A transformation-aware perceptual image metric

Petr Kellnhofer[a] Tobias Ritschel[a,b] Karol Myszkowski[a] Hans-Peter Seidel[a]

[a]Max-Planck-Institut für Informatik, Campus E1.4, Saarbrücken, Germany;
[b]Saarland University, Uni-Campus Nord, Saarbrücken, Germany

ABSTRACT

Predicting human visual perception has several applications such as compression, rendering, editing and retargeting. Current approaches however, ignore the fact that the human visual system compensates for geometric transformations, e. g., we see that an image and a rotated copy are identical. Instead, they will report a large, false-positive difference. At the same time, if the transformations become too strong or too spatially incoherent, comparing two images indeed gets increasingly difficult. Between these two extrema, we propose a system to quantify the effect of transformations, not only on the perception of image differences, but also on saliency. To this end, we first fit local homographies to a given optical flow field and then convert this field into a field of elementary transformations such as translation, rotation, scaling, and perspective. We conduct a perceptual experiment quantifying the increase of difficulty when compensating for elementary transformations. Transformation entropy is proposed as a novel measure of complexity in a flow field. This representation is then used for applications, such as comparison of non-aligned images, where transformations cause threshold elevation, and detection of salient transformations.

Keywords: Image metric, Motion, Optical flow, Homography, Saliency

1. INTRODUCTION

Models of human visual perception are an important component of image compression, rendering, retargeting and editing. Two typical applications are predicting if two images are perceived differently and detecting if a part of an image is salient. Such predictions are based on the perception of luminance patterns alone and ignore that a difference might also be well-explained by a transformation. As an example, the *Hamming distance* of the binary strings 1010 and 0101 is the same as between 1111 and 0000, however, the first pair is more similar in the sense of an *edit* distance, as 1010 is just a rotated i. e., transformed version of 0101. We apply this idea to images, e. g., comparing an image and its rotated copy.

In current models of visual perception, transformation is not represented, leading to several difficulties: For image similarity or quality evaluation approaches, it is typically assumed the image pair is perfectly aligned (registered), which is directly granted in image compression, restoration, denoising, broadcasting, and rendering. However, in many other applications such as visual equivalence judgement,[1] comparison of rendered and photographed scenes,[2] re-photography,[3] or image retargeting,[4] the similarity of images should be judged in the presence of distortions caused by transformations. Ecologically valid transformation[5] is a nonstructural distortion[6] and as such should be separated from others. However, current image difference metrics will report images that differ by such a transformation to be very dissimilar.[6] In the same vein, computational models of image saliency are based on luminance alone, or in the case of video, on the principle that motion has a "pop-up" effect.[7] However, for an image pair that differs by a spatially varying transformation some transformations might be more salient, not because they are stronger, but because they are distinct from others. We will show that all the difficulties in predicting the perception of transformed images can be overcome by an explicit model of human perception of transformations such as we propose.

In this work we assume the optical flow[5] of an image pair to be given, either by producing it using 3D graphics or (typically with a lower precision) using computer vision techniques and focus on how the human visual system represents transformations. We decompose the flow field into a field of elementary transformations,[8] a process which is likely to also happen in the dorsal visual pathway of the primate brain.[9] From this representation, we can model the effect of transformations on the perception of images. For comparing images, strong or incoherent transformations generally make the perception of differences increasingly difficult. We model this effect using a novel measure of transformation entropy. When given an image pair that differs by a transformation, we predict where humans will perceive differences and where not

Further author information: P. Kellnhofer: E-mail: pkellnho@mpi-inf.mpg.de

Human Vision and Electronic Imaging XX, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas, Huib de Ridder,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 9394, 939408 · © 2015 SPIE-IS&T
CCC code: 0277-786X/15/$18 · doi: 10.1117/12.2076754

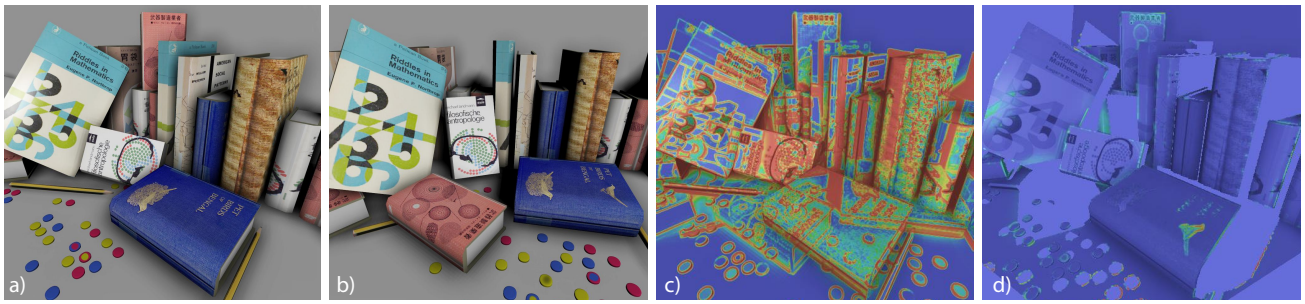

Figure 1: Given input image *(a)* that underwent deformations producing image *(b)*, common perceptual image metrics report unusable results *(c)* as they do not account for the human visual system's ability to compensate for transformations. Our transformation-aware approach models the ability to compensate for transformations *(d)* and its limitations when transformations are too strong (red book) or too incoherent (chips).

(Fig. 1). Using our representation, we can compare transformations and predict which transformations are salient compared to others.

In this work we make the following contributions:

- A perceptually motivated decomposition of optical flow
- A transformation-aware image difference metric
- Prediction of transformation saliency.

2. BACKGROUND

In this section we review the perceptual background of our approach. We will recall the idea of mental transformation and its relation to optical flow, saliency, as well as the basics of entropy in human psychology. The discussion of previous work for the two main applications we propose (image differences, saliency), is found in Sec. 4.

2.1 Mental transformation

Mental transformations of space play an important role in everyday tasks such as object recognition, spatial orientation, and motion planning. Such transformations involve both objects in the space as well as the egocentric observer position. Mental rotation is the best understood mental transformation,[10] where the time required to judge the similarity between a pair of differently oriented objects of arbitrary shape is proportional to the rotation angle both for 2D (image plane) and 3D rotations, irrespectively of the chosen rotation axis. Similar observations have been made for the size scaling and translation (in depth),[11] where the reaction time is shorter than for rotation. Moreover, in combined scaling and rotation[12] as well as translation and rotation[11] the response time is additive with respect to each component transformation. This may suggest that there are independent routines for such elementary transformations, which jointly form a procedure for handling any sort of movement that preserves the rigid structure of an object.[11] Another observation is that the mental transformation passes through a continuous trajectory of intermediate object positions, not just the beginning and end positions.[13]

A more advanced mental transformation is perspective transformation.[14] From own experience we know that observing a cinema screen from a moderately off-angle perspective does not reduce perceived quality, even if the retinal image underwent a strong transformation. One explanation for this ability is that humans compensate for the perspective transformation by mentally inverting it.[15]

Apparent motion in the forward and backward directions is induced when two 3D-transformed (e. g., rotated) copies of the same object are presented alternatively at proper rates. As the transformational distance (e. g., rotation angle) increases the alternation rate must be reduced to maintain the motion illusion. Again, this observation strongly suggests that the underlying transformations require time to go through intermediate stages and internally 3D representation is utilized,[16] and elementary transformations are individually sequential-additive.[17]

The human visual system (HVS) is able to recover depth and rigid 3D structure from two views (e. g., binocular vision, apparent motion) irrespectively whether the perspective or orthographic projection is used, and adding more views has little impact.[18] This indicates that the HVS might use some perceptual heuristics to derive such information as the structure-from-motion theorem stipulates that at least three views are needed in the case of orthographic projection (or under weak perspective).[19]

The 3D internal representation in the HVS and the rigidity hypothesis in correspondence finding, while tracking moving objects, is still a matter of scientific debate. Eagle et al.[20] have found a preference towards translation in explaining competing motion transformations in a two-frame sequence with little regard for the projective shape transformations.

2.2 Optic flow

The idea of optical flow dates back to Gibson[5] and has become an essential part of computer vision and graphics where it is mostly formalized as a 2D vector field that maps locations in one image to locations in a second image, taken from a different point in time or space. Beyond the mapping from points to points, Koenderink[8] conducted a theoretical analysis of elementary transformations such as expansion/contraction (radial motion), rotation (circular motion), and sheer (two-component deformation), which can be combined with translation into a general affine transformation. Such transformations map differential area to differential area. Electro-physiological recordings have shown that specialized cells in the primate brain are selective for each elementary transformation component alone or combined with translation[9] (refer also to [21, Ch. 5.8.4]). A spatially varying optical flow field does not imply a spatially varying field of transformations: a global rotation that has small displacements in the center and larger displacements in the periphery can serve as an example. For this reason, our perceptual model operates on a field of elementary transformations computed from homographies instead of a dense optical flow. Homography estimation is commonly used in the video-based scene 3D analysis and best results are obtained when multiple views are considered.[19]

In computer graphics, the use of elementary transformation fields is rare, with the exception of video stabilization and shape modeling. In video stabilization, spatially-varying warps of hand-held video frames into their stabilized version are performed with a desired camera path. Typically a globally reconstructed homography is applied to the input frame, before the optimization-driven local warping is performed,[22] which is conceptually similar to our local homography decomposition step (Sec. 3.2). Notably, the concept of subspace stabilization[23] constructs a lower-dimensional subspace of the 2D flow field, that is, a space with a lower number of different flows, i. e., lower entropy. In shape modeling, flow fields are decomposed into elementary transformations to remove all but the desired transformations, i. e., to remain as-rigid-as-possible when seeking to preserve only rotation.[24]

2.3 Visual attention

Moving objects and "pop-out" effects are strong attractors of visual attention.[7] The classic visual attention model proposed by Itti et al.[25] apart from the common neuronal features such as intensity contrast, color contrast, and pattern orientation can handle also four oriented motion energies (up, down, left, right). Differently, in our work, we detect saliency of motion which pops out not just because it is present and the rest is static, but because it is different from other motion in the scene, such as many rotating objects where one rotates differently. As humans understand motion in form of elementary transformations,[9] our analysis is needed to find those differences.

2.4 Entropy

Information entropy is a measure of complexity in the sense of how much a signal is expected or not.[26] If it is expected, the entropy is low, otherwise it is high. In our approach we are interested in the entropy of transformations, which tells apart a uniform transformations from incoherent ones, such as disassembling a puzzle. Assembling the puzzle is hard, not because the transformation is large, but because it is incoherent, i. e., it has a high entropy. This view is supported by studies of human task performance:[27,28] Sorting cards with a low entropy layout can be performed faster than sorting with high entropy. In computer graphics, entropy of luminance is used for the purpose of alignment,[29] best-view selection,[30] light source placement,[31] and feature detection, but was not yet applied to transformations.

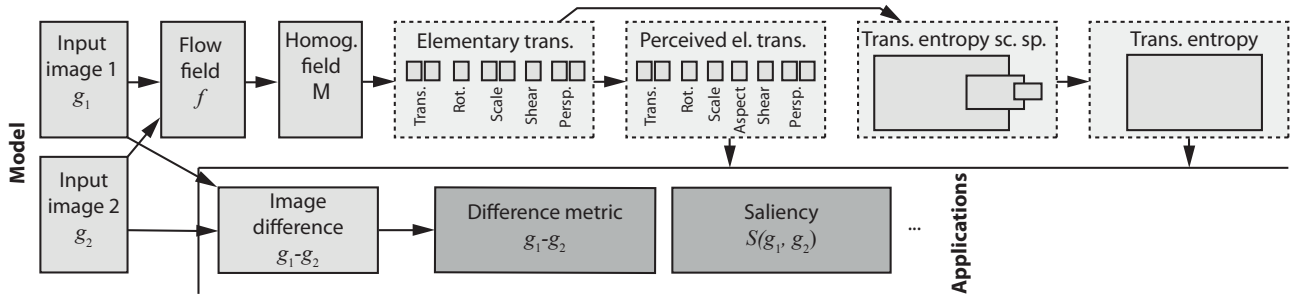

Figure 2: Flow of our approach *(Please see text)*.

3. OUR APPROACH

3.1 Overview

Our system consists of two layers (Fig. 2): A *model* layer described in this section and an *application* layer, described in Sec. 4. Input to the model layer are two images where the second image differs from the first one by a known mapping which is assumed to be available as a spatially dense optical flow field. This requires either to use optical flow algorithms that support large displacements[32] and complex mappings[23] or to use computer-generated optical flow (a.k.a. motion field). Output of our method is a field of perceptually scaled elementary transformations and a field of transformation entropy ready to be used in different applications.

Our approach proceeds as follows (Fig. 2). In the first step (Sec. 3.2), we convert the optical flow field, that maps positions to positions, into an over-complete field of local homographies, describing how a differential patch from one image is mapped to the other image. While classic flow only captures translation, the field of homographies also captures effects such as rotation, scaling, shear and perspective. Next, we factor the local homographies into "elementary" translation, scaling, rotation, shear and perspective transformations (Sec. 3.3). Also, we compute the local entropy of the transformation field, i. e., how difficult it is to understand the transformation (Sec. 3.4). Finally, the magnitude of elementary transformations is mapped to scalar perceptual units, such that the same value indicates roughly the same sensitivity (Sec. 3.5).

Using the information above allows for several applications. Most importantly, we propose an image difference metric (Sec. 4.1) that is transformation-aware. We model the threshold elevation due to transformation strength and entropy. The second application is a visual attention model, that can detect what transformations are salient (Sec. 4.2).

3.2 Homography estimation

Input is two images with luminances g_1 and $g_2(\mathbf{x}) \in \mathbb{R}^2 \to \mathbb{R}$ and $\mathbf{x}$ as spatial location as well as a flow $f(\mathbf{x}) \in \mathbb{R}^2 \to \mathbb{R}^2$ from g_1 to g_2. First, the flow field is converted into a field of homography transformations.[19] A homography maps a differential image patch into the second image while optical flow maps single pixel positions to other pixel positions. In human vision research this Helmholtz decomposition was proposed conceptually by Koenderink[8] and later confirmed by physiological evidence.[9] Examples of homographies are shown in Fig. 3, left. In our case homographies are two-dimensional projective 3×3 matrices. While 2×3 matrices can express translation, rotation and scaling, the perspective component allows for perspective foreshortening.

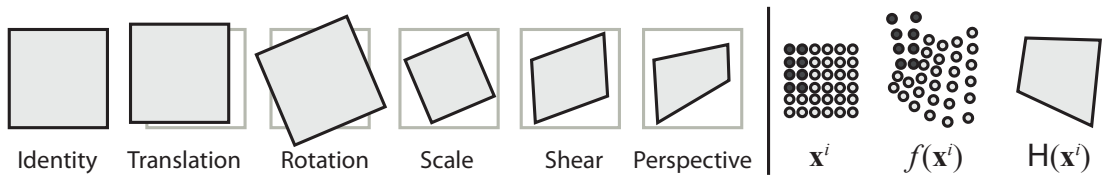

Figure 3: The effect of identity, translation, rotation, scale, shear and perspective transformations applied to a quad *(Left)*. Edge-aware moving least-squares estimation of a homography $\mathsf{M}(\mathbf{x})$ from a set of points $\mathbf{x}^i$ undergoing a flow f. Note how pixels from different image content *(dark pixels)* that undergo a different transformation are not affecting the estimation.

We estimate a field of homographies, i. e., a map that describes for every pixel where its surrounding patch is going. We compute this field $\mathsf{M}(\mathbf{x}) \in \mathbb{R}^2 \rightarrow \mathbb{R}^{3\times 3}$ by solving an edge-aware moving least-squares problem for every pixel using normalized eight-point algorithm.[33] The best transformation $\mathsf{M}(\mathbf{x})$ in the least squares sense minimizes

$$\int_{\mathbb{R}^2} w(\mathbf{x},\mathbf{y}) \left|\left| f(\mathbf{y}) - \phi\left(\mathsf{M}(\mathbf{x}) \begin{pmatrix}\mathbf{y}\\ 1\end{pmatrix}\right)\right|\right|_2^2 \mathrm{d}\mathbf{y}, \tag{1}$$

where $\phi(\mathbf{v}) = (v_1/v_3, v_2/v_3)^\mathsf{T}$ is a homogeneous projection and

$$w(\mathbf{x},\mathbf{y}) = \exp(-||\mathbf{x}-\mathbf{y}||_2^2/\sigma_\mathrm{d}) \exp(-||f(\mathbf{x}) - f(\mathbf{y})||_2^2/\sigma_\mathrm{r})$$

is a bilateral weight function[34] that accounts more for locations that are spatially close (domain weight) and have a similar flow (range weight). The parameters σ_r and σ_d control the locality of the weight. The range-weighting assures to not mix different flows into one wrong estimate of the homography, but to keep them separate (Fig. 3, right) resulting in a pixel-accurate, edge-aware field.

For one flow direction $\mathbf{f}$ at position $\mathbf{y}$ and a matrix M, we see, that

$$\mathbf{f} - \phi\begin{pmatrix} y_1 m_{11} + y_2 m_{12} + m_{13} \\ y_1 m_{21} + y_2 m_{22} + m_{23} \\ y_1 m_{31} + y_2 m_{32} + m_{33} \end{pmatrix} =$$
$$\mathbf{f} - \begin{pmatrix} (y_1 m_{11} + y_2 m_{12} + m_{13})/(y_1 m_{31} + y_2 m_{32} + m_{33}) \\ (y_1 m_{21} + y_2 m_{22} + m_{23})/(y_1 m_{31} + y_2 m_{32} + m_{33}) \end{pmatrix}$$

which leads to two equations linear in the elements of M

$$y_1 f_1 m_{31} + y_2 f_1 m_{32} + f_1 m_{33} - y_1 m_{11} - y_2 m_{12} - m_{13} = 0$$
$$y_1 f_2 m_{31} + y_2 f_2 m_{32} + f_2 m_{33} - y_1 m_{21} - y_2 m_{22} - m_{23} = 0.$$

We write $\mathbf{a}_1^\mathsf{T}\mathbf{m} = 0$ and $\mathbf{a}_2^\mathsf{T}\mathbf{m} = 0$ with

$$\mathbf{m} = (m_{11}, m_{12}, m_{13}, m_{21}, m_{22}, m_{23}, m_{31}, m_{32}, m_{33})$$
$$\mathbf{a}_1 = (-y_1, -y_2, -1, 0, 0, 0, f_1 y_1, f_1 y_2, f_1)^\mathsf{T}$$
$$\mathbf{a}_2 = (0, 0, 0, -y_1, -y_2, -1, f_2 y_1, f_2 y_2, f_2)^\mathsf{T}.$$

In the discrete case of Eq. 1, for pixel $\mathbf{x}$ we find one M that minimizes

$$\sum_{i\in\mathcal{N}} w_i \left|\left| \mathbf{f}^i - \phi\left(\mathsf{M}\begin{pmatrix}\mathbf{y}^i\\ 1\end{pmatrix}\right)\right|\right|_2^2, \tag{2}$$

where $\mathcal{N}$ is a 5×5 neighborhood around pixel location $\mathbf{x}$, and $\mathbf{f}^i$ and w_i are the flow and the bilateral weight of neighbor pixel i. For every neighbor i, we compute a vector $\mathbf{a}^i_{\{1,2\}}, i \in (1, |\mathcal{N}|)$. Let $\mathbf{b}^i_{\{1,2\}} = \mathbf{w}_i \mathbf{a}^i_{\{1,2\}}$ be a weighted version of the error vector and B the 9×50 matrix that stacks all those error vectors $\mathbf{b}^i_{\{1,2\}}$. Finally, the homography $\mathbf{m}$ that minimizes $||\mathsf{B}\mathbf{m}||_2^2$ is found by solving a homogeneous linear system (HLS) of the form $\mathsf{B}\mathbf{m} = \mathbf{0}$. Pseudo-inversion would only lead to trivial solution for HLS, therefore it cannot be used to solve the problem. Instead singular value decomposition in combination with preconditioning by translation of matched areas to the origin and normalization of their scale is commonly used.[33]

The procedure is similar to fitting of a single homography in computer vision.[19] It is more general, as our flow field is not explained by a rigid camera but needs to find one homography in each pixel. To ensure a consistent and piece-wise smooth output we combine a regularizing smooth kernel with an edge-aware component.

We implement the entire estimation in parallel over all pixel locations using graphics hardware (GPUs) allowing to estimate the homography field in less than 3 s for an high-definition (HD) image.

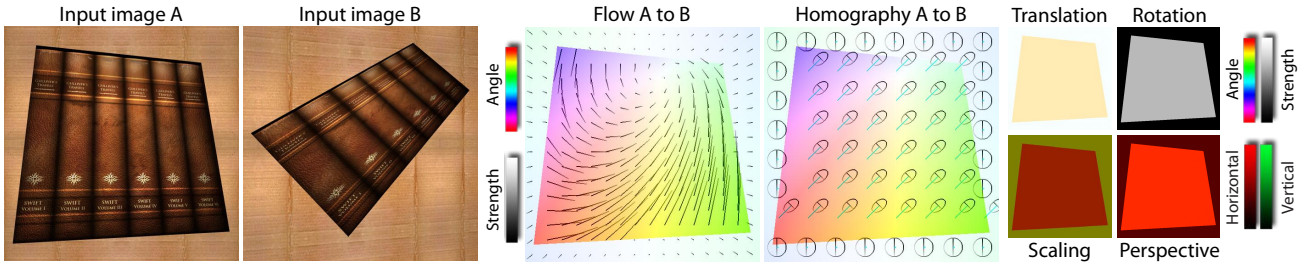

Figure 4: Conversion of an image pair into elementary transformations. Optical flow from A to B (polar representation) is fitted by local homography matrices (here locally applied on ellipses for illustration) to get five elementary transformations (shear left out for simplicity) having 8 channels in total.

3.3 Transformation decomposition

For perceptual scaling the per-pixel transformation M is decomposed into multiple elementary transformations: translation, rotation, scaling, shear and perspective (cf. Fig. 4). The relative difficulty of each transformation will later be determined in a perceptual experiment (Sec. 3.5).

We assume that our transformations are the result of a 2D transformation followed by a perspective transformation. This order is arbitrary, but we decided for it, as it is closer to usual understanding of transformations of 3D objects in a 2D world. This is motivated by the fact, that it is more natural to imagine objects to live in their (perspective) space and move in their 3D oriented plane before being projected to the image plane than to understand them as 2D entities undergoing possibly complex non-linear and non-rigid transformations in the image plane.

The decomposition happens independently for the matrix M at every pixel location. As M is unique up to a scalar, we first divide it by one element, which is chosen to be m_{33}. In the next five steps each elementary component T will be found first by extracting it from M, and then removing it from M by multiplying with T^{-1}.

First, perspective is extracted by computing horizontal and vertical focal length as $\mathbf{d} = (\mathsf{M}_\mathrm{a}^\mathsf{T})^{-1} \cdot (m_{31}, m_{32}, 0)$ where M_a is the affine part of M. The multiplication removes dependency of $\mathbf{d}$ on other transformations in M. We later convert the focal length into the field of view $\alpha = 2\,\mathrm{atan}(\mathbf{d}/2)$. To remove the perspective from M we multiply it by the inverse of a pure perspective matrix in the form $\left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ d_x & d_y & 1 \end{smallmatrix}\right)$. Second, a two-dimensional translation vector (m_{13}, m_{23}) is found. It is removed from M by multiplying with the inverse of a translation matrix. Next, we find the rotation angle $\mathrm{atan2}(m_{21}, m_{11})$ and remove it by multiplying with an inverse rotation matrix. Two-dimensional scaling power is recovered as $\log(m_{11}, m_{22})$ and removed from the matrix. The last component is scalar shearing angle found as $\mathrm{atan}(m_{12})$.

3.4 Transformation field entropy

Ease and difficulty of dealing with transformations does not only depend on the type and magnitude of a transformation on its own, but as it is often the case in human perception it depends on a context. Compensating for one large coherent translation might be easy compared to compensating for many small and incoherent translations. We model this effect using the notion of *transformation entropy* of an area in an elementary transformation field. Transformation entropy is high if many different transformations are present in a spatial area and it is low if it is uniform. Note how entropy is not proportional to the magnitude of transformations in a spatial region, but to the incoherence in their probability distribution.

We define the transformation entropy H of an elementary transformation at location $\mathbf{x}$ in a neighborhood s using standard entropy equation as

$$H(\mathbf{x}, s) = -\int_\Omega p(\omega|\mathbf{x}, s) \log p(\omega|\mathbf{x}, s) \mathrm{d}\omega,$$

where $p(\omega|\mathbf{x}, s)$ is the probability of observing transformation ω in a neighborhood of size s around $\mathbf{x}$. The type of Ω and ω depend on the type of elementary transformation. It is the real plane for translations, scaling, shear and perspective and the real circle with toroidal wrap-around for rotation. Examples of high and low transformation entropy are shown in Fig. 5.

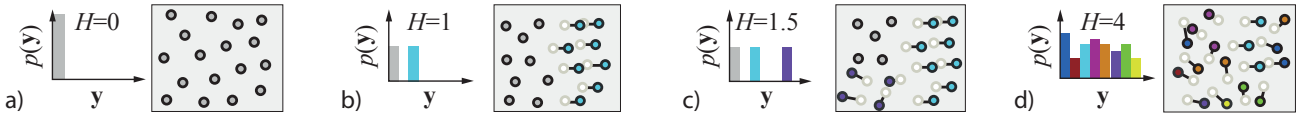

Figure 5: Entropy for different transformation fields applied to circular items. One transformation maps to one color. *a):* For identity, the histogram has a single peak and entropy is 0. *b):* For two transformations the histogram has two peaks and entropy is larger. *c):* For three transformations, there are three peaks. Entropy is even larger. *d):* In increasingly random fields the histogram flattens and entropy increases.

The probability distribution $p(\omega|\mathbf{x}, s)$ of elementary transformations at neighborhood $\mathbf{x}, s$ has to be computed using density estimation, i. e.,

$$p(\omega|\mathbf{x}, s) = \int_{\mathbb{R}^2} K(\omega - t(\mathbf{y}))\mathrm{d}\mathbf{y},$$

where K is an appropriate kernel such as the Gaussian and $t(\mathbf{x}) \in \Omega$ is a field of elementary transformations of the same type as ω.

Depending on the size of the neighborhood s, entropy is more or less localized. If the neighborhood size is varied, entropy changes as well, resulting in a scale space of entropy,[35] studied in computer vision as an image structure representation. For our purpose, we pick the entropy scale space maximum as the local entropy of each pixel and do not account for the fact at what scale it occurred. The entropy of all elementary transformation is summed into a single scalar entropy value as their difficulty also was found to sum linear.[10]

The decomposition into elementary transformations is key to the successful computation of entropy: Without it, a rotation field would result in a flat histogram as all directions are presented. This would indicate a high entropy, which is wrong. Instead, the HVS would explain the observation using very little information: a single rotation with low entropy.

Implementation In the discrete case, the integral to compute the entropy of the i-th pixel becomes:

$$H_i = -\sum_{j=0}^{n_\mathrm{b}} \sum_{k \in \mathcal{N}(s)} K(t_k - \omega_j) \log \sum_{k \in \mathcal{N}(s)} K(t_k - \omega_j), \tag{3}$$

where ω_j is the center of the j-th bin. The inner sums compute the probability of the the j-th transformation, essentially by inserting the k-th transformation from a neighborhood $\mathcal{N}(s)$ of size s into one of $n_\mathrm{b} = 32$ bins. An example result is shown in Fig. 6.

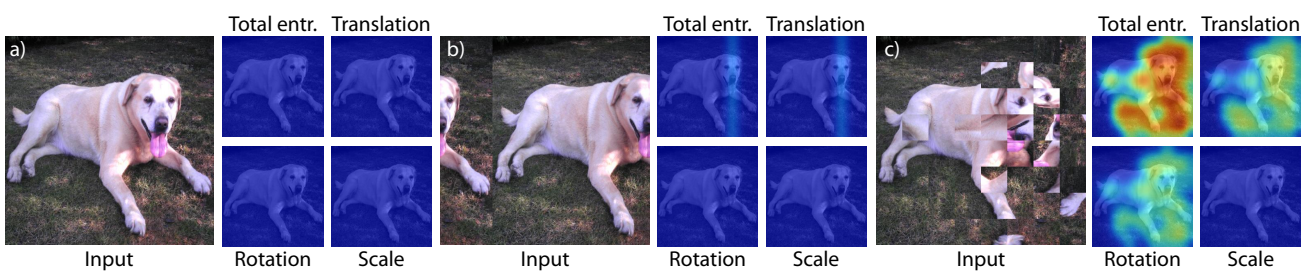

Figure 6: Entropy of an image under a puzzle-like flow field. *(a)* No-transformation yields zero entropy. *(b)* Low entropy is produced by two coherent transformations despite of a high average flow magnitude (≈192 px). *(c)* Transformations with increasing incoherence due to pieces that rotate and swap, lead to horizontally increasing entropy despite of a low flow magnitude (≈32 px).

Due to the finite size of our n_b histogram bins and the overlap of the Gaussian kernel K we systematically overestimate the entropy: Even when only a single transformation is present it will cover more than one bin, creating a non-zero entropy. To address this, we estimate the bias in entropy due a single Dirac pulse. We know that 0.99 of the area under a Gaussian distribution is within 3.2 standard deviations σ. That means that a conservative estimate of response is a uniform distribution

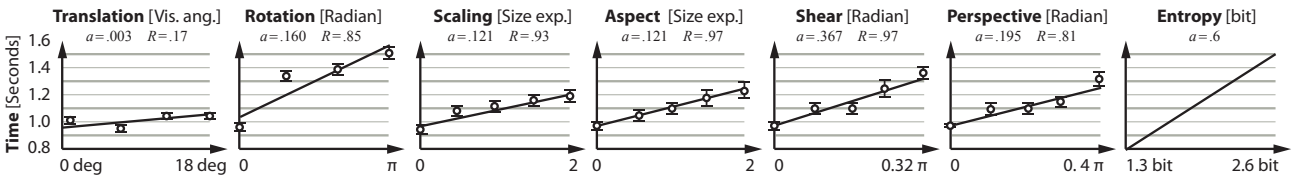

Figure 7: Perceptual scaling: x-axis: increasing elementary transformations and entropy. y-axis: increasing difficulty / response time.

of the value between 3.2σ bins. That yields the entropy $H_{\text{bias}} \approx -3.2\sigma \frac{1}{3.2\sigma} \log \frac{1}{3.2\sigma} = -\log \frac{1}{3.2\sigma}$. For our $\sigma = 0.5$ this evaluates to $H_{\text{bias}} = 0.2$. We approximate the entropy by subtracting this value.

Computing the entropy (Eq. 3) in a naïve way would require to iterate a large neighborhoods $\mathcal{N}(s)$ (up to the entire image) for each pixel $\mathbf{x}$ and every scale s. Instead, we use smoothed local histograms[36] for this purpose. In the first pass, the 2D image is converted into a 3D image with n layers. Layer i contains the discrete smooth probability of that pixel taking this value. Histograms of larger areas as well as their entropy can now be computed in constant time, by constructing a pyramid on the histograms.

3.5 Perceptual scaling

All elementary transformations as well as the entropy are physical values and need to be mapped to perceptual qualities. Psychological experiments indicate that elementary transformations such as translation, rotation and scaling require time (or effort) that is close to linear in the relevant x-axis variable in Fig. 7[10,12,16,17,37] and that the effect of multiple elementary transformations is additive.[12,17] A linear relation was also suggested for entropy in Hick's law.[28] Therefore, we scale elementary transformation and entropy using a linear mapping as seen in Fig. 7 and treat them additive.

Transformation To find the scaling an experiment was performed similar to the one that Shepard and Metzler[10] conducted for rotation but extended to all elementary transformations, including shear and perspective. The two-dimensional scaling power is further decomposed into isotropic scaling power scalar which is the minimum of absolute value in dimensions X and Y, and aspect ratio change which is the difference of values in X and Y. The assumption behind is that anisotropic scaling requires more effort to undo than simple isotropic size change. Objective of the experiment is to establish a relationship of transformation strength and difficulty, measured in response time increase in the mental transformation tasks.

Figure 8: Stimuli used in our mental transformation experiment.

Subjects were shown two abstract 2D patterns (Fig. 8). The two patterns were either different or identical. One out of the two patterns was transformed using a single elementary transformation of a certain strength x. Subjects were asked to indicate as quickly as possible if the two patterns are identical by pressing a key. Auditory feedback was provided to indicate if the answer was correct. The time $t(x)$ until the response was recorded for all correct answers where the two patterns were identical up to a transformation. The choice of pattern to transform (left or right), the elementary transformation and x were randomized in each trial.

21 subjects (17 M / 4 F) completed 414 trials of the experiment in 3 sessions. For each elementary transformation, we fit a linear function to map strength to response time (Fig. 7). We found a good fit of increasing linear functions of x for all transformations except translation. This agrees with findings for rotation[10] or scaling,[11] and our different bias or slope is likely explained by the influence of stimulus complexity also found by Shepard and Metzler.[10]

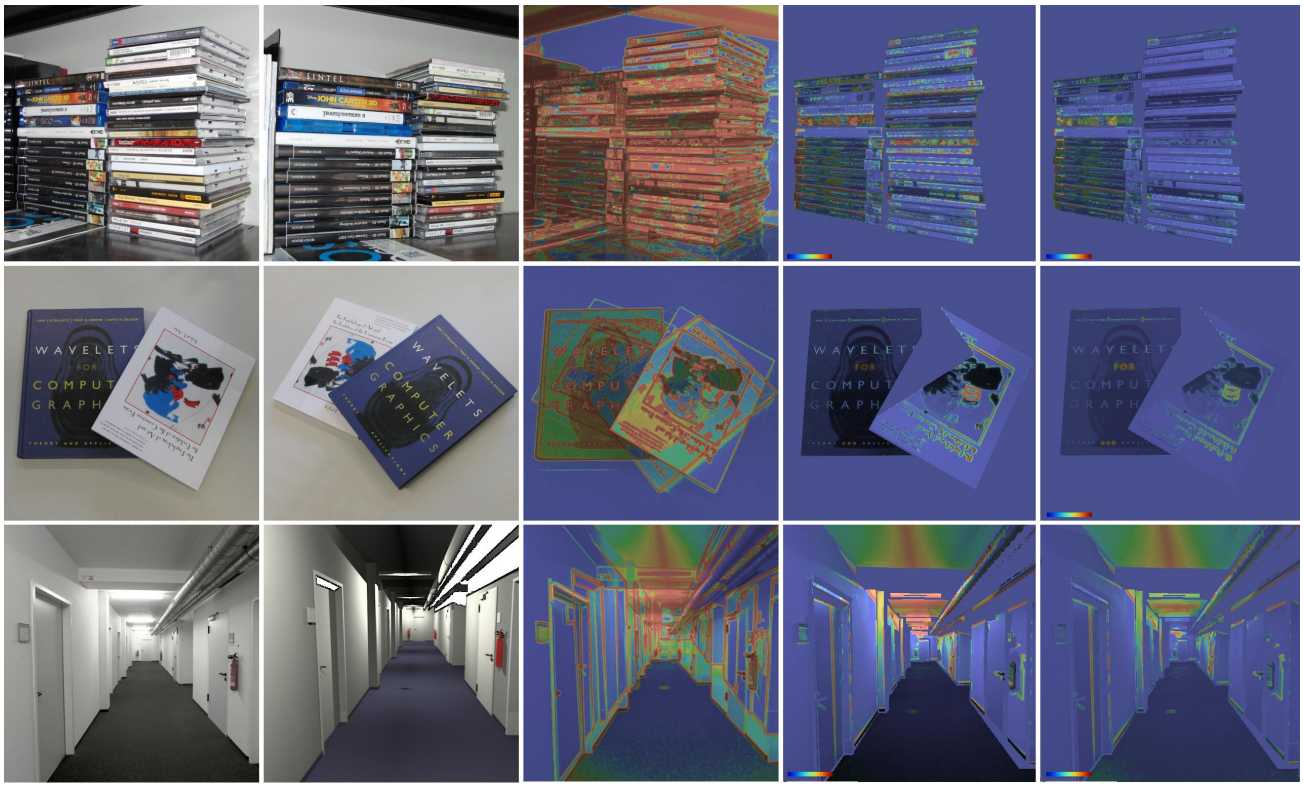

Figure 9: Results of image difference application using the SSIM index. *Rows (up to down): (I)* A real scene photograph with lot of entropy in the right shuffled CD stack. *(II)* A simple real scene. *(III)* A comparison of a photograph and a re-rendering of a corridor. *Columns (left to right): (a, b)* Two input images. *(c)* A naive quality metric without alignment results into false positive values everywhere. *(d)* Image alignment itself does not account for high entropy which would prevent observer to easily compare individual objects. *(e)* Our metric predicts such behavior and marks differences there as less visible.

Entropy We assume the effect of entropy can be measured similar as in the task of Hick,[28] where a logarithmic relationship between the number of choices (blinking lamps) and the response time (verbal report of count) was found. He reports a logarithmic time increase with a slope of 0.6 when comparing a visual search task with ten choices to a single choice-task. The negative logarithm of the inverse number of choices with equal probability is proportional to entropy, so entropy can be directly used for scaling (Fig. 7). Note, that while the increase in difficulty is bound for transformation strength, it is not bound for transformation entropy.

4. APPLICATIONS

The key applications of our model are an image metric (Sec. 4.1) and an image saliency (Sec. 4.2).

4.1 Image difference metric

Using the above building blocks, we can create a transformation-aware image metric (Fig. 1 and Fig. 9). Initially, the second image g_2 is aligned to the first one using the inverse flow f^{-1}. Next, the images can be compared using an arbitrary image metric (we experiment with SSIM[6]), with the only modification, that occluded pixels are skipped from all computations. Result is a map δ, that contains abstract visual differences (a unitless quality measure in the range from 0 to 1 for SSIM). This map does not account for the effect of the transformation strength and entropy while we have seen from our experiments that large or incoherent transformations make comparing two images more difficult. To this end, the perceived error of each pixel δ is adjusted by the increase of difficulty (response time minus optimal response time i. e., with no transformation or entropy) due to each elementary transformation: translation (d_t), rotation (d_r), scale (d_s), shear (d_h) and perspective (d_p) as well as the

entropy (d_e) at this pixel, resulting in a transformation-aware perceived difference $\delta' = \delta(1+d_\mathrm{t}+d_\mathrm{r}+d_\mathrm{s}+d_\mathrm{h}+d_\mathrm{p}+d_\mathrm{e})^{-1}$. The summation is motivated by the finding that response time besides being linear also sums in a linear fashion (if scaling adds one second and rotations adds another one, the total time is two seconds). As difficulty is in units of time, the resulting unit is visual difference per time. If the original map δ differed by 3 units and was subject to a transformation that increased response time by 1 second (e. g., a rotation by ca. $180\,\mathrm{deg}$), the difference per unit time is $3/(1+1) = 1.5$, whereas a change increasing response time by 3 seconds (e. g., a shuffling with high entropy) the difference per unit time is $3/(1+3) = 0.75$. In Fig. 9 we show the outcome of correcting the SSIM index by considering our measure of transformation strength and entropy.

Image transformations that contain local scaling power larger than 0 (zooming) might reveal details in g_2 that were not perceivable or not represented in the the first image g_1. Such differences could be reported as indeed they show something in the second image that was not in the first. However, we decided to not consider such differences as a change from nothing into something might not be a relevant change. This can be achieved by blurring the image g_1 with a blur kernel inversely proportional to the scaling. Occlusions are handled in the same way: No perceived difference is reported for regions only visible in one image.

Validation We validate our approach by measuring the correlation of human performance in perceiving differences in an image pair and transformation magnitude and entropy. Subjects were shown image pairs that differed by a flow field as well as a change in content. Two image pairs show 3D renderings of 16 cubes with different textures. The transformation between the image pairs included a change of 3D viewpoint and a variety of 3D motion for each cube. The images were distorted by adding near-threshold noise, and color quantization to randomly chosen textured cubes, so that the corresponding cubes could differ either by the presence of distortion or its intensity. 10 subjects were asked to mark the cubes that appear different using a 2D painting interface in unlimited time. We record this "heat map" from every trial. It localizes, where users gave wrong answers, i. e., where there was an image difference that they did not mark and where they marked a distortion while there was none. Difficult areas have a heat of 0.5 (chance level), while areas where users were confident have a value up to 1.0. The heat map is averaged over all subjects for one distortion and one scene.

We analyze the correlation of heat and transformation amplitude and entropy and found an average Pearson's r correlation of 0.701, which is is significant according to the t test with $p < 0.05$. The transformation magnitude has a lower correlation of $r = 0.441$ compared to transformation entropy $r = 0.749$. We conclude, that both transformation magnitude and entropy have an significant correlation with the ability to detect of distortions: In the presence of strong or complex transformation the increase in human detection error can be fit using a linear model.

The final performance of our approach is limited by the image metric used. The correlation of image metrics and quantitative user responses is low and difficult to measure[38] even without transformations. Therefore, the evaluation of the full metric, in particular for supra-threshold conditions, is relegated to future work.

Discussion Here we discuss the relation of our and existing image and video metrics, in particular how they deal with transformations. For more general survey of image quality metrics we refer to.[6]

Standard image difference (fidelity) metrics such as per-pixel MSE, per-patch structure similarity (SSIM) index,[6] or the perception-based visible differences predictor (VDP),[39] are extremely sensitive for any geometric distortions [6, Figs. 1.1 and 3.8]. The requirement of perfect image registration is lifted for the pixel correspondence metric (PCM),[40] closest distance metric (CDM),[41] or point-to-closest-point MSE, which account for edge distances. Natural images can be handled by the complex wavelet CW-SSIM index but mostly small translations can be supported [6, Ch.3.2.3]. All these approaches model local deformation invariance which is a low-level (C1 cortical area) process. Our transformation-aware quality metric can compensate for transformations of much larger magnitude which occurs at higher levels[9] including perspective transformation.

Video quality assessment (VQA) typically considers the temporal domain as a straightforward extension of 2D spatial filter banks into 3D.[42] This precludes any reasonable motion analysis based on its direction and velocity, which requires the optical flow computation. A notable exception work where optical flow is used to guide the local orientation of space-time, 3D Gabor filters.[43] Dominant motion increases the perception uncertainty and suppresses distortion visibility, while relative motion can make video degradations easier to notice.[42] Our homography decomposition enables to analyze dominant and relative motion, and our transformation entropy accounts for their local complexity, which we utilize in our transformation-aware quality metric.

The visual image equivalence[1] measures whether a scene's appearance is the same rather than predicting if a difference is perceivable. Perceivably different scenes can result in the same impression, as the HVS compensates for irrelevant changes. Our method can be considered another form of visual equivalence, as we model compensation for transformations. Comparing two aggregates of objects[44] is also related to entropy but goes beyond, if the aggregates differ by more than a transformation i. e., deletion or insertion of objects.

4.2 Saliency

Figure 10: Transformation-aware saliency for an input image *a)* being deformed into image *b)*. One patch is salient, as it deforms differently. Only transformation saliency is shown. This "differently" however is non-trivial and can only be detected with a transformation representation that captures the human ability to compensate for certain transformations including perspective, such as ours. Other motion saliency methods do not capture this effect *c)*, but instead, consider other parts more salient, following local variations in the flow.

Our decomposition into elementary transformations can be used to predict what part of an image pair is salient (Sec. 2.3) due to its transformation and content in combination (Fig. 10). Different from common image saliency, our approach takes two instead of one image as an input. It outputs saliency, e. g., how much attention an image region receives. We largely follow the basic, but popular, model of Itti et al.,[25] and replace its motion detection component by our component that detects salient transformations. The resulting model decomposes the input into a set of spatially varying luminance, color, and orientation features at multiple scales. Each feature map is then normalized to one, and multiplied by the squared difference of the global maximum and the average of all local maxima to suppress numerous comparable maxima and boost distinctive (conspicuous) ones. The same procedure is applied to our elementary transformations, where we independently consider the feature maps with the magnitude of translation, rotation angle, field of view, scaling power and shear angle. These additional conspicuousness maps are combined in the same linear way as luminance, color, and orientation is combined in[25] into a transformation-aware saliency map.

Discussion We compare our approach for scenes that contain complex elementary transformations to approaches by Le Meur et al.,[45] Zhai and Shah[46] and Itti et al.[25] (with the original motion detection component) in Fig. 10.

A vast majority of saliency models that handle temporal domain are focused on motion detection with respect to the static environment,[7] but motion pop-out may also arise from non-consistent relative motion. Therefore the global motion (e. g., due to camera motion) or consistent and predictable object motion should be factored out, and the remaining relative motion is likely to be a strong attention attractor. Along this line Le Meur et al.[45] derive the global motion in term of an affine transformation using robust statistics and remove it from the optical flow. The remaining outlier flow is compared to its median magnitude as a measure of saliency. Such per-pixel statistics make it difficult to detect visually consistent object transformations such as rotations, where the variability of the motion magnitude and direction might be high. Zhai and Shah[46] derive local homographies that model different motion segments in the frame. In this work, we compute transformation contrast similar to translation-based motion contrast in,[46] but we perform it for all elementary transformations and we account for neighboring homographies in a multi-resolution fashion, instead considering all homographies at once. This gives us a better locality of transformation contrast. Also, through decomposition into elementary transformations we

are able to account for the HVS ability to compensate for numerous comparable (non-salient) transformation components akin to camera or large object motion, and detect highly salient unique motion components. This way, instead of detecting local variations of optical flow, we are able to see more global relations between moving objects (as relative rotation in Fig. 10). The edge-stopping component of homography estimation enables us to find per-pixel boundary of regions with inconsistent motion, which further improves the accuracy of saliency maps. Finally, our saliency model is computationally efficient and can be performed at near-interactive rates.

5. CONCLUSION AND FUTURE WORK

We propose a model of human perception of transformations between a pair of images. The model converts the underlying optical flow into a field of homographies, which is further decomposed into elementary transformations that can be perceptually scaled and allows the notion of transformation entropy. Our model enables for the first time a number of applications. We extended perceptual image metrics to handle images that differ by a transformation. We extend visual attention models to detect conspicuous relative object motion, while ignoring predictable motion such as due to view changes or consistent object motion.

Our transformation-aware perceptual scaling may have other interesting applications, which we relegate as future work. In image change blindness[47] the same view has been considered so far, and our approach could be beneficial to predict the increased level of difficulty in the visual search task due to perspective changes. Also, the concept of visual equivalence[1] can be extended to handle different scene views, as well as minor deformations of the object geometry and their relocation. Our quality metric could be applicable to re-photography and re-rendering[3] allowing for a better judgement of structural image differences, while ignoring minor mis-registration problems. This is also the case in image retargeting, where all image distance metrics such as SIFT flow, bidirectional similarity, or earth mover's distance[4] account for some form of the energy cost needed to transform one image into another. While semantic and cognitive elements of image perception seem to be the key missing factors in those metrics, it would be interesting to see whether our decomposition of the deformation into elementary transformations and perceptual scaling of their magnitudes could improve the existing energy-based formulations.

In the end of the day the basic question is: "What is an image?". In most cases, "image" does not refer to a matrix of physical values but refers the mental representation of a scene. This mental representation is created by compensating for many variations in physical appearance. The ability to compensate for transformation as well as its limitations are an important part of this process and has been modeled computationally in this work.

REFERENCES

[1] Ramanarayanan, G., Ferwerda, J., Walter, B., and Bala, K., "Visual equivalence: towards a new standard for image fidelity," *ACM Trans. Graph. (Proc. SIGGRAPH)* (2007).

[2] Meyer, G., Rushmeier, H., Cohen, M., Greenberg, D., and Torrance, K., "An experimental evaluation of computer graphics imagery," *ACM Trans. Graph.* **5**(1), 30–50 (1986).

[3] Bae, S., Agarwala, A., and Durand, F., "Computational rephotography," *ACM Trans. Graph.* **29**(3), 24:1–24:15 (2010).

[4] Rubinstein, M., Gutierrez, D., Sorkine, O., and Shamir, A., "A comparative study of image retargeting," *ACM Trans. Graph.* **29**(6) (2010).

[5] Gibson, J. J., [*The perception of the visual world*], Houghton Mifflin (1950).

[6] Wang, Z. and Bovik, A. C., [*Modern Image Quality Assessment*], Morgan & Claypool Publishers (2006).

[7] Borji, A. and Itti, L., "State-of-the-art in visual attention modeling," *IEEE PAMI* **35**(1), 185–207 (2013).

[8] Koenderink, J., "Optical flow," *Vis. Res.* **26**(1), 161–180 (1986).

[9] Orban, G. A., Lagae, L., Verri, A., Raiguel, S., Xiao, D., Maes, H., and Torre, V., "First-order analysis of optical flow in monkey brain," *PNAS* **89**(7), 2595–99 (1992).

[10] Shepard, R. and Metzler, J., "Mental rotation of three dimensional objects," *Science* **171**(3972), 701–3 (1971).

[11] Bundesen, C., Larsen, A., and Farrel, J., "Mental transformations of size and orientation," in [*Attention and Performance IX*], 279–294 (1981).

[12] Sekuler, R. and Nash, D., "Speed of size scaling in human vision," *Science* **27**(2), 93–94 (1972).

[13] Cooper, L., "Demonstration of a mental analog of an external rotation," *Perception & Psychophysics* **19**(4), 296–302 (1976).

[14] Cutting, J. E., "Rigidity in cinema seen from the front row, side aisle.," *J Exper. Psych.: Human Perception and Performance* **13**(3), 323 (1987).

[15] Vishwanath, D., Girshick, A. R., and Banks, M. S., "Why pictures look right when viewed from the wrong place," *Nature Neuroscience* **8**(10), 1401–10 (2005).

[16] Robins, C. and Shepard, R., "Spatio-temporal probing of apparent motion movement," *Perception & Psychophysics* **22**(1), 12–18 (1977).

[17] Bundesen, C. and Larsen, A., "Visual apparent movement: transformations of size and orientation," *Perception* **12**, 549–558 (1983).

[18] Todd, J. T., "The visual perception of 3D structure from motion," in [*Perception of Space and Motion*], Epstein, W. and Rogers, S., eds., 201–226, Academic Press (1995).

[19] Szeliski, R., [*Computer vision: algorithms and applications*], Springer (2011).

[20] Black, M. and Anandan, P., "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Comput. Vis. Image Understand.* **63**(1), 75–104 (1996).

[21] Howard, I. P. and Rogers, B. J., [*Perceiving in Depth*], I. Porteous, Toronto (2012).

[22] Liu, F., Gleicher, M., Jin, H., and Agarwala, A., "Content-preserving warps for 3D video stabilization," *ACM Trans. Graph. (Proc. SIGGRAPH)* **28**(3), 44 (2009).

[23] Liu, C., Yuen, J., and Torralba, A., "SIFT Flow: Dense correspondence across scenes and its applications," *IEEE PAMI* **33**(5), 978–94 (2011).

[24] Igarashi, T., Moscovich, T., and Hughes, J. F., "As-rigid-as-possible shape manipulation," *ACM Trans. Graph. (Proc. SIGGRAPH)* **24**(3), 1134–41 (2005).

[25] Itti, L., Koch, C., and Niebur, E., "A model of saliency-based visual attention for rapid scene analysis," *IEEE PAMI* **20**(11), 1254–59 (1998).

[26] Bevenesee, R. M., [*Maximum entropy solutions to scientific problems*], Prentice-Hall (1993).

[27] Grossman, E., "Entropy and choice time: The effect of frequency unbalance on choice-response," *Quarterly J Exp. Psych.* **5**(2), 41–51 (1953).

[28] Hick, W. E., "On the rate of gain of information," *Quarterly J Exp. Psych.* **4**(1), 11–26 (1952).

[29] Pluim, J. P. W., Maintz, J., and Viergever, M., "Mutual-information-based registration of medical images: a survey," *IEEE Trans. Med. Imaging* **22**(8), 986–1004 (2003).

[30] Vázquez, P.-P., Feixas, M., Sbert, M., and Heidrich, W., "Automatic view selection using viewpoint entropy and its applications to image-based modelling," *Comput. Graph. Forum* **22**(4), 689–700 (2003).

[31] Gumhold, S., "Maximum entropy light source placement," in [*IEEE Visualization*], 275–282 (2002).

[32] Brox, T., Bregler, C., and Malik, J., "Large displacement optical flow," in [*Proc. CVPR*], 41–48 (2009).

[33] Hartley, R. I., "In defense of the eight-point algorithm," *IEEE PAMI* **19**(6), 580–93 (1997).

[34] Tomasi, C. and Manduchi, R., "Bilateral filtering for gray and color images," in [*Proc. Computer Vision*], 839–846 (1998).

[35] Ferraro, M., Boccignone, G., and Caelli, T., "On the representation of image structures via scale space entropy conditions," *IEEE PAMI* **21**(11), 1199–1203 (1999).

[36] Kass, M. and Solomon, J., "Smoothed local histogram filters," *ACM Trans. Graph. (Proc. SIGGRAPH)* **29**(4), 100 (2010).

[37] Huttenlocher, J. and Presson, C. C., "Mental rotation and the perspective problem," *Cog. Psych.* **4**(2), 277–299 (1973).

[38] Čadík, M., Herzog, R., Mantiuk, R., Myszkowski, K., and Seidel, H.-P., "New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts," *ACM Trans. Graph.* **31**(6), 147:1–147:10 (2012).

[39] Daly, S., "The Visible Differences Predictor: An algorithm for the assessment of image fidelity," in [*Digital Images and Human Vision*], 179–206 (1993).

[40] Prieto, M. S. and Allen, A. R., "A similarity metric for edge images," *IEEE PAMI* **25**(10), 1265–1273 (2003).

[41] Bowyer, K., Kranenburg, C., and Dougherty, S., "Edge detector evaluation using empirical ROC curves," *Comput. Vis. Image Understand.* **84**(1), 77–103 (2001).

[42] Wang, Z. and Li, Q., "Video quality assessment using a statistical model of human visual speed perception," *JOSA A* **24**(12), B61–B69 (2007).

[43] Seshadrinathan, K. and Bovik, A., "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Proc.* **19**(2), 335–350 (2010).

[44] Ramanarayanan, G., Bala, K., and Ferwerda, J. A., "Perception of complex aggregates," *ACM Trans. Graph.* **27**(3), 60:1–60:10 (2008).

[45] Le Meur, O., Le Callet, P., Barba, D., and Thoreau, D., "A spatio-temporal model of the selective human visual attention," in [*Proc. ICIP*], 1188–1191 (2005).

[46] Zhai, Y. and Shah, M., "Visual attention detection in video sequences using spatiotemporal cues," in [*Proc. Multimedia*], 815–824 (2006).

[47] Ma, L. Q., Xu, K., Wong, T. T., Jiang, B. Y., and Hu, S. M., "Change blindness images," *IEEE Trans. Vis. Comp. Graph.* **19**(11), 1808–19 (2013).