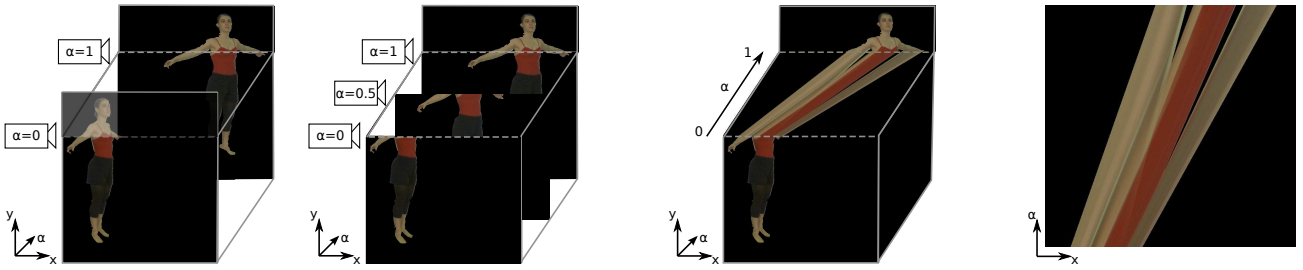
Wide-baseline object interpolation using shape prior regularization of epipolar plane images



(a) Reference views (b) Reference views + 1 intermediate view (c) Reference views + continuum of intermediate views (d) Epipolar Plane Image

Fig. 1: Two reference views (a) are available to generate intermediate views (b). For that purpose, our method estimates the Epipolar Plane Image Volume (c) made of the set of Epipolar Plane Images (d).

***Abstract*—This paper considers the synthesis of intermediate views of an object captured by two calibrated and widely spaced cameras. Based only on those two very different views, our paper proposes to reconstruct the object Epipolar Plane Image Volume [1] (EPIV), which describes the object transformation when continuously moving the viewpoint of the synthetic view in-between the two reference cameras. This problem is clearly ill-posed since the occlusions and the foreshortening effect make the reference views significantly different when the cameras are far apart. Our main contribution consists in disambiguating this ill-posed problem by constraining the interpolated views to be consistent with an object shape prior. This prior is learnt based on images captured by the two reference views, and consists in a nonlinear shape manifold representing the plausible silhouettes of the object described by Elliptic Fourier Descriptors. Experiments on both synthetic and natural images show that the proposed method preserves the topological structure of objects during the intermediate view synthesis, while dealing effectively with the self-occluded regions and with the severe foreshortening effect associated to wide-baseline camera configurations.**

***Index Terms*—Wide-baseline matching, image-based rendering, 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 general arbitrary view synthesis problem to the interpolation of images observed by a virtual camera located in an arbitrary position along the baseline connecting two reference cameras. The graceful transition between two reference viewpoints is a demanded feature, especially in the field of video production [2]. For example, in the rendering of cultural or 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 synthesize 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 interpolation of dynamic foreground object, situated relatively close to the pair of cameras, is the most complex question among both [4] [5], because the background can be easilly reconstructed through trivial projection of its 3D geometry [6]. Typically, the background 3D geometry can reasonnably be acquired, based on state-of-the-art active 3D acquisition systems [7] [8] if it is still, or based on piecewise planar 3D geometry approximations [9] [6] [10] when it is far from the cameras. The fundamental issues encountered to reconstruct a foreground object lie in (1) the availability of only two reference views, and (2) the object proximity to the cameras, compared to the distance between those cameras. The first factor prevents dense 3D estimation for the dynamic object, while the second causes many projective discrepancies between the two views (occlusions, foreshortening effects, etc.), which hamper the computation of dense correspondences and lead to holes in the interpolated views [5].

This paper focuses on the reconstruction of foreground objects and assumes that the object silhouette can be extracted from the reference views[1], as generally assumed by state-of-the-art foreground synthesis methods [12] [13] [6]. 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 (see Figure 1(b)) along the baseline by reconstructing the object's Epipolar Plane Image Volume [1] (see Figure 1(c)), composed on the set of Epipolar Plane Images (see Figure 1(d)).

The specificities of the proposed method lies in the regularization of the ill-posed reconstruction of the Epipolar Plane Images (EPIs) based on a sequence of plausible intermediate object silhouettes. As illustrated in Figure 4, this sequence is derived from a low-dimensional manifold, learnt from the

[1]In this paper, the foreground is generally extracted based on a simple thresholding of the $\mathcal{L}_2$ color distance with a gaussian mixture model of the background [11].

previous observations of the dynamic object by the wide-baseline stereo pair. Interestingly, the priors are used not only to disambiguate the matching, but also to determine how occluded parts vanish/appear while moving from one reference view to the other. To the best of our knowledge, our paper is the first one to reconstruct topologically consistent images from only two widely separated cameras, even for their occluded parts, while dealing effectively with the self-occluded regions and with the severe foreshortening effect associated to wide-baseline camera configurations.

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 constraint 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

The view synthesis techniques are generally categorized into two groups in the literature, namely model-based rendering and image-based rendering.

In model-based rendering, a 3D shape model of the observed scene is explicitly reconstructed from multi-view images. Adequate texture is then mapped on the model, and projected onto any arbitrary viewpoint. Methods such as projective grid space [14] [15], visual-hull [16] [17] [18] [19], 3D model adjustment [20], and shape from video [12] belong to this category. Those methods have the advantage to synthesize 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 [21]. To obtain an accurate 3D model, the model-based rendering methods therefore rely on a dense coverage of the scene, which requires a large number of precisely calibrated video cameras [22]. The trade-off between the accuracy of the reconstruction and the amount of cameras is often relaxed when the distance between the object and the reference cameras is important compared to the baseline distance separating these cameras [23]. In this particular case, a simple (set of) planar model(s) (called *billboards*) permits 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 [6]. This makes model-based rendering inappropriate to render close (dynamic) scenes between wide-baseline cameras.

In contrast, image-based rendering (IBR) methods [24] create the virtual view 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 [13], or the ones of the reference views [6]) and/or the virtual image's color [25]. However, the dense estimation of a depth map is only possible 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 [26] or morphing [27]. 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 with light-field reconstruction approaches [1], which require tens or hundreds of narrow-baseline[2] cameras/lenses [28] 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 to 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 reason for the failure of rendering methods in wide-baseline stereo setups is that the more different the viewpoints, the more important the geometrical deformations (including projective distortions and occlusions), and the more difficult it is to find correspondences between images from different cameras. More precisely, the three following issues are specific to wide-baseline configurations:

- *The foreshortening effect* causes a distance or an object to appear shorter/wider than it is because it is angled toward the viewer (see Figure 2). Because the compaction ratio depends on the viewpoints, a given 3D object will be represented by a totally different number of pixels in different views.

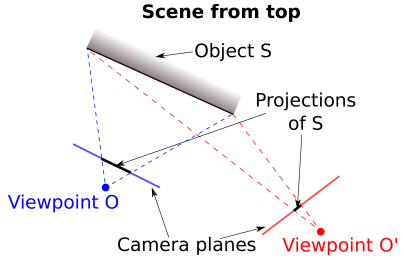


Fig. 2: Illustration of the foreshortening effect. The projection of the object S is more compact in O' than in O.

This implies that finding correspondences with fixed-template matching methods fails [29]. The same holds

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

when a pixel correspondence is optimized by graph-cut [30], belief propagation [31], or dynamic programming [32] 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 a part of an object hides another region of the same object. Parts of the object can thus be observed in only one of the camera views, so that no correspondence can be found with the other reference views. This problem drastically limits the correspondence-based interpolation methods [27] in a wide-baseline configuration.
- The lack of *sparse correspondences* and 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 [33] [34] [35], but they are either based 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 [36].

Our proposed method explicitly addresses those issues by computing correspondences between a continuous set of image segments (from which dense correspondences can be inferred, *e.g.*, through linear interpolation), and by constraining those correspondences to be consistent with a plausible deformation of the projected object silhouette between the reference views (guides the occlusion of segments, or their shrinkage/elongation due to the foreshortening effect).

III. Wide-baseline interpolation algorithm

This paper adopts an EPI interpolation formalism to reconstruct the image of a foreground object between two widely spaced cameras. As depicted in Figure 1, the transformations of images between different viewpoints can be described by the object Epipolar Plane Image Volume [37] [38] (EPIV) (see Figure 1(c)). By definition, an EPIV is obtained by arranging in a 3D stack the images captured by a dense array of cameras that are uniformly distributed along a line with their image plane coplanar and vertically aligned. This is performed through epipolar rectification [39] of the reference images, which associates each horizontal line in one image to a row with the same ordinate in the other image, as illustrated in Figure 3. Roughly speaking, it implies that two corresponding pixels must belong to the same horizontal plane in the EPIV, and that any transverse cross-section of this 3D cube, *i.e.*, an Epipolar Plane Image (EPI) (see Figure 1(d)), describe 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 that their slopes are inversionally proportional to the scene's depth. The EPIV is much richer than the depth information generally estimated by state-of-the-art wide-baseline stereo methods. Indeed, the EPIV additionally englobes the appearing/vanishment of occluded parts. However, its estimation has been limited so far to very narrow-baseline setups, which only permits to generate intermediate views in a very narrow range.

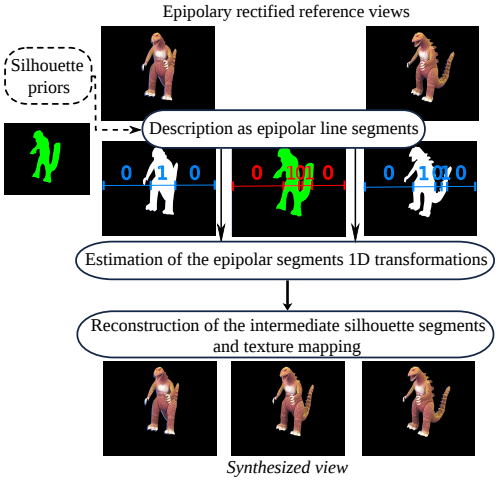


Fig. 3: Our view interpolation method overview: foreground object silhouette segments are matched between the epipolar lines of two reference views, based on the prior about plausible silhouettes in intermediate views.

We adopt a new object-based approach to reconstruct the EPIV. After epipolar rectification of the reference views, we:

1) Learn a low-dimensional silhouette manifold. It describes prior plausible transformations of the object silhouette when changing the viewpoint along the baseline (see Figure 4, left side).
2) Use a sequence of plausible silhouettes to define how the object silhouette epipolar line segments are transformed (*i.e.*, through scaling/translation/vanishing) between the two reference views (Figure 4, right side).
3) Interpolate object textures based on the transformations, vanishments or appearance of the silhouette epipolar line segments.

The different blocks of our novel view interpolation algorithm are despicted in Figure 3 and described in details in the next sections.

IV. Object silhouette priors

This section describes the construction of priors on the plausible deformations undergone by an object silhouette during a viewpoint change. It aims at providing a sequence of approximated object silhouettes that defines *a priori* a plausible transition from the left reference silhouette to the right one. This sequence is then used to regularize the reconstruction of the set of EPIs.

We propose to generate these silhouette priors in four steps:

1) Learning a low-dimensional space representing the plausible silhouettes of the object.
2) Locating, in this low-dimensional space, the silhouettes observed in the reference views.
3) Interpolating, in this low-dimensional space, a sequence of low-dimensional silhouettes that likely represent the deformation of the object silhouette in-between the reference views.
4) Converting these low-dimensional intermediate silhouette representations into high-dimensional images for view synthesis.

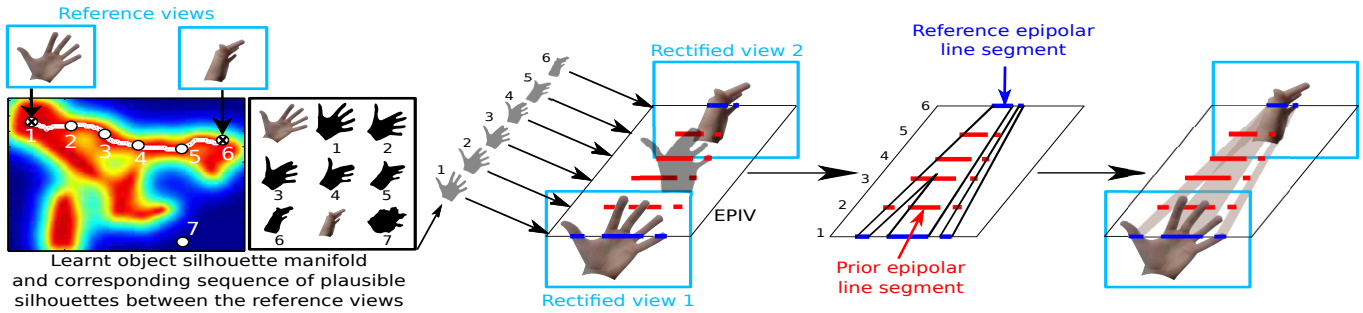


Fig. 4: We propose to regularize the ill-posed problem of reconstructing the set of EPIs by incorporating prior knowledge about the plausible deformations of the object silhouette. This prior knowledge is learnt beforehand and is described by a low-dimensional space, from which intermediate 2D prior silhouettes can be extracted in-between the projected reference ones (left part on the figure, each point of this manifold represents a silhouette, while the color scale refers to the confidence about its plausibility). These intermediate 2D priors are then adequately placed in the EPIV (middle part of the figure) and are converted into a set of 1D priors to disambiguate the reconstruction of the set of EPIs (right part).

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. We propose to follow the approach of [40], by splitting the low-dimensional manifold construction into two parts: the first part describes the shape of a silhouette as a set of high-dimensional features, and the second part maps those high-dimensional descriptors to a lower dimensional latent space. The different steps of the construction of the prior silhouettes are described in detail in the rest of this Section.

A. *High-dimensional silhouette description*

We first propose to use Elliptic Fourier Descriptors (EFD) [41] as high-dimensional features for object silhouettes. Elliptic Fourier Descriptors represent the shape of a silhouette, given as a set of 2D coordinates $(x(t), y(t))$, 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 exchanging Δx_p by Δy_p. Each harmonic is thus described by four coefficients, which have an intuitive geometrical interpretation: a_n (b_n) corresponds to the projection on the X axis of the semi-major (minor) axis of the n^{th} elliptic harmonic and c_n (d_n) to their projections on the Y axis. We thus propose to 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).

B. *Learning a silhouette manifold using GPLVM*

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) [42]. This technique is used because the shape spaces are often nonlinear. Moreover, since GPVLM makes no assumption about the distribution of the latent space, it permits to work with a low dimension, while still capturing most of the shape variance.

In more details, GPLVM represents 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 $d << D$. GPLVM can be considered as a generalization of the probabilistic PCA [43] to less restrictive covariance functions, by replacing the inner product kernel with nonlinear functions. Generally, the popular radial basis function kernel is used for the nonlinear mapping. GPLVM represents this mapping as a Gaussian process and determines the parameters of the mapping function in such a way that the distribution of the corresponding target data can be optimally approximated as a normal distribution. This mapping is optimised from the latent space to the original data space. As a result, GPLVM keeps apart in the latent space the points that are far apart in the data space, but nothing guarantees that points that are close in the data space will also be close in the latent space. Hence, to push GPLVM to also preserve local distances, we impose *back-constraints* [44] in the computation of the latent

variables. In particular, we constrain each latent variable to be a smooth mapping from its high-dimensional counterpart. 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 silhouettes transition in the high-dimensional space.

As an example, on the left part of Figure 4, the GPLVM optimization 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 35 elliptic harmonics. The colormap of Figure 4 represents the optimum approximation precision of this learnt latent space (computed as the certainty in the positions of the reprojected low-dimensional points $\mathbf{X}$ in the data space including $\mathbf{Y}$ [42]), where the regions with the warmest colors are more likely to represent the shape of a hand.

The set of silhouettes used for training are captured by one of the reference cameras before[3] the time at which the intermediate view synthesis is generated. In practice, the approach only requires a small amount of training samples; around 100 samples are used on average in our validation.

C. Interpolating intermediate silhouettes on the manifold

To obtain 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 on the left part of Figure 4), based on the mapping function learnt by GPLVM [42]. Then, we 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 back-projection of this path, from the latent space to the image space. The black silhouettes on the left of Figure 4 illustrates the silhouettes obtained by back-projection (from the latent space to the shape space) of the latent points represented in white, on the left part of the figure. More precisely, because the transition in the intermediate views must represent the motion of a 3D object, we constrain it to the latent points for which the approximation precision is the highest. Practically, the shortest geodesic path (white path in Fig. 4) is computed using Dijkstra's algorithm [45] with a transition cost c_{ij} from node i to j that is inversely proportional to the precision of j ($c_{ij} = -\log(\text{precision}_j + \epsilon)$, where ϵ avoids numerical instabilities).

D. Registering the silhouette priors with the reference ones

The set of prior foreground silhouettes obtained in the previous section represent a smooth and topologically consistent interpolation between the projections of the two reference silhouettes on the latent space. However, these priors describe the 2D shapes of the silhouettes, but not their position, scale and rotation. To exploit them during the EPIV reconstruction, we have thus to approximatively register them in the EPIV. This alignment is performed in three consecutive steps by:

1) Orientating the prior shapes with respect to the silhouettes observed in the reference views. The orientation of each silhouette is approximated by the angle of the first principal component of its PCA decomposition. Each shape is then rotated in such a way that its relative angle coincides with the linear interpolation of the angles of the two reference silhouettes.
2) Translating the oriented prior 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.
3) Scaling the translated and orientated prior shapes, based on the linear interpolation of the height of the object between the two reference silhouettes.

Figure 5(a) shows some aligned versions of the prior shapes (in white) extracted from a linear sampling along the shortest path in a latent space representing a dinosaur. As illustrated by the red segments on this figure, the resulting silhouette priors provide, for a given epipolar line, a set of silhouette borders. Hence, they describe *a priori* a smooth transition of the reference epipolar line segments, up to the alignment inaccuracies between the blue and red segments. In the following sections, those alignment inaccuracies are considered explicitely by using translation-robust metrics when comparing the reference epipolar line segments with the prior ones.

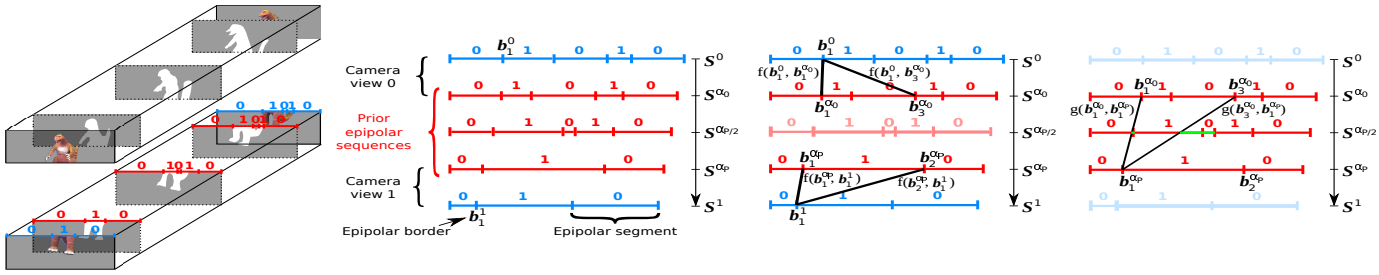
V. Transformations of epipolar line segments

This section explains how to disambiguate the ill-posed reconstruction of the object EPIs based on a sequence of $2D$ silhouette priors, as obtained in the previous section. As illustrated on the right side of Figure 4, our approach estimates how the object epipolar line segments evolve when moving the viewpoint from one reference view to the other. Due to the epipolar rectification of the reference images, the set of possible geometric transformations of a foreground (background) epipolar line segment is restricted to the combination of an horizontal translation, a 1D scaling and a potential split-up or merge with other foreground (background) epipolar line segments. In the following, 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 $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_S]$ 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 Fig. 5(a). For more clarity, in the following, these reference (blue) and prior (respectively red) epipolar line segments will be represented as a front view, as shown in Figure 5(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 if 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 $\mathbf{B} = [\mathbf{b}_0, \mathbf{b}_1, \cdots, \mathbf{b}_S]$ to each epipolar sequence $\mathbf{S}$, where $\mathbf{b}_{k-1}$ and $\mathbf{b}_k$ respectively represent the beginning and the end of the epipolar 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 the kind of transition

[3]The actual time-windows used in our validation are specified in Sec. VII.



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

Fig. 5: (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 minimizing 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)).

(foreground to background or background to foreground) that it supports, *i.e.*, $m(\mathbf{b}_k) = v(\mathbf{s}_k)$ if $k \in \{1, 2, \cdots, \mathcal{S}\}$, and $m(\mathbf{b}_k) = 0$ otherwise.

To determine how the 2D object silhouette, 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 5(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 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 details 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 Fig. 5(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 algorithm of Needleman and Wunsch [46] and adapt its underlying cost functions to account for our problem specificities.

The Needleman and Wunsch (NW) algorithm has been extensively used to compare sequences of characters [46]. 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 [47], (2) matches the most similar characters together by minimizing the sum of dissimilarities between matched characters and (3) tolerates unmatched characters at the cost of some skipping penalty $u(.)$. Its optimization scheme, which determines the associations and unmatched characters based on the matching cost $d(.,.)$ and skipping cost $u(.)$, is described in the supplementary material[4].

[4]http://infoscience.epfl.ch/record/200492

We now define the borders matching and skipping costs ($d(.,.)$ and $u(.)$ respectively), so as to capture the specificities of our problem, as well 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 to be 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 a different foreground/background value, the modes of consecutive borders are different. Skipping borders by pairs is thus equivalent to constraining each border to only match borders having the same modality. Hence, the distance between two borders with different modality 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.

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 represent thus *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 5(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 it 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 ($0 < \alpha_p < 1$).

Precisely, the **first metric**, illustrated in Fig. 5(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 share similar neighborhood. Because $\mathbf{S}^{\alpha_0}$ and $\mathbf{S}^0$ are seen from the same camera viewpoint, the foreshortening effect does not influence the length of their epipolar segments. This cost can be measured by the complementary of the normalized Hamming correlation (detailed in the supplementary material), *i.e.*, the number of positions in 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 on Fig. 5(d). We assume a linear displacement between two corresponding borders. With rectified cameras, the linearity is strictly verified when the silhouette borders correspond to the same physical 3D point [38], independently of the viewpoint. In other cases, since the actual 3D point supporting the silhouette border generally does not move a lot when changing the viewpoint, 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 color codes in Figure 5(d). The formal derivation of the prior deformation cost $g(\mathbf{b}_k^{\alpha_0}, \mathbf{b}_l^{\alpha_P})$ are given in supplementary material.

Finally, the cost $d(\mathbf{b}_i^0, \mathbf{b}_j^1)$ to match 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 [45]. By construction, 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 a 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 borders 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 analyzing how occluded parts vanish or appear when changing the viewpoint. As one of the most original contribution of this paper, we now show that it is possible to estimate how occluded parts vanish/appear when changing the viewpoint in-between the reference views. Since we know in which reference view the occluded epipolar segment[5] is visible, 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 6, 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 linearly proportional to α [38], we only have 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 Fig. 6(a)).
2) Fitting linear trajectories to these prior borders (Fig. 6(b)).
3) Associating the slope (vanishing speed) of each of these linear trajectory to one of the occluded reference border (Fig. 6(c)).

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

1) Identification of prior borders defining the occluded prior segments: Obviously, only the prior borders that do not support one of the association/matching 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 have used a simple heuristic threshold, set to 5% of the image width, to decide whether a prior border is sufficiently far from the linearly interpolated trajectories. This may however lead to many false positive. Hence, the following section proposes a robust way to estimate the vanishing/appearing paths from this initial set of occluded prior borders.

[5]An occluded epipolar segment is defined by two consecutive occluded epipolar borders.

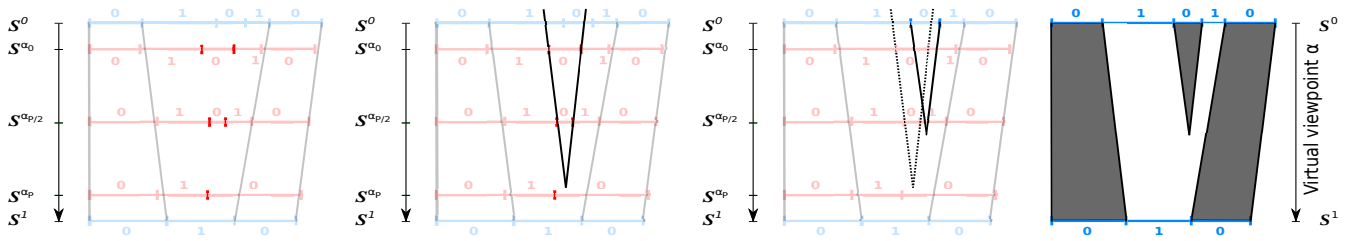


Fig. 6: The vanishing trajectories are estimated by (a) identifying the occluded prior borders (in dark red), (b) fitting linear trajectories to these prior borders and (c) associating the slope of these trajectories (dotted lines) to the occluded reference borders (plain lines). Finally, by (d) adding the vanishing trajectories to the set of tranjectories describing the transitions between the associated reference borders, the EPI of the object silhouette is reconstructed.

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. Precisely, 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 sequentially applying l_0 times (respectively l_1 times) the RANSAC algorithm [48], *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 the 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 as inliers to the trajectory model. This simple greedy algorithm appears to work well in practice, due to the relatively small amount of outliers in the set of occluded prior borders. The linear model that maximizes the amount of inliers is considered as 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 slope, *i.e.*, the constant speed at which the borders move along the EPI when the viewpoint index α changes. The process is illustrated in Figure 6(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 with 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 [49], so as to assign one and only one trajectory to each unmatched border while minimizing the sum of assignment costs. Finally, as illustrated on Fig. 6(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 a view is synthesized based on the estimation of the trajectories followed by the reference epipolar borders. We propose to synthesize the intermediate views by combining the textures of matched epipolar line segments and by propagating the texture of occluded line segments from the reference view in which those segments are visible.

Texturing an intermediate view by combining the textures of its corresponding elements in the both views has been deeply investigated in the past [27] [50]. By favoring the piecewise smoothness of the intermediate texture, most of these state-of-the-art methods permit to generate pleasant intermediate views despite corrupted matches. In contrast, in order to fairly validate our contribution, *i.e.*, the estimation of the geometric transformations of the epipolar line segments, we propose to simply rely on view morphing [27], which is not robust to corrupted matches. Indeed, it does not impose piecewise smoothness of the texture, so that any wrong border match results in highly noticeable discontinuities in textures.

More precisely, view morphing relies on epipolar rectification to synthesize the intermediate textures by linear interpolation of the reference textures, such 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 $\mathbf{I}_0$, y its fixed ordinate (studied scanline). The pixel abscissa x_α and x_1 are computed as follow:

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

with $\mathbf{s}_i^0$ denoting the epipolar line segment including x_0, $\mathbf{s}_j^1$ 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.

The proposed method propagates the texture for occluded segments with a similar principle, although the interpolation is done between a pixel in a reference view and the segment's vanishing point determined by the intersection (x_v, α_v) of the vanishing trajectories surrounding this occluded segment. This synthesis is 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). In contrast to conventional morphing strategies, the synthesized images represent both the parts that are visible in the two reference views, and the parts that are visible in a single reference view.

VII. Results

In this section, we demonstrate the performance of our approach on well-known datasets, namely the synthetic *Kung-Fu Girl* sequence [51] and the real *Ballet* sequence [35]. Due to the page limit, other validations on the multi-view *Dino* dataset [52] are given in the supplementary material. Although these multi-view datasets contain numerous images acquired by multiple (small-baseline) cameras, we only consider a pair of widely separated cameras from these sets to learn our shape priors model, and to reconstruct the intermediate views.

For each dataset, we interpolate five intermediate views uniformly sampled in-between the left and right reference views[6]. To show the advantage of using epipolar line segments as basis matching element, 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 camera (or right camera) is shown on the left (respectively right) of the first row in Figure 7. The image shown in-between corresponds to the ones captured by a camera situated approximatively at the middle ($\alpha \simeq 0.5$) in-between these two reference views and represents thus the ground-truth.

The second row in Figure 7 represents the intermediate views generated by a conventional visual-hull reconstruction [17], 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 [26]. The reconstructed intermediate views perfectly show the limitations of model-based approaches in our wide-baseline stereo, namely the requirement of observing the object with a large amount of reference cameras to avoid an imprecise 3D model, leading to corrupted intermediate views.

The third row in Figure 7 represents the intermediate views generated when morphing [27] a dense (pixel) correspondence obtained by dynamic programming [32] [47] on corresponding epipolar lines. The matching cost is simply defined as the $\mathcal{L}_2$ norm of the pixels' colors and the skipping penalty $u(.)$ is arbitrary set to 0.5. Two kinds of artefacts can be observed on these reconstructed views. First, they are topologically incoherent. This can be observed in-between the legs of the Kung Fu girl, near her neck and on her left hand, where some members get apart 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 [24], at the price of a slow matching process.

To explicitly impose the smoothness along the epipolar lines, in the 4^{th} row of Figure 7, we use epipolar line segments as matching elements. The method extends the one of [32] by considering epipolar line segments (and not pixels) as basic image elements. It corresponds to the approach we have introduced in Section V-A, but without prior silhouettes knowledge. Hence, each epipolar border is matched, by NW [47], 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 7 that the reconstructed intermediate views are smoother, but still exhibiting some topologically incoherent transitions, such as shown at the level of her head.

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 60 silhouettes captured by the two wide-baseline cameras, and observed uniformly in a time-window starting from the first frame of the sequence to 20 frames before the required transition. These silhouettes have been described using 70 elliptic harmonics, and the fifth row in Figure 7 illustrates this latent space.

The advantage of considering these priors is illustrated on the last row in Fig. 7, where intermediate views have been generated by the method proposed in this paper. Only 5 intermediate priors have been used to reconstruct the EPIs. We observe that the intermediate views reconstructed by our method shows 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* [35], has been captured using eight cameras placed along a $1D$ arc spanning about $30°$ end-to-end. While two neighbor cameras of 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

[6]We encourage the reader to refer to videos provided in the supplementary material to observe the continuous transition from the left to the right cameras.

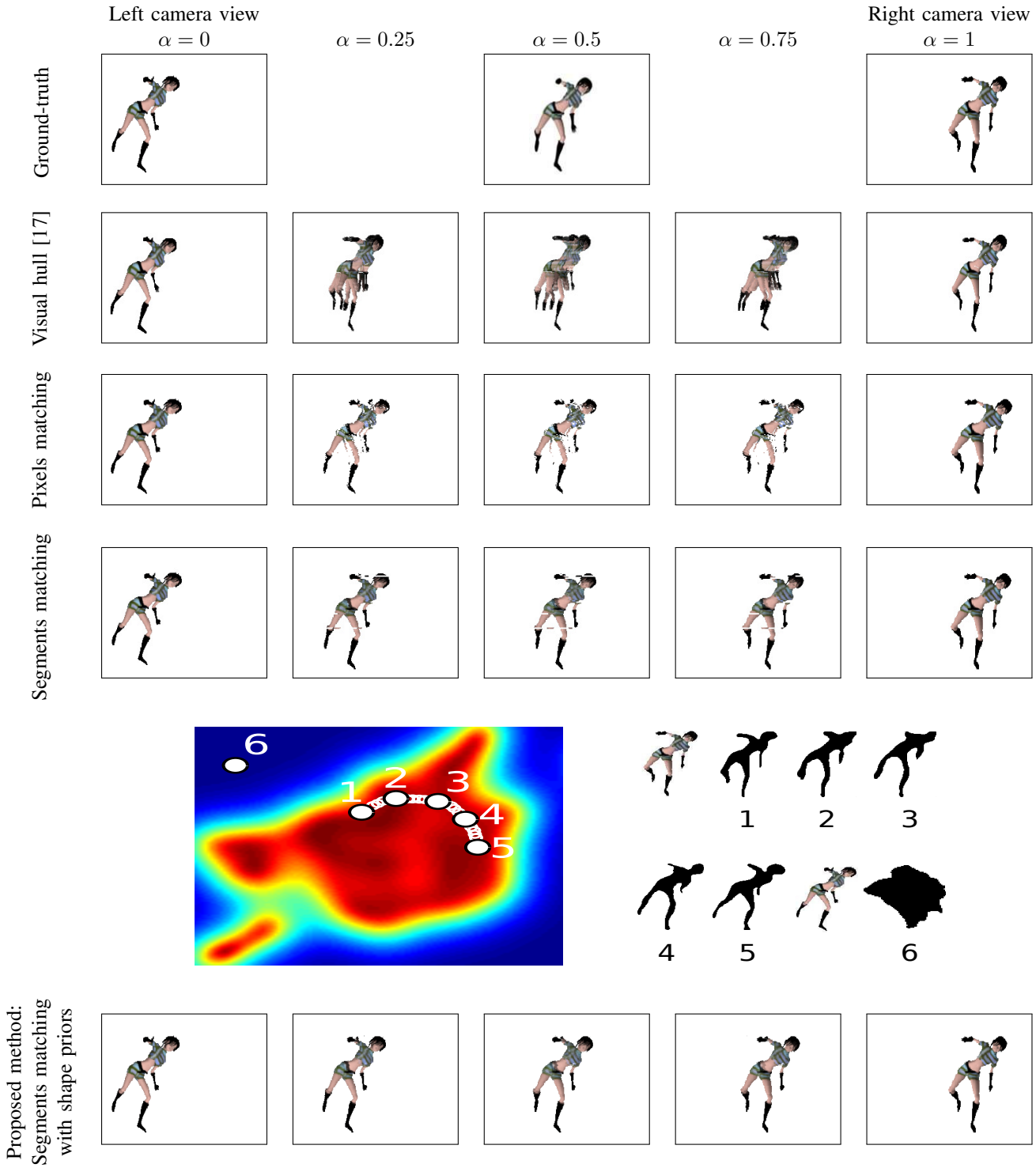


Fig. 7: Instead of projecting an estimated 3D model [17] (second row) or determining a dense (pixel) match (third row), epipolar line segments are used as basic matching elements (fourth row). In the last row, our method regularizes the epipolar segments matching so that the shapes of the intermediate silhouettes are topologically consistent with the plausible deformations of the object silhouette, learnt and described by a low-dimensional latent space (fifth row).

and foreshortening effects can be observed between these two external viewpoints (especially on the dancer's arms). In Figure 8, we compare the reconstructed images at intermediate viewpoints with 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 [35]). On the first row, the intermediate views are generated by view morphing, based on multiple depth maps as proposed in [35][7]. 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 neighbor 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 of Figure 8 illustrates the intermediate views obtained by a state-of-the-art stereo method [53], top-ranked in February 2015 in the well-known Middlebury Stereo Evaluation [54] [55] [47]. By combining a cost-filtering approach, especially adapted to manage the occlusions, with a global (fully connected Markov Random Field) optimization, 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 [47]) on pixels, as done in [32]. We observe that the strong foreshortening effect produces holes in the reconstructed intermediate views. By applying dynamic programming on the segment representation, dense correspondences have been found, but topological inconsistencies subsist (see fourth row on Figure 8). Because of the ill-posed nature of the problem, the lowest cost match does not necessarily give the optimal match in terms of topological consistency of the silhouette. The last row in Figure 8 illustrates the result obtained by our complete method using silhouette priors. The latent space has been learnt on 40 silhouettes observed by the two outermost cameras in a time-window 20 frames away from the transition time. These silhouettes have been described with 50 harmonics of Elliptic Fourier Descriptors, and 6 intermediate priors are used to regularize the determination of the segments' transformations.

C. Discussion

In contrast to the previous methods, we obtain topologically coherent intermediate views, thanks to the additional silhouette prior obtained from the latent space. Our method also efficiently deals with the foreshortening effect that are typical in wide-baseline configurations, as it can be seen on the front part of the chest of the dancer, which is severly slanted in the left reference view, while almost fronto-planar in the right one. Finally, to the best author's knowledge, this paper is the first one to infer the trajectories of occluded parts, allowing to interpolate their content in intermediate views, as shown by the left shoulder of the Kung-Fu girl or the space in-between the legs of the dancer. Next to those very encouraging results, two limitations of our approach however deserve to be pointed. The first one can be observed on the reconstructed fingers of the dancer's right hand in Fig. 8. Indeed, when changing the viewpoint from the left to the right reference view, her fingers detach from her hand, showing a topologically incoherent transition of her right hand. 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 the high frequencies details, such as the dancer's fingers, are not represented by the priors, their matching can not be regularized, and their transition might become topologically incoherent.

The second weakness of our method comes from the choice of using epipolar line segments as matching units. Indeed, while it permits to explicitely take into account the foreshortening effect and the sparse correspondence problem, the precise correspondence of their inner pixels is not known, and can only be infered from the knowledge of the matches of their borders. The simple linear interpolation of inner textures, as detailed in Section VI, may result in wrong matches of the inner pixels if the surface described by this epipolar line segment is not planar. This limit can be observed on the pixels representing the straps on the chest of the dancer, which are not correctly matched by linearly interpolating inside the (correctly matched) borders of the curved chest or on her face, and results in the ghosting artefact. This artefact could be reduced by generalizing the texture interpolation to convex surfaces, *e.g.*, based on floating textures [50].

Finally, we note that the processing time of our algorithm (Matlab implementation, Intel I5 CPU 2.4GHz and 8Gb of RAM) shows encouraging performances (on average 4.2s to describe a 768×1024 image into epipolar line segments, 0.06s to match all the epipolar lines independently and 0.16s to render an intermediate view). Our code is available online (see the supplementary material). Moreover, because the epipolar lines are processed independently, real-time implementation is within reach, *e.g.* based on GPU parallelisation.

VIII. Conclusion

In this paper, we have proposed a new and original interpolation technique for intermediate view synthesis between cameras in wide-baseline configurations. We also notice that although this coherence is imposed independently on each epipolar line, the fact that these constrains are derived from 2D priors favors consistency along the epipolar lines. Our method relies on prior information about the silhouettes of objects in the intermediate views to guarantee consistency between the synthesized silhouettes and the ones present in the two reference viewpoints. As a first contribution, these silhouette priors are learnt by reducing the dimensionality of Elliptic Fourier shape Descriptors, accumulated over a training set of representations of the objects under consideration, typically from earlier observations of the object moving in front of

[7]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.

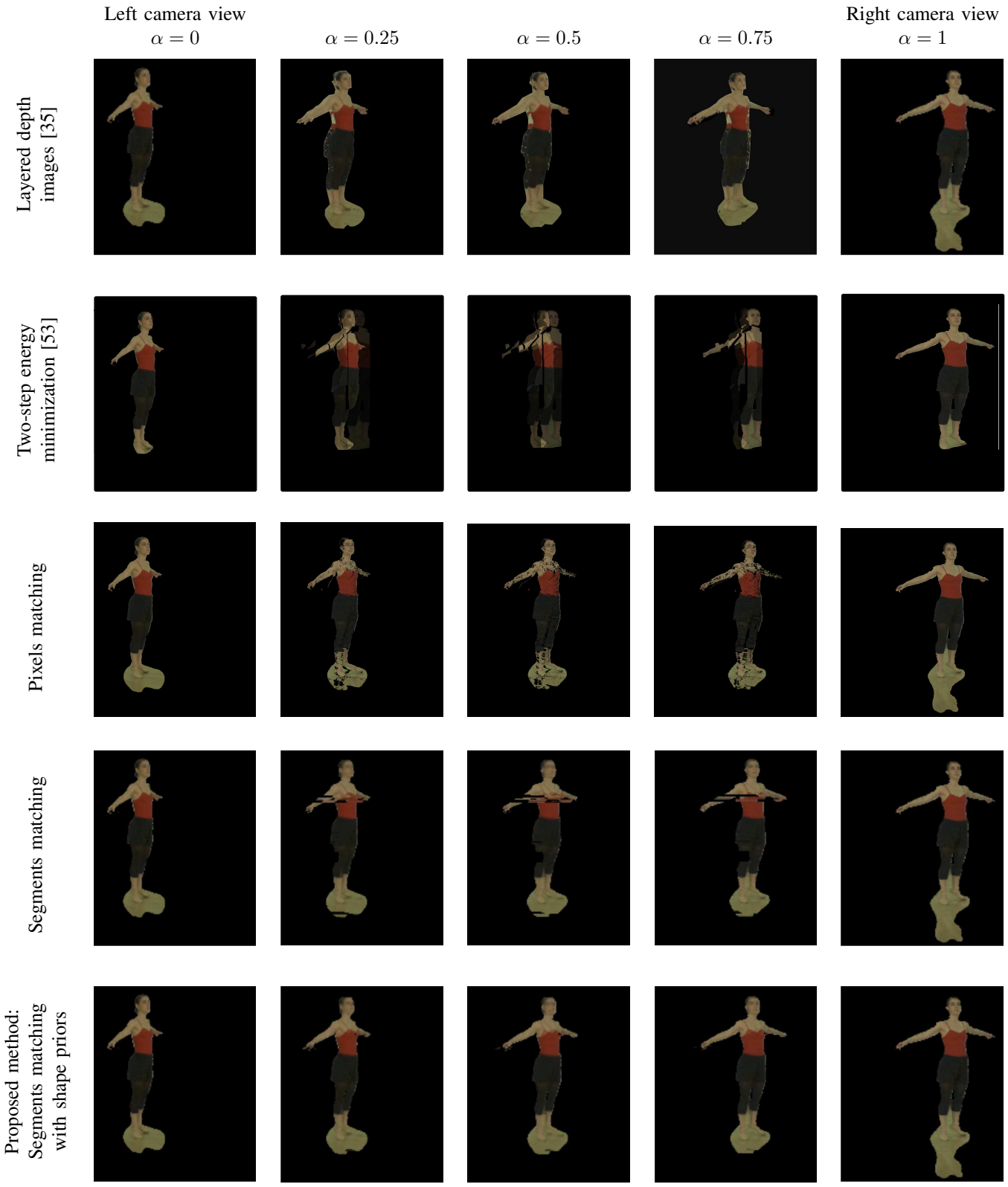


Fig. 8: Comparison between the interpolated intermediate views generated based on matching of layered representation [35] (first row), which exploits intermediate depth maps in addition to external views, a state-of-the-art method [53] which is top-ranked in the Middlebury Stereo Evaluation [54] (second row), dynamic programming on pixels [32] (3^{rd} row), dynamic programming on our proposed epipolar line segment representation (4^{th} row) and our method (5^{th} row).

the wide-baseline camera pair. This additional information is then exploited to determine the 1D transformation of epipolar line segments when moving from one view to the other. As a second contribution, this new framework has not only the advantage of generating consistent and smooth virtual transitions between the viewpoints where correspondences can be found in the two basis images, but it can also handle the vanishing of occluded informations. Finally, we have demonstrated that our method outperforms state-of-the-art view interpolation methods by generating topologically coherent intermediate views of an object, despite the multiple occlusions and severe foreshortening effect that are typical in wide-baseline configurations.

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] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

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

[15] 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.

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

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

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

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

[20] 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.

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

[22] 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.

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

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

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

[26] 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.

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

[28] 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.

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

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

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

[32] 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.

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

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

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

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

[41] 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.

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

[43] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 61, no. 3, pp. 611–622, 1999.

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

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

[46] 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.

[47] 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.

[48] 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.

[49] H. W. Kuhn, "The hungarian method for the assignment problem," Naval research logistics quarterly, vol. 2, no. 1-2, pp. 83–97, 1955.

[50] 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.

[51] 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/

[52] 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.

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

[54] 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.

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