Perception-Motivated Interpolation of Image Sequences

TIMO STICH, CHRISTIAN LINZ, TU Braunschweig
CHRISTIAN WALLRAVEN, Max Planck Institute Biological Cybernetics
DOUGLAS CUNNINGHAM, University of Tübingen
MARCUS MAGNOR, TU Braunschweig

We present a method for image interpolation that is able to create high-quality, perceptually convincing transitions between recorded images. By implementing concepts derived from human vision, the problem of a physically correct image interpolation is relaxed to that of image interpolation which is perceived as visually correct by human observers. We find that it suffices to focus on exact edge correspondences, homogeneous regions and coherent motion to compute convincing results. A user study confirms the visual quality of the proposed image interpolation approach. We show how each aspect of our approach increases perceived quality of the result. We compare the results to other methods and assess achievable quality for different types of scenes.

Categories and Subject Descriptors: I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Motion*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Perception, morphing, image interpolation

ACM Reference Format:

Stich, T., Linz, C., Wallraven, C., Cunningham, D., and Magnor, M. 2011. Perception-motivated interpolation of image sequences. ACM Trans. Appl. Percept. 8, 2, Article 11 (January 2011), 25 pages.
DOI = 10.1145/1870076.1870079 http://doi.acm.org/10.1145/1870076.1870079

1. INTRODUCTION

Physically impossible and stunning effects such as frozen-time view interpolation and extreme slow motion can be created from regularly recorded video sequences and images using image interpolation. Unfortunately, the problem of retrieving the true motion field, necessary to create the physically correct interpolated image, from images alone is often ill-posed due to inherent ambiguities, such as the aperture problem. Despite these ambiguities, humans are able to easily navigate through complex, dynamic environments based on what they see with their eyes. The brain constantly interprets changing visual input in terms of plausible motion of the viewpoint and/or of the observed scene and apparently does not rely on exactly solving the laws of physics for its judgements. Cartoon animations, for example, frequently contain motion which, while physically inaccurate, is perceptually totally acceptable.

In this work, we introduce an image interpolation algorithm that attempts to adapt to processes of the human visual system to understand seen images, relaxing geometric constraints. This approach to

Author's address: T. Stich; email: stich@cg.tu-bs.de.

DOI 10.1145/1870076.1870079 http://doi.acm.org/10.1145/1870076.1870079

image interpolation makes it possible to solve motion estimation as a much better conditioned problem. In a user study, we confirm the perceptual validity of each part of the proposed image interpolation algorithm. We quantify changes in perceptual quality introduced by parameter changes within our proposed approach, compare the results against other approaches, and investigate whether there is a perceptual difference between results on real-world and synthetic image material. Earlier results have been published in Stich et al. [2008a, 2008b]. The contribution of this article over Stich et al. [2008a, 2008b] manifests in an in-depth discussion of the perceptual and geometric interpretations of our approach and how they are related, the implementation of a multiscale formulation of the approach presented in Stich et al. [2008] and its evaluation with a new user study.

2. RELATED WORK

Many graphics areas can benefit from related literature of perception both in improving the overall quality as well as in focusing on important features. A very good survey of interesting perceptual models that have been applied successfully in computer graphics has been published by O'Sullivan et al. [2004].

Recently Vangorp et al. [Vangorp et al. 2007; Vangorp and Dutré 2008] researched the influence of material of objects to shape perception showing interesting dependencies between differentiability of material properties and surfaces in extensive user studies. Ramanarayanan et al. [2007; 2008] conducted user studies on visual equivalence of rendered images and the equivalence of the rendering of groups of objects depending on shape and texture. Both approaches open new insights on where computational power can be saved, in the form of simplified materials and/or shape properties or reduced object numbers, if the human observer is understood as the consumer of the output of the rendering pipeline.

Image morphing, the interpolation between images depicting different objects from user-defined correspondences, is another concept related to our method. Algorithms like Beier and Neely [1992], are often used in the movie industry to create visual effects as for example smooth transitions between actor appearances during performance. Other warping techniques have been discussed by Wolberg [1998], including the popular thin-plate spline interpolation, which is based on point correspondences. A computationally more complex method based on line features was recently proposed by Schaefer et al. [2006]. In general, image morphing methods are solely based on user specified features and are thus work-intensive when interpolating image sequences. Additionally, when motion discontinuities need to be taken care of, an additional segmentation of the image into the differently moving layers is necessary.

The optical flow refers to the flow field created by the spatiotemporal trajectories of image patches during an image sequence, and was first described by the psychologist James J. Gibson [1955]. Since the pioneering work on local and global optical flow reconstruction by Lucas and Kanade [1981] and Horn and Schunck [1981], respectively, a multitude of computational approaches have been devised and applied in a variety of fields [Barron et al. 1994; Baker and Matthews 2004]. While the input to typical optical flow algorithms is similar to the input to our proposed method, the approaches and goals are fundamentally different. We find that a physically correct motion field between images is not necessary to create perceptually convincing interpolation results but that the interpolation algorithm must focus on the properties important to human perception. Using our approach, we additionally achieve a robust and perceptually correct handling of motion discontinuities, which are even more challenging to solve physically correctly.

Image-based rendering (IBR) methods achieve highly realistic rendering results, using a collection of timely synchronized calibrated photographs. While some IBR methods rely solely on the number of images to minimize aliasing artifacts [Levoy and Hanrahan 1996; Matusik and Pfister 2004],

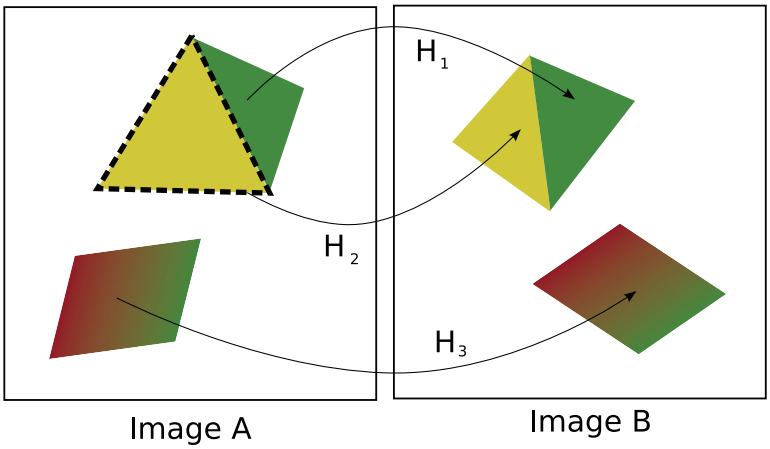


Fig. 1. Correspondences for views of a dynamic 3D scene consisting of planar surfaces can be described in image space by homographies. As a *translet*, we define an image segment of a 3D plane and its corresponding homography. For example, a translet of image A is the outlined image segment showing the bright face of the pyramid and the corresponding homography H_2, which defines its correspondence to image B.

most IBR approaches make additional use of epipolar constraints [McMillan and Bishop 1995; Seitz and Dyer 1996; Matusik et al. 2000; Vedula et al. 2005], scene depth [Chen and Williams 1993; Gortler et al. 1996; Isaksen et al. 2000; Buehler et al. 2001; Zitnick et al. 2004], or full 3D geometry information [Debevec et al. 1998; Wood et al. 2000; Carranza et al. 2003; Snavely et al. 2006]. The quality of IBR techniques is strongly dependent on accurate camera calibration, scene geometry, and/or time-synchronized acquisition. These limitations make data acquisition for IBR a time-consuming and delicate endeavor, which typically requires a controlled environment and expensive equipment.

In the image-processing and video-coding field, the temporal interpolation of images is also of interest. Here, the motion between consecutive images is computed block-wise. From this information, images can be predicted from the previous image such that the residual difference to the original image can be efficiently compressed (motion compensation) [Jain and Jain 1981; Tourapis et al. 2001; Fu et al. 2002]. In contrast to our goal, this motion estimation is geared toward achieving high compression ratios with as little computational complexity as possible rather than to create plausible in-between images. The images obtained from motion compensation without additional correction using the residual image, which is unknown in our scenario, often shows artifacts, especially at block boundaries.

3. GEOMETRIC CONSTRAINTS FOR IMAGE INTERPOLATION

The relation between two projections of a 3D plane can be directly described via a homography in image space [Hartley and Zisserman 2000]. Such homographies, for example, describe the relation between a 3D plane seen from two different cameras, the 3D rigid motion of a plane between two points in time seen from a single camera or a combination of both. Thus, the interpolation between images depicting a dynamic 3D plane can be achieved by a per-pixel deformation according to the homography directly in image space without the need to reconstruct the underlying 3D plane, motion and camera parameters explicitly (Figure 1). The relation between the corresponding pixels of images from a typical dynamic real-world scene, on the other hand, is far more complex. However, many graphics approaches have been very successful in creating photorealistic images from approximations of natural scenes and

objects with meshes consisting of simple planar triangles. For each such triangles, the relation of the corresponding pixels is, again, exactly described via local homographies.

Our proposed image deformation model is motivated by these observations. We assume that natural images can be decomposed into such regions by Gestalt rules [Wertheimer 1938], for which the deformation of each element is sufficiently well described by a homography. Specifically, we introduce *translets*, which are homographies that are spatially restricted. That is, a translet is described by a 3×3 matrix H and an image segment. To obtain a dense deformation, we enforce that the set of all translets is a complete partitioning of the image and thus each pixel is part of exactly one translet. Note that since the deformation model is defined piecewise, it can well describe motion discontinuities as, for example, resulting from occlusions.

4. PERCEPTUAL CRITERIA FOR IMAGE INTERPOLATION

Human vision is a very powerful system adept at extracting meaningful patterns in sometimes very underconstrained situations allowing us to understand, navigate through, and interact with our surroundings rapidly and efficiently. The importance and complexity of this task is perhaps reflected by the fact that approximately half of our brain is dedicated to processing visual input. While the human visual system as a whole is very complex and has many not yet fully understood aspects, some parts are well researched.

In one of the earliest works on neural networks, Exner [1894] described a "Centrum der optischen Bewegungsempfindungen" (center of optical motion sensitivity). Subsequently, based on his work with flies and beetles, Reichardt [1961] mathematically and neurally described a local-correlator motion detector. The detector, which explicitly relies on the fact that real-world objects tend to move rather smoothly, matches small image patches across small spatial and temporal distances. Although Reichardt's local-correlator can be said to respond more to temporal frequency than velocity [Reichardt 1961; van Santen and Sperling 1985], low-level motion processing in humans is rather well described by this detector [Qian and Andersen 1997; Heeger et al. 1999]. Moreover, global optical flow fields can easily be constructed from banks of local-correlator motion detectors.

In addition to the well-known roles of optical flow in human vision, it has been shown that the human visual system takes advantage of the fact that neighboring areas on an object tend to have the same motion (referred to as *common fate* in gestalt psychology). Local smoothness constraints help to compensate for noise and aid in image segmentation. Additionally, the common motion of neighboring patches and differently oriented edges are used to help solve the aperture problem [Wallach 1935].[1]

Interestingly, there are cases where local-correlators fail. For example, as an object moves over a textured background, it will successively hide and reveal the background texture. Thus, while the surface of the object gives rise to optical flow through common fate, there will be local motion discontinuities in the flow field at its edges. Far from being noise, these failures are of great importance to human perception [Shipley and Cunningham 2001]. First, all the appearances and disappearances are on the background texture, and thus, the relative rates of texture accretion and deletion on different sides of the discontinuity specify the relative depth of the two surfaces [Gibson et al. 1969]. More interestingly, since the accretion and deletion occur at (and only at) the edges of the object, they not only describe where the edges are, they actually specify the silhouette of the object [Kellman and Shipley 1992] and the degree of transparency of the nearer surface [Cunningham et al. 1998]. The use of the pattern of local motion discontinuities to define the shape of an object is referred to as spatiotemporal

[1]In optical flow estimation, only the motion component in the direction of local gradient of the image intensity can be estimated. The motion of a homogeneous region cannot be recovered optically, since the image gradient does not provide any information. This is known as the aperture problem.

boundary formation (SBF). Cooke et al. [2004] designed a simple computational simulation of SBF, yielding results very consistent with human data.

From these findings on the human motion perception, we conclude that to achieve perceptually plausible image interpolation results, local discontinuities in optic flow field are critical, and that it is important to transform corresponding edges defined there exactly onto each other while transforming homogeneous regions within the images coherently. Our algorithm can also be interpreted in terms of parallels to the motion processing of the human visual system. Actually, the different steps in the derivation of the motion field improve the quality of the interpolation in a perceptually plausible sense. The matching of the edglets ensures that edges are transformed exactly onto each other. Then, the estimation of local, short-range transformations and the final per-pixel smoothing of the motion field optimizes the coherence of the motion field.

5. ESTIMATING THE IMAGE DEFORMATION

In this section, we discuss how to robustly estimate dense correspondences between two images based on the proposed model. Therefore, both a suitable partitioning of the images into regions that can be approximated by 3D planes and sparse point correspondences have to be established. Then, homographies from the point correspondences are estimated for each region to form a translet. While the optimal partitioning of the images into such regions is not known a priori, this has great influence on the solution.

A small number of regions will result in a very robust but restrictive solution, while a larger number increases the flexibility at the cost of decreased robustness against outliers in the match. To obtain an optimal result, we follow a bottom-up approach. We start from a large number of regions and merge neighboring translets in a greedy manner until the optimal ratio is achieved. In the following, we discuss the steps in estimating the image deformation between two images in detail (see also Figure 3 for an overview of the proposed image interpolation approach).

5.1 Matching of Edge Pixels

The first step in estimating the parameters of the deformation model is to find a set of point correspondences between the images. These will be used in the second step to estimate the homographies. Additionally, for a plausible solution in terms of perceived motion quality, it is also necessary that these point features are also considered important for human vision. Edges and corners are thus especially suited point feature candidates. They are both relatively stable over time and viewpoints, are the image parts where motion is most apparent, and are also the features that the human visual system is known to measure very early [Marr 1983; Adelson and Bergen 1991; Hubel 1995]. For the detection of edge pixels, we can also resort to a large body of previous work. Specifically, we used the Compass operator [Ruzon and Tomasi 1999] in our experiments, as it has the advantage to directly make use of color information and often outperforms the Canny operator [Canny 1986]. After nonmaximal suppression, we obtain a set of edge pixels or *edglets* (Figure 2). Depending on the scene, between 2,000 and 20,000 pixels are edglets (Figure 2).

To establish a good match between the edglets of the two images to interpolate, it is necessary to match them as completely as possible and to consider the spatial context of each edglet to preserve local structure. We will ensure completeness by posing the matching problem accordingly and address local structure preservation with a stable descriptor that captures spatial context. The shape context descriptor [Belongie et al. 2001] has been shown to perform very well at capturing the spatial context of the nearest k neighbor edglets and is robust against the expected deformations. Completeness of the matching of the edglets based on Euclidean distance and shape context is then achieved by solving a maximum weighted bipartite graph matching problem. The additional advantage is that this problem

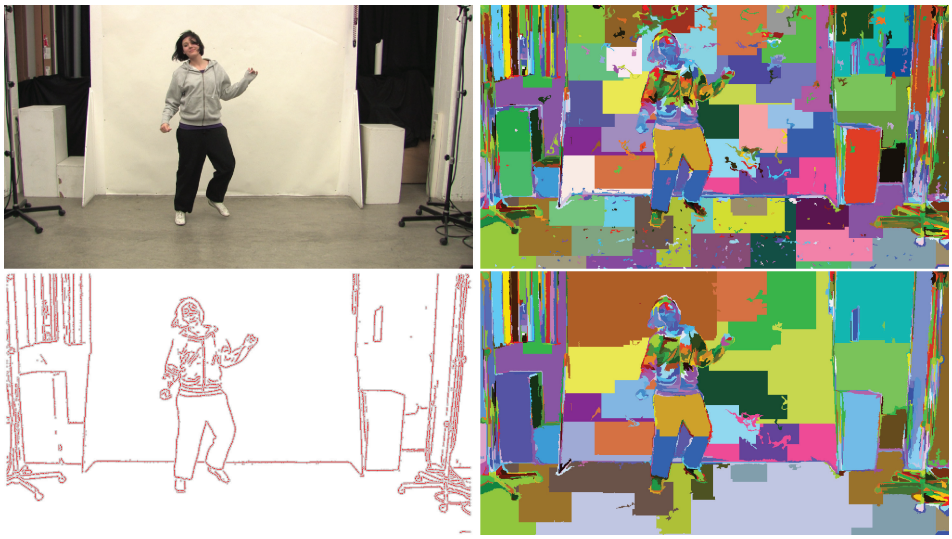


Fig. 2. An image (upper left) and its decomposition into its homogeneous regions (upper right): since the transformation estimation is based on matched edglets, only superpixels that contain actual edglets (lower left) are of interest. Superpixels with no edglets are merged with their neighbors (lower right).

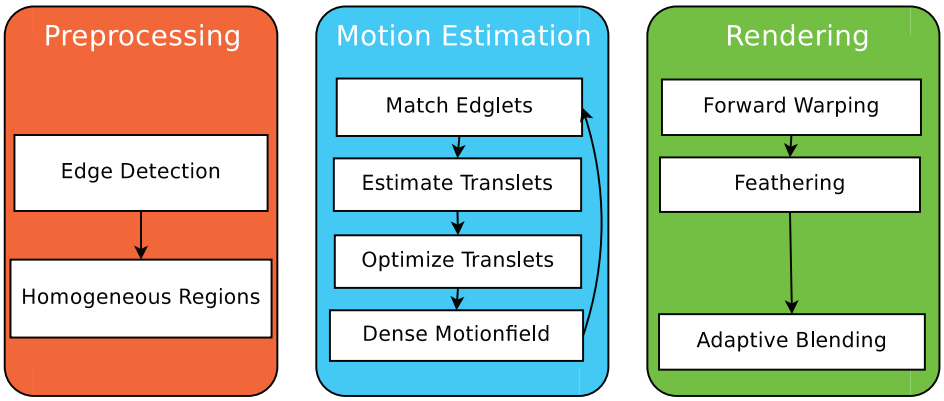


Fig. 3. Overview of our image interpolation approach. First, the images are preprocessed to find edges and homogeneous regions. These are then used to determine dense correspondences. Finally, we use this correspondence field for image interpolation rendering in real time.

can be solved globally optimal in a matter of seconds for the problem sizes we are facing [Bertsekas 1992]. One prerequisite for the reformulation is that for each edglet in the first set, a match in the second set exists; otherwise, the completeness cannot be achieved. While this is true for most edglets, some will not have a correspondence in the other set due to occlusion or small instabilities of the edge detector at faint edges. However, this is easily addressed by inserting virtual occluder edglets for each edglet in the first edglet set. The graph for the matching problem is then built as depicted in Figure 4. Each edge pixel of the first image is connected by a weighted edge to its possibly corresponding edge pixels in the second image and additionally to its virtual occluder edglet. The weight or cost function

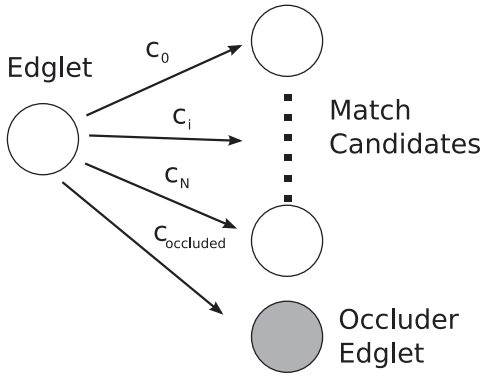


Fig. 4. Subgraph of the weighted bipartite graph matching problem for a single edglet. Each edglet has an edge to its possible match candidates and an additional edge to its virtual occluder edglet.

for edglet $\mathbf{e}_i$ in I_1 and $\mathbf{e}'_j$ in I_2 is then defined as

$$C(\mathbf{e_i}, \mathbf{e}'_j) = C_{dist}(\mathbf{e_i}, \mathbf{e'_j}) + C_{shape}(h_{\mathbf{e_i}}, h_{\mathbf{e'_j}}), \tag{1}$$

where the cost for the shape is the χ^2-test between the two shape contexts $h_{(.)}$ [Belongie et al. 2001] computed as

$$C_{shape}(h_{\mathbf{e_i}}, h_{\mathbf{e'_j}}) = \frac{1}{2}\sum_{k=1}^{K}\frac{[h_{\mathbf{e_i}}(k) - h_{\mathbf{e'_j}}(k)]^2}{h_{\mathbf{e_i}}(k) + h_{\mathbf{e'_j}}(k)}, \tag{2}$$

and the cost for the distance is defined as

$$C_{dist}(\mathbf{e_i}, \mathbf{e}'_j) = \frac{a}{(1 + e^{-b\,||\mathbf{e}_i - \mathbf{e}'_j||})}, \tag{3}$$

with $a, b > 0$ such that the maximal cost for the Euclidean distance is limited by a. The edge to the virtual occluder edglet is weighted by a user-defined cost $C_{occluded}$. This cost controls how aggressively the algorithm tries to find a match with an edglet of the second image. The lower $C_{occluded}$, the more conservative the resulting matching will be, as more edges will be matched to their virtual occluder edglets. For our experiments, we set $C_{occluded} = 5$.

5.2 Estimating the Local Homographies

According to the proposed motion model, we assume that the scene can be approximated with piecewise planes for interpolation purposes. For each such region, we would assume that the motion is described by the relation of projections of a 3D plane, as discussed in Section 3. From the Gestalt theory [Wertheimer 1938], it is known that for natural scenes, these regions share not only a common motion but in general also share other properties such as similar color and texture. Felzenszwalb and Huttenlocher [2004] proposed to partition images into so called superpixels based on neighboring pixel similarities (Figure 2). We resort to their algorithm to find an initial partitioning of the image into regions to become translets. Then, local homographies H_t for each superpixel are estimated from the set of edglet matches that lie within its spatial support (Figure 5). Edge pixels matched to the corresponding virtual occluder edglet are treated as having no match and thus do not influence the homography estimation. Since the least-squares estimation based on all matched edglets of a translet is sensitive to outliers and often more than the minimal number of four matched edge pixels is available, a RANSAC

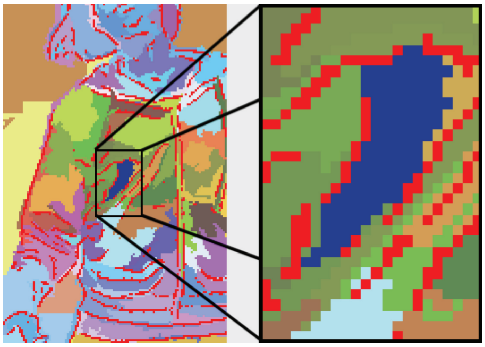


Fig. 5. The translets of an image are found by partitioning the image according to a superpixel segmentation and computing local homographies from point correspondences to the target image.

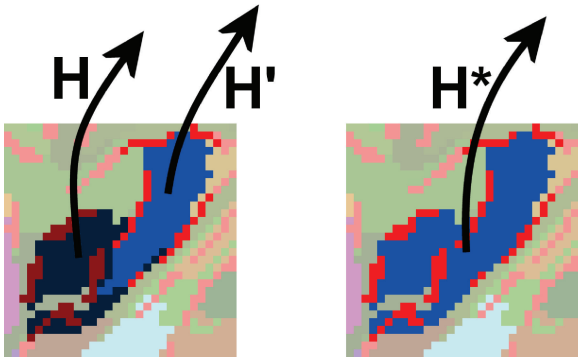


Fig. 6. During optimization, similar transformed neighboring translets are merged into a single translet. After merging, the resulting translet consists of the combined spatial support of both initial translets (light blue and dark blue) and their edglets (light red and dark red).

approach to obtain a robust solution and filter match outliers is preferred instead [Hartley and Zisserman 2000]. Matches are considered as outliers if the distance $||H_t\mathbf{e}_i - \mathbf{e}'_j||$ exceeds a threshold of 3 pixels.

5.3 Translet Optimization

From the point correspondences, we have established dense correspondences between the images using our deformation model. However, in our experiments we observed that between 20% and 40% of the computed matches are outliers, and thus some translets will have wrongly estimated transformations. We address this problem by optimizing the number of translets of our image deformation model to increase the robustness against these outliers. The initial solution of our model is generally very flexible and suffers from numerical instability, because the spatial support of the translets can be too small for a reliable estimation. Using a greedy approach, we iteratively merge the most similar transformed neighboring translets into one, as depicted in Figure 6, until the ratio of outliers to inliers is lower than a user defined threshold. When two translets are merged, the resulting translet then contains both edglet sets and has the combined spatial support. The homographies are re-estimated based on the new edglet set, and the influence of the outliers is again reduced by the RANSAC filtering.

Table I. Timing Results of Our Method for Pairs of Images on an AMD Athlon(tm) 64 ×2 Dual Core Processor 4800+, 4GB RAM, NVIDIA GeForce 7800 GTX to Compute the Dense Correspondences

Scene	Edglets	Res.	Matching	Optim.
Dancer	2,570	960 × 540	1.94s	5.67s
Dimet.	8,604	584 × 388	11.27s	16.54s
Hair	17,560	960 × 540	27.77s	35.57s

5.4 Per-Pixel Correspondences

So far, motions and discontinuity are handled on the translet level. However, when only a part of a translet boundary is at a true motion discontinuity, noticeably incorrect discontinuities still produce artifacts along the rest of the boundary. For example, the motion of an arm in front of the body is discontinuities along the silhouette of the arm, while the motion at the shoulder changes continuously. Additionally, small deviations from the planar motion are not sufficiently well handled by the general approach. Thus, we address these issues on a per-pixel basis. Since the translets partition the image, each pixel in the image is uniquely associated with a translet t. The deformation vector for a pixel $\mathbf{x}$ is thus computed from the translet's homography H_t as

$$d(\mathbf{x}) = H_t \cdot \mathbf{x} - \mathbf{x}. \tag{4}$$

We can then resolve the per-pixel smoothing by an anisotropic diffusion [Perona and Malik 1990] on this vector field using the diffusion equation

$$\partial d/\partial t = div(\, g(min(|\nabla d|, |\nabla I|)\, \nabla d), \tag{5}$$

which is dependent on the image gradient ∇I and the gradient of the deformation vector field ∇d whichever is smaller in magnitude at the observed pixel. The function g is a simple mapping function as defined in Perona and Malik [1990]. Thus, the deformation vector field is smoothed in regions that have similar color or similar deformation, while discontinuities that are both present in the color image and the vector field are preserved. This improves the smoothness of the deformations on a per-pixel level while preserving important motion discontinuities. During the anisotropic diffusion, edglets that have an inlier match, meaning they are only slightly deviating from the planar model, are considered as boundary conditions of the PDE. This results in exact edge transformations handling also nonlinear deformations for each translet and significantly improves the achieved quality. The total timings for the computation of the deformation field for different resolutions and scenes are listed in Table I.

5.5 Multiple Iterations, Multiscale and User Interaction

Since our matching energy function (Equation (1)) is based on spatial proximity and local geometric similarity, we can introduce a motion prior d by prewarping the edglets with a given deformation field i.e., we compute a transformed edglet by offsetting it according to the deformation field, substituting $\mathbf{e}_i$ with $\mathbf{e}_i + d(\mathbf{e}_i)$ in Equation (1)). The estimated dense correspondences described in the last sections can be used as such a prior. We can then implement a coarse to fine iterative approach to overcome local matching minima, as, for example, depicted in Figure 7, as follows: In the first iteration, we optimize the number of translets until we obtain the coarsest-possible deformation model with only one translet and thus approximate the underlying motion by a single perspective transformation. During consecutive iterations, the threshold is decreased to allow for more accurate deformations as the number of final translets increases. Using the previous solution as motion prior significantly reduces the risk to getting stuck in local matching minima (Figure 7).

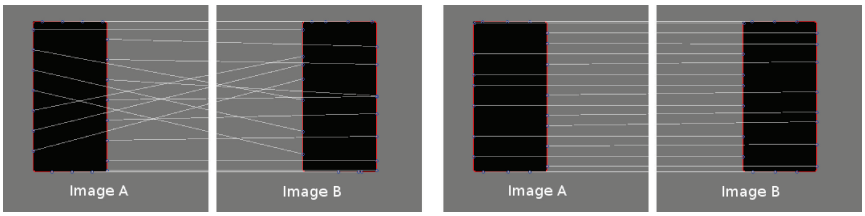


Fig. 7. Local matching minima (left) can be avoided by multiple iterations. In a coarse to fine manner, in each iteration, the number of translets increases, avoiding local matching minima by using the previous result as prior (right).

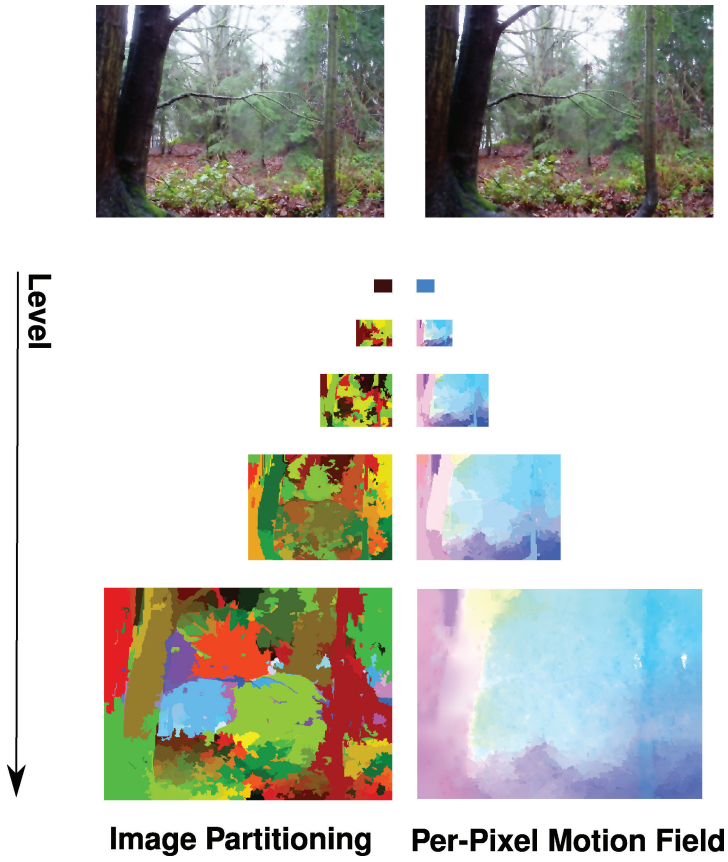


Fig. 8. Solving on small image scales first and upsampling the solution significantly improves quality especially for cluttered scenes. Input images are courtesy of Larry Zitnick [2005].

Additionally, solving on different image resolutions similar to scale-space [Yuille and Poggio 1986] further improves robustness. Especially in cluttered images, the obtained solution significantly profits from first solving on small scale and upsampling the solution to the next level as prior (Figure 8).

In rare cases, some scenes still cannot be matched sufficiently well automatically. For example, when similar structures appear multiple times in the images the matching can get ambiguous and only be addressed by high level reasoning. To resolve this, regions can be selected in both images by the user

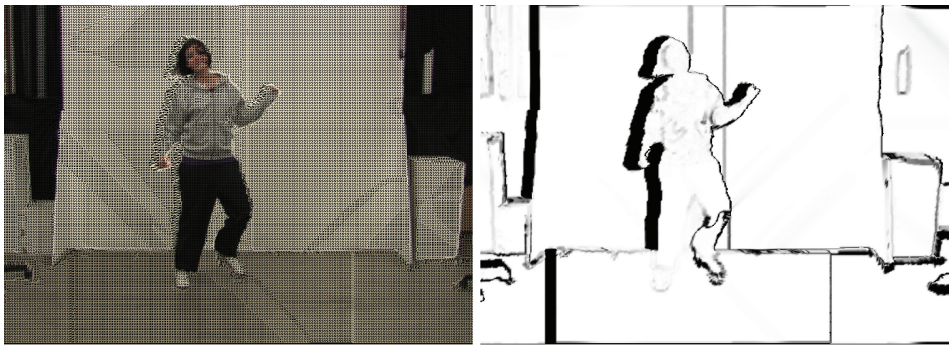


Fig. 9. Left: Per-vertex mesh deformation is used to compute the forward warping of the image, where each pixel corresponds to a vertex in the mesh. The mesh is depicted at coarser resolution for visualization purposes. Right: The connectedness of each pixel used during blending to avoid the, possibly incorrect, influence of missing regions.

and the automatic matching is computed again only for the selected subset of edglets. Due to this restriction of the matching, the correct match is found and used to correct the solution.

6. INTERPOLATION RENDERING

Rendering in-between images is achieved by applying the correspondence field estimated with our image deformation model to the images and blending these warped images. This can be implemented on graphics hardware using per-vertex mesh deformation and alpha blending with real-time rendering performance. To get the deformations for the in-between images, we linearly interpolate the deformation vector field.

6.1 Warping with Occlusions

We implemented the forward warping by a per-vertex deformation of a regular planar triangle mesh of the image plane, where each pixel in the image is represented by a quad with appropriate texture coordinates. Two problems arise with forward warping at motion discontinuities: fold-overs and missing regions.

Fold-overs occur when two or more pixels in the image end up in the same position during warping. This is the case when the foreground occludes parts of the background. Consistent with motion parallax, we assume that the faster moving pixel is closer to the viewpoint to resolve this conflict. When, on the other hand, regions get disoccluded during warping, the information of these regions is missing in the image and must be filled in from the other image. This leaves two options in this case: cutting the mesh at the motion discontinuities before warping, or detecting triangles that span over these discontinuities after rendering. Mark et al. [1997] pointed out that the second approach performs better and proposed a connectedness criterion evaluated on a per-pixel basis after warping. We adapt this measure and compute it directly from the divergence of the deformation vector field such that

$$c_A = 1 - div(d_{AB})^2, \tag{6}$$

with c_A the connectedness and d_{AB} vector field between the images A and B (Figure 9). The connectedness is computed on the GPU during blending to adaptively reduce the alpha values of pixels with low connectedness. Thus, in missing regions, only the image that has the local information has an influence on the rendering result.

In principle, having both images and the motion fields at hand, backward interpolation of the intermediate images would also be possible. However, this would require to splat the motion fields to the

Fig. 10. Jaggy artifacts due to aliasing artifacts can become visible at motion discontinuities. These are easily discriminated by thresholding on the motion field. In a second rendering pass, we correct for previously detected artifacts.

intermediate time, which potentially results in holes in the motion field that have to be filled. Further on, similar depth heuristics have to be applied when motion information is splatted to the same target position. Finally, when interpolating between more than two images, backward interpolation is no longer possible, whereas forward warping extends to this case as well.

6.2 Feathering

At fold-overs, the warped images can have jaggy artifacts due to aliasing problems of the rendering. Opposed to recordings with cameras, rendered pixels at the boundaries are not a mixture of background and foreground color but are either foreground or background color. However, these artifacts occur only at large motion discontinuities, which can be robustly discriminated by the local change in the motion vectors by simple thresholding (Figure 10). In a second rendering pass, we model the color mixing of foreground and background at boundaries using a small selective low-pass filter applied only to the detected motion boundary pixels. This effectively removes the artifacts with a minimal impact on rendering speed and without affecting rendering quality in the nondiscontinuous regions.

7. RESULTS

First, we compared our results to interpolation results based on state-of-the-art optical flow methods using the Middlebury examples and applying their evaluation method, that is, computing interpolation, normalized interpolation and angular errors according to the formulae given in Baker et al. [2007] (Table II, Figure 11).[2] Since these methods do not allow for user interaction, we compare the results of our unimproved automatic results. As shown, our approach is best when looking at the interpolation errors and best or up to par in the sense of the normalized interpolation error. We also like to point out that from a perception point of view, the normalized error is less expressive than the unnormalized error, since discrepancies at edges in the image (e.g., large gradients) are dampened. Interestingly, relatively large angular errors are observed with our method, emphasizing that the requirements of optical flow estimation and image interpolation are different.

[2]While Baker et al. [2007] list six different test sequences. Unfortunately, only the two listed in the table are still publicly available along with ground truth motion fields needed for comparison of the angular error.

Table II. Interpolation, Normalized Interpolation and Angular Errors Computed on the Middlebury Optical Flow Examples by Comparison to Ground Truth with Results Obtained by Our Method and by Other Methods (see Baker et al. [2007])

Venus	Interp.	Norm. Interp.	Ang.
Our Method	**2.88**	**0.55**	**16.24**
Pyramid LK [J.Y. Bouguet 2000]	3.67	0.64	14.61
Bruhn et al. [2005]	3.73	0.63	8.73
Black and Anandan [1996]	3.93	0.64	7.64
Mediaplayer [Microsoft Corporation]	4.54	0.74	15.48
Zitnick et al. [2005]	5.33	0.76	11.42
Brox et al. [2004]	9.16	0.94	6.10
Dimetrodon	Interp.	Norm. Interp.	Ang.
Our Method	**1.78**	**0.62**	**26.36**
Pyramid LK [J.Y. Bouguet 2000]	2.49	0.62	10.27
Bruhn et al. [2005]	2.59	0.63	10.99
Black and Anandan [1996]	2.56	0.62	9.26
Mediaplayer [Microsoft Corporation]	2.68	0.63	15.82
Zitnick et al. [2005]	3.06	0.67	30.10
Brox et al. [2004]	4.61	0.62	5.73

Fig. 11. Results on the Middlebury dataset. Left column: In-between image automatically computed with our method. Right column: Contrast-stretched difference to ground truth.

7.1 Limitations

Our algorithm is limited to scenes for which the decomposition into piecewise planar patches is valid. The current model thus cannot cope with transparencies, shadows or grid-like structures such as a fence. Those entities would have to be modeled by different layers. One further limitation is the

Fig. 12. Illustration of failure cases. The left and right column show the input images, the middle column shows the interpolated frame halfway between the two input frames. Top row: wrong shadow interpolation. Bottom row: failure of depth heuristics.

assumption that faster moving objects are in the foreground, which is used to resolve occlusions during warping. While this assumption works well in practice, it is not true in general and can thus result in artifacts. An example for incorrect shadow interpolation as well as a failure case of the depth heuristics is shown in Figure 12.

8. USER STUDY

In order to assess the perceptual quality of our interpolation algorithm, we ran a validation study that had three major goals:

(1) to quantify changes in perceptual quality introduced by parameter changes *within* the proposed approach to image interpolation,

(2) to compare the results of the proposed algorithm against other approaches to image interpolation,

(3) to investigate whether there would be a perceptual difference between results on real-world and synthetic image material.

8.1 Stimuli

Guided by our goals, we selected a total of 10 different approaches for creating interpolated image sequences. The input consisted of several sequences depicting rotations around objects. From these, we kept every third frame and used the algorithms for interpolating the missing two intermediate frames. For the used scenes, this was the largest gap that the automatic approaches could interpolate with reasonable quality—note also that this corresponds to changes in viewing angle of approximately 10 degrees, on average. We selected predominantly diffuse surfaces, as specularities must be treated as transparent entities to model their motion correctly. The following list describes the algorithms

in more detail[3]:

Original: as the baseline, we compared all algorithms against the original video sequence showing the full, smooth motion
Blend: a simple blending algorithm which creates intermediate frames by blending between two consecutive key-frames
Of_brox: a state-of-the-art optical-flow algorithm [Brox et al. 2004] was used to compute the motion field for the interpolation
Nooptim: our automatically computed initial transformation solution after the second iteration without further optimization of the translets
Optim100: the result of our method including translet optimization but without per-pixel diffusion after the second iteration (see Section 5.2)
Nofeathering: the result of our method but without the feathering at motion discontinuities during rendering (see Section 6.2)
Firstit: the result of our method after the first iteration with optimization of the correspondence field and subsequent diffusion
Full: the result of our proposed automatic perception-based image interpolation algorithm after the second iteration with optimization of the correspondence field and diffusion
Corrected: the result of our automatic approach with additionally manually corrected local errors, as discussed in Section 5.1
Multiscale: the result of the advanced version which seeks correspondences across multiple scale, as discussed in Section 5.5

The first three conditions together with the *full, corrected, multiscale* conditions address the goal of comparing different approaches to interpolation, whereas conditions *firstit, nooptim, optim100, nofeathering* were designed to compare the perceptual quality of different parameter settings.

In order to address our third goal of comparing performance differences of the algorithm on real-world and synthetic images, we used the two different types of scenes shown in Figure 13. The two real-world scenes were recorded by sweeping a handheld camera around a plant and several books on a table. Both scenes were recorded with a digital video camera and contained a considerable amount of jerky motion that could introduce potential problems for automatic correspondence finding. In addition, the plant sequence contains many object boundaries that might make mismatches between figure and ground during the interpolation more visible. The objects in the book sequence in addition contain text elements which also make correspondence artifacts potentially easy to spot. Despite these features, however, the sequences overall contain a lot of irregular texture, which in combination with the camera motion, could mask interpolation artifacts as the motion field gets more complex than can be predicted well by the perceptual system. Four scenes showed computer-generated sequences of objects rotating around the vertical axis for 180 degrees. We hypothesized that due to the regular, noise-free rendering and the smooth rotational motion, any glitches and jitters introduced by the algorithms through interpolation artifacts would be potentially more noticeable in these sequences than in the real-world sequences. Figure 14 shows the visual improvements for different configurations of the proposed algorithm on three of our test scenes.

[3]Compared to the previous article [Stich et al. 2008b], we thus included the new, multiscale method in our validation study and changed the optic-flow algorithm to a state-of-the-art method.

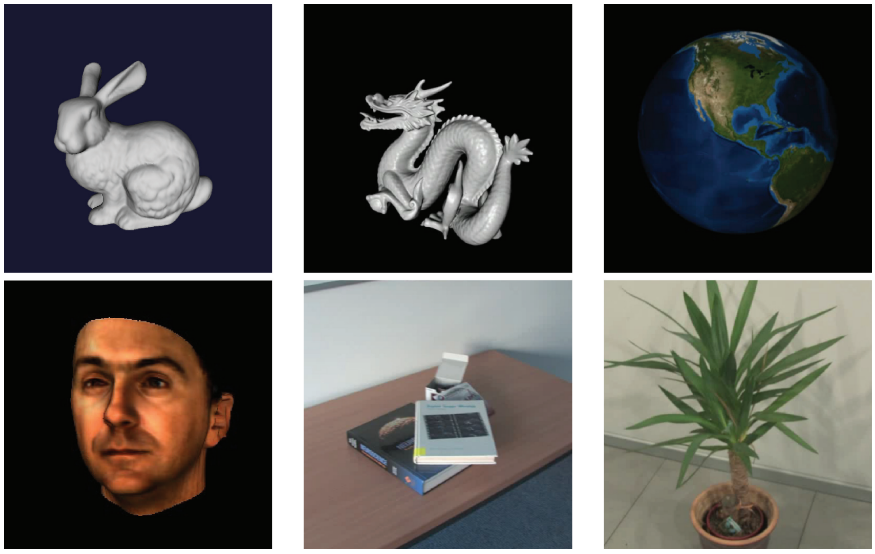


Fig. 13. Images from the different scenes used in our validation study. The first four scenes consist of computer-generated 3D objects, whereas the fifth and the sixth scene were recorded indoors with a standard handheld camera.

8.2 Experimental design

Rather than using a standard rating task in which participants would be shown a sequence and be asked to rate its quality, we opted for a more systematic approach. In the psychophysical study, we used a two-alternative-forced-choice task in which two video sequences were shown successively and participants were asked to indicate which sequence contained more visual artifacts. Such a direct comparison allows for a more fine-grained analysis of the data, as rating tasks are often subject to scaling problems [Wallraven et al. 2007]. For each of the 6 different scenes, we compared all 10 different interpolation algorithms against each other (only doing AB and AA, not BA comparisons), yielding a total of $6 \cdot (9 \cdot \frac{10}{2} + 10) = 330$ trials.

All scenes were rendered at 500 $\times$ 500 pixels with 25 frames per second and were 3 to 5 seconds long. Sequences were presented on a black background on a CRT monitor using a pixel resolution of 1,024 $\times$ 768 at 75Hz. Participants viewed the stimuli at a distance of roughly 50cm while sitting in a dark room. Each trial consisted of a fixation cross shown for 1 second, followed by the first sequence, a second fixation cross for 0.5 seconds, and the second sequence. After this, the screen was blanked and participants were asked to indicate by a key press which sequence contained more visual artifacts. Participants were briefed before the experiment that in this case artifacts were defined as “any visual disturbances resulting in non smooth motion.” Examples of two consecutive frames for different test conditions are given in Figures 17 and 18. All participants completed three test trials before the experiment, which were used to familiarize them with the task. Neither during the test trials nor during the experiment was any feedback given and none of the participants reported any difficulty accomplishing the task. The whole experiment lasted approximately 90 minutes. Our test group consisted of 10 participants who had *none to little* computer graphics-related experience.

8.3 Analysis

For the first analysis, we determined a perceptual quality score for each algorithm by counting how many times it was chosen as producing fewer visual artifacts when compared to one of the other

(a) nooptim

(b) optim100

(c) firstit

(d) full

(e) nooptim

Fig. 14. Visual improvements for different configurations of our algorithm on different scenes. From top to bottom: nooptim, optim100, firstit, full, multiscale.

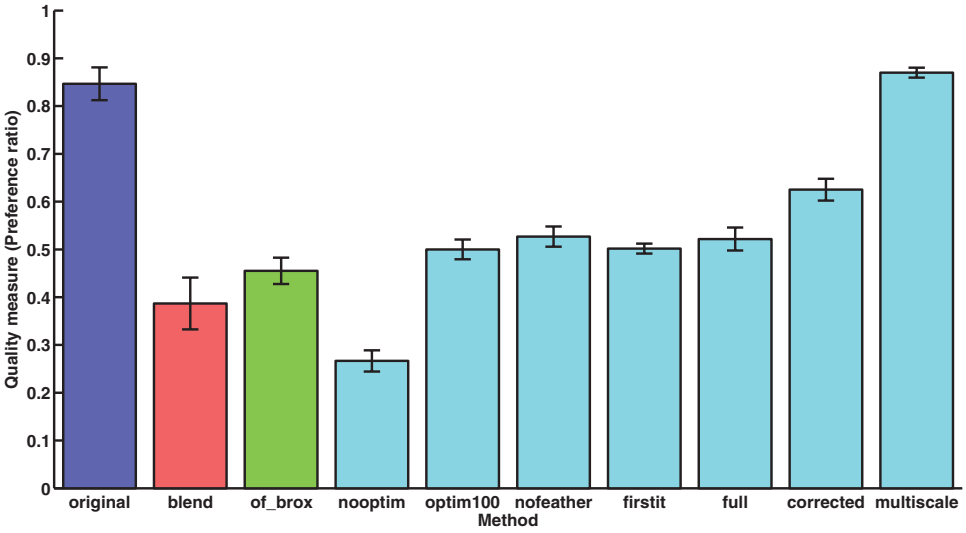


Fig. 15. Perceptual quality scores for ten different test conditions (image interpolation schemes).

algorithms. The normalized scores are shown in Figure 15 for all ten approaches. As can be seen immediately, the approaches vary a lot in terms of their perceptual quality, as confirmed by the highly significant one-way anova ($F(9) = 51.96$, $p < 0.001$). With Tukey-Kramer corrected multiple comparisons, we then identified which of the algorithms differed significantly[4] in their scores which allowed us to address the first two experimental questions.

Original. The original sequences are rated as having the best perceptual quality which can be seen as a sanity check for the experimental instructions.

Blend. Despite the technical simplicity of this condition, the quality score is still reasonably high. Whereas this might be surprising at first glance, the perceptual impression of the resulting motion is that of a jerky, but rather consistent motion. Indeed, several participants noted that while this was a rather noticeable artifact, compared to the other approaches, these errors were more predictable and thus, sometimes, preferable.

Of_brox. Whereas in our previous article the optical flow algorithm fared worst of all tested methods, the advanced implementation used here yields better perceptual quality. Compared to the Horn-Schunck approach [Horn and Schunck 1981], the Brox approach [Brox et al. 2004] handles motion discontinuities much better and thus is able to avoid local instabilities in the computed motion field to a certain degree.

Nooptim. Within the parameter changes of our approach [Stich et al. 2008b], this is the worst condition, and it also comes in last among all tested conditions. As the motion field in this condition is computed from only the matches, this often results in sharp spikes and discontinuity errors. Several participants mentioned in the debriefing that this condition was particularly unpleasant in terms of its artifacts. This finding underlines the importance of increasing the motion coherence during the optimization (see, Section 5.2), which already the improved optical flow method seems to handle much better.

[4]In the following, all p values are smaller than 0.01. The p value is the probability of obtaining a result at least as extreme as the one that was actually observed, given that the null hypothesis is true. This means, for a given significance level, typically 0.05 or even stronger 0.01, the comparison with the p value decides if the null hypothesis can be safely rejected.

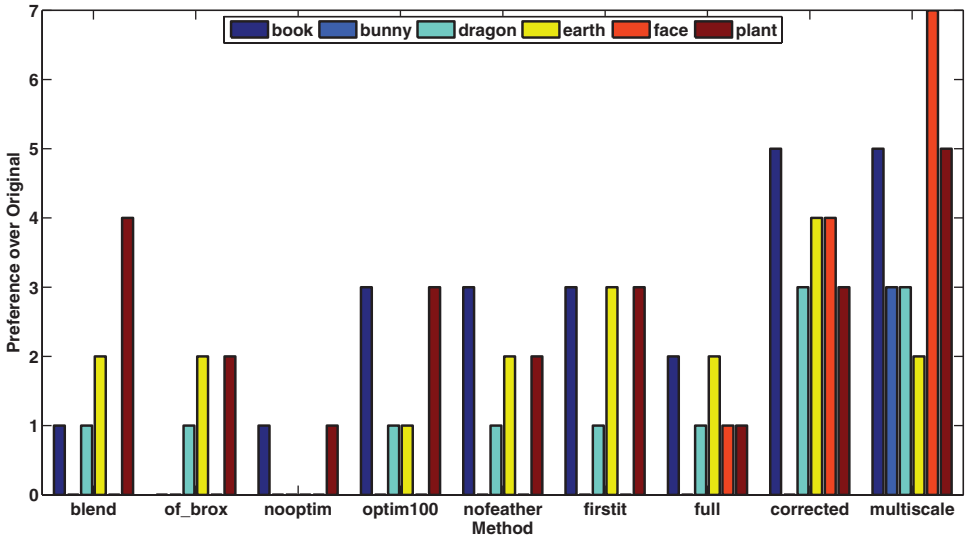


Fig. 16. Preference of corrected over original condition, broken down by test scene. Values around 5 denote that both conditions are of equal perceived quality.

optim100. Compared to the *nooptim* condition, the increase in perceptual quality due to the optimization of the correspondence field demonstrates the importance of producing locally consistent motions for perceptual fidelity.

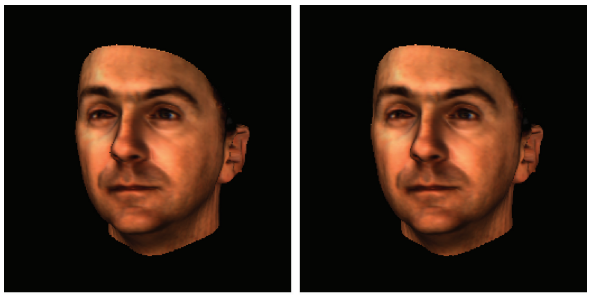
Nofeathering, firstit, full. These three conditions all produce a large increase over the *nooptim* condition without being able to produce a noticeable, perceptual increase as algorithmic features are added. The reason for this is discussed in the following text.

Corrected. Similar to the previous article, the scores show that a small amount of user interaction results in significantly improved perceptual results (see Section 5.1).

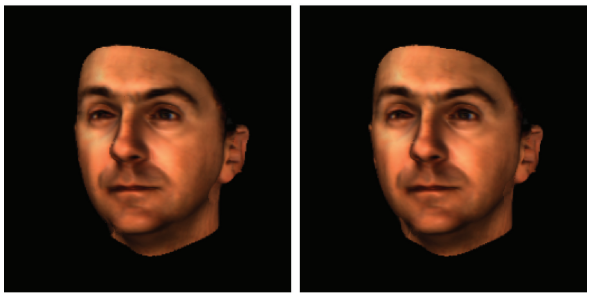
Multiscale. The main result of this experiment, however, is that the improved multiscale matching algorithm has by far the largest perceptual quality score. Most interestingly, this score is on par with the *original* condition, showing that the quality of the interpolated sequences meets real-world standards.

The reason for the nonsignificant differences between the four within-parameter conditions lies less in their perceptual difference (which was demonstrated in the previous study [Stich et al. 2008]), but rather in the fact that the introduction of the *multiscale* condition shifts the balance toward this condition. To illustrate this fact and to further compare the different conditions, we replot the data in Figure 16 to show how often participants would chose any other algorithm over the *original* condition, that is, how many times the perceptual quality of the sequence was at least equally good as in the original, noninterpolated case. This analysis is also broken down by object, which allows us to address the third experimental question of quality differences between real-world and synthetic scenes.

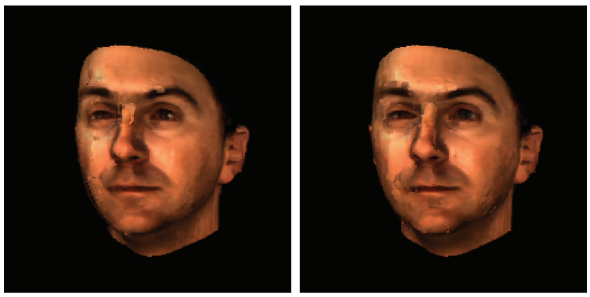
First of all, Figure 16 confirms the results outlined earlier: None of the first seven methods are selected more than one or two times, on average, (out of a maximum of ten) over the original sequence. The only significant contributions are by the *corrected* and *multiscale* conditions with the latter method scoring highest. The resolution of the experiment with only 10 participants in this case is too limited to differentiate between the small improvements introduced by the *nofeather, firstit, full* conditions in



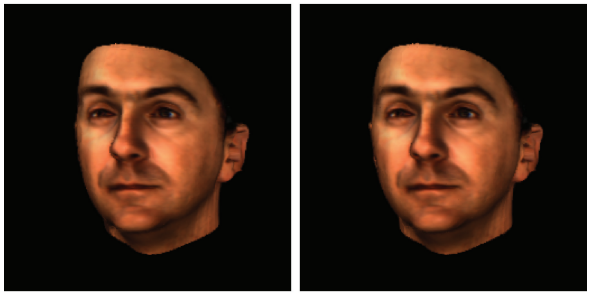
(a) original



(b) opticalflow

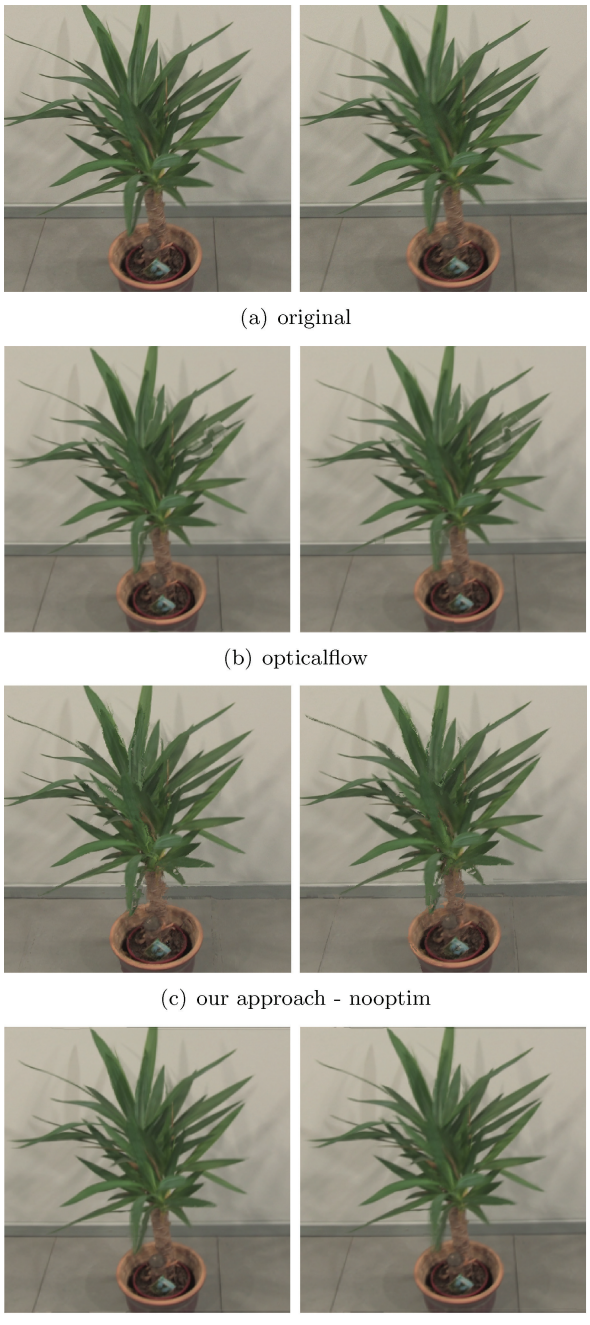


(c) our approach - nooptim



(d) our approach - full

Fig. 17. Two consecutive images from the face scene sequences. (a) ground truth, (b) Brox optical flow, (c) without optimizing motion coherence, (d) automatic interpolation result obtained with our multiscale approach.



(a) original

(b) opticalflow

(c) our approach - nooptim

(d) our approach - full

Fig. 18. Two consecutive images from the plant scene sequences. (a) ground truth, (b) Brox optical flow, (c) without optimizing motion coherence, (d) automatic interpolation result obtained with our multiscale approach.

the presence of the other two, clear winners (see Stich et al. [2008] for an experiment which shows the improvements). Furthermore, the plot already suggests a difference between the real-world and artificial sequences: it seems that the interpolated real-world sequences (book and plant) are confused more often with the original sequences than artificial sequences. This is confirmed by a posthoc, two-tailed t-test comparing the two types of sequences (corrected for unequal variance as the group sizes are different): On average, for real-world data, interpolated sequences are confused with the original sequence 2.5 times, versus only 1.3 times for artificial sequences ($t(36) = 3.07$, $p < 0.01$). Finally, we note that within the artificial stimuli, the criteria for a successful interpolation seemed to be most stringent for the face and bunny sequences, as these only gained significant votes for the *corrected* and *multiscale* conditions.

Taken together, these results have shown that our proposed approach to image interpolation produces perceptually plausible, high-quality interpolations, which approaches real-world fidelity in the case of the latest improved version. Whereas there is still some room for improvement in some cases—especially for identifying invalid correspondences and improving robustness against outliers—the quality of the sequences is surprisingly good given that no prior knowledge about camera calibrations, scene geometry, or object identity was used. Additionally, the results confirm and extend the perceptual approach to computer graphics—that our visual system has evolved to deal with natural *image statistics* (things tend to move smoothly; objects have well-defined, stable boundaries, etc.) rather than to explicitly and accurately reconstruct the 3D world from visual input (simple image morphing can be enough). Finally, the results have shown that whatever small artifacts are produced by the interpolation algorithms, they are best hidden in natural, complex scenes, suggesting that there is a limit to the complexity of the motion field within which the perceptual system can reliably predict the motion of all scene elements and depth discontinuities.

9. SUMMARY

In this article, we presented an interpolation method that can be related to both human motion perception and relaxed geometrical constraints. This method has several advantages over physically motivated approaches. First, by focusing on the properties important to visual motion perception, we solve for the better conditioned problem of computing images that are visually convincing, circumventing the often ill-posed problem of computing the physically correct interpolation. Second, motion discontinuities are handled correctly by our approach without the need for high-level information such as layers or figure-ground segmentation.

In a user study, we validated the overall visual quality of our results and evaluated the contribution of each part of our method. Perceived quality of the results significantly improved by optimizing motion coherence while correctly handling motion discontinuities. In comparison to the rated quality of the original sequences (ground truth), our achieved visual quality is already quite close and outperforms other tested approaches. Surprisingly, for the book and the plant sequence, subjects could not decide if ground truth or our result is better (preference was at 50%) and they even preferred our result over the ground truth for the face sequence, which presumably hints at an interference with the “Uncanny Valley” phenomenon.

Put into context, the ability to create high-quality image interpolations is beneficial to a wide field of interesting applications: visual effects created with standard cameras, historic movies improved in quality by increasing frame-rate to modern standards, and new possibilities to create stimuli for psychological questionnaires. By looking at concepts of human vision, we can identify what is necessary to make the human observer accept the results as physically correct and find new algorithms that are inspired by perception to compute such solutions.

REFERENCES

ADELSON, E. AND BERGEN, J. 1991. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, M. Landy and J. Movshon Eds., MIT Press, Cambridge, MA, 3–20.

BAKER, S. AND MATTHEWS, I. 2004. Lucas-Kanade 20 years on: A unifying framework. *Int. J. Comput. Vision 56,* 3, 221–255.

BAKER, S., SCHARSTEIN, D., LEWIS, J., TH, S. R., BLACK, M., AND SZELISKI, R. 2007. A database and evaluation methodology for optical flow. In *Proceedings of the International Conference on Computer Vision (ICCV '07)*. IEEE, Los Alamitos, CA, 1–8.

BARRON, J., FLEET, D., AND BEAUCHEMIN, S. 1994. Performance of optical flow techniques. *Int. J. Comput. Vision 12,* 1, 43–77.

BEIER, T. AND NEELY, S. 1992. Feature-based image metamorphosis. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. ACM, New York, 35–42.

BELONGIE, S., MALIK, J., AND PUZICHA, J. 2001. Matching shapes. In *Proceedings of the International Conference on Computer Vision (ICCV'01)*. IEEE, Los Alamitos, CA, 454–461.

BERTSEKAS, D. 1992. Auction algorithms for network flow problems: A tutorial introduction. *Comput. Optim. Appl. 1*, 7–66.

BLACK, M. J. AND ANANDAN, P. 1996. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Comput. Vision Image Understand. 63,* 1, 75–104.

BOUGUET, J. Y.. 2000. Pyramidal implementation of the lucas-kanade feature tracker: Description of the algorithm. Tech. rep., Intel Microprocessor Research Labs.

BROX, T., BRUHN, A., PAPENBERG, N., AND WEICKERT, J. 2004. High accuracy optical flow estimation based on a theory for warping. In *Proceedings of the European Conference on Computer Vision (ECCV'04)*. Springer-Verlag, Berlin, 25–36.

BRUHN, A., WEICKERT, J., AND SCHNÖRR, C. 2005. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *Int. J. Comput. Vision 61,* 3, 211–231.

BUEHLER, C., BOSSE, M., MCMILLAN, L., GORTLER, S., AND COHEN, M. 2001. Unstructured Lumigraph Rendering. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'01)*. ACM, New York, 425–432.

CANNY, J. 1986. A computational approach to edge detection. *Trans. Patt. Anal. Mach. Intell. 8*, 679–714.

CARRANZA, J., THEOBALT, C., MAGNOR, M., AND SEIDEL, H. P. 2003. Free-viewpoint video Of human actors. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'03)*. ACM, New York, 569–577.

CHEN, S. AND WILLIAMS, L. 1993. View interpolation for image synthesis. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'93)*. ACM, New York, 279–288.

COOKE, T., CUNNINGHAM, D., AND BÜLTHOFF, H. 2004. The perceptual influence of spatiotemporal noise on the reconstruction of shape from dynamic occlusion. In *Proceedings of the 26th DAGM Symposium on Pattern Recognition*. Springer-Verlag, Berlin, 407–414.

CUNNINGHAM, D., SHIPLEY, T., AND KELLMAN, P. 1998. The dynamic specification of surfaces and boundaries. *Perception 27,* 4, 403–416.

DEBEVEC, P., BORSHUKOV, G., AND YU, Y. 1998. Efficient view-dependent image-based rendering with projective texture-mapping. In *Proceedings of the Eurographics Rendering Workshop (EGRW '98)*. Springer-Verlag, Berlin, 105–116.

EXNER, S. 1894. *Entwurf zu einer physiologischen Erklärung der psychischen Erscheinungen*. F. Deuticke, Leipzig Wien.

FELZENSZWALB, P. AND HUTTENLOCHER, D. 2004. Efficient graph-based image segmentation. *Int. J. Comput. Vision 59*, 167–181.

FU, M. F., AU, O., AND CHAN, W. C. 2002. Temporal interpolation using wavelet domain motion estimation and motion compensation. In *Proceedings of the International Conference on Image Processing (ICIP '02)*. IEEE, Los Alamtios, CA, 393–396.

GIBSON, J. 1955. *The Perception of the Visual World*. Cambridge University Press, Cambridg, UK.

GIBSON, J., KAPLAN, G., JR., H. R., AND WHEELER, K. 1969. The change from visible to invisible: a study of optic transitions. *Percept. Psychophys. 5,* 2, 116–131.

GORTLER, S., GRZESZCZUK, R., SZELISKI, R., AND COHEN, M. 1996. The Lumigraph. In *Proceedings Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'96)*. ACM, New York, 43–54.

HARTLEY, R. AND ZISSERMAN, H. 2000. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK.

HEEGER, D., BOYNTON, G., DEMB, J., SEIDEMANN, E., AND NEWSOME, W. 1999. Motion opponency in visual cortex. *J. Neurosci. 19*, 7162–7174.

HORN, B. AND SCHUNCK, B. 1981. Determining optical flow. *Artif. Intell. 17*, 185–203.

HUBEL, D. 1995. *Eye, Brain, and Vision* 2nd Ed. W. H. Freeman, New York.

ISAKSEN, A., MCMILLAN, L., AND GORTLER, S. 2000. Dynamically reparameterized light fields. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'00)*. ACM, New York, 297–306.

JAIN, J. AND JAIN, A. 1981. Displacement measurement and its application in interframe image coding. *IEEE Trans. Comm. 29,* 12, 1799–1808.

KELLMAN, P. J. AND SHIPLEY, T. F. 1992. Perceiving objects across gaps in space and time. *Curr. Directions Psych. Science 1,* 6, 193–199.

LEVOY, M. AND HANRAHAN, P. 1996. Light Field Rendering. In *Proceedings Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. 31–42.

LUCAS, B. AND KANADE, T. 1981. "an iterative image registration technique with an application to stereo vision". In *Proceedings of the International Joint Conference on Artificial Intelligence*. 674–679.

MARK, W., MCMILLAN, L., AND BISHOP, G. 1997. Post-Rendering 3D Warping. In *Proceedings of the Symposium on Interactive 3D Graphics*. 7–16.

MARR, D. 1983. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman.

MATUSIK, W., BUEHLER, C., RASKAR, R., GORTLER, S., AND MCMILLAN, L. 2000. Image-Based Visual Hulls. In *Proceedings Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. 369–374.

MATUSIK, W. AND PFISTER, H. 2004. 3D TV: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. In *Proceedings Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'04)*. ACM, New York 814–824.

MCMILLAN, L. AND BISHOP, G. 1995. Plenoptic modeling: an image-based rendering system. In *Proceedings Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '95)*. ACM, New York, 39–46.

MICROSOFT CORPORATION. Media player 9 video quality demos. http://www.microsoft.com/windows/windowsmedia/demos/video quality demos.aspx.

O'SULLIVAN, C., HOWLETT., S., MCDONNELL, R., MORVAN, Y., AND O'CONOR, K. 2004. Perceptually adaptive graphics. In *Eurographics, State-of-the-art-Report 6*.

PERONA, P. AND MALIK, J. 1990. Scale-space and edge detection using anisotropic diffusion. *Trans. Patt. Anal. Mach. Intell. 12,* 7, 629–639.

QIAN, N. AND ANDERSEN, R. 1997. A physiological model for motion-stereo integration and a unified explanation of Pulfrich-like phenomena. *Vision Res. 37*, 1683–1698.

RAMANARAYANAN, G., BALA, K., AND FERWERDA, J. 2008. Perception of complex aggregates. In *Proceedings Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'08)*. ACM, New York, 1–10.

RAMANARAYANAN, G., FERWERDA, J., WALTER, B., AND BALA, K. 2007. Visual equivalence: Towards a new standard for image fidelity. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'07)*. ACM, New York, 654–663.

REICHARDT, W. 1961. Autocorrelation, A principle for the evaluation of sensory information by the central nervous system. In *Sensory communication*, W. Rosenblith Ed., MIT Press-Wiley, New York, 303–317.

RUZON, M. AND TOMASI, C. 1999. Color edge detection with the compass operator. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'99)*. IEEE, Los Alamitos, 160–166.

SCHAEFER, S., MCPHAIL, T., AND WARREN, J. 2006. Image deformation using moving least squares. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'06)*. ACM, New York, 533–540.

SEITZ, S. AND DYER, C. 1996. View morphing. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'96)*. ACM, New York, 21–30.

SHIPLEY, T. AND CUNNINGHAM, D. 2001. Perception of occluding and occluded objects over time: Spatiotemporal segmentation and unit formation. In *From Fragments to Object: Segmentation and Grouping in Vision.* Elsevier Science Ltd., Oxford, UK, 557–585.

SNAVELY, N., SEITZ, S., AND SZELISKI, R. 2006. Photo tourism: Exploring photo collections in 3D. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '06)*. ACM, New York, 835–846.

STICH, T., LINZ, C., ALBUQUERQUE, G., AND MAGNOR, M. 2008a. View and time interpolation in image space. *Comput. Graphics Forum 27,* 7, 1781–1787.

STICH, T., LINZ, C., WALLRAVEN, C., CUNNINGHAM, D., AND MAGNOR, M. 2008b. Perceptionmotivated interpolation of image sequences. In *Proceedings of the Symposium on Applied Perception in Graphics and Visualization (APGV'08)*. ACM, New York, 97–106.

TOURAPIS, A., CHEONG, H., LIOU, M., AND AU, O. 2001. Temporal interpolation of video sequences using zonal based algorithms. In *Proceedings of the International Conference on Image Processing (ICIP '01)*. IEEE, Los Alamitos, CA, 895–898.

VAN SANTEN, J. P. H. AND SPERLING, G. 1985. Elaborated reichardt detectors. *J. Opt. Soc. Am. A, 2*, 300–320.

VANGORP, P. AND DUTRÉ, P. 2008. Shape-dependent gloss correction. In *Proceedings of the Symposium on Applied Perception in Graphics and Visualization (APGV'08)*. ACM, New York, 123–130.

VANGORP, P., LAURIJSSEN, J., AND DUTRÉ, P. 2007. The influence of shape on the perception of material reflectance. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'07)*. ACM, New York, 1–9.

VEDULA, S., BAKER, S., AND KANADE, T. 2005. Image based spatio-temporal modeling and view interpolation of dynamic events. *ACM Trans. Graphics 24,* 2, 240–261.

WALLACH, H. 1935. Ueber visuell wahrgenommene Bewegungsrichtung. *Psychologische Forschung 20*, 325–380.

WALLRAVEN, C., BÜLTHOFF, H. H., FISCHER, J., CUNNINGHAM, D. W., AND BARTZ, D. 2007. The evaluation of real-world and computer-generated stylized facial expressions. *ACM Trans. Appl. Percept. 4,* 3, 1–24.

WERTHEIMER, M. 1938. Laws of organization in perceptual forms. In *A Source Book of Gestalt Psychology*, W. Ellis, Ed. Kegan Paul, Trench, Trubner & Co. Ltd., London, 71–88.

WOLBERG, G. 1998. Image warping: a survey. *Visual Comput. 14*, 360–372.

WOOD, D., AZUMA, D., ALDINGER, K., CURLESS, B., DUCHAMP, T., SALESIN, D., AND STUETZLE, W. 2000. Surface light fields for 3D photography. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'00)*. ACM, New York, 287–296.

YUILLE, A. L. AND POGGIO, T. A. 1986. Scaling theorems for zero crossings. *IEEE Trans. Pattern Anal. Mach. Intell. 8,* 1, 15–25.

ZITNICK, C., JOJIC, N., AND KANG, S. B. 2005. Consistent segmentation for optical flow estimation. In *Proceedings of the International Conference on Computer Vision (ICCV '05) 2*, 1308–1315.

ZITNICK, C., KANG, S., UYTTENDAELE, M., WINDER, S., AND SZELISKI, R. 2004. High-quality video view interpolation using a layered representation. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'04)*. ACM, New York, 600–608.

Received February 2009; revised February 2010; accepted April 2010