AN INTEGRATION ALGORITHM FOR STEREO, MOTION AND COLOR IN REAL-TIME APPLICATIONS

Minoru ETOH and Hiroshi ARAKAWA
Central Research Laboratories, Matsushita Electric Industrial Co., Ltd.
3-4 Hikaridai, Seika, Kyoto, 619-02 Japan
E-mail: etoh@crl.mei.co.jp and arakawa@crl.mei.co.jp

ABSTRACT

This paper describes a statistical integration algorithm for color, motion and stereo disparity, and introduces a real-time stereo system that can tell us where and what objects are moving. Regarding the integration algorithm, motion estimation and depth estimation are simultaneously performed by a clustering process based on motion, stereo disparity, color, and pixel position. As a result of the clustering, an image is decomposed into region fragments. Each fragment is characterized by distribution parameters of spatiotemporal intensity gradients, stereo difference, color and pixel positions. Motion vectors and stereo disparities for each fragment are obtained from those distribution parameters. The real-time stereo system can view the objects with the distribution parameters over frames. The implementation shows that we can utilize the proposed algorithm in real-time applications such as surveillance and human computer interaction.

INTRODUCTION

Detecting a moving object is important for a wide range of applications from surveillance to human-computer interaction. In this paper, we introduce a real-time stereo system that can tell us where and what objects are moving. As depicted in Fig. 1, with our system, an object of attention (e.g., object B) can be easily discriminated from background objects (object A) by color, motion, or depth. The system can provide "what-where" information, in real time, associated with color, motion and depth of segmented regions. Those are strong clues for selectively detecting a moving object. One state-of-the-art approach to "vision interface" uses multiple templates[DP93, TLT93]. Although the multiple templates method allows the machine to perform sophisticated tasks such as gesture recognition and face identification, it requires restrictive conditions for background, views and bootstraps. On the other hand, we believe that extracting the clues in a robust bottom-up manner can perform typical visual tasks in the above applications.

An algorithm proposed in the next section is an extension of the integration algorithm[ES93] that one of the authors has already proposed for 2D motion estimation. We introduce an expansion for dealing with stereo image sequences.

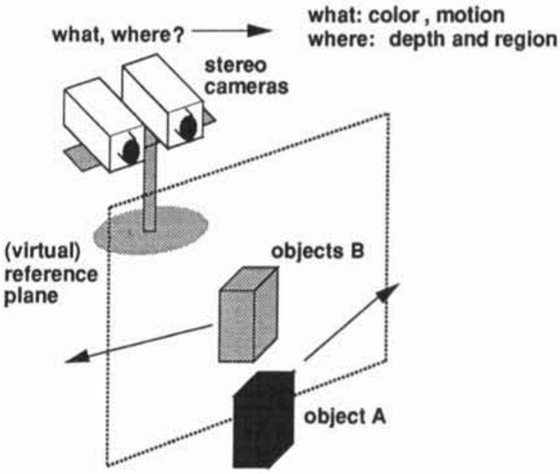


Figure 1: Stereo system.

The issues discussed in this paper are how we integrate stereo, motion and color, and how we implement a real-time system. In the following sections, we will discuss these issues.

INTEGRATION ALGORITHM

In our approach, an image is decomposed into region fragments by a clustering process (see Fig. 2). Each fragment is characterized by distribution parameters of colors, pixel positions, stereo difference and spatiotemporal intensity gradients. Then assuming the uniformity of the color, disparity and motion in a region fragment, a 2D motion vector and a stereo disparity (i.e., inverse of depth) for each fragment are obtained from the distribution parameters of the multidimensional features.

The features of our approach depicted in Fig.2 are:

statistical integration of color, motion and stereo : Segmentation, motion estimation and depth estimation are simultaneously performed as parameter estimation of joint probability densities of color, motion, disparity and pixel positions.

on-line processing and dynamics : The clustering process is realized as a competitive learning[RZ86,

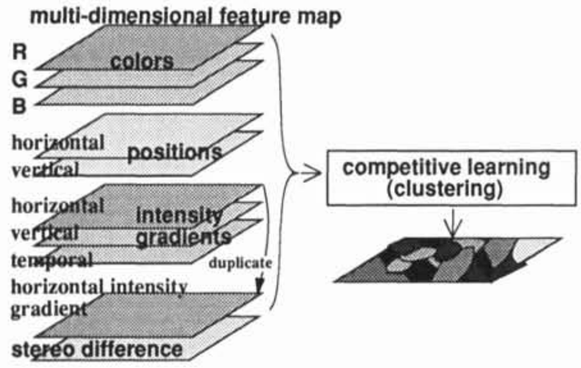


Figure 2: Clustering with nine-dimensional feature map.

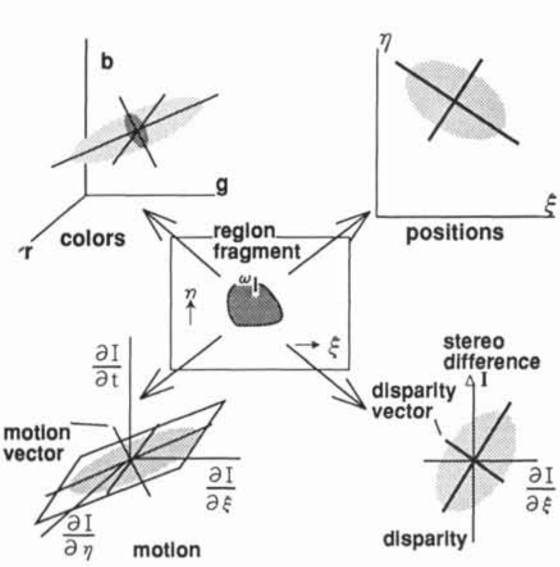


Figure 3: Four feature spaces.

Koh90]. *Competitive learning* provides a novel way for real time on-line computation of the parameter estimate. Moreover, sample-wise optimization ensures dynamics of the parameters over successive frames.

Firstly we briefly discuss assumptions and definitions of our method. Image features at each pixel consist of color vectors $\mathbf{x}$ (i.e., RGB values), pixel positions $\mathbf{p} = (\xi, \eta)^t$ (i.e., column,row) , spatiotemporal intensity gradients $\mathbf{g} = (\frac{\partial I}{\partial \xi}, \frac{\partial I}{\partial \eta}, \frac{\partial I}{\partial t})^t$, and $\mathbf{w} = (\frac{\partial I}{\partial \xi}, \Delta I)$ where I and ΔI denote intensity of left image, intensity difference between left and right images. $(\cdots)^t$ is matrix transpose operator. We use the left image for the parameter estimation. Let the left image $\mathbf{R}$ that is to be described be represented by c classes of pixels (i.e., region fragment) such that $\mathbf{R} = \{\omega_1, \omega_2, \ldots, \omega_c\}$, where $\omega_i, i = 1, 2, \ldots, c$ indicate the region fragments.

At each region fragment, the following distributions are assumed independently (see Fig. 3).

color: We model the distribution of measured color vectors $\mathbf{x}$ for each region fragment ω_i with multivariate normal density $N_3(\bar{\mathbf{x}}_i, \mathbf{X}_i)$ where $\bar{\mathbf{x}}_i, \mathbf{X}_i$ are the mean vector and the 3×3 covariance matrix of the color vectors.

position : We model the distribution of pixel positions $\mathbf{p}$ for each region fragment with multivariate normal density $N_2(\bar{\mathbf{p}}_i, \mathbf{P}_i)$ where $\bar{\mathbf{p}}_i, \mathbf{P}_i$ are the mean vector and the 2×2 covariance matrix of the distribution. The region fragment is approximated by an ellipse that has the same parameters $\bar{\mathbf{p}}$ and $\mathbf{P}_i$. We use the $N_2(\bar{\mathbf{p}}_i, \mathbf{P}_i)$ to approximate the density of $\mathbf{p}$ within the ellipse.

motion: On the constant intensity assumption, expansion of the total derivative of intensity I leads to the well-known *gradient constraint equation*[HS81] described by

$$u\frac{\partial I}{\partial \xi} + v\frac{\partial I}{\partial \eta} + \frac{\partial I}{\partial t} = 0, \tag{1}$$

where u, v denote the image component velocity. Let $\mathbf{m}_i = (u_i, v_i, 1)^t / \sqrt{u_i^2 + v_i^2 + 1}$ denote the 2D motion vector of the fragment ω_i, Eq. 1 can be written as $\mathbf{m}_i^t \mathbf{g} = 0$. By assuming the 2D homogeneous image motion, we model the distribution of $\mathbf{m}_i^t \mathbf{g}$ for each region fragment with univariate normal density $N(0, s_i^2)$ described by

$$p(\mathbf{m}_i^t \mathbf{g} | \omega_i) = \frac{1}{\sqrt{2\pi} s_i} \exp[-\frac{1}{2s_i^2}(\mathbf{m}_i^t \mathbf{g})^2]. \tag{2}$$

stereo disparity: We treat disparity estimation as a kind of motion estimation[LK82] with the constraint that the camera moves along the ξ axis, and assumes the equation:

$$q_i \frac{\partial I}{\partial \xi} + \Delta I = 0, \tag{3}$$

where q is a stereo disparity. Let $\mathbf{h}_i = (q_i, 1)^t / \sqrt{q_i^2 + 1}$ denote the disparity vector of the fragment ω_i, Eq. 3 can be written as $\mathbf{h}_i^t \mathbf{w} = 0$. We model the distribution of $\mathbf{h}_i^t \mathbf{w}$ for each region fragment with univariate normal density $N(0, c_i^2)$ described by

$$p(\mathbf{h}_i^t \mathbf{w} | \omega_i) = \frac{1}{\sqrt{2\pi} c_i} \exp[-\frac{1}{2c_i^2}(\mathbf{h}_i^t \mathbf{w})^2]. \tag{4}$$

Let $\mathbf{G}_i$ and $\mathbf{W}_i$ be the covariance matrices of $\mathbf{g}$ and $\mathbf{w}$ in ω_i. Owing to the above assumption, as depicted in the Fig.3 , $\mathbf{m}_i$ and $\mathbf{h}_i$ are obtained by normalizing the third eigenvector of $\mathbf{G}_i$ and the second eigenvector of $\mathbf{W}_i$ respectively. Details are described in [ES93].

Here, the competitive learning in this work is roughly as follows:

1. Assume a sequence of statistical samples of the features $(\mathbf{x}(t), \mathbf{p}(t), \mathbf{g}(t), \mathbf{w}(t))$, and their parameters $\{\boldsymbol{\theta}_i(t) : \boldsymbol{\theta}_i = (\bar{\mathbf{x}}_i(t), \mathbf{X}_i(t), \bar{\mathbf{p}}_i(t), \mathbf{P}_i(t), \mathbf{G}_i(t), \mathbf{W}_i(t)), i = 1, 2, \ldots, c\}$, where t is the learning step.

2. $\boldsymbol{\theta}_i(0), i = 1, 2, \ldots, c$ have been initialized by random selection.

3. Sample $(\mathbf{x}(t), \mathbf{p}(t), \mathbf{g}(t), \mathbf{w}(t))$ is randomly selected from the feature map, and is simultaneously compared with each $\boldsymbol{\theta}_i(t)$ at each successive instant of

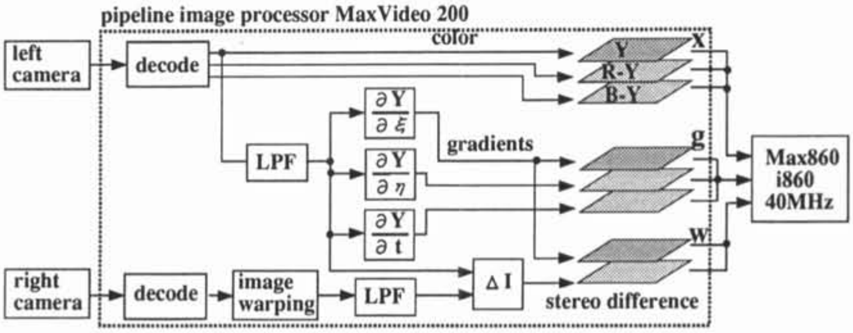


Figure 4: System configuration.

learning step, $t = 1, 2, 3, \ldots$, and then the best-matching $\boldsymbol{\theta}_i(t)$ is updated by the *delta rules*[RZ86] to match even closer to the current sample.

Once the parameter sets are initialized, the step 3 is iterated over successive frames.

Assuming that the probability densities of $\mathbf{x}, \mathbf{p}, \mathbf{m}_i^t\mathbf{g}$, and $\mathbf{h}_i^t\mathbf{w}$ are jointly normal and independent, we use a distance metric based on a log-likelihood as:

$$d_i(\mathbf{x}, \mathbf{p}, \mathbf{g}, \mathbf{w}) = (\mathbf{x} - \bar{\mathbf{x}}_i)^t \mathbf{X}_i^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i) + (\mathbf{p} - \bar{\mathbf{p}}_i)^t \mathbf{P}_i^{-1} (\mathbf{p} - \bar{\mathbf{p}}_i) + (\mathbf{m}_i^t\mathbf{g})^2/s_i^2 + (\mathbf{h}_i^t\mathbf{w})^2/c_i^2 + \ln|\mathbf{X}_i| + \ln|\mathbf{P}_i| + \ln s_i^2 + \ln c_i^2 \text{ for } i = 1, 2, \ldots, c\,. \quad (5)$$

Using this metric, the pixel assignment to the region fragment is performed *without setting any heuristic weighting coefficients. This metric enforces similarity of color, locality of each region fragment, and the constraint satisfaction to the same 2D motion and stereo disparity.* Altering $\boldsymbol{\theta}_i$ must be such that, if $i = k$ is the index of the best-matching region fragment ω_k, then the $d_k(\mathbf{x}(t), \mathbf{p}(t), \mathbf{g}(t), \mathbf{w}(t))$ is decreased, and all the other parameter vectors $\boldsymbol{\theta}_i$ with $i \neq k$ are left intact. Note that our delta rules update not only the mean vectors but also the covariance matrices such as $\mathbf{G}_i(t), \mathbf{W}_i(t)$. In this way, parameters of the different region fragments tend to become specifically *tuned* to input samples over successive frames. The system produces the current estimates of the parameter sets every moment.

IMPLEMENTATION

Fig. 4 illustrates the implementation on a "DataCube"[1], real-time image processor. In this implementation, color vector $\mathbf{x}$ consists of (Y, R-Y, B-Y) signals and the stereo difference and the intensity gradients are calculated from Y signals.

We apply an image warping to the right image so that the stereo difference can be forced to be zero on the virtual reference plane as depicted in Fig.1. This is because Eq. 3 is vulnerable to large disparities. Prior to the competitive learning, we can preset the transformation(warping) parameters by putting a real panel and minimizing the sum of absolute errors as depicted in Fig. 5. Since the image warper of the DataCube can perform second order transformations, the stereo difference can be set exactly to zero at any projection from the reference plane. A downhill simplex method[PFTV88, pp.305] searches the minimal error parameters by controlling the image warper iteratively. *Using this image warping, we can qualitatively separate foreground objects from background objects with reference to a sign of disparity.* In Fig.1, for example, the region of object B has a positive disparity while the the region of object A has a negative disparity. DataCube performs the image warping and all differential operations at about 20Hz using an image pipeline, Max Video 200. However, the competitive learning on a digital signal processor i860 takes rather long time. Consequently, the frame rate of feature map is 2 frames per second for 200 random sampling per frame.

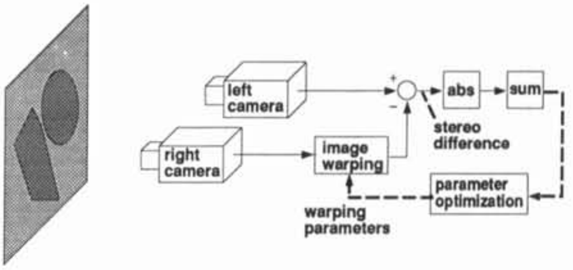


Figure 5: Calibration procedure.

EXPERIMENTS

Fig.6(a) shows a typical indoor scene consisting of a chair and a background wall. Prior to the experiment, a panel was put just in front of the chair for the zero stereo disparity calibration (See Fig.5). Fig. 6 (b) shows the clustering result at the beginning of this experiment. The figures from 6(b) to 8(b) were produced from the system's screen snap-shots. The ellipses illustrate the clustered pixels, and their second moments are equal to $\mathbf{P}_i$. The ellipses have two colors, white and black. The white

[1] †product of Datacube Inc., U.S.A.

ellipses indicate the region with positive disparity. On the other hand, as shown in this figure, the black ellipses indicate the region with negative disparity. That means the objects are farther away than they were in the reference plane. Note that ellipses without a confident depth estimate are not drawn here (confidence is determined by examining the eigenvalues of $\mathbf{W}_i$[ES93]). In Fig.7, a person enters the scene. He sits at the chair and swings from left to right. The system detects him as foreground objects and indicates him as white ellipses as shown in (b). In addition, the system outputs the 2D motion vectors for each region. The short thick white lines starting from the centered dots indicate the 2D motion vectors. The mean motion vector of foreground regions is also indicated at the top of this figure. In Fig. 8, another person gets into the scene. This time, however, he sits in the background. Although those two persons are wearing the same color jackets, the system can discriminate the foreground person from the other person.

CONCLUSION

This paper has presented our new description method for moving stereo images. What we propose in this paper is not only an integration algorithm for color, motion and stereo disparity but also a real-time implementation scheme that enables us to utilize the algorithm in real-time applications such as surveillance and human computer interaction. Our experiments show the advantages of that algorithm.

References

[DP93] T. Darrell and A. Pentland. Space-time gestures. In *Proc. CVPR'93*, pp. 335–340. IEEE Computer Society, June 1993.

[ES93] M. Etoh and Y. Shirai. Segmentation and 2D motion estimation by region fragments. In *Proc. 4th ICCV*, pp. 192–199, Berlin, May 1993. IEEE Computer Society.

[HS81] B. Horn and B. Schunk. Determining optical flow. *Artificial Intelligence*, Vol. 17, pp. 185–203, 1981.

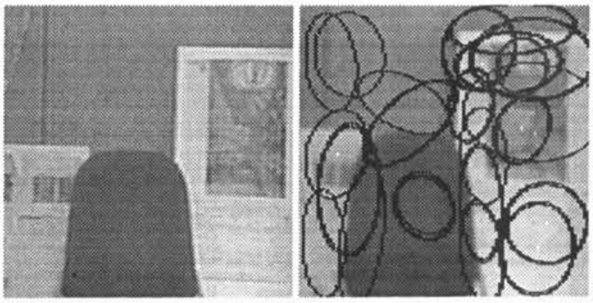
[Koh90] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, Vol. 78, No. 9, 1990.

[LK82] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. DARPA Image Understanding Workshop*, pp. 121–130, 1982.

[PFTV88] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.

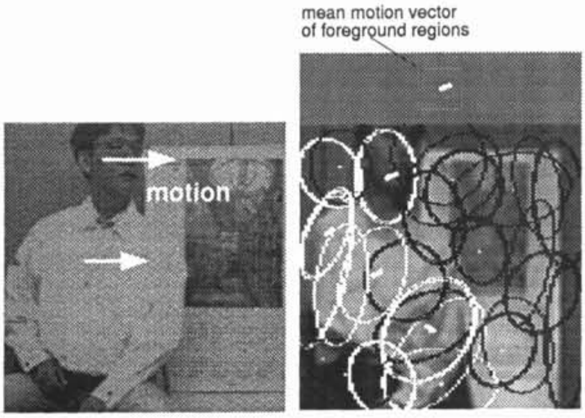
[RZ86] D. Rumelhart and D. Zipser. *Parallel Distributed Processing*, chapter Feature Discovery by Competitive Learning. MIT Press, 1986.

[TLT93] A. Tsukamoto, C-W. Lee, and S. Tsuji. Detection and tracking of human face with synthesized templates. In *Proc. ACCV'93*, pp. 183–186, Osaka, Nov. 1993. IEICE Japan.



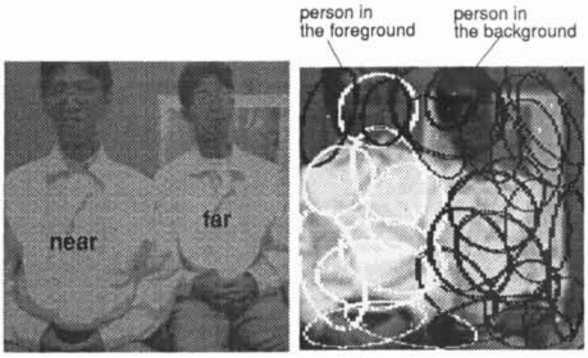
(a) input image (b) system output

Figure 6: Early stage of the experiment.



(a) input (b) system output

Figure 7: A swinging person.



(a) input (b) system output

Figure 8: Two persons in different distance.