Clausius Normalized Field-Based Stereo Matching for Uncalibrated Image Sequences

Eunjin Koh, Jaeyeon Lee, and Junseok Park

We propose a homology between thermodynamic systems and images for the treatment of time-varying imagery. A physical system colder than its surroundings absorbs heat from the surroundings. Furthermore, the absorbed heat increases the entropy of the system, which is closely related to its disorder as given by the definition of Clausius and Boltzmann. Because pixels of an image are viewed as a state of lattice-like molecules in a thermodynamic system, the task of reckoning the entropy variations of pixels is similar to estimating their degrees of disorder. We apply this homology to the uncalibrated stereo matching problem. The absence of calibrations alleviates user efforts to install stereo cameras and enables users to freely modify the composition of the cameras. The proposed method is also robust to differences in brightness, white balancing, and even focusing between stereo image pairs. These peculiarities enable users to estimate the depths of interesting objects in practical applications without much effort in order to set and maintain a stereo vision setup. Users can consequently utilize two webcams as a stereo camera.

Keywords: Computer vision, uncalibrated stereo matching time-varying imagery, HCI.

Manuscript received Mar. 15, 2010; revised July 22, 2010; accepted Aug. 2, 2010.

This work was supported by the IT R&D program of MKE/KEIT [KI002096, Contact-free Multipoint Realistic Interaction Technology Development].

Eunjin Koh (phone: +82 42 860 1842, email: eikoda@gmail.com) is with the Software Research Laboratory, ETRI, Daejeon, Rep. of Korea, and is also with the University of Science and Technology, Daejeon, Rep. of Korea.

Jaeyeon Lee (email: leejy@etri.re.kr) is with the IT Convergence Technology Research Laboratory, ETRI, Daejeon, Rep. of Korea.

Junseok Park (email: parkjs@etri.re.kr) is with the Software Research Laboratory, ETRI, Daejeon, Rep. of Korea.

doi:10.4218/etrij.10.1510.0067

I. Introduction

Stereo matching has been extensively studied in recent years because of its important applications in the areas of surveillance [1], reconstruction of 3D coordinates [2], human-computer interactions (HCIs) [3], 3D television (3DTV) [4], 3D video coding [5], and face recognition [6]. However, most stereo vision techniques have not been successfully utilized in many practical applications. There may be several reasons for this, but the main problem is the difficulty of installation. To use stereoscopic vision, users must rectify and calibrate stereo image pairs and maintain the composition of calibrated stereo cameras. To adjust the composition, users must fulfill the calibration task each time they wish to make an adjustment. In addition, even an accidental microscopic change in composition totally spoils the stereo matching. Though there have been various efforts to obviate the need for explicit epipolar lines [7]-[10], they cannot be applied to real-time applications because the efforts make the stereo matching task extremely slow.

To make a real-time stereo vision system without calibration, we employ a feature-based approach (FBA) rather than an area-based approach (ABA) because an FBA is often faster than an ABA and is robust to environmental factors, such as differences in brightness, white balance, and lens focus between stereo image pairs. Although an FBA is known to be unable to make dense disparity maps [10]-[12], we devised a way of making maps semi-dense enough to be used for practical applications, and we can adjust the density level by optimizing the various parameters.

Many of the current best-performing techniques [13]-[16] are based on formulations of either Markov random fields (MRFs) [17] or conditional random fields [18]. However, most

of them need to estimate the global parameters by using the maximum a posterior estimate, simulated annealing [17], [19], improved iterative scaling [20], or the Gibbs sampler [17]. Because these iterative processes require considerable computational power, general systems have difficulty processing them in real time. Moreover, most random fields are used as postprocessing steps to reduce the noise in the quantized results of disparity maps. Quantization necessarily accompanies a reduction of information, which may lead to a generation of incorrect results because the reduced information cannot be successfully conveyed to the postprocessing steps. Thus, we need a strategy that can reduce the number of quantizing operations.

To solve these problems, we adopt a thermodynamic approach and propose a Clausius normalized field (CNF). The essential concepts of the CNF derive from the entropy definition of Clausius [21] and a thermodynamic system that exchanges heat with its surroundings. We assume a solid plane, O_a , that consists of many cliques of molecules. The solid plane, O_a, is hotter than the surrounding atmosphere and is periodically contacted by a different plane, O_b, which is hotter than O_a. Because of temperature differences among the planes and atmospheres, O_a absorbs heat from O_b and emits heat to its surroundings. The heat absorbed from O_b raises the entropy of O_a. Consequently, the molecules of O_a become more unstable due to increased disorder [22].

With the thermodynamic properties, we make a homology with images, which can be roughly described as follows:

- Let O_a be an ideal image plane made of many cliques C, where, as shown in Fig. 1, each clique is comprised of many molecules. Each molecule of O_a corresponds to each pixel of an image. Each clique can have a different temperature.
- Let O_b be the input image from an image sequence made up of cliques of the same size as O_a. Only selected feature points are activated molecules of O_b. Deactivated molecules are ignored in the following steps.
- Each molecule of O_a absorbs heat from each corresponding molecule of O_b in every frame, and the entropy of the molecule of O_a is increased by the absorbed heat as given by the entropy definition of Clausius. A molecule that momentarily has great increments in entropy is likely to be a part of an outlier because of its increased level of disorder.
- Cliques of O_a emit heat to the surrounding atmosphere in every frame until the temperatures are equal to the temperature of the surroundings. The temperatures decrease when the emitted heat is greater than the absorbed heat, and vice versa.
- Increments in entropy are smaller when heat is supplied to the molecules with higher temperature states and vice versa as given by the entropy definition of Clausius.
- Note that CNF does not estimate absolute values of entropy but variations of entropy. Because the amounts of variation in entropy can be summated [23], the spatial and temporal information can naturally be considered together by adding entropy variations produced by neighboring cliques and previous frames. This property enables the proposed uncalibrated stereo matching system to avoid local maxima.
- Note also that there is only one quantization step per frame as outliers are eliminated.

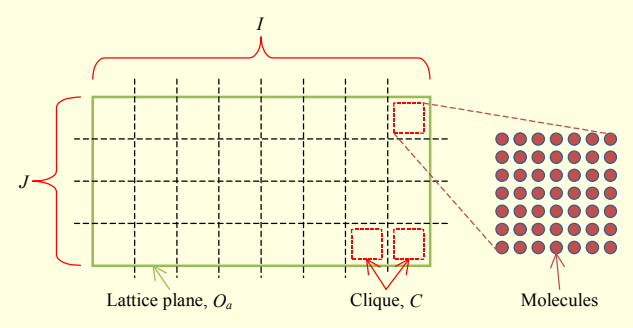


Fig. 1. Ideal lattice plane made of $I \times J$ cliques. Each clique is made up of many molecules.

We emphasize that the purpose of this paper is not to propose a specific matching algorithm that performs better than existing techniques. Rather, we introduce a method that boosts the performance and accuracy of existing algorithms for time-varying applications, uncalibrated stereo matching in particular, by optimizing various parameters along with the elapse of time. Although we utilize a matching algorithm based on optical flow techniques [24], any feature matching algorithm could be utilized instead.

II. CNF

In thermodynamics, entropy is a measure of the energy of a system that is unavailable to do work with heat. Entropy is also a measure of disorder [22]. In this section, we introduce the CNF by focusing on the relations among heat, entropy, and disorder.

1. Entropy Definitions of Clausius and Boltzmann

In the thermodynamic domain, entropy S is not defined directly but rather by the following equation of relative variation with the exchanged heat of a system by Rudolf Clausius [23]:

$$\Delta S = \frac{\Delta Q}{T}, \tag{1}$$

where ΔQ is the amount of heat exchanged in an isothermal and reversible reaction, and T is the absolute and constant temperature. If the temperature is not a constant value, (1) becomes a differential equation:

$$dS = \frac{dQ}{T}. \tag{2}$$

Because the exchanged heat generally changes the temperature of the system, the total entropy change [23] can be defined as

$$\Delta S = \int \frac{1}{T(Q)} dQ, \tag{3}$$

where $T(Q)$ is a temperature function of heat, Q.

According to the entropy definition of Ludwig Boltzmann, entropy is also related to probability as a measure of randomness or disorder in the statistical thermodynamics. Higher entropy indicates higher disorder because entropy is proportional to the natural logarithm of the number of possible microstates corresponding to the macroscopic state of a system, W, as

$$S = k_{\mathrm{b}} \ln W, \tag{4}$$

where k_{b} is Boltzmann's constant. This formula is considered as the elemental definition of entropy. In 1896, Boltzmann proved that (4) provides a measure for the entropy of classical thermodynamics, (1), by showing that (4) gives a measure of entropy for systems of molecules in a gas phase. Actually, all other entropy definitions can be mathematically derived from (4).

Let W_a be the number of microstates of macrostate a and W_b be the number of microstates of macrostate b. Thus, the number of total accessible states, W_{a+b}, is $W_a \times W_b$. Let $S(\rho)$ be the amount of entropy of ρ. Then, (4) proves the following equations:

$$S(W_{a+b}) = S(W_a) + S(W_b), \tag{5}$$

$$dS(A \cup B) = dS(A) + dS(B), \tag{6}$$

where A and B are exclusive systems comprised of different molecules.

Equation (6) means that the amount of variation in the entropy of a specific system can be added to the amount of variation in the entropy of other systems. Furthermore, because dS is not the absolute amount of entropy but the amount of variation in entropy over a unit period of time (dt), the addition of the amounts of entropy variations during a certain period (Δt) gives the entire amount of entropy variation ΔS during Δt. This phenomenon can also be analogized by (3). Thus, (3) and (6) give a natural way of considering both spatial and temporal information in images. In other words, the amount of variation in the entropy of a pixel can be added to the amount of variation in the entropy of the pixel in the same location of previous frames. Similarly, the amount of variation in the entropy of a pixel can also be added to neighboring pixels of the present frame.

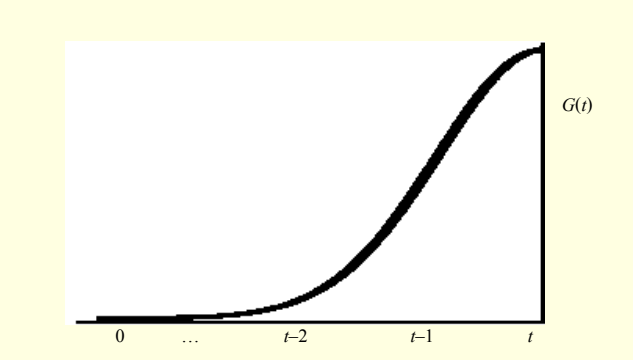


Fig. 2. Example of $G(t)$. A well-defined $G(t)$ can decide Δt. The weight of t is heaviest and the weight becomes less as time approaches zero.

2. Definition of the CNF

The probability of the CNF in images is defined as

$$p(x, y, t) = \frac{\int_0^t G(t) \cdot dS_{\mathrm{net},t}\, dt}{\int_0^t G(t) \cdot dS_{\mathrm{netMax},t}\, dt}, \tag{7}$$

where x and y are the horizontal and vertical pixel indices, $G(t)$ is a temporal function that refers to time, and dS_{net} and dS_{netMax} are given as

$$dS_{\mathrm{net}} = \sum_{i,\,j} N(i,j) \cdot dS_{i,\,j}\ , \tag{8}$$

$$dS_{\mathrm{netMax}} = \sum_{i,\,j} N(i,j) \cdot dS_{\mathrm{max},\,i,\,j}\ , \tag{9}$$

where i and j are indices of cliques, $N(i, j)$ is a spatial function that refers to neighboring cliques, $dS_{i,j}$ is the entropy variation produced by corresponding cliques, and $dS_{\mathrm{max},ij}$ is the maximum amount of entropy variation during a unit of time (dt). However, we can ignore $dS_{i,\,j}$ in most cases. Because $\Delta S/\Delta T$ is ultimately necessary for using CNF rather than $p(x, y, t)$, we redefine (7) as a simpler and general form as in (10) by using the appropriate function $G(t)$ as shown in Fig. 2. The function, $G(t)$, is treated as a weight function affected by time in (10). In Fig. 2, the weight of t is the heaviest, and the weight lessens as the time approaches zero. Thus, Δt can be decided by $G(t)$.

$$p'(x, y, t) = \frac{\Delta S}{\Delta t} = \int_0^t G(t) \cdot dS_{\mathrm{net},\,t}\, dt. \tag{10}$$

3. Advantages of the CNF

The CNF has some beneficial features. First, it can naturally consider spatial and temporal information together to avoid local maxima in conformity with the rules of thermodynamics. Because a moving object in an image sequence is represented as a group of feature points and the position of the group changes with the elapse of time, spatial and temporal information needs to be considered. Traditional optimization techniques, such as SA or MRFs, usually consider only one function to be optimized or spatial information. An MRF is sometimes modeled as a method that considers temporal information [25]. However, this is a forced approach in physics because the potential energy noted by Ising or Gibbs [17] is not related to the temporal changes of one molecule but to the spatial states among neighboring molecules.

Second, the CNF is simple and entails a small calculation load because it needs no additional iterative processes, for example an MAP, SA, IIS, or Gibbs sampler, to estimate values that are difficult to measure [17]. Instead, the CNF iteratively adapts some parameters along with specific rules with the elapse of time. This way of adaptation is sufficient to produce fast and correct decisions because the CNF is designed for time-varying imagery.

Third, the quantization is minimized. The quantization of intermediate results is unnecessary because, in contrast to other random fields, the CNF is not a postprocessing step. Accordingly, there is no loss of information, and the results are more accurate.

Fourth, whereas the temperature of SA only decreases, the CNF offers information regarding the degree of disorder as a temperature by increasing or decreasing. This information can be utilized to produce applications that are adaptive to environmental conditions. For example, in this stereo matching application, we can use the information to adjust the density levels of the disparity maps. The details are presented in III.5.

Fifth, the CNF is easy to normalize. The results are represented as a sum of dS and automatically normalized because a well-defined dQ and T always yield a normalized dS. Thus, we can simply use $p'(x, y, t)$ instead of $p(x, y, t)$. The details are also presented in III.6.

III. CNF-Based Uncalibrated Stereo Matching

To use the CNF for uncalibrated stereo matching, we need to define the following four features:

- an actual system of a target application,
- a virtual system that exchanges heat,
- a function that estimates a specific measure, dQ,
- a rule for altering temperatures, $T(Q)$.

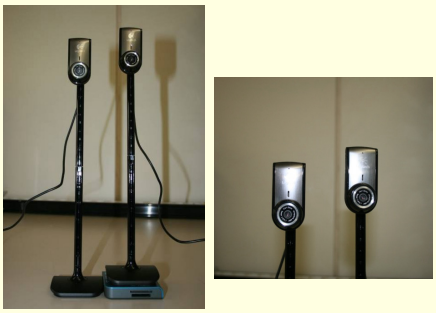


Fig. 3. Camera composition used in this paper.

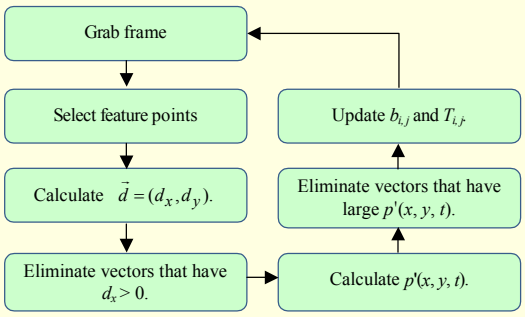


Fig. 4. Block diagram of proposed method.

A system defined in this way generates $p'(x, y, t)$ through updating its parameters along with some physical rules. With a higher variation in the entropy, $p'(x, y, t)$, of a feature point, the point is more likely to be classified as a part of the outliers because a greater variation in the entropy means that there is a higher instantaneous variation in the disorder of the feature points.

1. Defining an Actual System

Because stereo visions with non-parallel camera compositions require calibration tasks [26], we assume that two cameras have approximately the same intrinsic parameters and are positioned manually so that they are roughly parallel with each other as shown in Fig. 3. Then, the uncalibrated stereo matching model is known to be the same as the following equation because there is no epipolar line:

$$I_{\mathrm{L}}(x + d_x, y + d_y) = I_{\mathrm{R}}(x, y), \quad d_x < 0, \quad d_y = m \pm r \cdot \sigma, \qquad (11)$$

where I_{L} and I_{R} are intensity functions of the left and right images, respectively; d_x and d_y are displacement variables; m and σ imply mean and standard deviation of vertical displacement, respectively; and r is an arbitrary value which can be chosen by according to a normal distribution.

The block diagram of the method is depicted in Fig. 4. In every frame, we select feature points from the left image and

match the points to the right image. Each feature point may have a displacement vector, $\vec{d} = (d_x, d_y)$, assigned by a specific matching algorithm [24]. Vectors that have a d_x value greater than 0 are eliminated first because those vectors are certainly outliers. Of the remaining vectors, some can be eliminated whenever they show a greater increase in entropy during a unit period of time which is greater $p'(x, y, t)$. The reckoning of $p'(x, y, t)$ is performed using a virtual system. All the processes are automatically operated according to the predefined model and parameters.

2. Defining a Virtual System

We assume an ideal lattice plane, O_a, as shown in Fig. 1, where O_a is made up of $I \times J$ cliques, C, and each clique is comprised of many molecules that can absorb heat. There are some activated molecules that correspond to feature points and some deactivated molecules that correspond to the remaining pixels of the actual system. Each clique can have a different temperature, $T_{i,j}$. Each clique has a virtual variable, $b_{i,j}$, which, as described in III.3, is used to calculate the amount of absorbed heat, dQ. The lattice plane, O_a, also emits heat to the surrounding atmosphere at every frame, and cliques with a higher temperature emit more heat in accordance with the laws of physics. If the absorbed heat is greater than the emitted heat, the temperature increases, and vice versa.

3. Defining dQ

The amount of absorbed heat with $C_{i,j}$ of a molecule (a feature point) that is located at (x, y) is defined as

$$\begin{aligned} dQ_{x,y,i,j} &= (d_y^{x,y} - b_{i,j})^2 \quad , \\ d_y^{x,y} &\rightarrow d_y \quad of \quad \vec{d}_{x,y} \quad , \end{aligned} \tag{12}$$

where $\vec{d}_{x,y}$ is a displacement vector of a feature point located at (x, y). When different objects meet each other in a real physical system, the greater the difference in temperature is between the objects, the greater is the heat that is exchanged during a unit period of time. We similarly define dQ so that the larger difference between d_y and b yields a larger dQ.

The actual system can be adapted to the elapse of time because $b_{i,j}$ is updated at every frame as

$$\begin{aligned} b_{i,j,t} &= (1-\eta) \cdot b_{i,j,t-1} + \eta \cdot \mu(d_y^{\in i,j}), \\ d_y^{\in i,j} &\rightarrow d_y \in C_{i,j}, \end{aligned} \tag{13}$$

where η is the updating rate of b, t is a time or frame index, and $\mu(\bullet)$ is the mean of $\bullet$. A large η value enables the actual system to adapt to new variations of d_y in a short period.

To calculate dS_{net}, we define $N(i, j)$ as

$$N(i, j) = D_{i,j} / D_{\text{total}} \ , \tag{14}$$

where

$$D_{\text{total}} = \sum_{i}^{I} \sum_{j}^{J} D_{i,j}, \tag{15}$$

and

$$D_{i,j} = (h - E_{i,j}) \ , \tag{16}$$

where h is the horizontal resolution of input images, and $E_{i,j}$ is the Euclidean distance between the feature point and each center position of $C_{i,j}$.

4. Defining $T(Q)$

The absolute temperature of a clique is proportional to the average energy of the molecules that comprise the clique [27]. Thus, if we let κ be the rate constant between the absorbed heat per molecule and the temperature, the amount of temperature variation caused by the absorbed heat is

$$dT_{i,j,t} = \kappa \frac{\sum^{n} dQ_{i,j,t}}{n}, \tag{17}$$

where n is the number of activated molecules that comprise a clique. In addition, the rule for altering the temperature, $T(Q)$, is defined as

$$T(Q)_{i,j} \rightarrow T_{i,j,t} = T_{i,j,t-1} - \rho \cdot T_{i,j,t-1} + dT_{i,j,t} \ , \tag{18}$$

where ρ is an updating rate of T related to naturally emitted heat. According to (17) and (18), the amount of emitted heat, dQ_{E}, in every frame is

$$dQ_{\text{E}} = \sigma \frac{n \cdot \rho T}{\kappa}, \tag{19}$$

where σ is the rate constant. Thus, $T(Q)$ obeys the law of nature, namely, that a molecule under a higher temperature emits more heat. If $d(Q)$ is larger than dQ_{E}, the temperature gradually increases. A high temperature means that the average kinetic energy of a clique is high, which in turn indicates that the variation of $d_y^{\in i,j}$ is also high. At a displacement vector in such a clique, although the difference between d_y and b is large, dS is abated due to the higher temperature.

The term $T_{i,j,t}$ in (18) is an absolute temperature and cannot take a negative value [28]. If we let $\gamma (\gamma \geq 1)$ be the absolute temperature of the surrounding atmosphere, then $T_{i,j,t}$ cannot be smaller than γ because it is in a state of thermal equilibrium when $T_{i,j,t} = \gamma$. In a virtual system without γ, the temperatures may decrease to a near-zero value. This decrease means that dS can take an extremely high value even if dQ is small. Thus, the

actual system could reach the wrong conclusion, and the system may produce very sparse disparity maps.

5. Fabrication of a Semi-dense Disparity Map

The density of a disparity map is directly related to the number of activated molecules. Any method of selecting feature points may be used by an actual system. However, the method must be able to distinguish between feature points that have strong characteristics (strong feature points) and those that have weak characteristics (weak feature points). Ideally, the method should be capable of adjusting the level between the strong and weak characteristics. Fortunately, we found a traditional feature selection method [29] that is expressed as

$$\min(\lambda_1, \lambda_2) > \lambda \ , \tag{20}$$

where λ_1 and λ_2 are the two eigenvalues of G, and λ is a predefined threshold. All these terms are mentioned in [29]. This method can distinguish strong feature points from weak feature points and adjust the level pursuant to the value of λ.

To enable the actual system to adjust the density level of a disparity map, we ignore the predefined λ but define $\lambda_{i,j}$ as

$$\lambda_{i,j} = \varepsilon \cdot T_{i,j}^{\delta}, \quad 0 < \varepsilon < 1, \quad \delta > 1, \tag{21}$$

where ε and δ are constant values that determine the behavior of the actual system. If the initial value of the temperature, $T_{i,j,0}$, is 4, and if $\varepsilon = 0.0000001$ and $\delta = 10$, then the initial value of $\lambda_{i,j}$ is approximately 0.1. As time goes by, $T_{i,j}$ may dwindle; and, if $T_{i,j}$ reaches to a minimum, 1 (or γ), $\lambda_{i,j}$ is approximately equal to ε (or $\varepsilon \cdot \gamma^{\delta}$). This outcome indicates that the actual system selects a few strong feature points in the early frames of the stereo sequence for a higher matching accuracy and that more and more feature points can be selected as time goes by. As a result, the level of density of the disparity maps gradually increases.

6. Automatic Normalization

If γ is chosen as a value higher than one, dS_{net} is normalized in a specific range even if there is no additional normalization step. The reasons for this behavior are as follows:

- the total number of cliques is constant,
- the sum of $N(i,j)$ is constant,
- the range of $(d_y^{x,y} - b_{i,j})^2$ is 0 ~ $(2h)^2$, where h is the vertical resolution of input images.

Note that the sum of the normalized values is also a normalized value. Thus, $p'(x, y, t)$ in (10) is always automatically normalized.

IV. Experiments

In the experiments, we used two kinds of test image pairs. The first pairs were image sequences captured from cameras, and the second pairs were image pairs from the Middlebury stereo database [30]. We present qualitative and quantitative results together for this stereo database. However, for the image sequences captured from cameras, we present only the qualitative results because we have no ground truth. For clarity, we eliminated some disparity points that were either too near or too far from the results. In addition, to show the potential and usefulness of the proposed method, we implemented a real-time HCI application and describe the application at the end of this section.

1. Image Sequences

In the case of utilizing a stereo camera, we tested 640×480 raw image pairs without using any of the rectification or calibration functions. The experiments were run on a 2.83 GHz quad core PC. The running time of Bumblebee2 [31] was about 0.102 s for a calibrated pair, and the running time of the proposed method was about 0.053 s for an uncalibrated pair.

Figure 5(a) shows a raw image pair captured by Bumblebee2. The lines indicated by the white arrows in the background are actually straight lines, but they are distorted by the wide-angle lenses used. Figure 5(b) shows the results of the calibrated stereo matching of Bumblebee2, and Fig. 5(c) shows the results of the uncalibrated stereo matching of the proposed method.

The image of Fig. 6(a) contains some areas that lack texture, especially in the background. Figure 6(a) also contains the displacement vectors and a disparity map of the raw matching

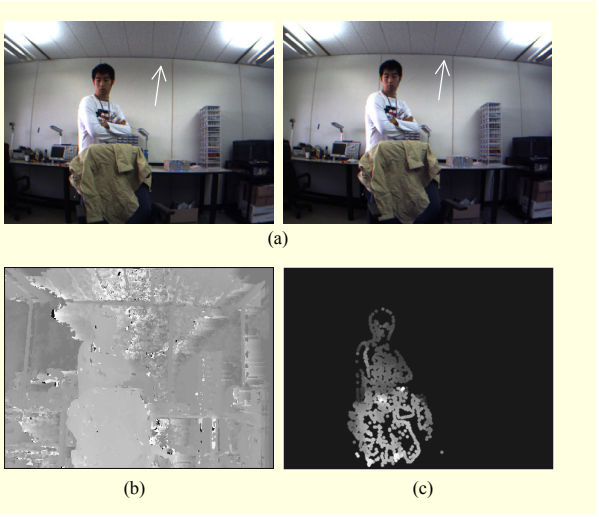


Fig. 5. Results from an unrectified image pair: (a) input stereo image pair, (b) disparity map of Bumblebee2, and (c) disparity map of proposed method.

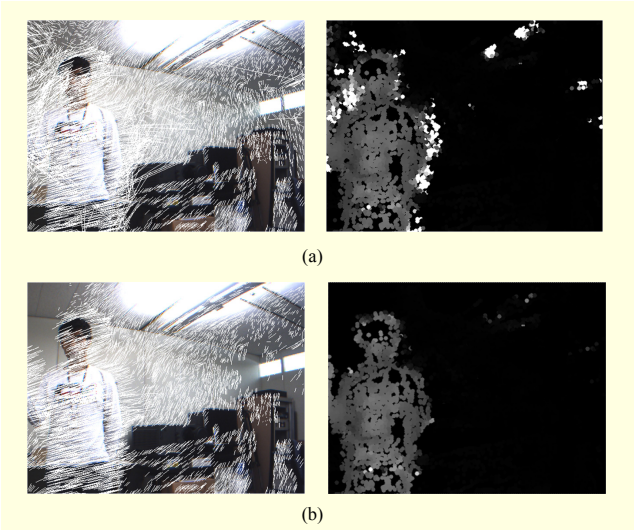


Fig. 6. Results using an image pair that has some areas lacking texture: (a) matching result and disparity map using raw matching algorithm of [28] without CNF and (b) matching result and disparity map with CNF.

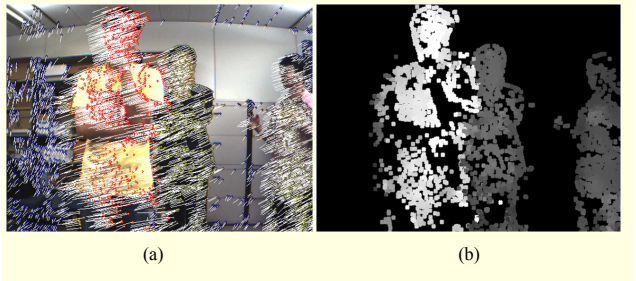


Fig. 7. Results using an image pair that contains some areas that have occlusions and depth discontinuity: (a) a left image with displacement vectors and (b) a corresponding disparity map.

algorithm of [24] without the CNF. Figure 6(b) shows the results with the CNF. There are some noises produced by the backlighting, but most of the noises were successfully eliminated.

The image in Fig. 7 contains some areas that have occlusions and depth discontinuities. One person occludes another person standing at the rear. Figures 7(a) and 7(b) show displacement vectors and the corresponding disparity map.

In the case of utilizing USB webcams, we used the auto white balance, auto brightness, and auto focus options of the cameras. There were short time gaps between the stereo image pairs because we successively captured two images. The camera positions were located manually so that the cameras were aligned roughly parallel to each other as shown in Fig. 3. The resolution of the tested images was 320×240.

The images of Fig. 8 were taken from different height settings of the two cameras. The left image was taken from a

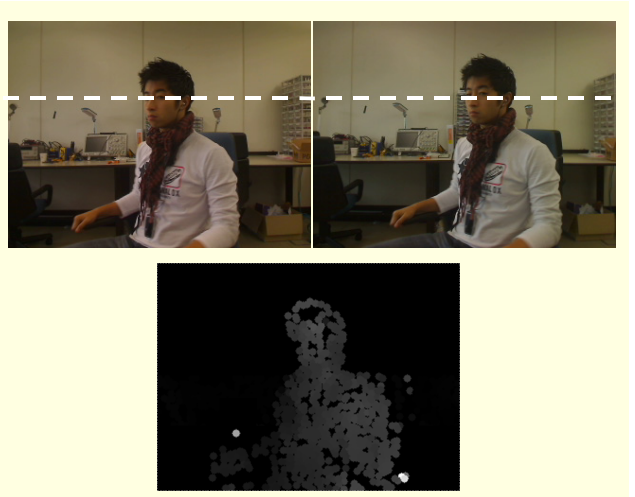


Fig. 8. Results from image pair taken from different camera heights.

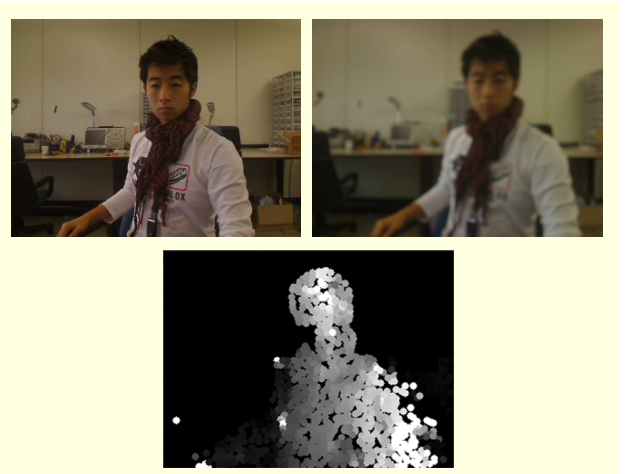


Fig. 9. Stereo image pair that has different focuses and a corresponding disparity map.

higher position than the right image so the heights of the subject's eyes are different in each picture. Because we used an auto white balance option, the color temperatures are also slightly different from each other. The upper garment of the person in the left image is slightly redder than that of the right image.

The images in Fig. 9 are under the same conditions. The colors of the left images are slightly redder than those of the right. In addition, the focuses of the images also are different because we used an auto focus option.

The experiments were run on a 2.79 GHz dual core notebook. The running time of the proposed method was about 0.092 s for the uncalibrated images.

2. Middlebury Datasets

In the case of the Middlebury stereo database, we tested the

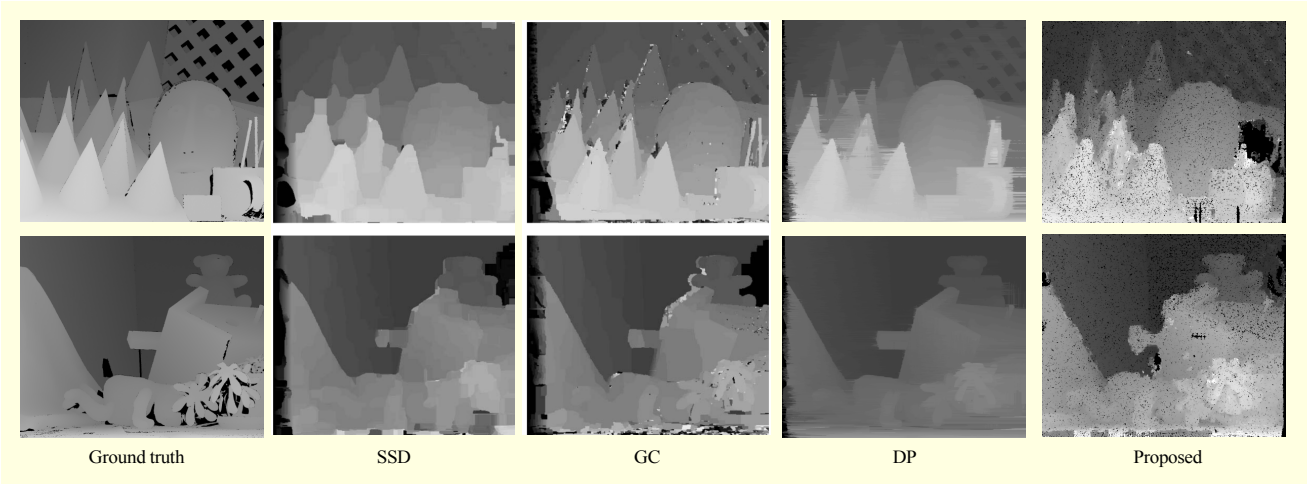


Fig. 10. Ground truth images and stereo results of four algorithms: SSD, GC, DP, and the proposed method.

Table 1. Results of matching for calibrated images.

(*Ae*)	SSD	GC	DP	Proposed
Cones	9.69 (1.03%)	9.23 (0.92%)	8.68 (0.41%)	9.66 (5.80%)
Teddy	14.01 (1.29%)	12.06 (1.06%)	9.56 (0.65%)	9.83 (4.31%)

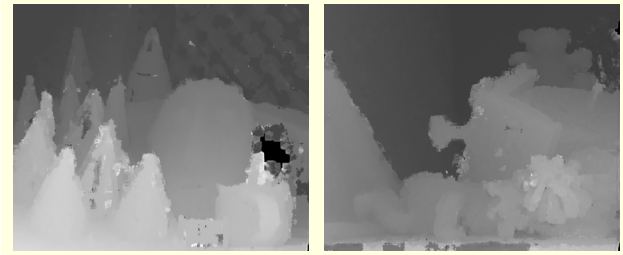


Fig. 11. Interpolated results.

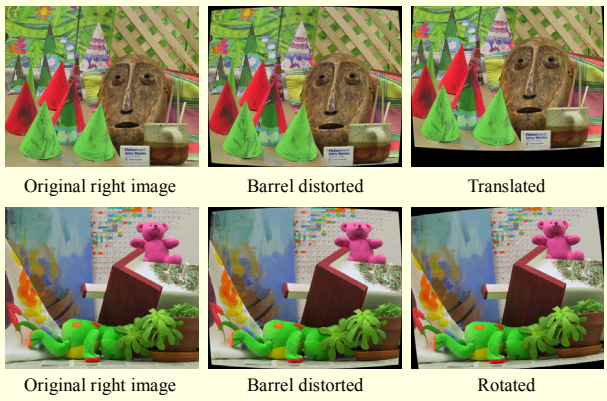


Fig. 12. Examples of distorted right images.

distorted and original images together because the images were already rectified and calibrated. We tested two full-sized datasets, 'cones' and 'teddy,' that have resolutions of 1800 × 1500.

Figure 10 shows the results of using calibrated pairs with four algorithms: sum of squared differences (SSD), graph cut (GC), dynamic programming (DP), and the proposed method. The results of the proposed method are for the 20th frame because the method is for image sequences and needs a short adaptation time to produce semi-dense disparity maps. To show the quantitative performance results, we define the average error (*Ae*) as follows:

$$Ae = \mu(\|I_G - I_D\|), \quad (22)$$

where I_G and I_D are the intensity values of the ground truth and disparity map, respectively. We ignore points whose disparities are undefined or unknown; that is, I_G =0 or I_D=0. Table 1 shows the quantitative performance of the algorithms. The percentages in parentheses indicate the portions occupied by black spots or regions of the disparity maps in Fig. 10. We appropriately interpolate two disparity maps of the proposed method to make dense maps, which are shown in Fig. 11. However, we do not interpolate the rest results in Figs. 13 and 14 to show the raw matching results.

Figure 12 shows examples of the distorted images. We distorted the images by means of a Barrel distortion [32] with K_1=0.00000001. To make the conditions similar to practical applications, we translated only the right images toward the upper direction (y_t) as shown in the upper row in Fig. 12. We rotated only the right images counterclockwise (θ) as shown in the bottom row in Fig. 12. Tables 2 and 3, respectively, show the quantitative performance of two algorithms as y_t and θ are altered, which confirms that the proposed method is stable and robust to distortions.

We illustrate the qualitative results along with the parameters

Table 2. Results of matching by DP and proposed method for distorted images (y_t).

	Cones		Teddy	
y_t	DP (*Ae*)	Proposed (*Ae*)	DP (*Ae*)	Proposed (*Ae*)
0	20.54	11.58	31.97	11.45
4	27.50	11.80	32.01	11.51
8	37.19	12.02	36.42	12.06
12	45.25	12.11	48.84	12.04
16	53.46	12.16	65.17	12.22
20	61.44	12.22	80.20	12.61

Table 3. Results of matching by DP and the proposed method for distorted images (θ).

	Cones		Teddy	
θ	DP (*Ae*)	Proposed (*Ae*)	DP (*Ae*)	Proposed (*Ae*)
0.3	22.64	12.95	34.14	12.00
0.5	23.19	16.63	38.53	15.94
0.7	26.41	20.79	44.16	20.47
0.9	32.15	25.15	49.91	25.18
1.1	38.60	29.67	56.84	30.04
1.3	45.52	34.41	61.91	38.83

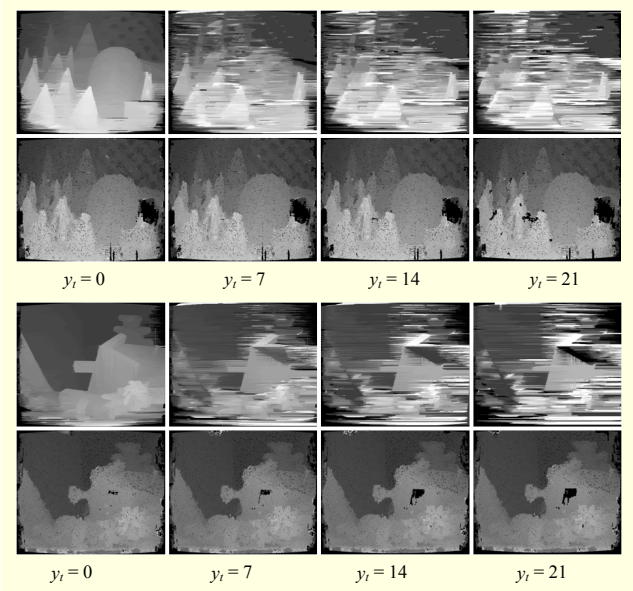


Fig. 13. Stereo results of two algorithms: upper row depicts results of DP and bottom row contains results of proposed method for cones and teddy.

y_t and θ in Figs. 13 and 14, respectively. The upper rows show the results of DP, and the bottom rows show the results of the proposed method.

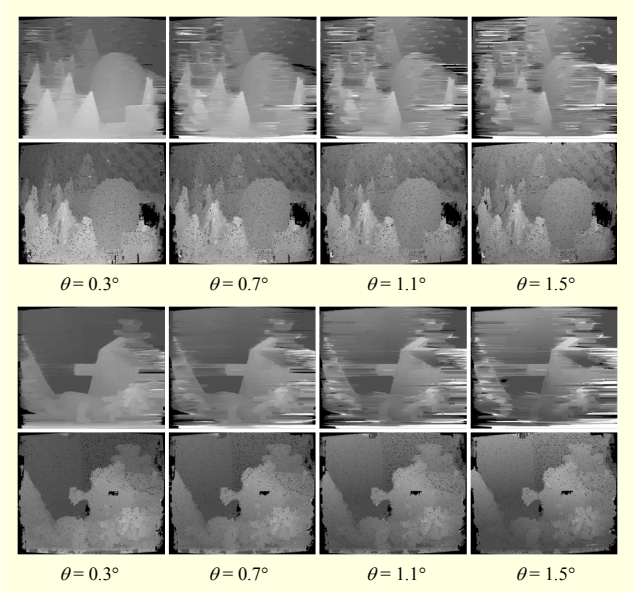


Fig. 14. Stereo results of two algorithms: upper row depicts results of DP and bottom row contains the results of proposed method for cones and teddy.

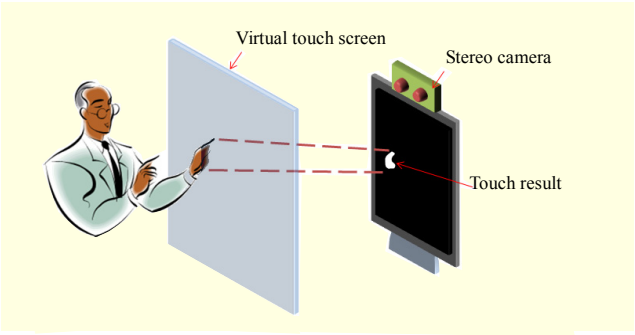


Fig. 15. Concept of virtual touch screen and practical setup.

In this case, the experiments were run on a 2.83 GHz quad core PC. The running time of DP was about 5.88 s, and the running time of the proposed method was about 1.87 s.

3. Application

We implemented a real-time HCI application to show the potential and usefulness of the proposed method. As illustrated in Fig. 15, the application was a virtual touch screen. A user stood in front of a large screen, and a stereo camera took an image sequence of the user. The application transparently displayed the user's body image taken by the camera as a proxy. When a disparity map of the user was made, the user's body became an interface itself. If the user stretched out a hand to the front, the system recognized the position of the hand as well as the relative depths between the user's hand and body and

moved a mouse pointer to the position of the hand. When the user's hand position changed, the system moved the mouse pointer to the new position. As a result, the user could utilize the application and select items as if carrying a mouse.

This application was operated on a PC equipped with a 2.83 GHz quad core CPU and a stereo camera with a focal length of 3.8 mm. We used uncalibrated 230×380 resolution images so that the images contained the full body of the user, and the system ran the application at approximately 15 fps. The depth resolution perceived by both the user and the system was 10 cm, and the minimum size of an item that could be stably selected by the user was 40×40 pixels at a distance of 2 m from the camera.

V. Conclusion

We have introduced a new theoretical processing method for time-varying imagery or uncalibrated stereo matching. Largely inspired by thermodynamics, the method has some advantages that enable it to yield more accurate results with less operating power than existing approaches. First, the CNF models are tailor-made for representing the dependencies of nearby spatial and temporal information so that exogenous information can be incorporated into a feature point. Second, the CNF needs only a small amount of processing power because it requires no additional iterative processes, such as a MAP, IIS, or Gibbs sampler, for estimating some parameters. Third, the CNF obviates the need to quantize intermediate values because, in contrast to other random fields, it is not a postprocessing step. Fourth, the normalization process is easy.

The approach is very flexible because the CNF can be tailor-made for each application by defining the actual and virtual systems. We have designed a CNF model and applied it to a stereo matching system. In addition to effectively eliminating noise, the system is robust to various distortions, for example, the Barrel distortion, as well as differences in height, rotation, brightness, white balance, and focus. Moreover, it is fast enough to be used for practical real-time applications, and users can consequently use two webcams as a stereo camera.

There are several possible directions for future research. We are interested in applying the CNF to other time-varying imagery applications and designing well-defined models that optimally fit the applications. Developing algorithms that improve the performance of our method is another area of potential future work.

In addition, we plan to apply the disparity maps of users to applications that support a multiuser, multitouch interface. Because the proposed approach can already express depth information of multiple users, we expect to easily attain successful results in the near future.

References

[1] S.J. Krotosky and M.M. Trivedi, "Person Surveillance Using Visual and Infrared Imagery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, Aug. 2008, pp. 1096-1105.

[2] R. Anchini et al., "A Comparison Between Stereo-Vision Techniques for the Reconstruction of 3-D Coordinates of Objects," *IEEE Trans. Instrum. Meas.*, vol. 55, no. 5, Oct. 2006, pp. 1459-1466.

[3] Y. Matsumoto and A. Zelinsky, "An Algorithm for Real-Time Stereo Vision Implementation of Head Pose and Gaze Direction Measurement," *Autom. Face Gesture Recognition*, 2000. pp. 499-504.

[4] J.I. Park et al., "Virtual Control of Optical Axis of the 3DTV Camera for Reducing Visual Fatigue in Stereoscopic 3DTV," *ETRI J.*, vol. 26, no. 6, Dec. 2004, pp. 597-604.

[5] A. Smolic and P. Kauff, "Interactive 3-D Video Representation and Coding Technologies," *IEEE Proc.*, vol. 93, no. 1, 2005, pp. 98-110.

[6] C.D. Castillo and D.W. Jacobs, "Using Stereo Matching with General Epipolar Geometry for 2D Face Recognition across Pose," *IEEE Trans. PAMI*, vol. 31, no. 12, Dec. 2009, pp. 2298-2304.

[7] Y.V. Venkatesh, S.K. Raja, and A.J. Kumar, "On the Application of a Modified Self-Organizing Neural Network to Estimate Stereo Disparity," *IEEE Trans. Image Process.*, vol. 16, no. 11, 2007, pp. 2822-2829.

[8] L.H. Liu and P. Bhattacharya, "Uncalibrated Stereo Matching Using DWT," *ICPR*, 2000, pp. 114-118.

[9] J. Zhow, Y. Xu, and X. Yang, "Quaternion Wavelet Phase based Stereo Matching for Uncalibrated Images," *Patt. Recog. Lett.,* vol. 28, no. 12, Mar. 2007, pp. 1509-1522.

[10] W. Li, C.H. Leung, and Y.S. Hung, "Matching of Uncalibrated Stereo Images by Elastic Deformation," *Int. J. Imaging Syst. Technol.*, vol. 14, no. 5, Mar. 2005, pp. 198-205.

[11] S.D. Cochran and G. Medioni, "3-D Surface Description from Binocular Stereo," *IEEE Trans. PAMI*, vol. 14, no. 10, 1992, pp. 981-994.

[12] P. Premaratne and F. Safaei, "Feature Based Stereo Correspondence Using Moment Invariant," *Proc. IEEE Int. Conf. Inf. Automation for Sustainability*, 2008, pp.104-108.

[13] D. Scharstein and C. Pal, "Learning Conditional Random Fields for Stereo," *CVPR*, 2007, pp. 1-8.

[14] R. Gherardi et al., "Optimal Parameter Estimation for MRF Stereo Matching," *ICIAP*, *LNCS* 3617, 2005, pp. 818-825.

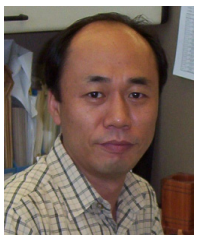
[15] L. Zhang and S.M. Seitz, "Parameter Estimation for MRF Stereo," *CVPR*, vol. 2, 2005, pp. 288-295.

[16] J. Sun, H.Y. Shum, and N.N. Zheng, "Stereo Matching Using Belief Propagation," *ECCV*, *LNCS 2351*, 2002, pp. 510-524.

[17] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs

Distributions, and the Bayesian Restoration of Images," *IEEE Trans. PAMI*, vol. 6, no. 6, 1984, pp. 721-741.
[18] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *ICML*, 2001, pp. 282-289.
[19] S Kirkpatrick, C.D. Gelatt, Jr., M.P. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, no. 4598, 1983, pp. 671-680.
[20] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing Features of Random Fields," *IEEE Trans. PAMI*, vol. 19, no. 4, 1997, pp. 380-393.
[21] L. Boltzmann, *Lectures on Gas Theory*, trans. S.G. Brush, New York: Perfamon Press, 1995. Available: http://openlibrary.org /books/OL1115209M/Lectures_on_gas_theory
[22] J. Daintith, *Oxford Dictionary of Physics*, Oxford University Press, 2005.
[23] P. Pierre, *A to Z of Thermodynamics*, Oxford University Press, 1998.
[24] J. Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the Algorithm." Available: http://robots.stanford.edu/cs223b04/algo_tracking.pdf.
[25] J. Migdal and W.E.L. Grimson, "Background Subtraction using Markov Thresholds," *WACV/MOTION*, 2005, pp. 58-65.
[26] Alan C. Bovik, *Handbook of Image and Video Processing*, Academic Press, 2000.
[27] H. Chang, *Inventing Temperature: Measurement and Scientific Progress*, Oxford University Press, 2004.
[28] Herbert B. Callen, *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed., New York: John Wiley & Sons, ISBN 0-471-86256-8, 1985.
[29] C. Tomasi and T. Kanade, "Detection and Tracking of Point Features," Technical Report CMU-CS-91-132, 1991.
[30] Middlebury Stereo Vision Page. Available: http://vision. middlebury.edu/stereo
[31] Barrel Distortion. Available: http://ptgrey.com/products/stereo.asp
[32] Bumblebee2 Camera. Available: http://en.wikipedia.org/wiki/ Distortion_(optics)

Eunjin Koh received his BS and MS in computer science from Inha University, South Korea, in 2005 and 2007, respectively. Since 2007, he has been a member of ETRI, Daejeon, South Korea and pursuing the PhD at the University of Science and Technology, Daejeon, South Korea. His main research interests include image processing, human-computer interaction, object tracking, and pattern classification.



Jaeyeon Lee received his PhD from the Tokai University, Japan, in 1996. He has been a research scientist at ETRI since 1986. His research interests include robotics, pattern recognition, computer vision, and biometric security.

Junseok Park received his BS in computer science from Inha University, South Korea, in 1984, MS in computer science from KAIST, South Korea, in 1999, and his PhD in information engineering from Inha University, South Korea, in 2006. Since 1987, he has been a member of the principal research staff at ETRI, Daejeon, South Korea. His main research interests include human computer interaction, mobile haptic interfaces, gestural computing, and personal area networks.