Intermediate view synthesis considering occluded and ambiguously referenced image regions[1]

Jeffrey S. McVeigh[*], M. W. Siegel[**] and Angel G. Jordan[*]

[*]Department of Electrical and Computer Engineering

[**]Robotics Institute / School of Computer Science

Carnegie Mellon University, Pittsburgh, PA 15213

Contact Address:

Jeffrey S. McVeigh

Department of Electrical and Computer Engineering

Carnegie Mellon University

5000 Forbes Avenue

Pittsburgh, PA 15213

jmcveigh@cmu.edu

Total number of pages: 16

Number of figures: 3

Number of footnotes: 2

Running headline: Intermediate view synthesis

Keywords: Stereoscopic image processing, view interpolation, disparity estimation, occlusion detection, motion parallax, viewer-specified inter-camera separation.

ABSTRACT

In this paper, we present an algorithm for synthesizing intermediate views from a single stereopair. The key contributions of this algorithm are the incorporation of scene assumptions and a disparity estimation confidence measure that lead to the accurate synthesis of occluded and ambiguously referenced regions. The synthesized views have been displayed on a multi-view binocular imaging system, with subjectively effective motion parallax and diminished eye strain.

1. INTRODUCTION

For three-dimensional television (3D-TV) to become feasible and acceptable on a wide scale, the added realism must outweigh any required increases in processing and system complexity, and the stereoscopic information must be comfortable to view. Both of these goals can be achieved if intermediate views of the scene are available. An intermediate view is defined as the image that would be obtained from a camera located between and on a straight-line connecting the given stereopair's cameras.

While binocular systems provide depth information through the sensation of stereopsis, a system consisting of only two views of a scene lacks the important depth cue of motion parallax, which provides the distinction between binocular and three-dimensional systems [14]. Motion parallax can be synthesized from multiple intermediate views of the scene by presenting the correct stereopair according to the observer's position.

Discomfort is often experienced when viewing stereoscopic images on two-dimensional displays [9]. As with any subjective assessment, this discomfort is viewer-dependent. The authors observe that viewers prefer varying degrees of depth perception from binocular imagery based on individual stereoscopic viewing ability and the range of depth present in the scene. A greater sense of depth is provided by a relatively large inter-camera separation, but the larger the separation the more difficulty marginal viewers have in fusing the images. If intermediate views of a scene are available, a viewer can dynamically select the inter-camera separation for comfort and preferred sense of depth[2].

After a brief discussion of prior work on image interpolation, we describe our algorithm for intermediate view synthesis. Synthesis is performed, in the standard manner, using the geometric relationship between the given views and estimated disparity values [1, 4, 6, 7, 10-12]. Depth information for occluded and ambiguously referenced regions are approximated through the development of reasonable scene assumptions and the utilization of the signal-

to-noise ratio (SNR) from the disparity estimation procedure as a confidence measure; these novel steps, which are discussed in detail, allow for the accurate synthesis of "difficult" regions, and are the key contributions of this paper. We conclude with experimental results and potential directions for future research.

2. PRIOR WORK

A common technique for the synthesis of intermediate views is through the generation of epipolar plane images (EPI) [4, 6]. Skerjanc and Liu base their synthesis on the mapping of feature points from three camera views [11]. While these techniques reduce the occurrence of occlusion and improve disparity estimation performance through the use of multiple views of the scene, they require increases in camera complexity and bandwidth to transmit the additional views. Other prior work has attempted to construct a three-dimensional model of the scene from two or more images and then utilize computer-graphics routines to synthesize intermediate views [2]. However, the construction of a 3D model may be an unnecessary and time-consuming step if the ultimate goal is only the intermediate view.

Intermediate view synthesis is closely related to work performed on motion compensated interpolation for frame rate conversion, de-interlacing and frame skipping, where views are generated from two temporally offset images [1, 7, 10, 12]. For these techniques, when a portion of the interpolated image is not mapped from the given displacement vectors, zero displacement is often assumed most probable and the missing region is mapped from the corresponding region in one of the two reference image. However, for stereopairs only a single plane in the three-dimensional scene will exhibit zero displacement, and using replication from one of the given views for these unmapped regions results in unsatisfactory synthesis performance. Ribas-Corbera and Sklansky addressed the problem of ambiguously referenced regions by assuming a static background and no inter-object occlusion [10].

Our technique attempts to handle potentially large occluded regions, without requiring multiple views, by inferring depth information for these regions from the estimated disparity of adjacent unoccluded regions. We attempt to handle ambiguously referenced regions more appropriately for stereopairs by evaluating the estimation performance for the disparity values in question.

3. SYNTHESIS ALGORITHM

We assume that the cameras used to capture a given stereopair were separated by a horizontal distance and arranged in the parallel camera configuration, as detailed by Grinberg, *et. al.* [3]. A point in the scene generates corre-

sponding points $\boldsymbol{p_l}$ and $\boldsymbol{p_r}$ in the left- and right-eye images, respectively. If the scene point is unoccluded (visible in both images), the disparity is defined as the distance, in pixels, between the corresponding points in the appropriately framed images. Due to the parallel camera axis geometry, the disparity is only in the horizontal direction and the relationship between $\boldsymbol{p_l}$ and $\boldsymbol{p_r}$ is given by,

$$\boldsymbol{p_r} = \begin{bmatrix} x_r \\ y_r \end{bmatrix} = \begin{bmatrix} x_l + d_{lr}(x_l, y_l) \\ y_l \end{bmatrix} = \boldsymbol{p_l} + \begin{bmatrix} d_{lr}(x_l, y_l) \\ 0 \end{bmatrix} \tag{1}$$

where $d_{lr}(x_l, y_l)$ is the disparity from the left pixel to the corresponding right pixel.

Figure 1 depicts a simple scene containing a background and a single foreground object, in one-dimension. In Fig. 1a, the object and background are located at the distances they are perceived to lie based on their disparities. Interesting rays from the background to the eyes are included in the figure, and the foreground object and background have been shaded according to whether a particular portion is visible to both, one, or neither of the eyes. In Fig. 1b, the overlaid images have been separated for clarity and the disparity vectors for unoccluded portions have been indicated.

Figure 1:

The geometric relationship between unoccluded pixels in the given stereopair and the desired intermediate view can be derived from sample points lying along a disparity vector in Fig. 1b. The distance between the left and right eye is normalized to one. The desired intermediate view is then parameterized by its relative location (v), where $0 < v < 1$. The mapping from the given left image to the position in the intermediate view is a function of disparity, as shown by the linear relationship,

$$\boldsymbol{p_i} = \begin{bmatrix} x_i \\ y_i \end{bmatrix} = \begin{bmatrix} x_l + v d_{lr}(x_l, y_l) \\ y_l \end{bmatrix} \tag{2}$$

The symmetrical relationship can be written for the right image. If the actual disparity values are known, the pixel intensity values for unoccluded regions are mapped easily from either given image to the point in the intermediate image. However, so far we have neglected two important considerations: 1) the effect of disparity vector estimation errors, and 2) the synthesis from occluded regions in the original stereopair. We next describe our solutions to these problems.

In Fig. 2a, two disparity vectors pass through the desired intermediate pixel (x_i). This situation may arise when

an object lies in front of another object and the difference in disparities between the two objects is greater than the width of the foreground object. However, ambiguously referenced regions are most likely the result of disparity estimation errors. We select the most probable disparity for ambiguously referenced pixels through the use of a disparity estimate confidence measure. In our synthesis algorithm, we perform block-based disparity estimation in both directions. We first attempt to eliminate false estimates due to occluded regions with a procedure similar to one proposed by Cafforio, *et. al.* [1]. If only a single vector was detected as coming from an unoccluded region, this vector is used to perform the mapping. Otherwise, the ambiguously referenced intermediate pixel is mapped using the disparity vector with maximum estimation confidence. We use the signal-to-noise ratio between the desired and estimated blocks from the disparity estimation procedure as the confidence measure, since this provides an easily calculated indication of estimation performance [5].

Figure 2:

Figure 2b illustrates generating an intermediate image pixel from an occluded region, where the occlusion is in the left image. Since disparity is undefined for occluded regions, we cannot use Eq. (2) directly to map pixel intensities from these regions. We handle the synthesis from occluded regions by making best reasonable assumptions of the scene. These assumptions are that occluded regions have constant depth, and that this depth is equal to the depth of the unoccluded region located *on the appropriate side* of the occlusion. If the occlusion is visible in the left image, the appropriate side is to the left of the occlusion, and vice versa for an occlusion in the right image (see Fig. 1a). These assumptions, when valid, allow for depth information to be inferred for occluded regions, resulting in accurate intermediate view synthesis for these regions.

With these tools to handle ambiguously referenced pixels and the synthesis from occluded regions in hand, we proceed with the complete intermediate view synthesis algorithm. Synthesis is performed separately from both left and right images, and then the individual intermediate views are combined into the final synthesized view. This seemingly redundant process allows us to minimize the effect of disparity estimation errors and to select the correct mapping for occluded regions.

Our occlusion detection procedure is similar to one proposed by Cafforio, *et. at.*; however, our procedure takes the form of a classic detection problem, and does not require a search step [1, 13]. For each pixel in both images, we sum the disparity vector in one direction with the corresponding disparity vector in the opposite direction. If the absolute value of the sum is greater than a threshold the pixel is assumed to be occluded. The threshold is derived from

statistical descriptions of disparity estimates for occluded and unoccluded regions.

The actual synthesis from both images then is performed using the estimated and processed disparity values and Eq. (2). All disparity values detected as coming from unoccluded regions are used to map the pixel intensity values to the intermediate image pixels. If the intermediate pixel is mapped by more than one disparity vector, the vector with maximum block SNR is used for the mapping. The image block SNR (in dB) is given by,

$$SNR = 10\log\frac{\sigma^2}{\sigma_e^2} \tag{3}$$

where σ^2 and σ_e^2 respectively denote the original image block variance and prediction block mean squared error. For the case of one or more valid mappings, the intermediate pixel is marked as being mapped from an unoccluded region.

If no disparity values map to a particular intermediate pixel, the intermediate pixel is marked as coming from an occluded region. After the intermediate view has been synthesized from all unoccluded regions, these missing pixels are mapped using the relationship established for the pixels on the appropriate side of the occlusion (e.g., the mapping on the left side of the occlusion for the synthesis from the left reference image). After mapping a set of intermediate image pixels resulting from occluded regions, the disparity values used for the pixels bounding this set are examined to verify that the occluded portion was in fact visible in the reference image. For example, if the synthesis is being performed from the left image and the bounding disparity values indicate that the object to the right of the occlusion lies in front of the object to the left, we conclude that the occluded portion was present in the left image. If the occluded region was mapped correctly it is marked as such, otherwise the pixels are marked as being incorrectly mapped from occluded regions. This information is used in the decision process when the two separately synthesized intermediate views are combined, which is the final step.

The selection between choosing the pixel intensity value from the view synthesized from the left or right image is based on the mapping performed and the relative signal-to-noise ratios. The order of preference for the selection is as follows: 1) pixels mapped from unoccluded regions over those from occluded regions, 2) pixels correctly mapped from occluded regions over those incorrectly mapped, and 3) pixels mapped using disparity values with higher SNR.

4. EXPERIMENTAL RESULTS

We have examined eight image pairs in varying degrees of detail; in this section we discuss the synthesis results of one exemplary stereopair. The initial disparity estimates were obtained from an exhaustive search, block-based displacement estimation technique, where the block sizes and disparity search ranges were manually defined. The disparity estimation included the realistic possibility of a vertical disparity component, which was handled by extending the occlusion detection procedure and the geometric mapping relationship to two dimensions.

Figure 3:

An “original” stereopair and two intermediate views were obtained from four frames of the *Flower Garden* monoscopic image sequence (Stereopair: Figs. 3a and 3b; Intermediate views: Figs. 3c and 3d). Since the motion in this sequence is almost entirely horizontal camera motion, two temporally offset frames provide an effective stereopair. A block size of 5 pixel widths and 8 pixel heights was selected, and the search range was ±18 horizontal and ±3 vertical pixels.

The occluded regions in the left-eye image are the regions to the left of the large foreground tree and the left image borders, and vice versa for the right-eye image. The occluded portions are estimated to occupy roughly 7% of the original stereopair. The reader is encouraged to inspect especially these areas in the following synthesized views to evaluate the performance of the algorithm.

Twenty equally spaced intermediate views were synthesized from the original stereopair (Figs. 3a and 3b). Assuming that Figs. 3(a, c, d, b) are equally spaced, synthesized Figs. 3(e, f) correspond to Figs. 3(c, d). A binocular imaging system equipped with a head-tracking device was used to display the appropriate stereopair based on the viewer’s horizontal position. A “stereo-dial” was also incorporated in the imaging system to allow the viewer to specify the inter-camera separation [8]. Viewers reported effective motion parallax and diminished eye strain with this system.

The error images for these two intermediate views are shown in Figs. 3g and 3h, where the absolute value of the error signal has been normalized to the range [0, 255]. As expected, the residuals are small, but relatively largest at the left and right edges of the foreground objects, especially the tree. Note that since we do not know that Figs. 3(a, c, d, b) are exactly equally spaced, the technique may actually be even better than the comparison of (c) against (e) and (d) against (f) suggests.

5. CONCLUSION

We have presented a technique for the synthesis of intermediate views from a single stereopair that can accurately handle occluded and ambiguously referenced regions. The gain in terms of improved synthesis for “difficult” regions is obtained with a relatively small increase in the number of processing steps compared with techniques that do not consider these regions. We have provided experimental results that illustrate the algorithm’s synthesis performance for both unoccluded and occluded regions. The majority of the imperfectly synthesized image regions were located at the discontinuity between occluded and unoccluded regions. The errors are most likely due to the blocking effects of the disparity estimation procedure and the inability of the occlusion detection scheme to precisely find the edges of occluded regions. Errors may also occur when objects are actually visible in the intermediate view, but are occluded from both of the images in the stereopair used to estimate this view.

Our future work will address improving the performance of the disparity estimation and occlusion detection steps. If successful, this will reduce the errors now present at occlusion discontinuities. We also plan to increase the computational efficiency of this algorithm to avoid the need to pre-compute the intermediate views used in the experiments.

6. REFERENCES

[1]. C. Cafforio, F. Rocca and S. Tubaro, “Motion compensated image interpolation”, *IEEE Trans. on Communications*, Vol. 38, No. 2, February 1990, pp. 215-222.

[2]. T. Fujii and H. Harashima, “Data compression of an autostereoscopic 3-D image”, Technical Report, University of Tokyo, 1994.

[3]. V. S. Grinberg, G. Podnar and M. W. Siegel, “Geometry of binocular imaging”, *Proc. SPIE Internat. Conf. on Stereoscopic Displays and Virtual Reality Systems*, Vol. 2177, San Jose, CA, 6-10 February 1994, pp. 56-65.

[4]. R. Hsu, K. Kodama and K. Harashima, “View interpolation using epipolar plane images”, *Proc. IEEE Internat. Conf. on Image Processing*, Vol. 2, Austin, TX, 13-16 November 1994, pp. 745-749.

[5]. A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs, NJ, 1989, Chapter 3, pp. 59.

[6]. A. Katayama, K. Tanaka, T. Oshino and H. Tamura, “A viewpoint dependent stereoscopic display using interpolation of multi-viewpoint images”, *Proc. SPIE Internat. Conf. on Stereoscopic Displays and Virtual Reality Systems II*, Vol. 2409, San Jose, CA, 7-9 February 1995, pp. 11-20.

[7]. J.-S. Kim and R.-H. Park, “Local motion-adaptive interpolation technique based on block matching algorithms”, *Signal Processing: Image Communication*, Vol. 4, No. 6, November 1992, pp. 519-528.

[8]. J. S. McVeigh, V. S. Grinberg and M. W. Siegel, “Double buffering technique for binocular imaging in a window”, *Proc. SPIE Internat. Conf. on Stereoscopic Displays and Virtual Reality Systems II*, Vol. 2409, San Jose, CA, 7-9 February 1995, pp. 168-175.

[9]. S. Pastoor, “3D-television: A survey of recent research results on subjective requirements”, *Signal Processing: Image Communication*, Vol. 4, No. 1, November 1991, pp. 21-32.

[10]. J. Ribas-Corbera and J. Sklansky, “Interframe interpolation of cinematic sequences”, *Journal of Visual Communication and Image Representation*, Vol. 4, No. 4, December 1993, pp. 392-406.

[11]. R. Skerjanc and J. Liu, “A three camera approach for calculating disparity and synthesizing intermediate pictures”, *Signal Processing: Image Communication*, Vol. 4, No. 1, November 1991, pp. 55-64.

[12]. R. Thoma and M. Bierling, “Motion compensating interpolation considering covered and uncovered background”, *Signal Processing: Image Communications*, Vol. 1, No. 2, October 1989, pp. 191-212.

[13]. H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, John Wiley & Sons, New York, 1968, Chapter 2, pp. 19-46.

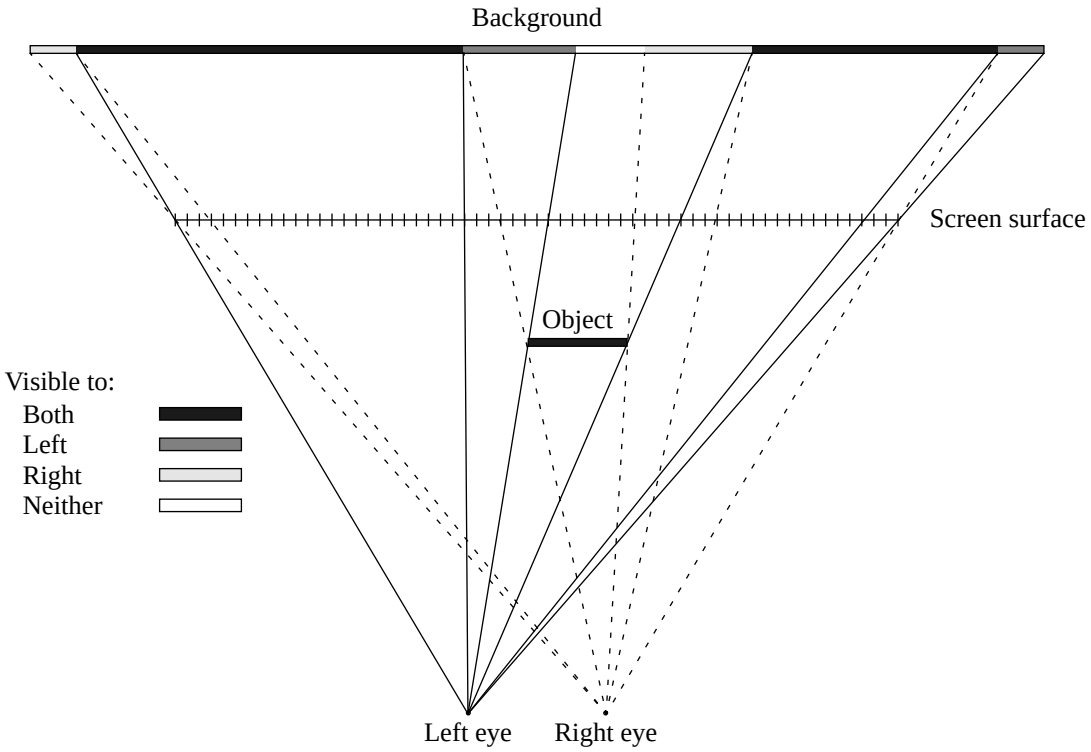
[14]. C. D. Wickens, “Three-dimensional stereoscopic display implementation: Guidelines derived from human visual capabilities”, *Proc. SPIE Internat. Conf. on Stereoscopic Displays and Applications*, Vol. 1256, Santa Clara, CA, 12-14 February 1990, pp. 2-11.

List of footnotes:

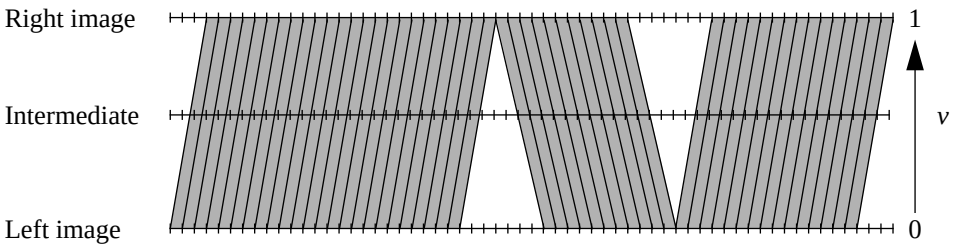
1. This research was supported by the Advanced Research Projects Agency under ARPA Grant No. MDA 972-92-J-1010.

2. We speculate that those viewers who have difficulty fusing stereopairs may derive as much stereopsis sensation from the reduced depth range image as other viewers who have no difficulty derive from increased depth range. These assessments are informal in the sense that designed and controlled human factors experiments were not conducted. They are our impressions based on perceptions reported by colleagues and the many visitors to our lab.

List of Figures:

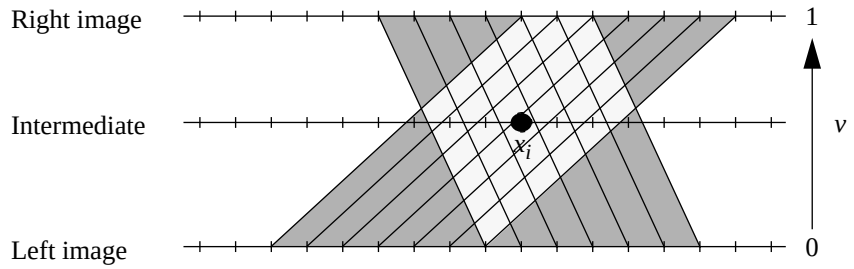

Background
Screen surface
Object
Visible to:
Both
Left
Right
Neither
Left eye
Right eye

a.

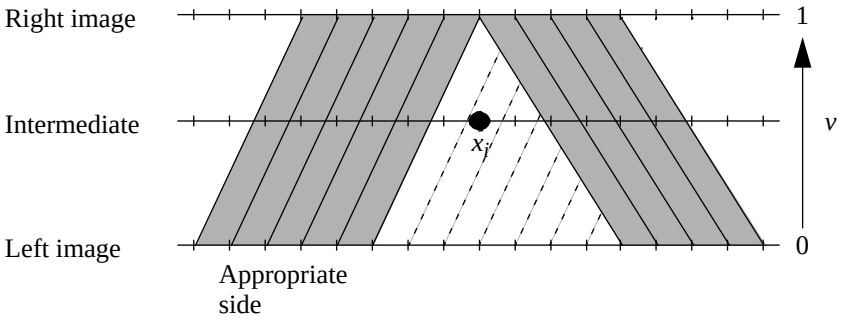

Right image
Intermediate
Left image
1
v
0

b.

Figure 1:

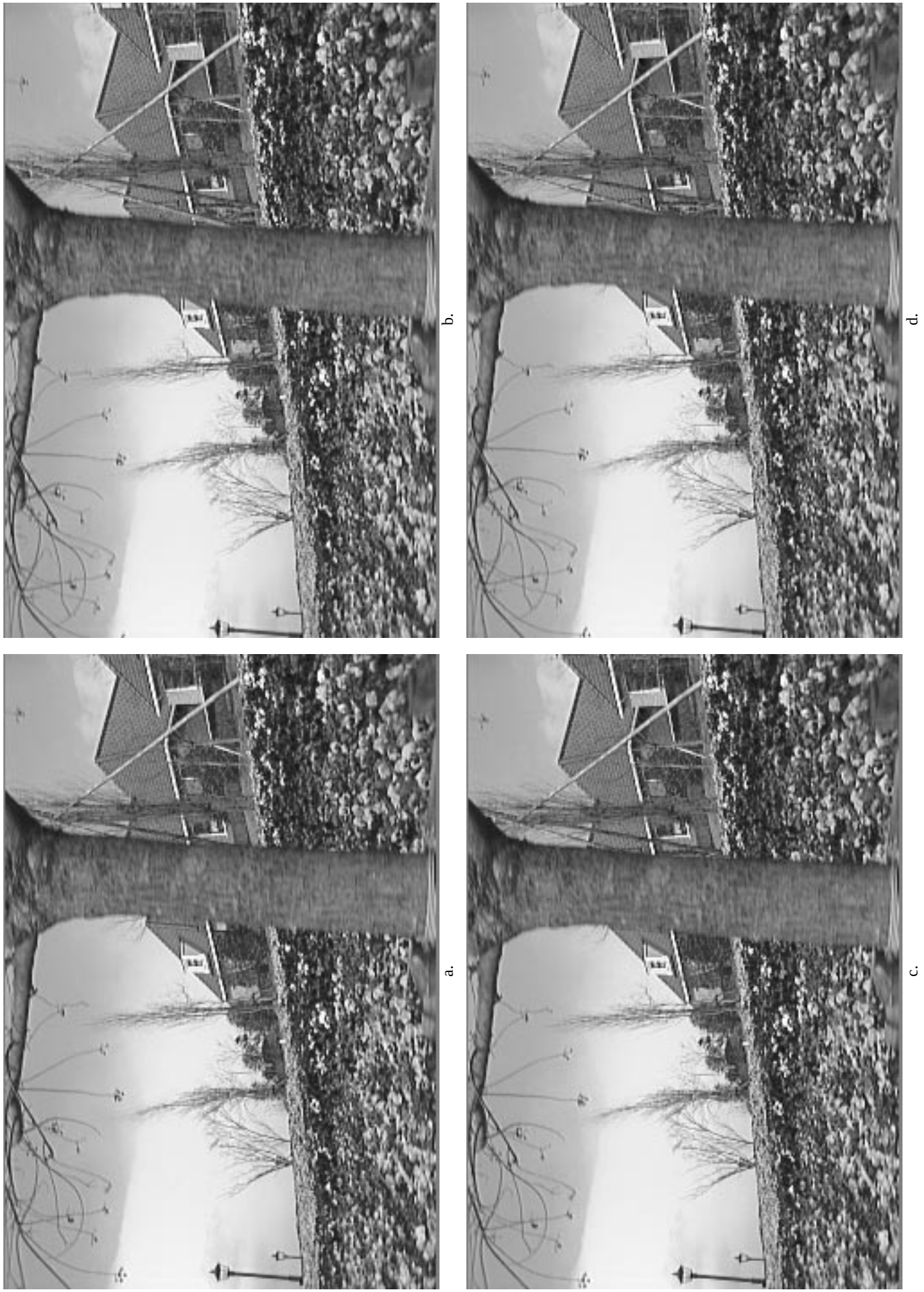

Right image
Intermediate
Left image
x_i
1
v
0

a.


Right image
Intermediate
Left image
x_i
1
v
0
Appropriate
side

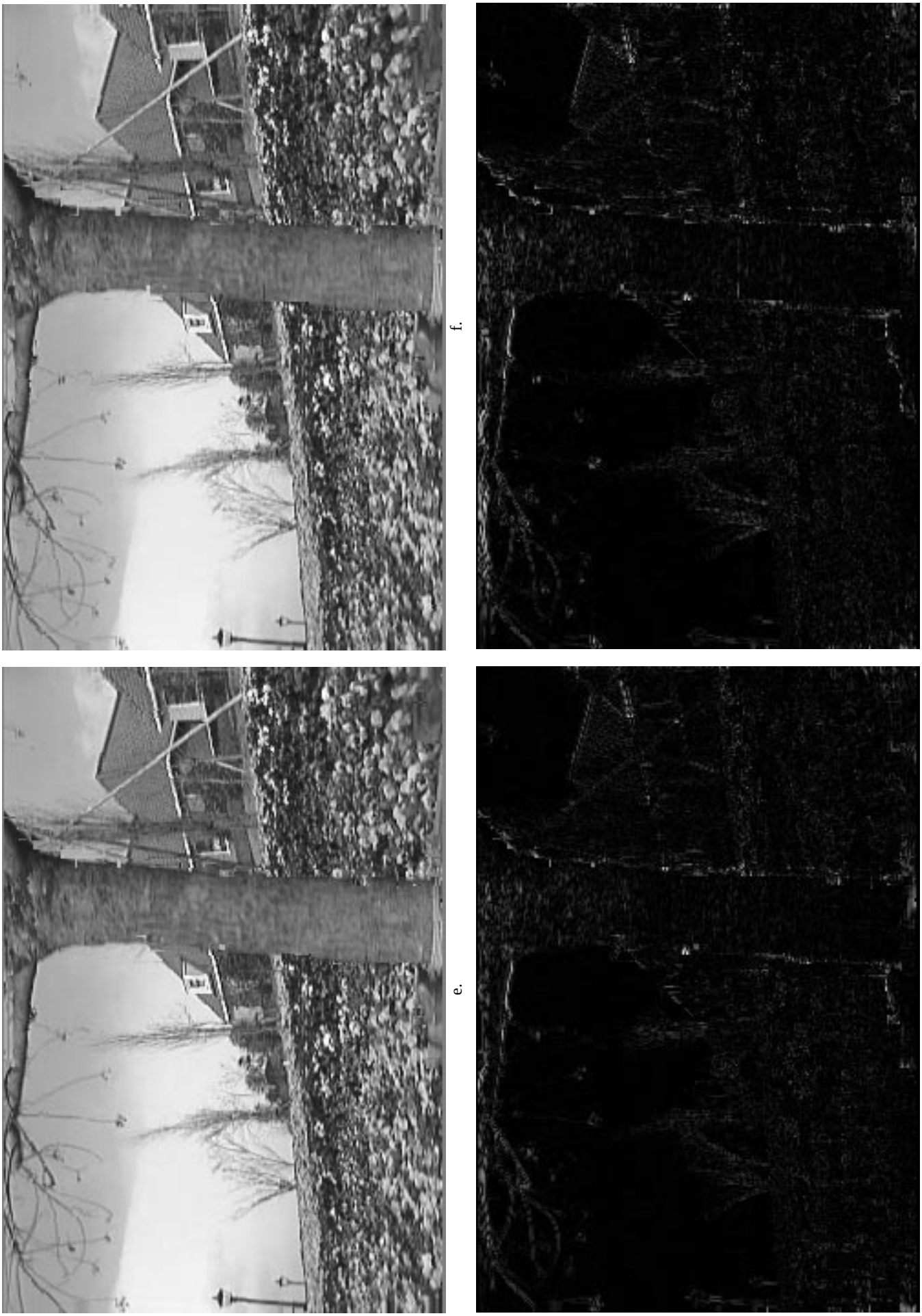
b.

Figure 2:



a. b. c. d.

Figure 3:



e. f. g. h

Figure 3: (cont.)