An integrated system for virtual scene rendering, stereo reconstruction, and accuracy estimation.

Marichal-Hernández J.G., Pérez Nava F*., Rosa F., Restrepo R., Rodríguez-Ramos J.M.
Universidad de La Laguna, Facultades de Física y Matemáticas.
{jmariher@ull.es, fdoperez@ull.es, frosa@ull.es, rrestrep@iac.es, jmramos@ull.es}

Abstract

This article depicts an integrated system capable of generating virtual stereo scenes, reconstructing a three-dimensional map of distances, and calculating the accuracy of the results obtained. This tool is particularly useful for verifying the validity of distance reconstruction techniques, and is robust enough to use in a wide variety of scenarios since it avoids the need for an image-capture system in laboratory simulations. The importance of the distance reconstruction problem lies in the range of its applications: industrial inspection and quality control, surveillance and security, autonomous vehicles, robotic systems, medical image analysis, image databases, virtual reality, telepresence, and telerobotics.

The system is based on the VRML standard for three-dimensional geometric scene representation, on OpenGL for rendering virtual scenes, and on the TRW algorithm for the problem of three-dimensional reconstruction.

Keywords- **3D modeling, rendering, stereo reconstruction, stereo accuracy and validation**.

1. Introduction

The goal of 3D scene reconstruction is to take a set of images and estimate the positions and orientations of the cameras that produced the images, as well as a representation of the scene that was imaged. This is an example of an *inverse problem* named the stereo reconstruction problem. The *forward* (or direct) problem is: given a scene and the position and orientation of a set of cameras, what is the expected image? This is the area of computer graphics known as rendering [1]. While both problems have their own difficulties, it is widely believed that the inverse problem is considerably harder than the direct problem.

Recent advances in the development of efficient optimization algorithms [2] have produced a dramatic impact on the stereo reconstruction problem, according to several benchmarks that evaluate stereo algorithms using real scenes with known reconstructions [3]. Almost all top-performing methods approach the 3D reconstruction problem from the point of view of energy function optimization.

In our system we provide a solution for both the direct (rendering) and inverse (stereo reconstruction) problem. This allows us to estimate the accuracy and robustness of the inverse solution. The need for a tested solution to the stereo problem is of great importance for several applications in diverse fields.

One of these applications is the Virtual Acoustic Space (VAS). The VAS is a multidisciplinary problem that aims to model a three-dimensional space by just using sounds. The project's goal is to capture a 3D scene via cameras and to obtain a set of elements, the acoustic projection of which will give the illusion of having the objects covered by a thin, sound-producing layer. A world of virtual sounds could thus be generated which directly corresponds to the real world.

Achieving the objective of the VAS requires techniques in artificial vision, 3D reconstruction, and sound spatialization. The 3D reconstruction of a scene yields elements such as distance maps, in which the sound sources are placed at given distances by using sound spatialization. The technique of spatializing sounds by way of a transfer function related to the subject who receives the sounds, called a Head-Related Transfer Function (HRTF), is sufficiently advanced to assure proper spatial perception.

Closely related with the VAS is the VSIDS (Virtual Scene Immersive Distance Simulator) module, which solves the direct problem thus allowing for the substitution of the image capture system by a simulation system which permits for easier testing of different scenes and for improved 3D reconstruction, sound generation, and spatialization. The simulator can be used to perform experiments that clarify both the audio channel's capacity to transmit the information of the scene's three-dimensional structure, as well as the ability of the employed stereo algorithm to resolve the inverse problem by measuring the degree of precision in retrieving the 3D distances. In this paper we present the results of such experiments, making use of stereo reconstruction techniques to calculate the distance map. Figure 1 shows a complete diagram of the modules already implemented. The TRW [4] algorithm, currently one of the most advanced, is used for the stereo problem.

*Corresponding author

This algorithm can also be generalized to multiview stereo. Its parallel nature lets it perform real-time image sampling as long as the hardware chosen is adequate, for example a Graphics Processing Unit (GPU).

2. The VSIDS Module

The VSIDS simulator as a whole consists of a system for acquiring the position and orientation of the head of the user to which it is connected. This information is then used to position a virtual camera in a virtual scene, as required by the experiment being performed. The simulator allows for a measurement of distance, color, and outline parameters for objects located within the field of view of the camera. When the virtual scene is made to match the reality that surrounds the user, while at the same time synchronizing his movements, the result is an immersive virtual reality system. The information output by the simulator is then used to generate the spatiated sounds that the user hears. The goal of this feedback is to achieve acoustic immersion without transmitting to the user any visual information.

The elements handled by the simulator are a virtual world on one hand, and the real world position and orientation of the user's head on the other. This module merges both elements, rendering with a virtual camera located within a virtual scene that which a real camera in the same position would see. The rendering does not yield what might be called a "real-world" image. Additionally, the renderer must supply distance information, distances measured from the camera to the objects, as well as a geometric outline of the edges in the world.

Since the communications channel with the user must be necessarily reduced, a strategy must be specified for selecting the most relevant information before it is sent to the generator. This selection takes place when deciding on a stereo-pixel or sound-voxel strategy. The next module, the sound generator, is capable of producing in the user the sensation that a sound proceeds from a specific point in space. If a pixel is an image element in the two-dimensional space represented on a screen, then a stereo-pixel is a sound element in the three-dimensional space of the user's acoustic perception. The stereo-pixel selection strategy reconstructs, calculates, and selects the attribute information generated by the simulator module which is to be sent to the next block.

2.1 Scene Loader (SL).

Of fundamental importance to the simulator is the potential and versatility it is given for representing the three-dimensional geometry of the worlds it simulates. These requirements hinge on two points: the representation of the world in the simulator's memory, and its file representation while keeping to a geometric description standard. The choice of a standard must conform to the needs of both the simulator and the creator of the virtual worlds, a 3D designer used to working with the highly specialized software products developed for this task: 3DStudio, LightWave, AutoCad, or Blender, among others. To this end we have implemented, after considering different options, a VRML world loader, subject to the VRML 97 norm (The Virtual Reality Modeling Language International Standard ISO/IEC 14772-1:1997). Said module will let us manipulate the geometry exported by the different 3D design platforms. Among the characteristics of the module implemented stand out the ability to organize and subdivide files (Inline nodes), geometric object reutilization (DEF-USE nodes), and triangular and square color geometry support with surface or node normals (Transform-Shape-geometry Indexed Face Set nodes). This module, combined with the one for hardware accelerated three-dimensional representation described later, achieves high-resolution interactive frame rates in excess of 18 frames per second in scenes with one million vertices.

2.2. Virtual world rendering (VWR)

Once a geometric representation of the virtual world is available in memory, the steps needed to immerse the user in the simulated environment require registering his position in the virtual world, adjusting this measurement to the scale of the virtual world, and placing a camera in the virtual world. Put another way, the role of the user in the virtual world is played by the camera in front of which the world is projected. This camera follows in the virtual scene the user's movements in the real world.

For rendering the virtual world, we have chosen the OpenGL standard for three-dimensional hardware acceleration. To improve performance we have used the advanced techniques contained in specification 1.3 of the standard (OpenGL 1.3 Specification), specifically the DrawElements extension along with generation lists, which allows for interactive manipulations in frames containing up to one million vertices.

Adequately formatted information from the world, as well as the camera's position, must be constantly relayed to the API OpenGL of the graphics card. The hardware-software platform consists of a ViewPoint nVidia Geforce 7800 SE graphics card and an AMD 3500+ with 1 GB of memory, running the GNU/Linux Debian 3.0 OS. The world represented is generated by the LightScape utility in test mode. Even if the worlds needed for experiments are not necessarily as sophisticated and complex as the ones shown, the utility is ready to support highly elaborate models.

2.3. Distance extractor (DE).

What distinguishes the simulator we implemented from other virtual world viewers is the emphasis it places on acquiring the distances from the objects to the point of

observation. To obtain these measurements we have resorted to state of the art 3D graphics acceleration hardware, which yielded two methods for manipulating distances. The first method, based on accessing the internal Z-buffer of the OpenGL machine, is slower but compatible with the majority of graphics cards on the market. The second method is based on the programming of the more advanced GPUs (Graphics Processing Units), only available on the most recent cards, such as the GeForce4 Ti. With the first method, interactive distance extraction is only possible with the simplest scenes. For more complex worlds, this method of distance extraction is only viable up to 10 frames per second. Using higher capacity hardware, however, allows for higher interactivity rates within complex sceneries.

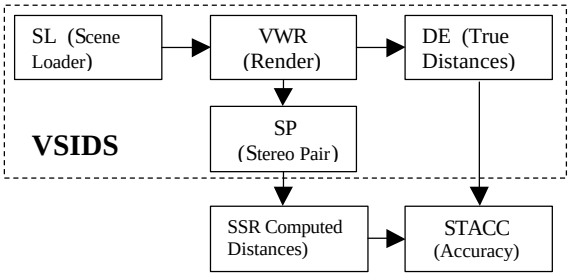

Figure 1: General Scheme.

3. Stereo Scene Reconstruction: The SSR module.

The distance maps generated by VSIDS are very useful for testing the viability of the user's acoustic immersion in the 3D scene under ideal circumstances. However, the distance map reconstruction in VSIDS is obtained by resolving the direct problem when in fact, for the acoustic immersion to be of practical use, the distance map must be obtained by resolving the inverse problem (obtaining the geometric model based on the images of the scene). This inverse problem is the stereo problem, of great importance in computer imaging. To solve this problem we could use several publicly available stereo algorithms. However, in our system we have implemented the Tree Reweighed Message Passing (TRW) [4] algorithm. There are two main reasons for this selection: first, it outperforms other well known stereo algorithms based on Loopy Belief Propagation or Iterated Conditional models [2]; and second, it can be paralellizable if real time performance is needed.

3.1 Stereo reconstruction with the TRW algorithm.

In this section we will present an overview of the TRW algorithm. Given a stereo pair composed of two images I and I', the stereo problem can be solved by assigning to every pixel p on image I of the stereo pair a disparity value (proportional to the inverse of the 3D distance), which we write as d_p. This disparity is equal to the difference in image coordinates between pixel p in I and its corresponding pixel in the other image I' of the stereo pair. Therefore, to obtain the disparity it is necessary to obtain for each pixel in I its corresponding pixel in the other image I'. This correspondence problem can be solved by a variety of approaches. State of the art methods solve this correspondence problem by obtaining the minimum of an energy function E.

This energy function E, which can also be viewed as the log likelihood of the posterior distribution of a Markov Random Field (MRF) [5], is composed of a data energy E_d and smoothness energy E_s, $E = E_d + \lambda E_s$, where the parameter λ measures the relative importance of each term. The data energy is simply the sum of a set of per-pixel data costs $c_p(d)$, $E_d = \Sigma_p c_p(d_p)$. In the MRF framework, the data energy comes from the (negative) log likelihood of the measurement noise. We assume that pixels in the image form a 2D grid, so that each p can also be written in terms of its coordinates $p = (i, j)$. If we use the standard 4-connected neighborhood system, then the smoothness energy is the sum of spatially varying horizontal and vertical nearest-neighbor smoothness costs, $V_{pq}(d_p, d_q)$, where if $p = (i, j)$ and $q = (s, t)$ then $|i - s| + |j - t| = 1$. If we let N denote the set of all such neighboring pixel pairs, the smoothness energy is:

$$E_s = \sum_{\{p,q\} \in N} V_{pq}(d_p, d_q) \qquad (1)$$

In this paper we will use:

- $c_p(d_p) = \| I(p) - I'(p+d_p) \|^2$, where $I(p)$ and $I'(p+d_p)$ denote image intensities in pixel p on image I and pixel $p+d_p$ on image I' (as vectors due to the use of color images).

- $V_{pq}(d_p, d_q) = 0$ if $d_p \neq d_q$
  $V_{pq}(d_p, d_q) = 1$ if $d_p = d_q$.

Tree-reweighed message passing (TRW) [4] is a message-passing algorithm that can be used to obtain an approximate minimum of the energy function E.

Let $M^t_{p \to q}$ be the message that pixel p sends to its neighbor q at iteration t; this is a vector of size m (the number of possible disparities). The message update rule is:

$$M^t_{p \to q}(d_q) = \min_{d_q} \left(\begin{array}{l} m_{pq}\left(c_p(d_p) + \sum_{s \in N(p)} M^{t-1}_{s \to p}(d_q) - \right. \\ \left. - M^{t-1}_{q \to p}(d_q) + V_{pq}(d_p, d_q) \right) \end{array} \qquad (2)$$

The coefficients m_{pq} are determined as follows: first, a set of trees from the neighborhood graph (a 2D grid in our case) is chosen so that each edge is in at least one tree; then a probability distribution r over the set of

trees is chosen; finally, m_{pq} is set to r_{pq}/r_p, i.e. the probability that a tree chosen randomly under r contains edge (p, q) *given* that it contains p.

An interesting feature of the TRW algorithm is that for any messages it is possible to compute a lower bound in the energy E. The original TRW algorithm used in this paper does not necessarily converge, and does not, in fact, guarantee that the lower bound always increases with time. These problems can be solved using damping in the TRW updates or by the use of a sequential version of the TRW. [6]

4 Stereo Accuracy Estimation: The STACC module

In this section we describe the quality metrics we use for evaluating the performance of the stereo correspondence algorithm in the simulated scenes. Two general approaches for validating stereo algorithms are to compute error statistics with respect to some ground truth data, or to evaluate the synthetic images obtained by warping the reference or unseen images by the computed disparity map. Since the VSIDS module provides us ground truth data of the scene, we have chosen the first approach. Following [3], we compute two quality measures based on known ground truth data:

1. Relative RMS (root-mean-squared) error (measured in distance units) between the computed distance map $z_C(i, j)$ and the ground truth map $z_T(i, j)$, i.e.,

$$R = \frac{\left(\frac{1}{N} \sum_{(i,j)} \left(z_C(i,j) - z_T(i,j) \right)^2 \right)^{1/2}}{zmax_T} \qquad (3)$$

where N is the total number of pixels and $zmax_T$ is the maximum true distance in the scene.

2. Percentage of bad matching pixels,

$$B = \frac{100}{N} \sum_{(i,j)} \left(\left| z_C(i,j) - z_T(i,j) \right| > d \; zmax_T \right) \qquad (4)$$

where d is a relative error tolerance. For the experiments in this paper we use $d = 0.1$.

4. Experimental Results.

In this section, we describe the experiments used to evaluate the individual modules of our system. Our main interest is to test the coherence between the direct and inverse estimations of the distance map.

4.1 Test data

To evaluate the system in complex environments we have selected two realistic indoor scenes. The first scene will be named the “armchair” scene and the second the “table” scene. The stereo pair rendered for both scenes by the VSIDS module is shown in Figures 2 and 4 respectively. Several difficulties arise when any stereo algorithm tries to find the correct reconstruction in both cases: there are many textureless regions (walls, floors) where it is very difficult to locally solve the correspondence problem, so the smoothness constraint has to be enforced. By contrast, there are abrupt discontinuities in the distance map (chairs, armchairs, table) where the smoothness constraint fails. Note also that these abrupt changes usually coincide with occluded points.

Both image pairs were generated as RGB 256x256 images. No color pre-processing was done and the stereo algorithm used the three components of the RGB image.

4.2 Accuracy Results

Once the image pair was generated we reconstructed the distance map with the TRW algorithm. In the "armchair" scene the disparities were in the range [-34, -6]. For the "table" scene the disparities were in a wider range: [-57, -8]. As distances are inversely related to disparities, an effect of the discrete nature of the disparities is that the distance map is more finely sampled at near distances and more coarsely sampled at far distances. This quantization effect is clearly shown in the presented inverse distance maps.

The true and computed distance maps for the “armchair” and “table” scenes are shown in Figures 3 and 5. Numerical results for the quality measures for optimal values of l (see Section 4.3) are given in Table 1. The distribution of errors is shown in Figures 6 and 7.

Figure 2: The “armchair” stereo pair

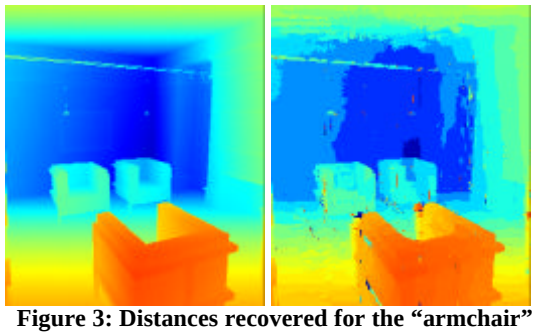

Figure 3: Distances recovered for the “armchair” stereo pair by the direct (rendering) and inverse (stereo reconstruction) methods.

Figure 4: The “table” stereo pair

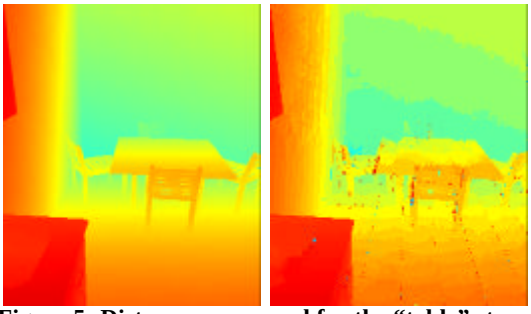

Figure 5: Distances recovered for the “table” stereo pair by the direct (rendering) and inverse (stereo reconstruction) methods.

Scene	l	R	B
"Armchair"	6.0	0.0514	2.29%
"Table"	5.1	0.0288	1.04 %

Table 1: Quality measures for the stereo reconstruction

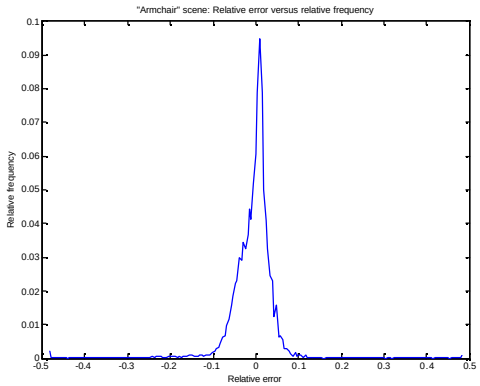

Figure 6: Error (direct estimation minus inverse estimation) for the “armchair” stereo pair

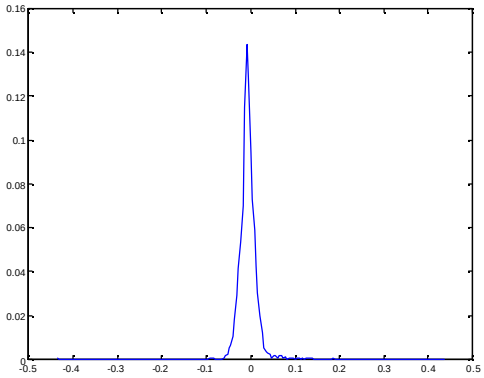

Figure 7: Error (direct estimation minus inverse estimation) histogram for the “table” stereo pair

Results show that a reliable 3D reconstruction is obtained with the TRW method. Relative RMS error for the table scene is very small and slightly larger for the armchair scene. The percentage of misclassified pixels is again very small for the table scene and larger for the armchair. However, reconstruction errors in the armchair scene are biased by distance errors in the front wall, while distances near the user are well reconstructed.

4.3 Optimal values for the parameters

Another type of experiment tried to determine the optimum value for l in the energy function $E = E_d + l E_s$. To obtain this optimum we computed the R quality measure for several values of l. In Figures 8 and 9 we can see a plot of the R measure versus l. Optimum values were l =6.0 and l=5.1 for the "armchair" and "table" scenes respectively.

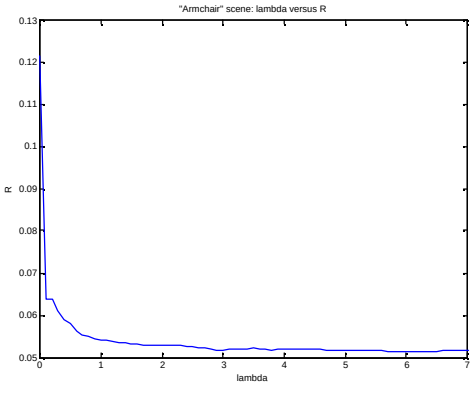

Figure 8: Determination of the optimum value for l in the “armchair” stereo pair. Plot of l versus the quality measure R.

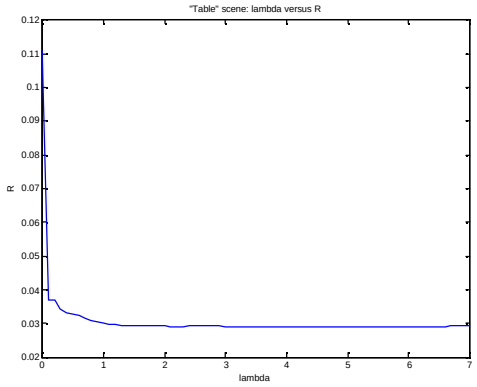

Figure 9: Determination of the optimum value for l in the "table" stereo pair. Plot of l versus the quality measure R.

As we can in Figures 8 and 9 there is a wide range of l values that produce approximately the same values for the R quality measure in both scenes. This is highly desirable since the computation of the optimal l value has a high computational cost. In fact, Figures 8 and 9 were obtained running the TRW algorithm for several partitions of the possible values of l in a cluster of 64 Intel Itanium 2 processors at the SAII Service of the University of La Laguna (ULL).

The last type of experiment tried to determine the optimum value for the number of iterations in the TRW algorithm. To obtain this optimum we computed the R quality measure at each iteration of the algorithm for the optimum value of l. In Figures 10 and 11 we can see a plot of the R measure against the number of iterations. The optimum values for the number of iterations *nit* were *nit* = 13 and *nit* =16 for the "armchair" and "table" scenes respectively.

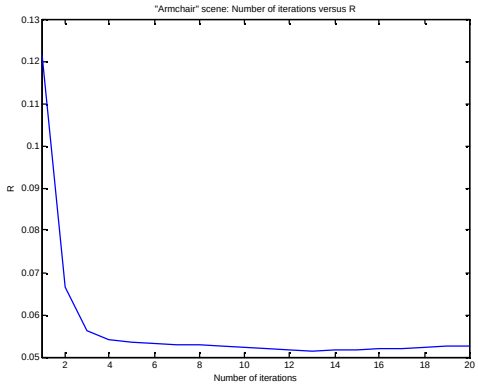

Figure 10: Plot of the number of iterations versus the quality measure R in the "armchair" stereo pair for the optimal l value.

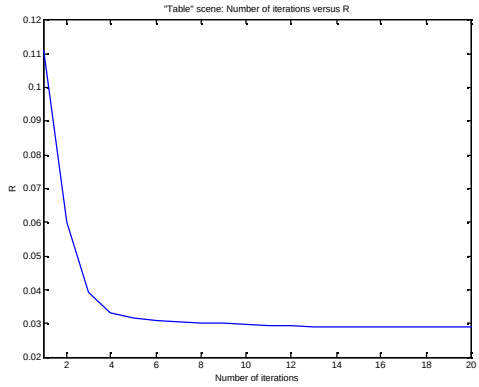

Figure 11: Plot of the number of iterations versus the quality measure R in the "table" stereo pair.

5 Conclusions and future extensions

We have implemented a robust tool for testing different stereo algorithms and scenarios running on powerful electronic hardware, namely, a GPU where real time developments and high sample rates can be successfully processed. The good results shown here can be improved by generalizing the stereo algorithm for multi-view stereo vision. Even more realistic scenes and added textures can be easy simulated by using GPUs.

Acknowledgements

This work has been partially funded by "Programa Nacional I+D+i" (Projects DPI 2003-09726 and TIC2001-3916) of the "Ministerio de Ciencia y Tecnología", by the "European Regional Development Fund" (ERDF), and by "Becas Convenio Cajacanarias-ULL (2005)". We thank the SAII Service of the ULL for the use of its computing facilities and to Sergio Ortiz Eirín for a critical reading of this paper.

References

[1] J. Foley, A. van Dam, S. Finer, and J. Hughes. *Computer Graphics, principles and practice*. Addison-Wesley, 2nd ed. edition, 1990.

[2] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, C. Rother. "A Comparative Study of Enegy Minimization Methods for Markov Random Fields", unpublished

[3] D. Scharstein and R. Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. IJCV 47(1/2/3):7-42, April-June 2002

[4] M. Wainwright, T. Jaakkola, A. Willsky, TRW belief propagation and approximate ML estimation by pseudo-moment matching. In: AISTATS. (2003)

[5] S. Geman, D. Geman: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6 (1984) 721–741

[6] V. Kolmogorov and M. Wainwright. On optimality of tree-reweighted max-product message-passing. Uncertainty in Artificial Intelligence, July 2005, Edinburgh, Scotland.