliers in the set of occluded prior borders. The linear model that maximises the amount of inliers is considered to be the optimal model of the l_i^{th} sequential application of RANSAC.

3) Assignment of linear trajectories to the reference occluded borders: We want to assign the trajectories computed from the prior occluded borders to the unmatched borders in the reference views so as to transfer their slopes, *i.e.*, the constant speeds at which the borders move along the EPI when the viewpoint index α changes. The process is illustrated in Figure 4(c). The cost of assigning a prior 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 prior in the reference view (compensated by a linear interpolation of the translation error indicated by the matches of the NW algorithm). 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 minimising the sum of assignment costs. Finally, as illustrated in Figure 4(d), these vanishing trajectories are added to the set of trajectories describing the transitions between the associated reference borders to form the EPI of the object silhouette.

VI. View Synthesis

This section describes how the intermediate view epipolar lines are synthesised based on the matching of reference silhouette epipolar borders.

For occluded epipolar segments, we simply propose to propagate the texture from the reference view, in which the segment is visible, to the vanishing point, using a linear morphing strategy, as described in Section VI-A.

For pairs of matched silhouette segments, we synthesise the intermediate segment by combining the textures of matched epipolar segments. Texturing an intermediate view by combining the textures of its corresponding elements in reference views has been thoroughly investigated [28] [53]. By favouring the piecewise smoothness of the intermediate texture, most of these previous methods make it possible to generate pleasant intermediate views despite corrupted matches. In our case, to validate our contribution, *i.e.*, the estimation of the geometric transformations of the epipolar line segments, fairly, we primarily rely on simple linear view morphing [28], so that any wrong association results in highly noticeable discontinuities in reconstructed images. This method is described in Section VI-B. To complete our validation and to show that our method can benefit from more accurate solutions to combine the textures of matched segments, Section VII also presents the result obtained based on a non-linear pixel-wise association betwen the matched segments when assuming the strict preservation of the order of pixels associated between the pairs of matched epipolar segments. This effective way of combining the textures is presented in Section VI-C.

A. Texture occlusion

The texture of an occluded segment is propagated to intermediate views based on the linear interpolation of pixels' positions between the reference view (where the occluded

segment is visible) and the segment's vanishing point, as determined by the intersection (x_v, α_v) of the vanishing trajectories surrounding this occluded segment. Formally, this synthesis is thus defined as follows:

$$\mathbf{I}_\alpha(\frac{\alpha}{\alpha_v} \cdot (x_v - x_1) + x_0, y) = \mathbf{I}_0(x_0, y)$$

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

$$\mathbf{I}_\alpha(\frac{(\alpha - 1) \cdot (x_v - x_1)}{\alpha_v - 1} + x_1, y) = \mathbf{I}_1(x_1, y)$$

if the vanishing segment belong to the right reference image (disocclusion).

B. View morphing

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

$$\mathbf{I}_\alpha(x_\alpha, y) = (1 - \alpha) \cdot \mathbf{I}_0(x_0, y) + \alpha \cdot \mathbf{I}_1(x_1, y)$$

with $\mathbf{I}_0$ and $\mathbf{I}_1$ the rectified reference images, $\mathbf{I}_\alpha$ the reconstructed intermediate image, x_0 the x coordinate of a pixel of $\mathbf{I}_0$, and y its fixed ordinate (studied scanline). The pixel abscissa values x_α and x_1 are computed as follow:

$$\begin{aligned} x_\alpha &= (1 - \alpha) \cdot x_0 + \alpha \cdot x_1 \\ x_1 &= \frac{l(\mathbf{s}_j^1)}{l(\mathbf{s}_i^0)} \cdot \big(x_0 - p(\mathbf{b}_i^0)\big) + p(\mathbf{b}_j^1) \end{aligned} \quad (2)$$

with $\mathbf{s}_i^0$ denoting the epipolar line segment including x_0 and $\mathbf{s}_j^1$ denoting the epipolar line segment matched to $\mathbf{s}_i^0$. Then $p(\mathbf{b}_i^0)$ is the position of the left border of the epipolar line segment including x_0 and $p(\mathbf{b}_j^1)$ 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 observers point of view. It is not strictly valid, especially in a wide baseline stereo setup, but it is usually valid within pairs of matched 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 [33]. Once the pixels have been associated, the interpolation of an intermediate segment is straightforward, and reads:

$$\mathbf{I}_\alpha(x_\alpha, y) = (1 - \alpha) \cdot \mathbf{I}_0(x_0, y) + \alpha \cdot \mathbf{I}_1(x_1, y)$$

with $\mathbf{I}_0$ and $\mathbf{I}_1$ the rectified reference images, $\mathbf{I}_\alpha$ the reconstructed intermediate image, x_0 and x_1 the x coordinates of a pair of associated pixels between two matched epipolar segments, and y its fixed ordinate (studied scanline). The reconstructed pixel abscissa x_α is $x_\alpha = (1 - \alpha) \cdot x_0 + \alpha \cdot x_1$.

VII. Results

In this section, we demonstrate the performance of our approach on well-known datasets, namely the synthetic *Kung-Fu Girl* sequence [54] and the real *Ballet* sequence [36].

Due to this article's page limit, other validations on the multi-view *Dino* dataset [55] are given in the supplementary material. Although these multi-view 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 particular, as explained in Section IV, the silhouettes used for training are captured by one of the reference cameras before the time at which the intermediate views are generated. 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 in the supplementary material (http://infoscience.epfl.ch/record/200492).

To show the advantage of using epipolar line segments as basis matching elements, we provide reconstructed views when pixels are chosen as basis matching elements. To demonstrate the benefit of the silhouette priors, we also provide the views that have been reconstructed without silhouette priors to disambiguate the epipolar segments' matching. We also compare the reconstructed intermediate views resulting from our method with the ones obtained by three other conventional and state-of-the-art methods.

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 $45°$. The view captured by the left (or right) camera) is shown on the left (respectively right) of the first row in Figure 5. The image shown in-between corresponds to the ground-truth one captured by a camera situated approximatively at the middle ($\alpha \simeq 0.5$) in-between these two reference views.

The second row in Figure 5 represents the intermediate views generated by a conventional visual-hull reconstruction [18] in which the two foreground silhouettes are projected back in the 3D world, forming two cones whose intersection defines the 3D boundary of the object. The intermediate views are obtained by projecting and texturing this 3D model onto an arbitrary viewpoint [27]. The reconstructed intermediate views show perfectly the limitations of model-based approaches in our wide-baseline stereo, namely the requirement of observing the object with a large number of reference cameras to avoid an imprecise 3D model, leading to corrupted intermediate views.

The third row in Figure 5 represents the intermediate views generated when morphing [28] a dense (pixel) correspondence obtained by dynamic programming [33] [50] on corresponding epipolar lines. The matching cost is simply defined as the l_2-norm of the pixels' colours and the skipping penalty $u(.)$ is arbitrarily set to 0.5. Two kinds of artefact can be observed in

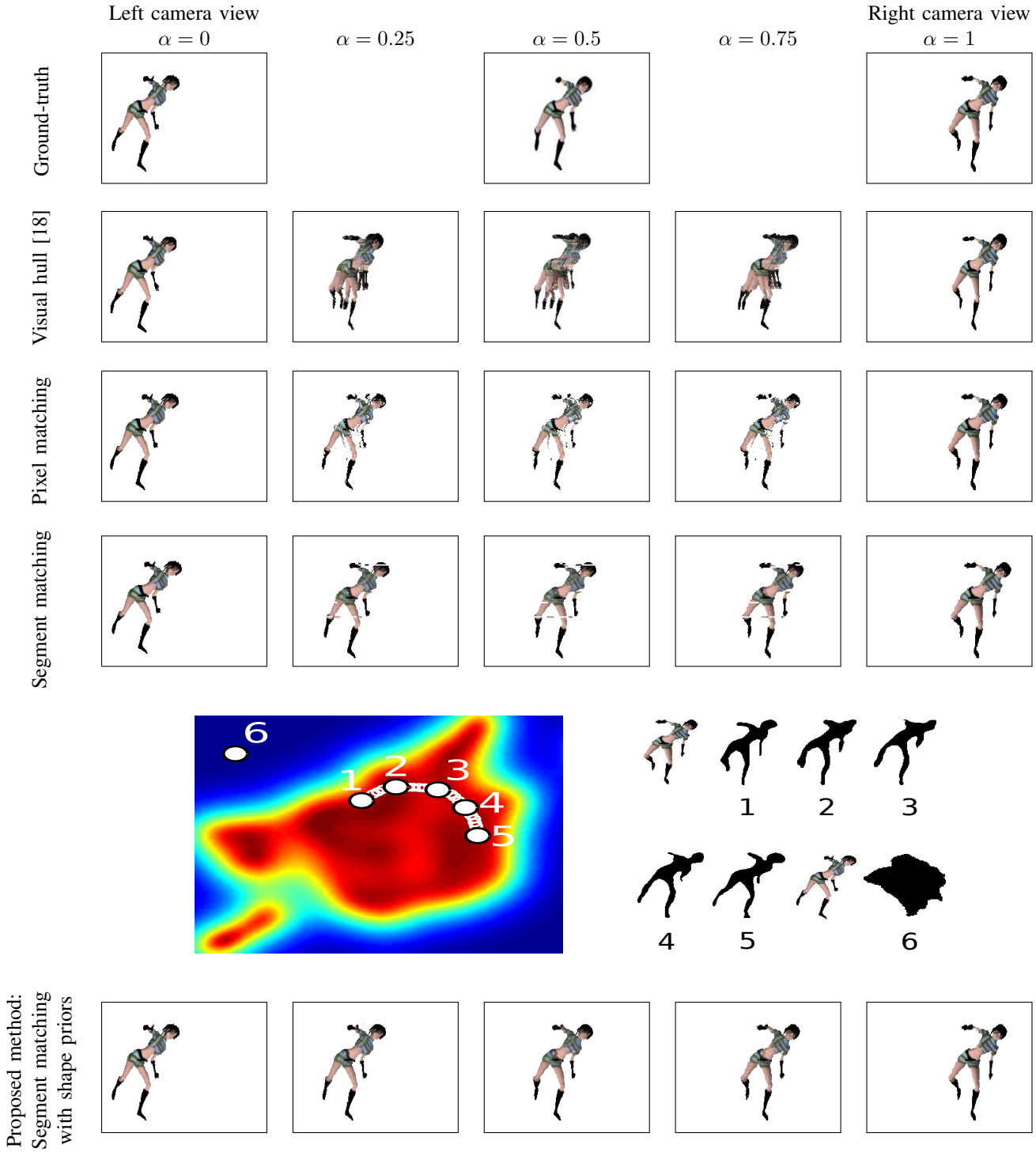


Fig. 5: Instead of projecting an estimated 3D model [18] (second row) or determining a dense (pixel) match (third row), epipolar line segments are used as basic matching elements (fourth row). In the last row, our method regularizes the epipolar segments' matching so that the shapes of the intermediate silhouettes are topologically consistent with the plausible deformations of the object silhouette, as learnt and described by a low-dimensional latent space (fifth row).

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 members separate from her body. This is due to the ill-posedness of the wide-baseline matching problem, leading to wrong pixel correspondences. Second, because the resulting matching is not smooth, holes appear in the reconstructed intermediate views. This artefact, caused by the foreshortening effect, is generally avoided by imposing a smooth disparity/depth-map [25], although this slows the matching process.

To impose smoothness along the epipolar lines explicitly, as shown in the 4^{th} row of Figure 5, we use epipolar line segments as matching elements. The method extends that of [33] by considering epipolar line segments (and not pixels) to be basic image elements. It corresponds to the approach we introduced in Section V-A, but without prior silhouettes knowledge. Hence, each epipolar border is matched by NW [50], only considering the $f(.,.)$ terms in the definition of $d(\mathbf{b}_i^0, \mathbf{b}_j^1)$ (see Equation 1). We observe in Figure 5 that the reconstructed intermediate views are smoother but still exhibit some topologically incoherent transitions (see head and legs). To regularize the reconstruction of the EPIs in such a way that they provide topologically coherent intermediate views, a latent space representing plausible silhouettes of the Kung-Fu girl has been learnt on a total of sixty silhouettes captured by the two wide-baseline cameras, and observed uniformly in a time-window starting from the first frame of the sequence to twenty frames before the required transition. These silhouettes have been described using seventy elliptic harmonics. The fifth row in Figure 5 illustrates this latent space.

The advantage of considering these priors is illustrated in the last row in Figure 5, where intermediate views have been generated by the method proposed in this paper, using linear view morphing for texture synthesis (Section VI-B). Only five intermediate priors have been used to reconstruct the EPIs. We observe that the intermediate views reconstructed by our method show a topogically coherent transition of the Kung-Fu girl from the left to the right reference view.

B. The Ballet sequence

The second sequence, called *Ballet* [36], has been captured using eight cameras placed along a $1D$ arc spanning about 30° end-to-end. While two neighbouring cameras in this array constitute a small-baseline stereo pair, the outer cameras represent a wide-baseline configuration. Indeed, because of the small depth of the foreground dancer, strong self-occlusions and foreshortening effects can be observed between these two external viewpoints (especially on the dancer's arms). In Figure 6, we compare the images reconstructed at intermediate viewpoints by five methods, using only the two wide-baseline reference views (in contrast to the use of the small-baseline multi-views pairs, as done in [36]). In the first row, the intermediate views are generated by view morphing, based on multiple depth maps as proposed in [36][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 neighbouring cameras (small-baseline configuration). Even with this additional information, the small depth inaccuracies (equivalently weak pixel correspondences) lead to merging of non-corresponding textures, *i.e.*, ghosting artefacts. The second row in Figure 6 illustrates the intermediate views obtained by a state-of-the-art stereo method [56] top-ranked in February 2015 in the well-known Middlebury Stereo Evaluation [57] [58] [50]. By combining a cost-filtering approach, especially adapted to manage the occlusions, with a global (fully connected Markov Random Field) optimisation, which imposes the smoothness of the disparity map, their method achieves impressive results on small-baseline stereo setups. However, as expected, the strong geometrical and photometric changes, as well as the foreshortening effects affecting our wide-baseline stereo setup make this algorithm pretty vulnerable, especially due to the oversmoothing of the disparity map. In the third row, we use only the external views and test wide-baseline stereo matching by applying the Needleman-Wunsch algorithm (dynamic programming [50]) to pixels, as done in [33]. We observe that the strong foreshortening effect produces holes in the reconstructed intermediate views. By applying dynamic programming to the segment representation, we find dense correspondences, but topological inconsistencies subsist (see fourth row on Figure 6). The last row in Figure 6 illustrates the result obtained by our complete method using silhouette priors and morphing-based texture interpolation (as described in Section VI-B). The latent space has been learnt on forty silhouettes observed by the two outermost cameras in a time-window twenty frames away from the transition time. These silhouettes have been described with fifty harmonics of Elliptic Fourier Descriptors, and six intermediate priors are used to drive the epipolar segment matching.

C. Discussion

Our method appears to generate topologically coherent intermediate views, thanks to the additional silhouette prior obtained from the latent space. It also deals efficiently with the foreshortening effect, as can for example be seen in Figure 6 on the front part of 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 (Figure 5) and the space between the dancer's legs (Figure 6). Next to these very encouraging results, two limitations of our approach, however, deserve to be mentioned. The first one can be observed on the reconstructed fingers of the dancer's right hand in Figure 6. Indeed, when the viewpoint changes from the left to the right reference view, her fingers separate from her right hand, 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 can not be

[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.

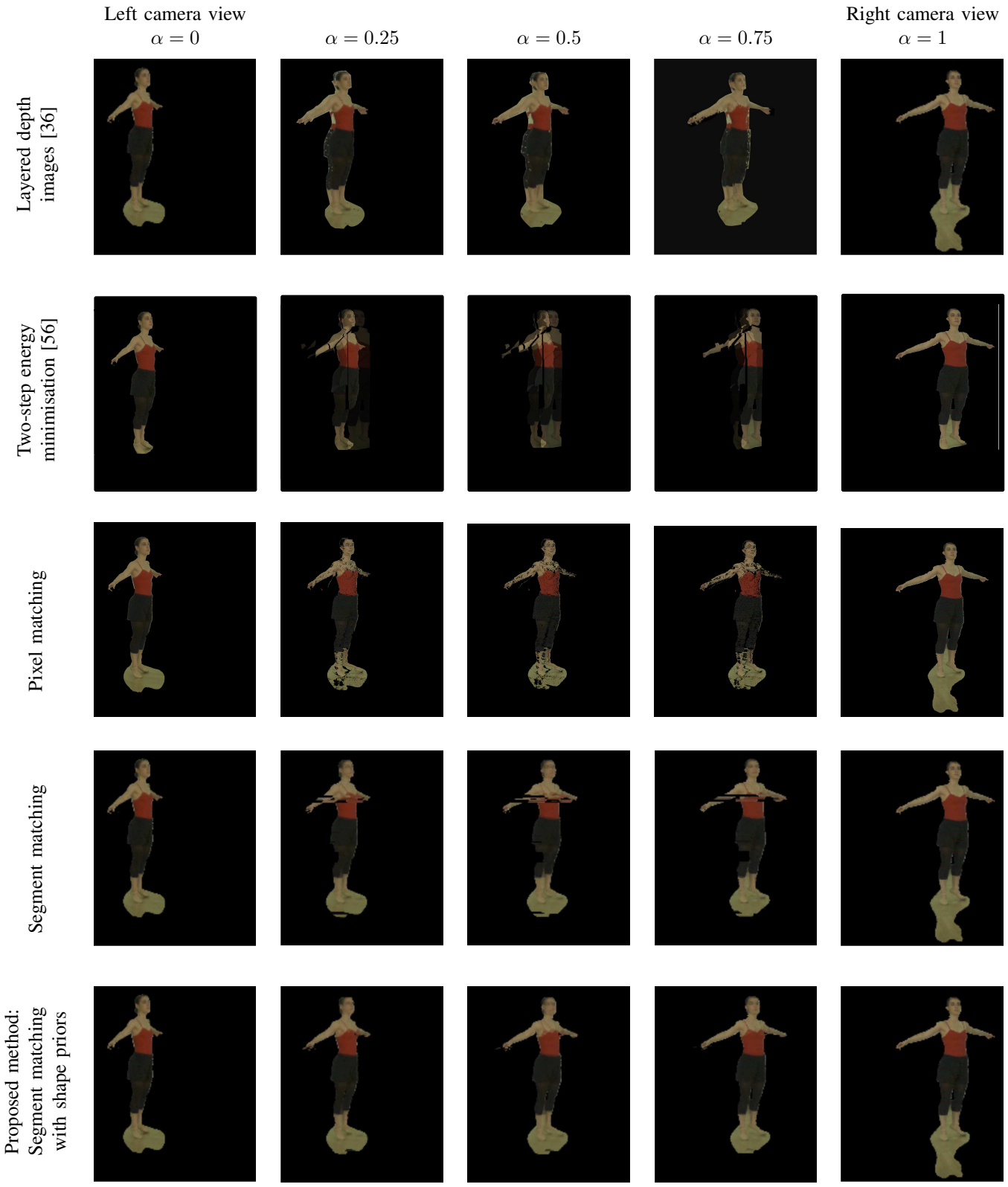


Fig. 6: Comparison between the interpolated intermediate views generated based on matching of layered representation [36] (first row), which exploits intermediate depth maps in addition to external views; a state-of-the-art method [56] that is top-ranked in the Middlebury Stereo Evaluation [57] (second row); dynamic programming on pixels [33] (3^{rd} row); dynamic programming on our proposed epipolar line segment representation (4^{th} row); and our method (5^{th} row).

regularized, and their transition may become topologically incoherent.

The second weakness of our method comes from the choice of a morphing-based inner texture combination method. Deriving the correspondences between inner pixels directly from the knowledge of the matches of the epipolar segments' borders, as defined by Equation (2), might result in wrong matches of the inner pixels if the surface described by this epipolar line segment is not planar. This limitation can be observed on the pixels representing the face of the dancer. Their wrong association results in a blurry ghosting artefact. To convince the reader that a more accurate inner texture interpolation would mitigate this issue, Figure 7 compares the images reconstructed using the morphing-based combination of inner texture with the ones obtained by the approach presented in Section VI-C, which strictly forces ordering preservation while associating the inner pixels of two matched epipolar segments, using the dynamic programming approach initially described for whole epipolar lines in [33]. We observe that, indeed, finer association of inner pixels significantly improves the sharpness of edges in the reconstructed image. Moreover, when considering the artefacts appearing in the third row of Figure 6, we conclude that our proposed matching of silhouette epipolar segments provides a worthwhile step, preventing the failure of conventional pixel-based matching techniques. Alternative inner texture interpolation could obviously be envisioned, *e.g.*, based on the floating textures concept [53].

Finally, 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 performances: 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. Moreover, because the epipolar lines are processed independently, real-time implementation is within reach, *e.g.*, based on GPU parallelisation.

VIII. Conclusions

This paper has proposed a novel interpolation technique for intermediate view synthesis between cameras in wide-baseline configurations. Our method relies on prior information about the silhouettes of objects in the intermediate views to guarantee consistency between the synthesised silhouettes and the ones present in the two reference viewpoints. As a first contribution, we propose to learn this prior from earlier observations of the object moving in front of one of the cameras in the reference stereo pair. This is done by reducing the dimensionality of Elliptic Fourier Shape Descriptors. The learnt 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. As a second contribution, we demonstrate that this new framework not only has the advantage of generating consistent and smooth virtual transitions between the viewpoints where correspondences can be found in the two basis images, but 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 the occlusions and severe foreshortening effect that are inherent to wide baselines.

References

[1] M. Levoy and P. Hanrahan, "Light field rendering," in ACM SIGGRAPH, 1996, pp. 31–42.

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

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

[4] M. Reynolds, J. Dobos, L. Peel, T. Weyrich, and G. J. Brostow, "Capturing time-of-flight data with confidence," in IEEE CVPR, 2011.

[5] 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 camera," in Proc. of the ACM User interface software and technology, 2011.

[6] C. Verleysen and C. De Vleeschouwer, "Piecewise-planar 3d approximation from wide-baseline stereo," in IEEE CVPR.

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

[8] L. Ballan, G. J. Brostow, J. Puwein, and M. Pollefeys, "Unstructured video-based rendering: interactive exploration of casually captured videos," ACM Trans. on Graphics, vol. 29, no. 4, pp. 1–11, July 2010.

[9] I. Kitahara and Y. Ohta, "Scalable 3d representation for 3d video display in a large-scale space," in IEEE VR, 2003, pp. 45–52.

[10] 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.

[11] C. D. Castillo and D. W. Jacobs, "Wide-baseline stereo for face recognition with large pose variation," in IEEE CVPR, 2011.

[12] L. Sun, Q. De Neyer, and C. De Vleeschouwer, "Multimode spatiotemporal background modeling for complex scenes," in EUSIPCO, 2012.

[13] 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 Proc. of Computer Graphics Forum, vol. 31. Wiley Online Library, 2012, pp. 325–333.

[14] 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 IEEE VISAPP, 2013.

[15] H. Saito and T. Kanade, "Shape reconstruction in projective grid space from large number of images," in IEEE CVPR, vol. 2, 1999.

[16] S. Yaguchi and H. Saito, "Arbitrary viewpoint video synthesis from multiple uncalibrated cameras," IEEE Trans. on Systems, Man, and Cybernetics, vol. 34, no. 1, pp. 430–439, 2004.

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

[18] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, "Image-based visual hulls," in ACM SIGGRAPH, 2000, pp. 369–374.

[19] T. Matsuyama and T. Takai, "Generation, visualization, and editing of 3d video," in IEEE 3DDPVT, 2002, pp. 234–245.

[20] G. Slabaugh, R. Schafer, and M. Hans, "Image-based photo hulls," in IEEE 3DDPVT, 2002, pp. 704–862.

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

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

[23] 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 IEEE ICME, 2006, pp. 2161–2164.

[24] B. Goldlucke and M. Magnor, "Real-time microfacet billboarding for free-viewpoint video rendering," in IEEE ICIP, vol. 3, 2003, p. 713.

[25] H.-Y. Shum, S.-C. Chan, and S. B. Kang, Image-based rendering. Springer Science and Business Media, 2007.

[26] A. W. Fitzgibbon, Y. Wexler, A. Zisserman et al., "Image-based rendering using image-based priors," in IEEE ICCV, vol. 3, 2003.

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

[28] S. M. Seitz and C. R. Dyer, "View morphing," in ACM SIGGRAPH, 1996, pp. 21–30.

[29] 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.

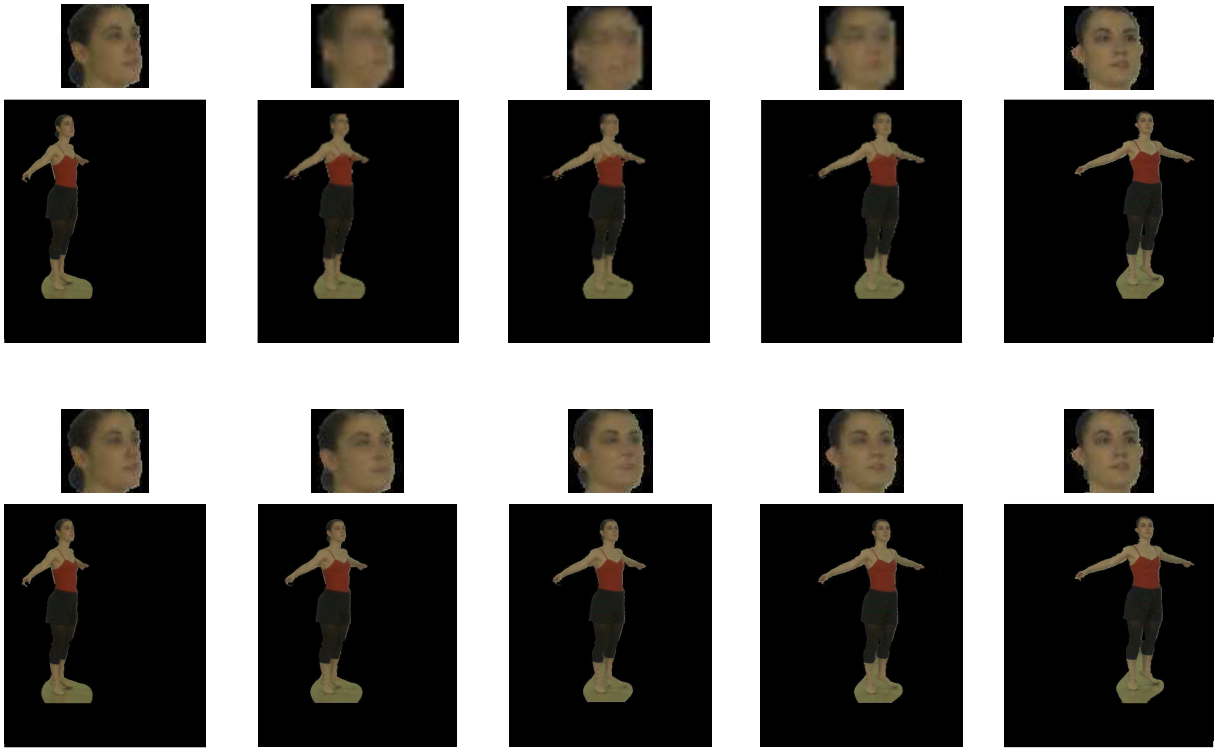


Fig. 7: Comparison between the interpolated intermediate views generated based on our proposed epipolar silhouette segment matching, when using view morphing (top, Section VI-B) or dynamic programming on pixels [33] (bottom, Section VI-C) to reconstruct the inner texture of each segment. Zoomed-in versions of the face area are provided for easier comparison.

[30] O. Veksler, "Fast variable window for stereo correspondence using integral images," in IEEE CVPR, vol. 1, 2003, pp. I–556.

[31] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," IEEE PAMI, vol. 23, no. 11, 2001.

[32] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," IJCV, vol. 70, no. 1, pp. 41–54, 2006.

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

[34] S. Mattoccia, "Accurate dense stereo by constraining local consistency on superpixels," in IEEE ICPR, 2010, pp. 1832–1835.

[35] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in IEEE ICPR, vol. 3, 2006, pp. 15–18.

[36] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in ACM Trans. on Graphics, vol. 23, 2004, pp. 600–608.

[37] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz, "Patchmatch belief propagation for correspondence field estimation," IJCV, pp. 1–12, 2012.

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

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

[40] R. Hartley and A. Zisserman, Multiple view geometry in computer vision. Cambridge University Press, 2000.

[41] V. A. Prisacariu and I. Reid, "Nonlinear shape manifolds as shape priors in level set segmentation and tracking," in IEEE CVPR, 2011.

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

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

[44] N. D. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," NIPS, vol. 16, no. 3, 2004.

[45] N. Lawrence, "Probabilistic non-linear principal component analysis with gaussian process latent variable models," Journal of Machine Learning Research, vol. 6, p. 17831816, 2005.

[46] N. D. Lawrence and J. Quiñonero-Candela, "Local distance preservation in the gp-lvm through back constraints," in ACM ICML, 2006.

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

[48] 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.

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

[50] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," IJCV, vol. 47, no. 1-3, pp. 7–42, 2002.

[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. 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. Wiley Online Library, 2008, pp. 409–418.

[54] 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/

[55] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in IEEE CVPR, vol. 1, 2006, pp. 519–528.

[56] M. Mozerov and J. van de Weijer, "Accurate stereo matching by two-step energy minimization," IEEE TIP, vol. 24, no. 3, 2015.

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

[58] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in IEEE CVPR, 2007, pp. 1–8.