

Toward Assessing and Improving the Quality of Stereo Images

Minwoo Park*, *Member, IEEE*, Jiebo Luo, *Fellow, IEEE* and Andrew Gallagher, *Member, IEEE*

E-mail: {minwoo.park, jiebo.luo, andrew.gallagher}@kodak.com

Phone: 1-585-588-5721, 1-585-722-7139, 1-585-722-2890

Abstract—Imaging systems have incorporated numerous technological innovations such as 3D television and handheld devices. Despite these advances, these techniques still require the human eyes to refocus until the sense of depth perception is achieved by the observer. The more time this takes, the more eye muscles become fatigued and the brain tires from confusion. However, the exact intricacies involved are far more complex. To alleviate these problems, we introduce a learning framework that aims to improve the quality of stereo images. Instead of attempting to cover all factors that affect the quality of stereo images, such as image resolution, monitor response, viewing glass response, viewing conditions, viewer differences, and compression artifacts, we first introduce a set of universally relevant geometric stereo features for anaglyph image analysis based on feature point correspondence across color channels. We then build a regression model that effectively captures the relationship between the stereo features and the quality of stereo images and show that the model performs on par with the average human judge in our study. Finally, we demonstrate the value of the proposed quality model in two proposed applications where it is used to help enhance the quality of stereo images and also to extract stereo key frames from a captured 2D video.

Index Terms—3D, Quality, Improvement, Stereo Keyframe.

I. INTRODUCTION

IN the first half of the 19th century, not many years after the dawn of photography, imaging systems began to capture scenes in stereo. The human viewer perceives depth when each eye perceives a scene from a slightly different viewpoint, corresponding to human physiology (i.e., the arrangement of the eyes on the face). A number of different systems (e.g., anaglyph images, stereoscopes, and, recently, shutter glasses and displays) have been proposed so that each eye receives its intended view.

While many problems related to stereo capture have been studied by researchers (e.g., stereo correspondence [29]), there are some areas of stereo image processing that have yet to receive much attention. This paper addresses one such topic: the automatic assessment of stereo image quality. Figure 1 illustrates the overview of this research. For an input 3D image, we intend to extract a set of features, mostly related to geometry and useful for characterizing the quality of the input stereo image. Next, we explore a regression model that can capture the connection between the quality ratings of stereo images and the extracted features in order to derive a model

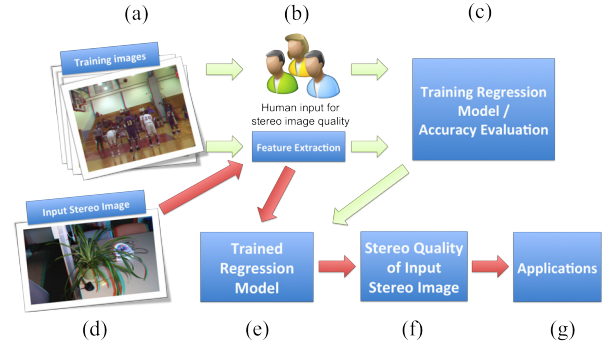


Fig. 1. Overview of the paper: (a) Training images. (b) Human input is collected and features are computed for the training images. (c) The regression model is trained and its performance is measured. (d) Input stereo image. (e) Extracted features for input stereo image (d) are input to the trained regression model. (f) Stereo quality of (d) is estimated. (g) Novel applications using (e) and (f) to improve the stereo image quality and to extract good 3D frame pairs from 2D video.

or metric for stereo image quality. Such a model can then find use in many applications that involve 3D images, such as stereo image enhancement.

In our work, we analyze anaglyph images to determine geometric stereo features and show their relationship to stereo image quality. The stereo image quality is a characteristic of an image that measures the quality of the perceived depth of the scene in the image. The geometric features should capture characteristics of the geometry of stereo cameras such as viewing angle and length of baseline between the two cameras. Although many other factors are in fact important to stereo image quality perception, these geometric features are expected to be universally relevant regardless of the viewing conditions. These features and a quality metric that is built upon them can be used for applications that require a ranking of stereo images by quality, or for improving the quality of stereo images. While our study currently focuses on stereo images in the form of anaglyphs due to its simplicity and low cost, we believe that our contributions will extend to other stereo image presentation methods (e.g., lenticular imagery or shutter glasses) because the model is built on universally relevant geometric features. We are also encouraged that the model can, to some extent, tolerate various viewing conditions (this will become more clear later).

Our main contributions are the following:

- We introduce a set of geometric stereo features for

measuring anaglyph image quality based on feature point correspondence across color channels.

- We produce a regression model that captures the relationship between the stereo features and the quality of anaglyph images.
- We present two proposed applications where the regression model proves instrumental.

In Section II, we review related work. In Section III, we define a four-level rating scale of stereo image quality. We then describe how we collect a stereo image dataset for our study, and the procedures for collecting human ratings as ground truth. In Section IV, we introduce a set of geometric features and describe the training and testing procedures for the proposed regression model of stereo image quality (Section III-A). After we evaluate the effectiveness of the regression model in comparison with human judges at the end of Section IV, we present two applications to demonstrate the promise of the model in Section V. Finally, we conclude the paper in Section VI.

II. RELATED WORK

This work relates to the fields of assessing and improving stereo image quality based on geometric features that capture characteristics of geometry of stereo cameras such as viewing angle and length of baseline between the two cameras. As such, various aspects of this work have been explored in different but related contexts. In the area of automatically assessing image aesthetic quality, there has been work to distinguish between amateur and professional photographers [18] as well as to rate the quality of the photo [7], [21], [35]. In general, these works extract low-level features from single images and use ratings gathered from the Internet. A recent approach uses high-level attributes to estimate image aesthetics [8]. However, these computational approaches do not consider the direct estimation of the quality of stereo images from analysis of the image content.

There is a great deal of research devoted to the analysis of stereo (or multi-view) captures of a scene. We refer the reader to [29] for a description of algorithms in this area. In general, this line of work is devoted to processing multiple images of a scene to compute either dense or sparse depth. Many of our features are inspired by the work in this area, but we use these features for a new purpose: determining the quality of a stereo image. In one particularly related work, [10] uses stereo analysis between channels of an image to determine *if* the image is a stereo anaglyph or not. In this work, we tackle the problem as more than a binary classification problem by estimating the quality of a stereo image, and then using this estimate to improve the stereo quality.

In the psychophysics of stereo images, researchers have investigated the impact of various modifications to images in a stereo pair on the overall quality or depth perceived by a viewer [2], [14]. For example, the effect of wavelet coding on stereo perception is investigated in [5]. Further, the effect on the quality of a stereo pair by filtering one image of a stereo pair is smaller than filtering both images [3], [34]. Seuntiëns et al. [31] have investigated the impact

of noise level on “viewing experience” and “naturalness” of 3D images. The finding in the work of Seuntiëns et al. [31] is that there is a significant difference in the assessment of 2D and 3D images. Solh and AlReigh [32] investigated effect of photometric and geometric distortion on the multi-camera image quality. The geometric distortion described in [32] is 2D perspective distortion simulating 3D distortion. However, our proposed geometric feature is one that enables the estimation of stereo camera configuration such as viewing angles and length of baseline between two cameras.

In another line of work, researchers have described techniques to enhance 2D image appearance using algorithms to modify attributes such as contrast, sharpness, and crop window [17], [22]. In processing stereo images to produce anaglyphs, researchers have found ways to improve the subjective appearance, such as by registering the images [13]. An automatic algorithm is presented in [25] to reduce eye strain in real stereoscopic images and sequences through horizontal image translation and corresponding image cropping, based on a statistical description of the estimated disparity within a stereo image pair. In [26] the mismatch between accommodation (focus) and convergence (fixation distance) in 3D displays is discussed, along with ways to remedy the problem. Lang et al. [20] address the problem of remapping the disparity ranges of stereoscopic images and video to remedy eye strain. Our work is intended to extend or complement these ideas into the realm of stereo image processing, with the ultimate goal to produce the most pleasing stereo anaglyph images.

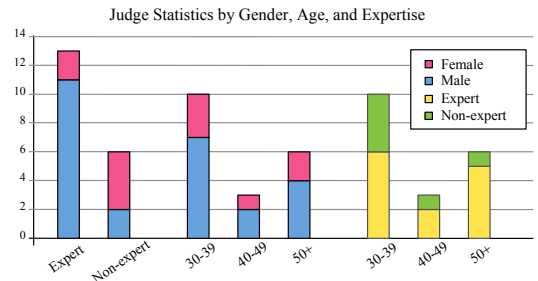


Fig. 2. Thirteen research scientists and technicians who had experience in assessing 2D and video quality in Kodak Research Laboratories and the other six non-experts participated in the rating. All of the participants had normal stereopsis for evaluations of static 3D images.

III. STEREO QUALITY RATINGS BY HUMANS

Assessing the objective quality of (2D and 3D) images with human input is still a somewhat open problem. The perception of quality is affected by the viewing environment (including display and illumination), image size, and even factors inherently associated with the human observer such as fatigue, color blindness, or visual acuity [19]. Designing a viewing experiment that controls each of these factors in a deterministic manner is an expensive endeavor. The situation is further complicated in stereo viewing scenarios, where perception of various depths depends on the viewer, and factors such as the mismatch between the focal plane of the

image and the plane of the display and stereo convergence (the distance at which convergence occurs) [12].

Rather than explicitly controlling each of these factors, recent work on 2D image quality prediction has relied on gathering input from many people (including through the Internet). For example, quality ratings were gathered from photo.net [7], dpchallenge.com [18], and solicited from Amazon Mechanical Turk [21]. As there are no standard datasets that have corresponding subjective quality ratings, we need to gather subjective quality ratings related to the perception of depth in anaglyph images. Standards have been suggested for assessing the subjective quality of stereo images [14], but again, these standards generally require dedicated lab space that cannot be deployed easily to the uncontrolled displays that a typical user may use to view anaglyph images. In addition, using a dedicated lab space for the experiments makes adding additional images at a later time more difficult than with our approach of performing the evaluations on the users own display.

Instead, we take a practical approach similar to the recent work on 2D quality prediction, and ask volunteers to rate the quality in the perceived depth in a set of stereo images on their own computer display. As in the case of 2D image quality assessment and prediction, we find that the human input we gather is useful for the task of 3D image quality prediction, despite the fact that we do not explicitly control all of the factors that could conceivably affect 3D perception.

In this Section, we describe our efforts to gather human rating input on the perceived quality of anaglyph stereo images. Such human input serves as the training and testing data for our quality assessment model.

A. Images

A total 4500 anaglyph images were gathered from a number of sources, including:

- using a query for “anaglyph images” from flickr.com. Note that for these images, we have only the composite anaglyph images and not the original left and right stereo pairs, and we do not know how these images were actually produced.
- using a stereo Fujifilm FinePix 3D[®] camera to capture stereo pairs and to produce anaglyph images.
- using sequential image capture [6] (i.e., capturing a first image with a standard single-lens camera, translating the camera horizontally, and then capturing a second image to form a stereo pair) and then producing anaglyph images from the pair.

We then took these 4,500 anaglyph images and first applied a crude quality classifier to roughly screen the stereo images. From the resulting crude scores, we selected 400 images such that the images were roughly uniformly distributed in the crude score space. This sampling is intended to provide an image dataset that spans the range from high to low quality and gives us a fairly balanced dataset.

A Crude Quality Classifier: We first train a SVM-based binary classifier to classify 3D images versus non-3D images.

We use a support vector machine (SVM) with a radial basis kernel $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$. We use the same feature, which will be introduced in Section IV. Training is achieved by minimizing the objective function given by equation (1) with respect to \mathbf{w} , b (support vector) and ξ (slack variable for non-separable data). For this purpose, we use the OpenCV Machine Learning toolbox [4].

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi_i \quad (1)$$

$$y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \xi_i \geq 0$$

The parameters C and γ are iterated on a logarithmic grid and selected based on a 10-fold cross validation estimate of error rate given by the ratio of the number of misclassified samples over the number of test samples.

Training Samples and Training Error: We collect a dataset that consists of positive stereo pair samples and negative samples. Positive samples came from: 1) stereo image pairs from the Middlebury stereo website [29] [30], 2) stereo image pairs captured using a Fujifilm FinePix 3D[®]¹, and 3) (pseudo) image pairs captured by a single-lens camera with mostly translational (horizontal) movements of static objects in the scene.

Negative samples came from: 1) image pairs captured by a single-lens camera that rotates slightly around its optical axis and 2) image pairs captured by a single-lens camera with slight vertical movements. The negative samples contain overlapping image content; however, they do not contain views of the scene from horizontally translated viewpoints and thus cannot invoke correct 3D perception for human viewing.

The number of positive and negative samples are 332 and 403, respectively. The 10-fold cross validation estimate of error rate is 1.54%.

Sampling Procedure: We run the trained classifier on 4500 images and divides 4500 images into four groups according to the classification result. The first group consists of images classified as non-3D images. The second, third, and fourth groups consist of images classified as 3D images and equally divided into three groups according to the variance of horizontal optical flow, $var(v_x)$. For this purpose, we compute $var(v_x)$ between left and right image pair, then sort stereo images by the $var(v_x)$, and set equally spaced two thresholds to divide all images classified as 3D images into three groups.

The higher $var(v_x)$ indicates the more structures in different depths, resulting in better 3D perception, since horizontal optical flow in stereo images originated from parallax and more variations of the horizontal flow means the more variations on the depth. Finally, we selected 400 images such that the images were roughly and uniformly distributed in the crude score space indicated by four groups.

B. Rating Scale

Our proposed learning-based framework can be used to learn a regression model that can quantify the quality of

¹http://www.fujifilm.com/products/3d/camera/finepix_real3dw1/

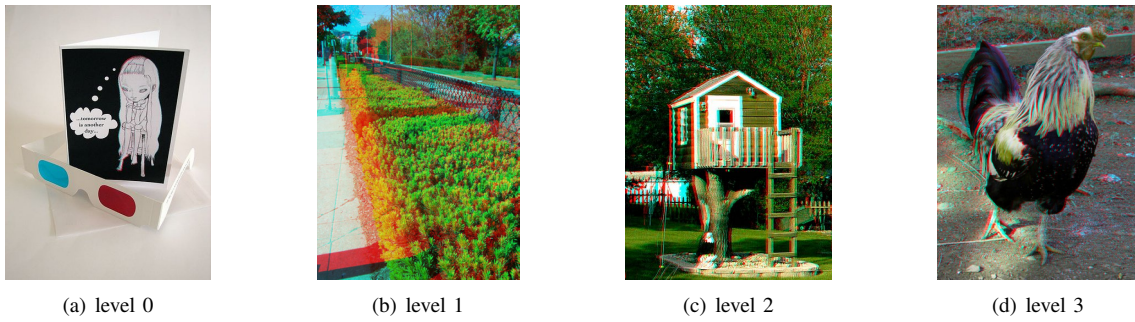


Fig. 3. From left to right, these example anaglyph images are based on human input, from the lowest (level 0) to the highest quality level (level 3).

stereo images by using features related to resolution, color rendition, motion portrayal, overall quality, sharpness, depth, depth resolution, depth motion, puppet theatre effect, and cardboard effect,² and so on. However, our current goal is to identify the relationship between the quality of stereo images and universally relevant geometric stereo features regardless of the viewing conditions of human observers. Therefore, design of the rating scale and definition should reflect the characteristics of geometric stereo features.

Although Section 1 in ITU-R BT.1438 [14] recommends following ITU-R BT.500 [15] for the conventional factors such as general picture quality, sharpness, and depth using the double-stimulus continuous quality scale (DSCQS) method, this cannot be applied to our case, as there is no reference image available because the best camera geometry for a given 3D scene is unknown.

We could possibly use the five-scale adjective categorical judgment methods according to Section 6.1.4.1 in ITU-R BT.500 [15]. However, our judges encountered difficulties distinguishing between the five points of the scale during the design of the study due to the ambiguity and subtlety of the task. Therefore, justified by Section 6.1.4.1 in ITR-R BT.500 [15] which indicates that categories may reflect judgments of whether or not an attribute is perceived, and Sections 1 and 2 in ITU-R BT.1438 [14] state that further studies are required to identify other factors to establish physical definitions and to assess particular factors of stereoscopic television systems, we propose the simplified four levels with definitions of certain attributes for each level as follows:

- level 0: No or slight depth perception.
- level 1: Inconsistent or difficult depth perception.
- level 2: Good depth perception.
- level 3: Very good depth perception.

The levels represent points on a continuous scale. Definitions and example images were used to calibrate the subjects. The level 0 images induce minimal or zero depth perception or have severe artifacts or very poor quality; with the level 1 images, the subject may perceive depth, but of poor quality (e.g., eye strains or other difficulties); the level 2 images should enable good depth perception without difficulties and eye strains; and the level 3 images should enable excellent depth perception without difficulty and eye strains.

Our use of the specific definition of each level is also supported by the results of Seuntiëns et al. [31] where human raters could take the added value of 3D over 2D images into account only when the raters are asked to rate the “viewing experience” and “naturalness” rather than the overall quality of stereo images. Although this scale might not be perfect, our subjects found it useful as a guide. It has the advantage that as artifacts increase, the perception of depth decreases, and both of these characteristics push the quality scores toward zero.

C. Human Judges and Rating Condition

A total of 19 human judges participated in the human rating study. Among all 19 human judges, 13 were imaging scientists who had experience in assessing 2D image and video quality in Kodak Research Laboratories, and 6 judges were non-experts. More detailed statistics on human judges by gender, age group, and expertise can be found in Figure 2.

Participants were supplied with standard red-cyan anaglyph glasses to view the anaglyph images and all had undergone a short training and Q&A session on how to assess stereo image quality using the defined four-level rating scale. Note that we screened the observers such that those who are incapable of perceiving 3D from stereo images or are not comfortable looking at stereo images were not part of the quality evaluation process. All judges had normal stereopsis for evaluating static stereo images.

As we said at the beginning of the Section III, rather than explicitly control the viewing environment, we follow the strategies of recent works [7], [18], [21] on 2D image quality prediction where the authors have relied on gathering input from many people (including through the Internet). We take a practical approach similar to the recent work on 2D quality prediction, and ask volunteers to rate the quality in the perceived depth in a set of stereo images on their own computer displays. We ask all judges whether they can perceive depth on the training set using the provided red-cyan anaglyph glasses. The judges are asked to rate the test set only when they can perceive depth on the training set. Each rated between 40 and 400 anaglyph images, producing about 8 ratings on any given stereo image in the dataset.

D. Human Rating Results

Figure 3 shows several example images from the human experiment for which there is good agreement between all

²ITU-R BT.1438 [14] lists factors affecting the quality of a stereo image.

human ratings. Overall, we were encouraged that the ratings have fair agreement by Fleiss' Kappa measure [9] with p-value less than 10^{-6} . For 77.00% of the images in our dataset, the rating is agreed upon by at least four people.

We use these 77.00% images (308 images) and their majority ratings to construct the final dataset with ground truth labels (i.e., the quality levels) of stereo image quality. Note that eventually the images are not uniformly distributed in terms of stereo quality levels. The numbers of images for levels 0 through 3 are 21, 132, 128, and 27, respectively, reflecting the fact that most images in the real world are in the medium range of the stereo quality scale. As in the case of 2D image quality assessment and prediction [7], [18], [21] where the human input is gathered without controlling the viewing environment, we find that the human input that we gather is useful for the task of stereo image quality prediction, despite the fact that we do not explicitly control all of the factors that could conceivably affect 3D perception.

We would like to point out that the measurement of 3D perception is still an evolving topic of research. In this work, we attempted to adapt the standard [14] to our scenario of distributed internet-based viewing and quality rating. Although we made some of the mentioned nonstandard choices, our human ratings are statistically significant (p-value less than 10^{-6}) and have fair agreement by Fleiss' Kappa measure [9]. It is possible that the methodology we designed to collect the human quality ratings will be improved upon as progress in 3D perception and quality rating is achieved. We would expect that improvements to this field should provide further improvements to our learning-based methods to assess and improve stereo image quality.

In Section IV, we will describe a regression model that incorporates *geometric* features to infer the stereo image quality.

IV. QUALITY RATINGS BY A REGRESSION MODEL

In this section, we introduce a set of geometric features. Next, we describe the training and testing procedures for the proposed regression model of stereo image quality. Last, we evaluate the effectiveness of the regression model in comparison with human judges.

A. Feature Extraction

We first resize images to the same width of 320 pixels for all images before we extract all of the 2D corner points then track these points over the image pair or different channels (from red channel to green channel if the input image is an anaglyph) to establish optical flow, and we compute additional geometric features from the computed optical flows.

The corners are local maximums of M given as:

$$M = I_x^2 I_{yy} + I_y^2 I_{xx} - 2I_x I_y I_{xy} \quad (2)$$

where I_x and I_y denote first image derivatives and I_{xx} and I_{yy} denote second image derivatives. Tracking is performed on the detected local maximums of M using the Lucas-Kanade tracking algorithm [24].

Although scale and rotation invariant keypoints such as SIFT [23], SURF [1], and Harris affine [27] can be used to compute geometric features introduced in this Section without loss of generality, the benefit of scale and rotation invariance are minimal in stereo images since the features in the left and right images have the same scale and are rotation free around the z axis (axis perpendicular to the image plane). Moreover, the computation of scale- and rotation-invariant features (e.g., SIFT) is much more expensive than computing simple corner. Our proposed corner tracking is in real time.

To compute geometric features, we then perform the RANSAC algorithm to estimate the epipolar geometry [11]. Estimation of the camera geometry by epipolar geometry is a standard method in computer vision [11]. Next, we group each tracked 2D point using the RANSAC algorithm into inliers and outliers. Inliers are tracked points that are consistent with the estimated epipolar geometry, and outliers are the remaining tracked points. Then, a collection of geometric features is computed from the tracked points to characterize the relative camera motion with respect to the scene. Since all of the computed optical flows in a stereo image pair is originated from the 3D structure of the scene and those flows are used to estimate the epipolar geometry, we call them collectively geometric features. A complete set of all computed features (81 features) and their descriptions are shown in Tables I and II. Symbols used in Table II are defined in Table I.

Let us explain some of these features. For example, we can infer from features $avg(\angle E)$ and $var(\angle E)$ if there is camera rotation only, translation only, or both. If the features $avg(\angle E)$ and $var(\angle E)$ are close to 0, there is only a horizontal camera translation. Also, the number of 3D points reconstructed using 2-view epipolar geometry and their mean and variance can indicate the existence and degree of depth, although the number can be affected by texture or error due to incorrect intrinsic camera parameters. Also, the degree of 3D can be inferred from $var(v_x^{(in)})$ and $var(v_y^{(in)})$ since a scene that does not contain objects at different depths would produce low variances for optical flows. Although some of the 81 features may be somewhat correlated to each other, they are all important in a sense that certain features are more reliable than others under certain conditions.

B. Building a Regression Model

Since the resulting number of images for each level is unbalanced, we resample in categories 0 and 3 to achieve a better balance and use modified random trees originally introduced by Leo Breiman and Adele Cutler [33]. The regression model using our modified random trees works as follows. The random trees classifier takes the input feature vector \mathbf{X} , classifies it with every tree $y_i = Tr_i(\mathbf{X})$ in the forest, and outputs the level label C that receives the majority of "votes." Next, we perturb C by all $y_i \neq C$. The motivation of this modification is to also account for votes of minorities, i.e., $y_i \neq C$, since the proposed stereo rating scale levels are meant to form a continuum of values. If there are more $y_i > C$ than $y_i < C$, C becomes $C + \epsilon$; otherwise C becomes $C - \epsilon$. We use $\epsilon = std(y_i \neq C)$ where $std(\cdot)$ is a standard deviation operator. For this purpose,

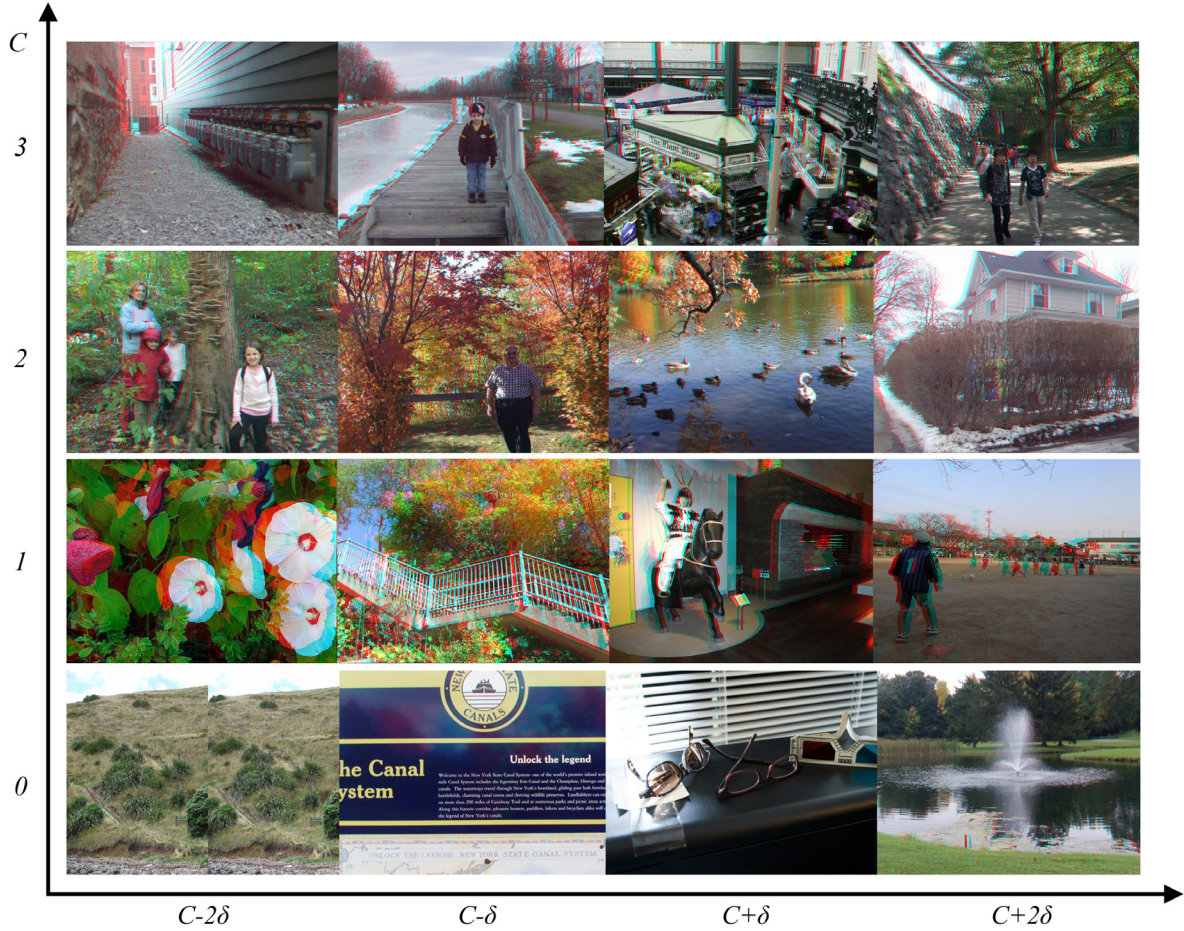


Fig. 5. Sample ratings by the regression model (We round the regression response to produce levels): The 1st, 2nd, 3rd, and 4th rows show stereo images with levels 3, 2, 1, and 0, respectively. The value of the y axis C is the rounded value of the regression model output and the value of the x axis varies from $C - 2\delta$ to $C + 2\delta$ where δ is the standard deviation of the regression response at that level. The bottom left image has the lowest quality rating by the model and the top right image has the highest quality measure.

we modify the OpenCV libraries [4] to add this perturbation feature. Formally, the trained regression model function with this perturbation is given as:

$$Q_{learned} = R_{forest}(\mathbf{X}) \quad (3)$$

C. Evaluating the Regression Model

We evaluate the trained regression model using 10-folds cross validation and we take the rounded value of $Q_{learned} = R_{forest}(\mathbf{X})$ to measure the classification accuracy. The overall accuracy of the trained regression model is 69.81% while the average accuracy of all 19 human judges is 65.89%. As shown in Figure 4, the trained regression model is comparable to the average human performance. Finally, it is noteworthy that the misclassification rate by more than one level is only 6.49% and the mean absolute error of estimation $avg(|GT - Q_{learned}|)$ is 0.53. The confusion matrix of the regression model is also shown in Figure 4.

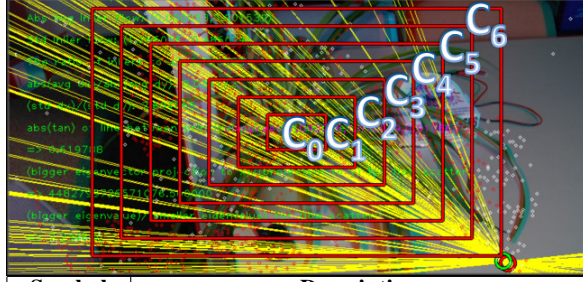
In addition, we also use Pearson linear correlation coefficient (CC), Spearman rank order correlation coefficient (ROCC), and outlier ratio (OC) to measure the prediction accuracy, monotonicity, and consistency of the learned model,

respectively, according to the VQEG recommendations [16]. Since there is no reference image in our study, we use mean opinion score (MOS) instead of difference mean opinion score (DMOS). Although the learned model rates the stereo image quality of a given image without access to the reference image, the measured prediction accuracy, monotonicity, and consistency are satisfactory. The measured accuracy by CC, monotonicity by ROCC, and consistency by OC are 0.59, 0.54, and 0.12, respectively.

Figure 5 presents sample ratings by the regression model (we round the regression response to produce levels). The first, second, third, and fourth rows show stereo images with levels 3, 2, 1, and 0, respectively. The value of the y axis C is the rounded value of the regression model output and the value of the x axis varies from $C - 2\delta$ to $C + 2\delta$ where δ is the standard deviation of the regression response at that level. The bottom left image has the lowest quality rating by the model and the top right image has the highest quality measure. Such ratings are generally consistent with the collective human ratings.

TABLE I

TOP IMAGE SHOWS (RED) IMAGE REGIONS C_i WHERE $0 \leq i \leq 7$ USED BY THE FEATURES AND (YELLOW) EPIPOLAR LINES (C_7 IS NOT SHOWN HERE). BOTTOM TABLE SHOWS SYMBOLS USED IN THE TABLE II.



Symbol	Description
$v_x^{(all)}$	All of the horizontal optical flows
$v_y^{(all)}$	All of the vertical optical flows
$v_x^{(in)}$	Horizontal optical flow of Epipolar inliers
$v_y^{(in)}$	Vertical optical flow of Epipolar inliers
$v_x^{(out)}$	Horizontal optical flow of Epipolar outliers
$v_y^{(out)}$	Vertical optical flow of Epipolar outliers
C_i	Center region of image
S_i	Surround region of C_i
$I(h, s, v)$	Image in HSV color space
$L(X)$	Logical operator if $X==true$ then $L(X)=1$ if $X==false$ then $L(X)=0$

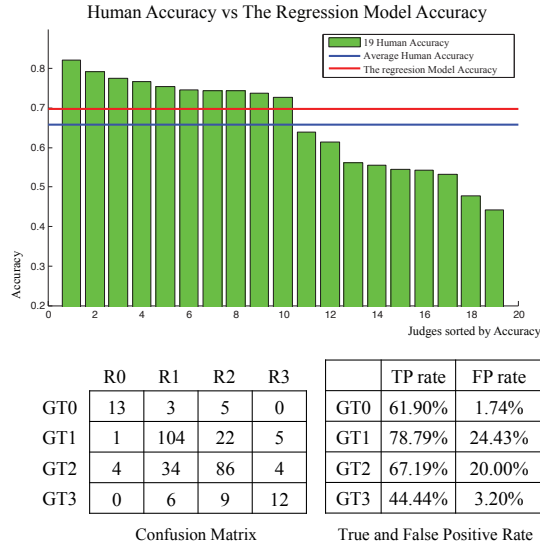


Fig. 4. Performance comparison of the trained regression model with the above and below average human performance. The red horizontal line and the blue line in the graph are the average accuracy of the trained regression model and the average accuracy of the human, respectively. Table shows the confusion matrix for the regression model using 10-fold cross validation and true positive and false negative rate. In the confusion matrix, entry at row GTx and col Ry indicates number of instances where the ground truth level x is classified as level y. The regression response is rounded to compute the confusion matrix.

D. Feature Importance

Moreover, we analyze the feature importance learned by the regression model. Figure 6 shows a sorted list of features with respect to their importance weights. Analysis of the top 10 important features shows that the regression model follows intuitively the theory governing a stereo vision system. Most

TABLE II

A COMPLETE SET OF 81 FEATURES AND THEIR DESCRIPTIONS. PLEASE REFER TO THE SYMBOL DEFINITION IN TABLE I.

Index	Feature	Description
1	$avg(\angle E)$	Average angle of Epipolar lines
2	$var(\angle E)$	Variance of angle of Epipolar lines
3	$\angle \mathbf{c}_1 \mathbf{e}_1$	Angle of line between centers of image and Epipoles.
4	$avg(v_x^{(in)})$	Average of $v_x^{(in)}$
5	$avg(v_y^{(in)})$	Average of $v_y^{(in)}$
6	$avg(v_x^{(out)})$	Average $v_x^{(out)}$
7	$avg(v_y^{(out)})$	Average $v_y^{(out)}$
8	$\lambda_{max}^{(in)}$	Eigen values of 2D scatter matrix of $v_x^{(in)}$ and $v_y^{(in)}$
9	$\lambda_{min}^{(in)}$	
10	$\epsilon_{max}^{(in)}$	Eigen values of 2D scatter matrix of $x^{(in)}$ and $y^{(in)}$
11	$\epsilon_{min}^{(in)}$	
12,13	$\mathbf{u}_{max}^{(in)}$	Eigenvectors of 2D scatter matrix of $x^{(in)}$, $y^{(in)}$
14,15	$\mathbf{u}_{min}^{(in)}$	
16,17	$\mathbf{v}_{max}^{(in)}$	Eigenvectors of 2D scatter matrix of $x^{(in)}$ and $y^{(in)}$
18,19	$\mathbf{v}_{min}^{(in)}$	
20 ~ 23	$\mathbf{e}_1, \mathbf{e}_2$	Locations of Epipole 1 and 2
24,25	$avg(v_x^{(out)}), avg(v_y^{(out)})$	Average $v_x^{(out)}$, Average $v_y^{(out)}$
26,27	$var(v_x^{(in)}), var(v_y^{(in)})$	Variance of $v_x^{(in)}$, Variance of $v_y^{(in)}$
28,29	$var(v_x^{(out)}), var(v_y^{(out)})$	Variance $v_x^{(out)}$, Variance $v_y^{(out)}$
30	R_1	$avg(v_x^{(in)})/avg(v_y^{(in)})$
31	R_2	$\lambda_{max}^{(in)}/\lambda_{min}^{(in)}$
32	R_3	$var(v_x^{(in)})/var(v_y^{(in)})$
33	R_4	$avg(v_x^{(all)})/avg(v_y^{(all)})$
34	$\#in / \#all$	$\frac{\text{the number of } v^{(in)}}{\text{the number of } v^{(all)}}$
35	$\#RC / \#pixels$	$\frac{\sum_{-20 < h < 20, s > 0.5, v > 0.5} L(I(h, s, v) > 0)}{width \times height}$
36	$\#N_{3D}$	The number of reconstructed 3D points
37	b_E	Is Epipole inside image?
38	$var(v_x^{(3D)})$	Variance of x, y, and z components in 3D points
39	$var(v_y^{(3D)})$	
40	$var(v_z^{(3D)})$	
41	T_x	x, y, and z components of second camera location
42	T_y	
43	T_z	
44	C_r	Max ratio of image compression for RGB channels [36]
45	C_g	
46	C_b	
47	C'_r	Variational max ratio of image compression for RGB channels [36]
48	C'_g	
49	C'_b	
50 ~ ~81	$avg(v_x^{(in)} \in C_i)$ $avg(v_y^{(in)} \in C_i)$ $avg(v_x^{(in)} \in S_i)$ $avg(v_y^{(in)} \in S_i)$	Average of $v_x^{(in)}, v_y^{(in)}$ inside C_i where $0 \leq i \leq 7$ Average of $v_x^{(in)}, v_y^{(in)}$ inside S_i where $0 \leq i \leq 7$

local maxima of the unsorted features are related to features that were computed from horizontal optical flows or epipolar geometry. In addition, the 6th feature shows that the epipolar inliers' flow in region C_4 is important. This suggests that the explicit use of salient region detection might be an important feature that we can use in the future. Also, the 5th important feature $\#RC / \#pixels$ initially does not appear to be related to the camera geometry but it makes sense that the small disparity in a stereo image would not produce much redness and cyaness unless the 2D scene itself is dominated by those colors. We can compute this feature for other 3D devices as well without loss of generality since other 3D devices have access to a pair of images; we can generate anaglyphs from image pairs.

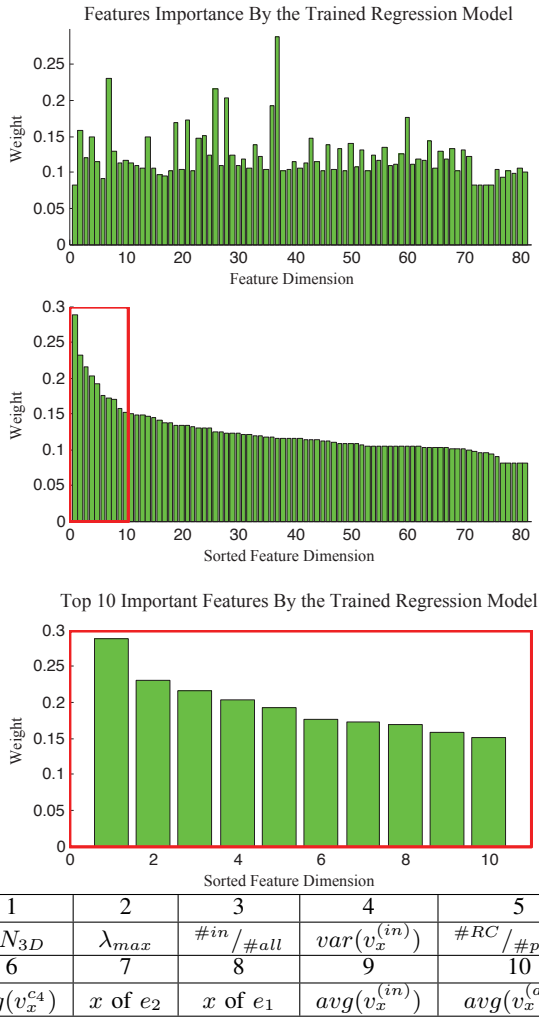


Fig. 6. Feature importance learned by the regression model: (Top) Feature importance for all 81 feature dimensions. Each dimension corresponds to the index in Table II. (Middle) All 81 feature dimensions sorted by their importance. (Bottom) Top 10 important features are shown along with their descriptions. Analysis of the top 10 important features shows that the regression model follows intuitively the theory governing a stereo vision system. Please refer to Table II for meaning of the symbols.

V. APPLICATIONS OF THE QUALITY MODEL

In this section, we apply the learned quality model to two applications. First, we use the model to improve stereo images by modifying the geometric characteristics of a given stereo image. As a first attempt, instead of applying full projective transform on the extracted optical flow, we apply simple 2D translation to the extracted optical flow and check if the recomputed features $\tilde{\mathbf{X}}$ can cause a better-quality rating by the model. We then re-align the second view using the found 2D translation. In addition, to show the versatility of the quality model, we also use it to select quality stereo key frames in a 2D video.

A. Stereo Improvement

The goal here is to improve stereo images initially in levels 1 and 2 by modifying optical flow v_x and v_y . As indicated by the feature importance in Figure 6, this is a reasonable

approach for those stereo images rated as a level 1 due to poor alignment of the stereo pairs. For this purpose, we formulate stereo improvement as finding a global maximum of the trained regression model $Q_{learned}$ (Section IV-B) for a given random variable space $\mathbf{X}^{(t)}$ where $t = 0$ initially. The feature $\mathbf{X}^{(t)}$ is modified by adding 2D offset of optical flow $\delta v_x^{(t)}$ and $\delta v_y^{(t)}$ where they are 0 when $t = 0$. Then new feature $\tilde{\mathbf{X}}^{(t)}$ is computed from $\mathbf{X}^{(t)}$ using $\delta v_x^{(t)}$ and $\delta v_y^{(t)}$ independently sampled from normal distributions with means of $\delta v_x^{(t)}$ and $\delta v_y^{(t)}$ and variances of $\sigma_x = 20$ and $\sigma_y = 2$, respectively. The new feature $\tilde{\mathbf{X}}^{(t)}$ is computed by adding the sampled $\delta v_x^{(t)}$ and $\delta v_y^{(t)}$ to the original optical flow correspondences. Then the $\tilde{\mathbf{X}}^{(t)}$ is evaluated by the regression model $\tilde{Q}_{learned} = R_{forest}(\tilde{\mathbf{X}}^{(t)})$. If $R_{forest}(\tilde{\mathbf{X}}^{(t)}) > R_{forest}(\mathbf{X}^{(t)})$ we accept $\tilde{\mathbf{X}}^{(t)}$ and set $\mathbf{X}^{(t+1)} = \tilde{\mathbf{X}}^{(t)}$; if not, we propose a new sample using the same normal distributions. Pseudo code for this improvement procedure is given in Figure 7.

Enhancement Algorithm

- 1 Start with initial values of $\delta v_x^{(t)} = 0, \delta v_y^{(t)} = 0$, $\sigma_x = 20, \sigma_y = 2$. Set $t = 0, it = 0$
- 2 while $it < maxit (= 1000)$
- 3 **Draw 2D sample** $\delta \tilde{v}_x^{(t)}, \delta \tilde{v}_y^{(t)}$
 from $\mathcal{N}(x; \delta v_x^{(t)}, 20), \mathcal{N}(x; \delta v_y^{(t)}, 2)$
- 4 **Modify feature** $\mathbf{X}^{(t)}$ w.r.t $\delta \tilde{v}_x^{(t)}, \delta \tilde{v}_y^{(t)}$
- 5 **Compute a response from** $R_{forest}(\tilde{\mathbf{X}}^{(t)})$
- 6 if $R_{forest}(\tilde{\mathbf{X}}^{(t)}) > R_{forest}(\mathbf{X}^{(t)})$
 Set: $\delta v_x^{(t+1)} = \delta \tilde{v}_x^{(t)}, \delta v_y^{(t+1)} = \delta \tilde{v}_y^{(t)}$
- 7 end
- 8 $t \leftarrow t + 1, it \leftarrow it + 1$
- 9 end
- 10 **Image warp using** $\delta v_x^{(t)}, \delta v_y^{(t)}$ **and generate anaglyph.**

Fig. 7. Stereo enhancement algorithm using our learned quality model $Q_{learned}$.

In Figure 8, we use two problematic anaglyph images from [26] to demonstrate the effectiveness of our stereo enhancement algorithm. The top image contains a scene with a large depth and a mismatch between the distance at which our eyes are focused (accommodated) and the distance at which they are fixated (converged). With the improvement by the automatic algorithm, the foreground meadow is moved closer (and in front of the display plane) while the background is also brought closer to (while still behind) the display plane, leading to a stereo image that is less straining for our eyes. The bottom image is an example of the window violation problem, which cannot occur with real windows and real-world physical objects. It is a result of the conflict between our ocular focus system and the binocular vergence system (convergence and divergence to fuse objects at various distances) such that an object that appears in front of the display plane by stereo cues is cropped by the image or display border so it also must be interpreted as being on the display concurrently. In this case, our algorithm properly moved the foreground objects (rocks) behind the display plane, thus alleviating the problem. We are greatly encouraged by our automated results that achieved the effect similar to that of the presumably manual operations shown in [26]. More results can be seen in Figure 9.

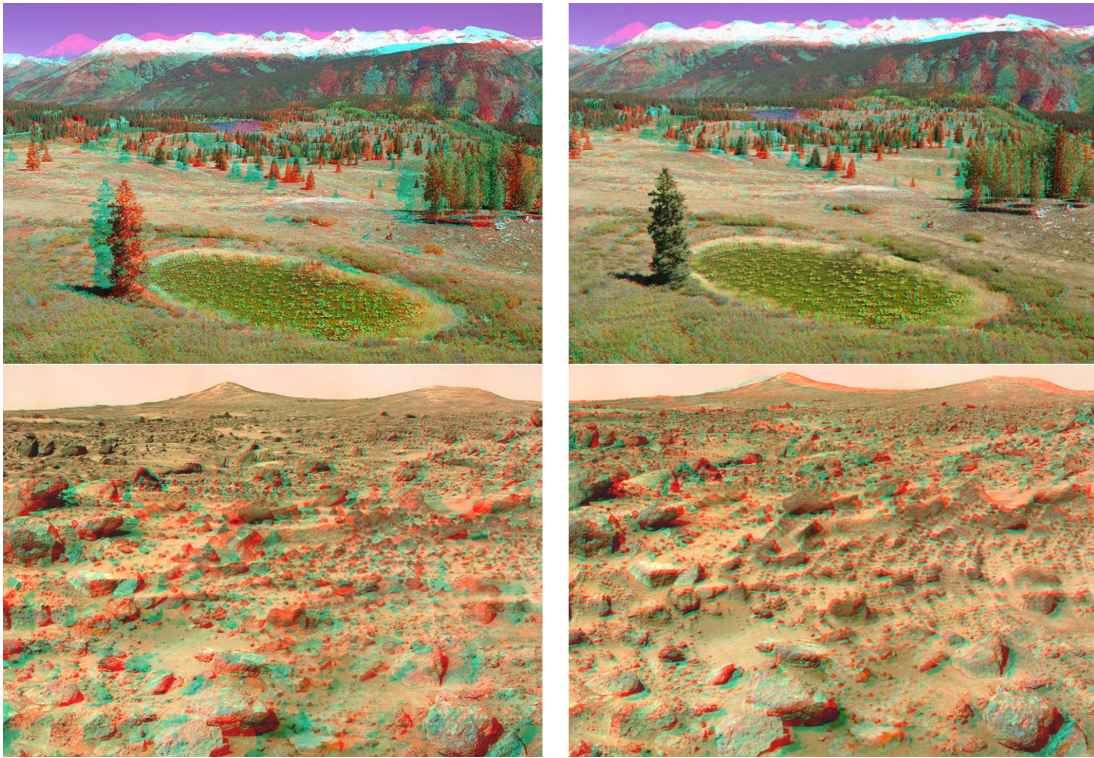


Fig. 8. Sample results of stereo enhancement using our algorithm and our learned quality model. Note that the enhanced image may be cropped slightly. Left column: original anaglyphs, Right column: improved anaglyphs.

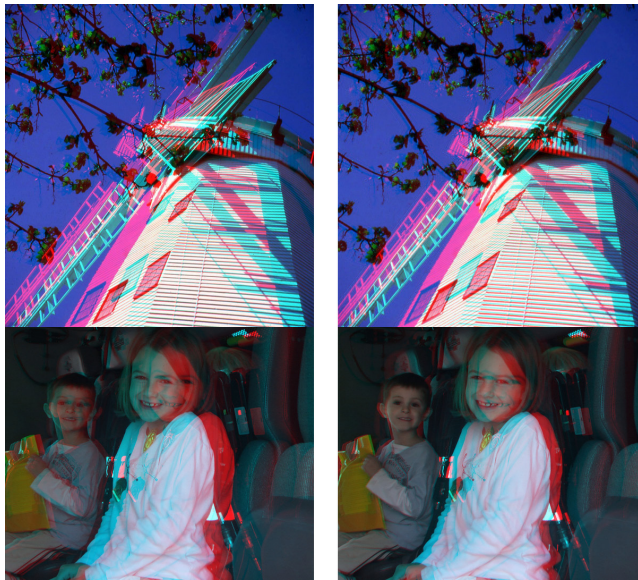


Fig. 9. More sample results of stereo enhancement. Note that the enhanced image may be cropped slightly. Left column: original anaglyphs, Right column: improved anaglyphs.

User Study: In the psychovisual study, a subject was presented with two different stereo versions of the same content, one from our proposed method and the other from the original anaglyph, in succession on screen (we choose not to display both stereo images side by side to eliminate inference; the subject can toggle back and forth between the two stereo

versions labeled “A” and “B”). All images were presented in a blind random order such that the subject cannot discern which version is produced by which method. The subject is asked to select the one that has higher stereo quality, or indicate that the both images have the same quality. Then they are required to indicate the quality difference for the preferred image on a scale of 1 to 3 with 1 being slight, 2 being moderate, and 3 being large difference, respectively. The sample display for the choices can be seen in Figure 10.

Which image has higher stereo quality?
☐ A ☐ B ☐ The Same
 What is the stereo quality difference?
☐ 1: Slight ☐ 2: Medium ☐ 3: Large
 If you consider the two images of the same quality, you will not need to specify the quality difference.
 Thank you for your participation

Fig. 10. A sample display for the study questions and choices.

A total of 13 judges who have experience in judging 2D and 3D image quality (a subset of the 19 judges) including imaging scientists rated five pairs of images in this study.

In analyzing the judges’ responses, the ratings were coded such that preferences in favor of our proposed method were given positive scores (Slight: +1, Medium: +2, Large: +3) while ratings favoring the original anaglyph images were given negative scores (Slight: -1, Medium: -2, Large: -3) and ratings with the same qualities were given zero score. Histogram

of the quality difference can be seen in Figure 11. This corresponds to 7-scale adjective categorical judgment methods in ITU-R BT.500 [15]. The 95% confidence interval for a quality difference and a sample mean of the quality difference were $[0.39, 1.18]$ and 0.78, respectively. Therefore, we can conclude that the judges preferred the improvement of the proposed algorithm.

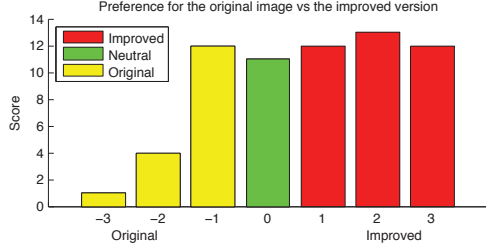


Fig. 11. Histogram of the user study where each of 13 experts is asked to express his or her preference and quality difference for five stereo image pairs. The 95% confidence interval for a quality difference and a sample mean of the quality difference were $[0.39, 1.18]$ and 0.78, respectively. Overall, the judges preferred the improvement of the proposed algorithm.

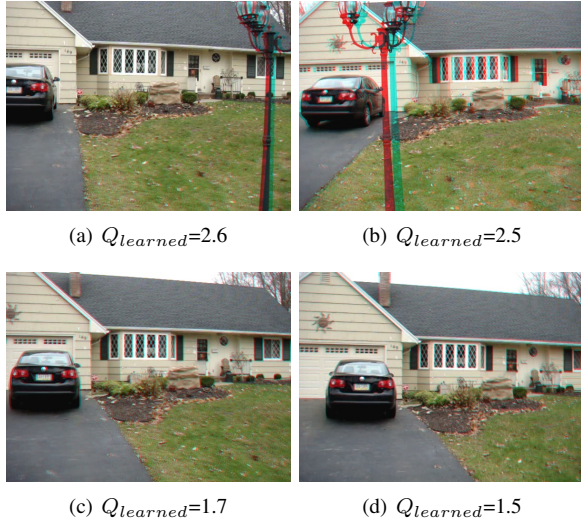


Fig. 12. Sample selected stereo key frames ($Q_{learned} \geq 1.5, level \geq 2$) from a home video. Note that the perceived 3D impression of the extracted stereo key frames increases with the computed quality value.

B. Stereo Keyframes from Video

To further demonstrate the usefulness of the proposed stereo quality metric, we apply it to a given 2D video where we sweep through entire time line t and compute the quality rating for frames pairs I_t and $I_{t+\hat{t}}$ where $\hat{t} = 1 \sim 4$ as it is proposed by Park et al. [28] where they proposed 42 geometric features to detect a stereo image pair from a captured 2D video. They train a classifier to determine good stereo frames in a captured 2D video. For detected stereo pairs, they evaluate the quality of stereo frames and select the best pairs over time t where the best pair is computed by:



Fig. 13. Sample selected stereo key frames ($Q_{learned} \geq 1.5, level \geq 2$) from a home video. Note that the perceived 3D impression of the extracted stereo key frames increases with the computed quality value.

$$Q = var(v_x^{(in)}). \quad (4)$$

We follow the same procedure proposed in [28] except that we use the proposed learned quality metric $Q_{learned}$ to determine concurrently both good stereo pairs and qualities based on the proposed 81 features. We select the best pairs over time t to produce stereo images from captured 2D video frames. Empirically, we found it adequate to search within the neighborhood of four consecutive frames since large inter-frame movement will cause large appearance variations of both rigid and non-rigid objects. The sample stereo key frames extracted from a 2D video are shown in Figures 12 and 13. Both provide good 3D perception.

VI. CONCLUSIONS

In this paper we present our efforts to assess and improve the quality of stereo images. We first describe our efforts to gather human rating input on the perceived quality of anaglyph stereo images. Although we made some of the nonstandard choices, our human ratings are statistically significant (p-value less than 10^{-6}) and have fair agreement by Fleiss' Kappa measure [9]. Next, we introduce a set of geometric features computable from a stereo image or anaglyph image based on feature point correspondence across the stereo pair or color channels. We then build a regression model that captures the relationship between these stereo features and the human consensus rated quality of anaglyph images. The performance of the regression model performs comparably to an average human judge. Finally, we present two proposed applications where the regression model is used to first notably improve the quality of captured stereo images, and second, effectively retrieve stereo key frames from a captured 2D video.

REFERENCES

- [1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, pp. 346–359, June 2008.

- [2] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Using Disparity for Quality Assessment of Stereoscopic Images," in *IEEE Proc. ICIP*, 2008.
- [3] A. Berthold, "The Influence of Blur of the Perceived Quality and Sensation of Depth of 2D and Stereo Images," ATR Human Info. Proc. Research Lab. TR, Tech. Rep., 1997.
- [4] G. Bradski, "The OpenCV Library," *Dr. Dobbs's Journal of Software Tools*, 2000.
- [5] P. Campisi, P. Le Callet, and E. Marini, "Stereoscopic Images Quality Assessment," in *Proc. European Signal Processing Conference*, 2007.
- [6] D. Cohan and E. Sadun, *Mac Digital Photography*. Sybex, 2003.
- [7] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying Aesthetics in Photographic Images," in *Proc. ECCV*, 2006.
- [8] A. Dhar, V. Ordonez, and T. Berg, "High Level Describable Attributes for Predicting Aesthetics and Interestingness," in *Proc. CVPR*, 2011.
- [9] J. Fleiss *et al.*, "Measuring Nominal Scale Agreement among Many Raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.
- [10] A. Gallagher, "Detecting anaglyph images with channel alignment features," in *Proc. ICIP*, 2010.
- [11] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.
- [12] I. Howard, *Seeing in Depth. Volume 1: Basic Mechanisms*. Oxford University Press, 2002.
- [13] I. Ideses and L. Yaroslavsky, "Three methods that improve the visual quality of colour anaglyphs," *J. of Optics A: Pure Applied Optics*, 2005.
- [14] ITU-R BT.1438, "Subjective assessment of stereoscopic television pictures," International Telecommunications Union, Tech. Rep., 2000.
- [15] ITU-R BT.500.12, "Methodology for the Subjective Assessment of the Quality of Television Pictures," International Telecommunications Union, Tech. Rep., 2009.
- [16] ITU-T Tutorial, "Objective perceptual assessment of video quality: full reference television," ITU-T Telecommunication Standardization Bureau, Tech. Rep., 2004.
- [17] S. Kang, A. Kapoor, and D. Lischinski, "Personalization of image enhancement," in *Proc. CVPR*, 2010.
- [18] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. CVPR*, 2006.
- [19] B. Keelan, *Handbook of Image Quality: Characterization and Perception*. CRC Press, 2002.
- [20] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3d," *ACM Transactions on Graphics*, vol. 29, no. 4, p. 1, 2010.
- [21] C. Li, A. Gallagher, A. Loui, and T. Chen, "Aesthetic visual quality assessment of consumer photos with faces," in *Proc. ICIP*, 2010.
- [22] C. Li, A. Loui, and T. Chen, "Towards aesthetics: A photo quality assessment and photo selection system," in *Proc. ACM MM*, 2010.
- [23] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vision*, vol. 60, pp. 91–110, November 2004.
- [24] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. of International Joint Conference on Artificial Intelligence*, vol. 1, 1981, pp. 674–679.
- [25] J. S. McVeigh, M. Siegel, and A. Jordan, "Algorithm for Automated Eye Strain Reduction in Real Stereoscopic Images and Sequences," in *Human Vision and Electronic Imaging*, vol. 2657, March 1996, pp. 307 – 316.
- [26] J. O. Merritt, "Not too far, not too close, just right: 3d tv for the home may provide an optimum viewing distance," in *WORKSHOP ON 3D IMAGING*, 2011.
- [27] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vision*, vol. 60, pp. 63–86, October 2004.
- [28] M. Park, J. Luo, A. Gallagher, and M. Rabbani, "Learning to produce 3d media from a captured 2d video," in *ACM Multimedia*, 2011.
- [29] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, 2002.
- [30] —, "High-accuracy stereo depth maps using structured light," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1, june 2003, pp. 195 – 202.
- [31] P. Seuntjens, I. Heynderickx, W. IJsselstein, P. van den Avoort, J. Berentsen, I. Dalm, M. Lambooy, and W. Oosting, "Viewing Experience and Naturalness of 3D Images," in *SPIE Optics East, IT 107 Three dimensional TV, Video, and Display IV*, B. Javidi and F. O. J.-Y. Son, Eds., vol. 6016, Boston, MA, 2005 2005, pp. 43 – 49.
- [32] M. Solh and G. AlRegib, "MIQM: A Multi-Camera Image Quality Measure," *IEEE Transactions on Image Processing (To appear)*.
- [33] L. B. Statistics and L. Breiman, "Random forests," in *Machine Learning*, 2001, pp. 5–32.
- [34] L. Stelmach and W. Tam, "Stereoscopic image coding: Effect of disparate image-quality in left- and right-eye views," in *Singal Processing: Image Communication*, 1998.
- [35] X. Sun, H. Yao, R. Ji, and S. Liu, "Photo assessment based on computational visual attention model," in *Proc. ACM MM*, 2009.
- [36] J. Luo, Q. Yiu, and M.E. Miller, "A triage method of determining the extent of JPEG compression artifacts," in *IEEE ICIP*, 2009.



Minwoo Park is a research scientist with Corporate Research and Engineering, Eastman Kodak Company, in Rochester, NY. His research area is computer vision, with current emphasis on understanding the theory and application of a probabilistic graphical model on computer vision problems. His particular interests are in automatic understanding of 3D from an image, perceptual grouping, event recognition, and an efficient inference algorithm. He received the B.Eng. degree in electrical engineering from Korea University, Seoul, in 2004, the M.Sc.

degree in electrical engineering from The Pennsylvania State University in 2007, and the Ph.D. degree in computer science and engineering from The Pennsylvania State University in 2010. He is a co-organizer of the Joint IEEE/IS&T Western New York Image Processing Workshop 2011 and a leadership committee of IEEE Signal Processing Society, Rochester Chapter. He routinely serves as a reviewer for IEEE, ACM, Springer, and Elsevier conferences and journals in the area of computer vision. He is a member of the IEEE, the IEEE Signal Processing Society, and the IEEE Computer Society.



Jiebo Luo is a Senior Principal Scientist with Corporate Research and Engineering, Eastman Kodak Company, in Rochester, NY. His research interests include image processing, computer vision, machine learning, social media data mining, medical imaging, and computational photography. Dr. Luo has authored over 180 technical papers and holds over 60 US patents. Dr. Luo has been actively involved in numerous technical conferences, including serving as the general chair of ACM CIVR 2008, program co-chair of IEEE CVPR 2012, ACM Multimedia

2010 and SPIE VCIP 2007, area chair of IEEE ICASSP 2009-2010, ICIP 2008-2010, CVPR 2008 and ICCV 2011, and an organizer of ICME 2006/2008/2010 and ICIP 2002. Currently, he serves on several IEEE SPS Technical Committees (IMDSP, MMSP, and MLSP) and conference steering committees (ACM ICMR and IEEE ICME). He is the Editor-in-Chief of the Journal of Multimedia, and has served on the editorial boards of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), the IEEE Transactions on Multimedia (TMM), the IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), Pattern Recognition (PR), Machine Vision and Applications (MVA), and Journal of Electronic Imaging (JEI). He is a Fellow of the SPIE, IEEE, and IAPR.



Andrew C. Gallagher earned the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University in 2009. He received his M.S. degree from Rochester Institute of Technology in 2000, and the B. S. degree from Geneva College in 1996, both in electrical engineering. Dr. Gallagher joined the Eastman Kodak Company, Corporate Research and Engineering in 1996, initially developing image enhancement algorithms for digital photofinishing. This effort resulted in the award of more than 80 U.S. Patents and Kodak's prestigious Eastman

Innovation Award in 2005. More recently, Dr. Gallagher's interests are in the arena of improving computer vision by incorporating context, human interactions, and image data. Further, Andrew enjoys working in the areas of graphical models and image forensics.