Video Analysis

LV 188.329, WS 2016/17

3D Scene Reconstruction for 3D Film/TV Applications

Ao. Univ.-Prof. Margrit Gelautz

margrit.gelautz@tuwien.ac.at

FAKULTÄT FÜR !NFORMATIK

Faculty of Informatics

Outline

- Introduction
- Principles of Stereo Processing
- 3D Displays
- Research Topics (Selection)
  - Stereo matching
  - Image matting
  - Novel view generation and inpainting

Introduction

- **Development in recent years**
  - **3D movies starting in the year 2009: „Ice Age 3“, „Avatar“, …**
  - **Tests with 3D TV started in Korea, Japan, …**

Japan hat drei Dimensionen auf dem Schirm

3-D steht im Mittelpunkt der asiatischen Unterhaltungselektronik-Messe Ceatec – Europa braucht sich nicht verstecken

Martin Kölling aus Tokio

Japans berühmte Elektronikmesse Ceatec bietet in diesem Jahr buchstäblich einen Blick in die Kristallkugel. Am Eingang der Halle für Konsumelektronik hat die kleine Erfinderzunft Holoart ein Glaskugel ausgestellt, die durch ein fotoreigniprojiziertes Bild zum fest ben Display mutiert. Der nächste Trend der Elektronikindustrie ist allerdings auch ganz ohne wahrsagerische Hilfsmittel offensichtlich: Im Im fünfzehnte dreidimensionale Flachbildfernseher.

Japans Hersteller Sony hat auf der Ceatec gleich seine Hauptbühne der neuen Technik zwischenbeben und zeigt auf einem halben Dutzend Flachbildfernseher Videospiele, Sport und Filme in drei Dimensionen. Doch auch die Rivalen Panasonic, Sharp und Mitsubishi Electric stellen die 3-D-TVs in den Mittelpunkt. Mit den neuen Geräten enthen sie von derzeitigen 3-D-Kinofilmen profitieren.

Auch wenn Sony den anderen Herstellern auf der Messe die Schau stiehlt, der Vorreiter bleibt Panasonic. Nachdem Panasonic 2008 den ersten Inchaufflüssigen 3-D-Fernseher der Welt vorgestellt hat, folgt dieses Jahr das erste massenmögliche Modell. Der präsentierte 50-Zoll-Plasmafernseher kommt 2010 auf den Markt. Sony wird eine Monate später folgen. Die

Vorspielte Idee: Kristallkugel, die durch ein hineinprojiziertes Bild zum fesdsen Display wird. Foto: ap/dpa

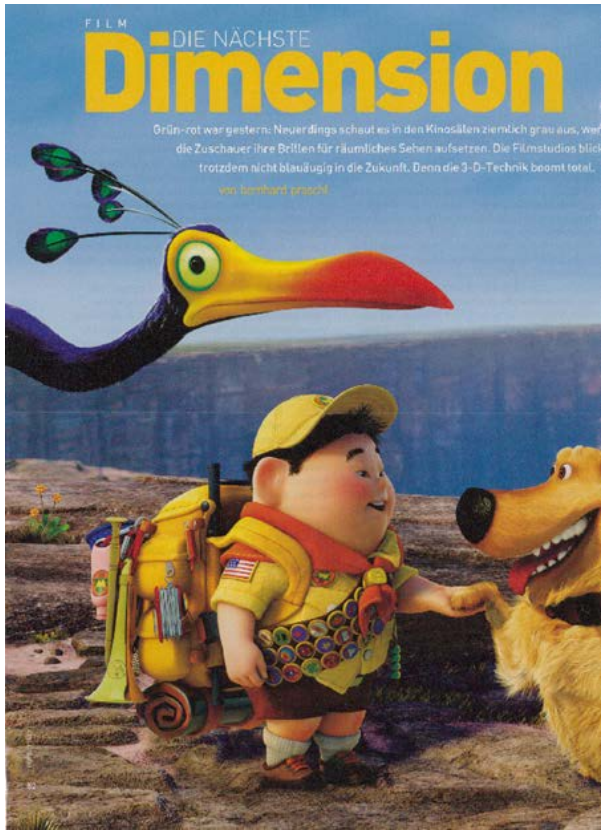
rattische Konkurrenz dürfte noch ein bisschen länger mit der anspruchsvollen Technik ringen.

Auf den ersten Blick gleicht die Technik den bisherigen Experimenten: Ohne Brille geht nichts. Noch immer werden zwei leicht versetzte Bilder gesendet, von denen die Brille je eines an je ein Auge durchlässt. Das Gehirn setzt sich aus den Informationen dann ein räumliches Bild zusammen. Doch die Pixel der kommenden TV-Generation reagieren zwischen so schnell, dass die Hersteller die zwei Bilder in voller Auflösung abwechselnd auf den Schirm schicken können.

Die Kinder tragen daher statt der bisher bekannten rot-grünen Brillen Hightechgeräte, die in Takt mit dem Bildwechsel abwechselnd das rechte und das linke Glas verdunkeln.

Ein Infrarotsender auf dem Schirm dient dabei als Taktgeber. „Die Technik ist anspruchsvoll, weil die Bilder sehr genau voneinander getrennt werden müssen“, erläutert ein Manager von Panasonic. Denn wenn es Überschneidungen gibt, wird das Fernsehen zum Augenschmaus zum Augengraus.

Die europäische Fahne im dreidimensionalen Kegel hält derweil das Heinrich-Hertz-Institut der deutschen Fraunhofer-Gesellschaft hoch. Die deutschen Forscher stellen ein nutzbares, brillenloses 3-D-Display vor. Schon, dessen Bilder sich durch Fingerzeig in alle Richtungen drehen lassen. Nur trägt es nicht zum Fernsehen, sondern nur zum Werbeoder Infodisplay.



FILM

DIE NÄCHSTE

Dimension

Grün-rot war gestern: Neuerdings schaut es in den Kinosälen ziemlich grau aus, wenn die Zuschauer ihre Brillen für räumliches Sehen aufsetzen. Die Filmstudios blicken trotzdem nicht blauäugig in die Zukunft. Denn die 3-D-Technik boomt total.

von bernhard praschl

What is 3D Film/Video?

- **Knowledge of the 3D scene structure is the basis for**
  - **Novel view generation**
    - Virtual camera view points
    - Interactive navigation through 3D scenes
  - **„Mixed Reality“ applications**
    - Combination of real and artificial scene content/video objects
    - Correct treatment of
      - Perspective distortions
      - Illumination and shadows
      - Occlusions
  - **3D displays**
    - Available 3D content needs to be adapted to different types of displays and viewing environments.

- **Generation of novel (virtual) views**

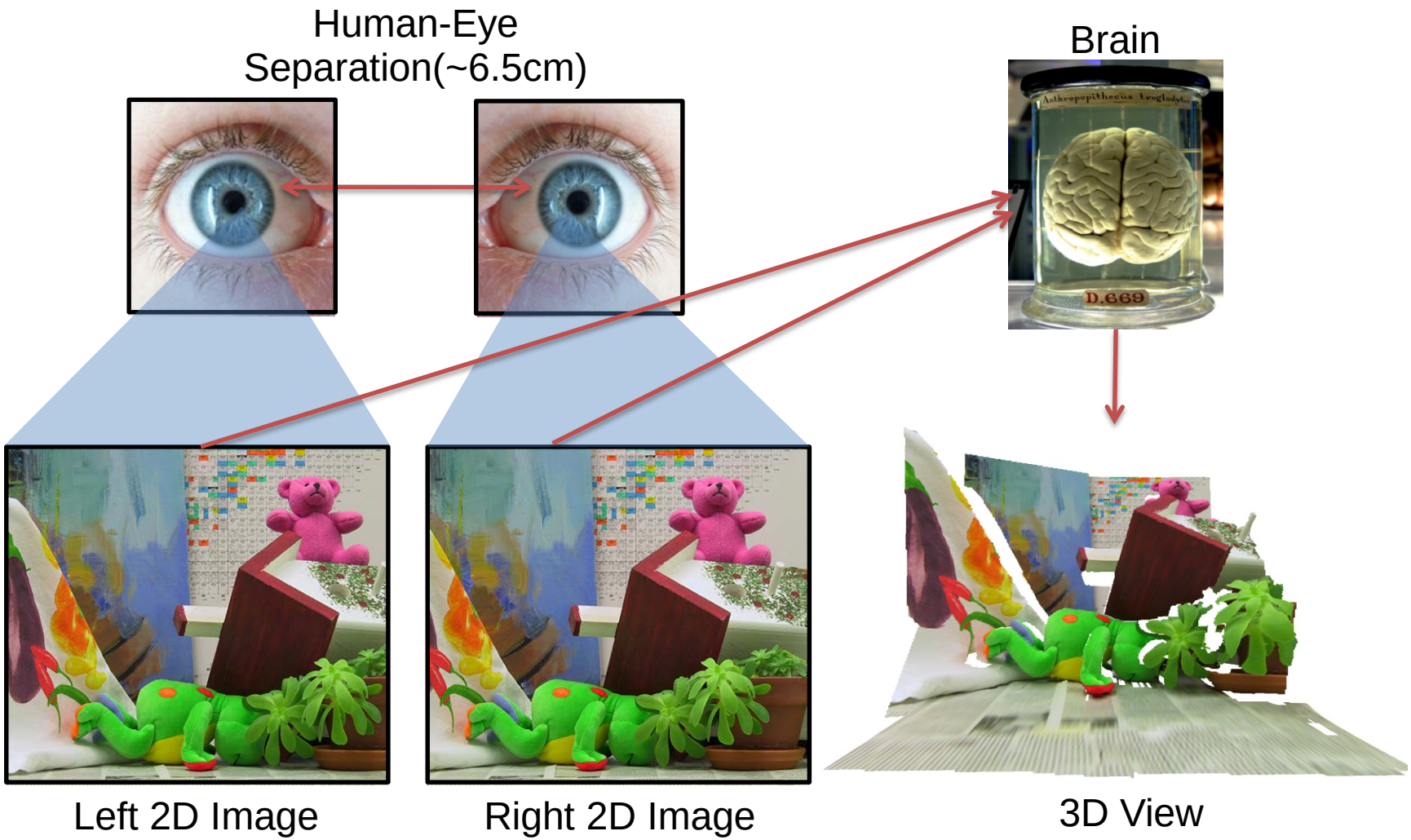
novel viewpoint

novel background

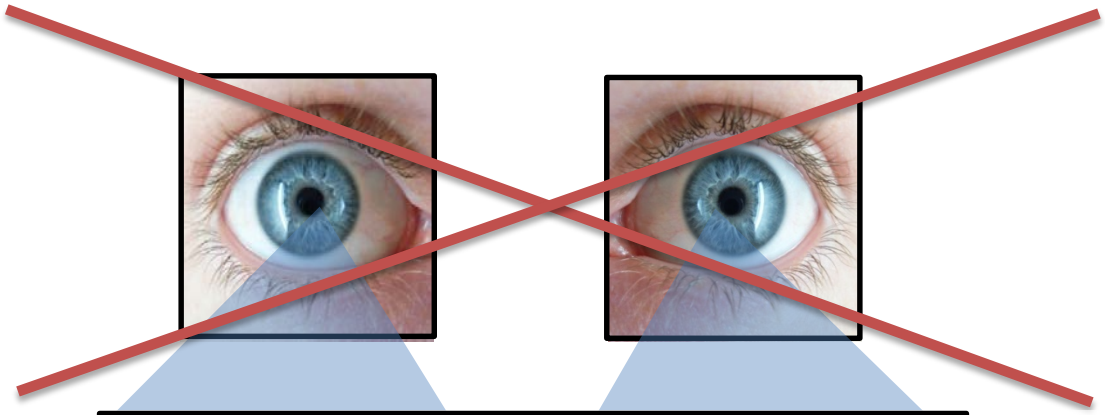
The original stereo pair (upper row) needs to be processed to generate novel views (bottom row).

Principles of Stereo Processing

Human 3D Vision

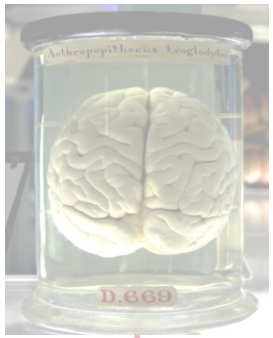


Computer Stereo Vision

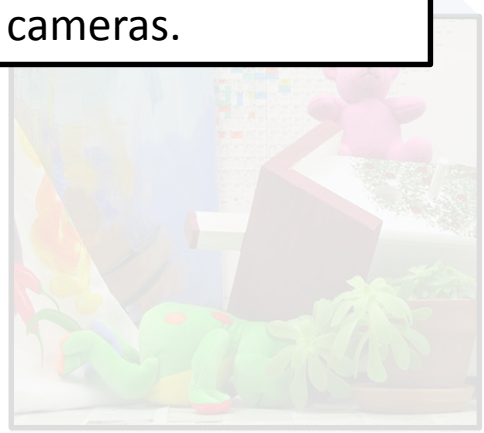


Replace human eyes with a pair of slightly displaced cameras.

Brain



Left 2D Image

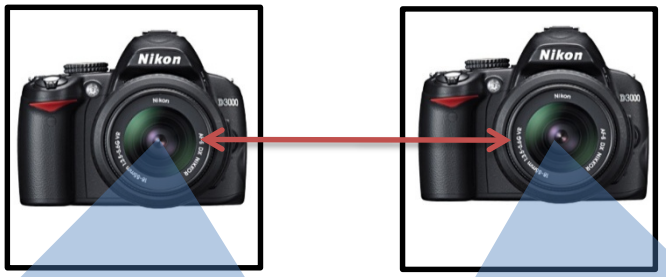


Right 2D Image

3D View

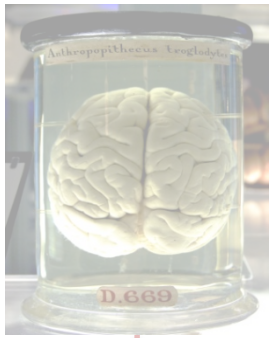
Computer Stereo Vision

Displacement
(Stereo Baseline)



Replace human eyes with a pair of slightly displaced cameras.

Brain

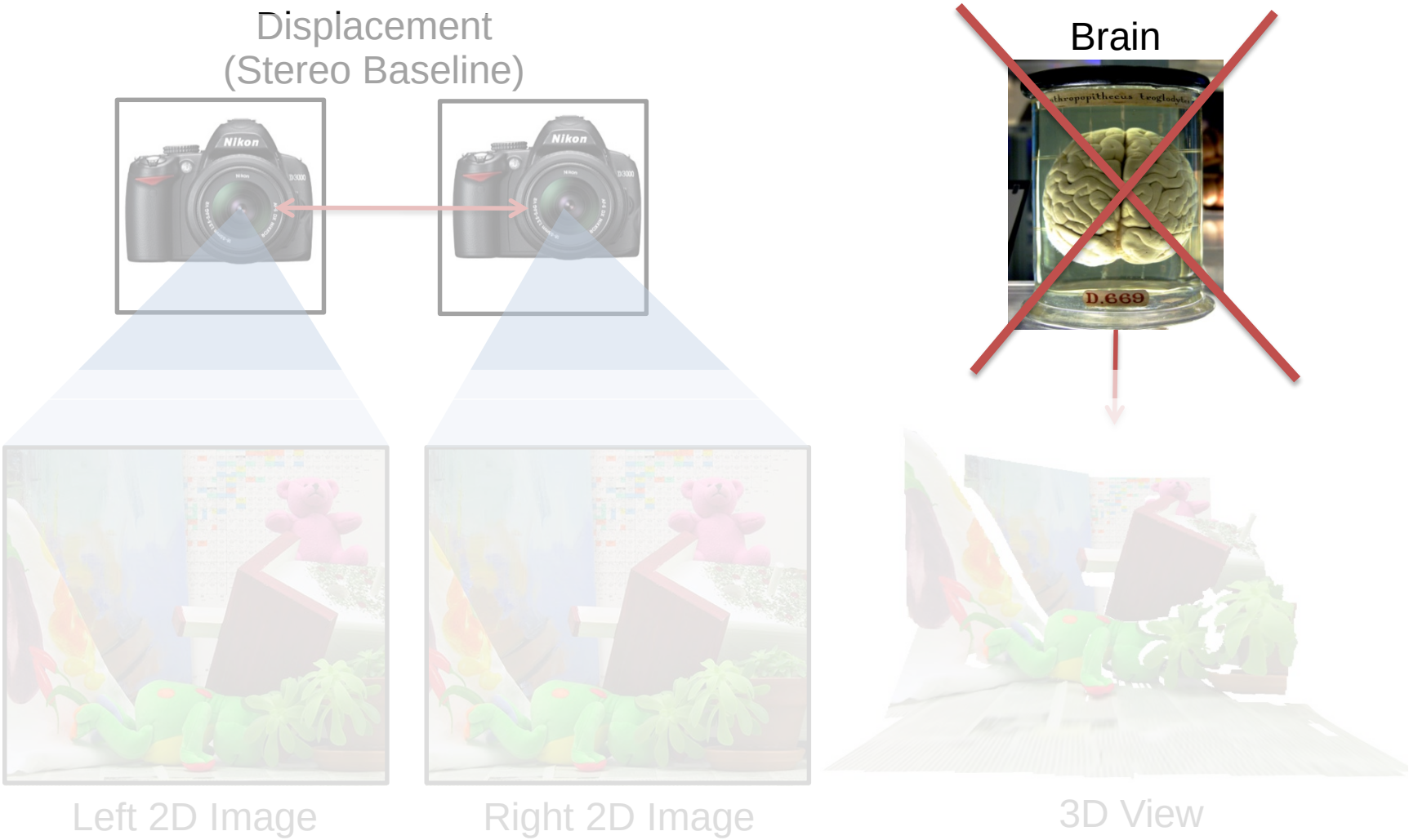


Left 2D Image

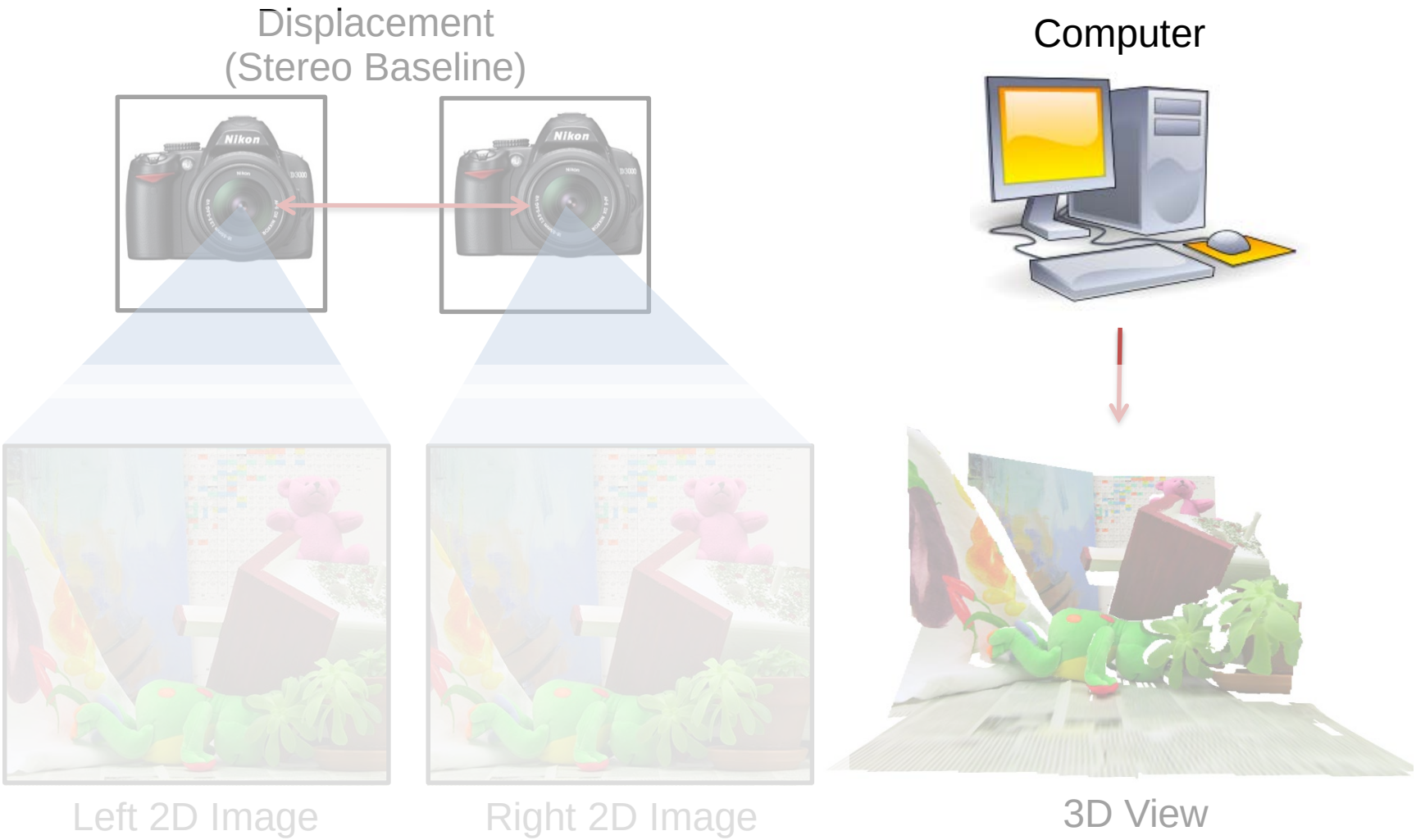
Right 2D Image

3D View

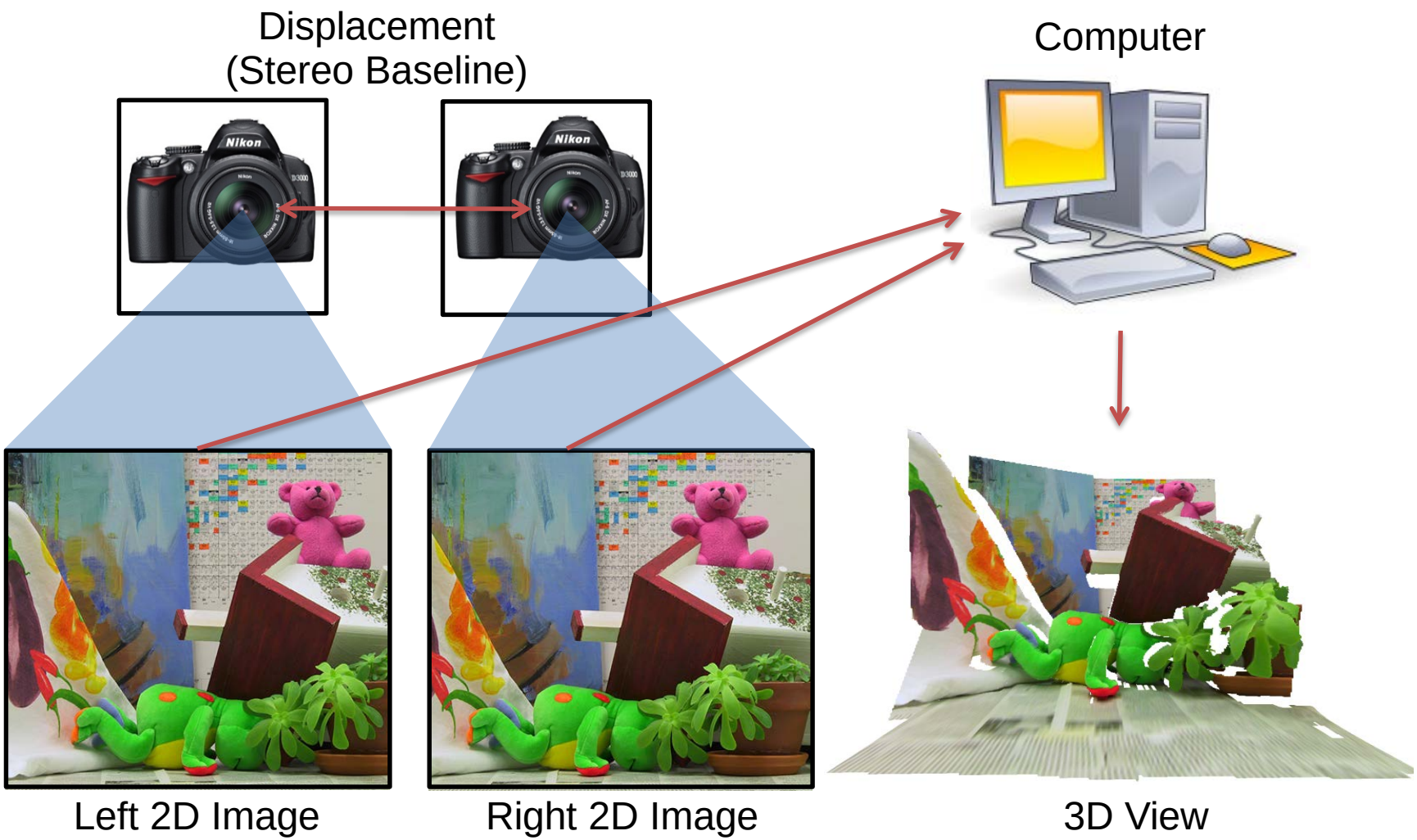
Computer Stereo Vision



Computer Stereo Vision



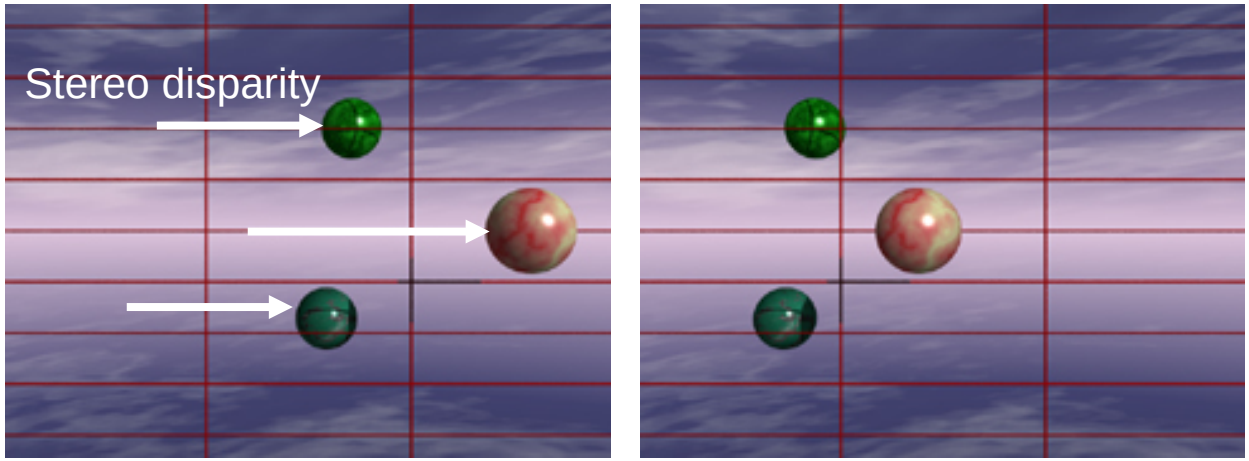
Computer Stereo Vision



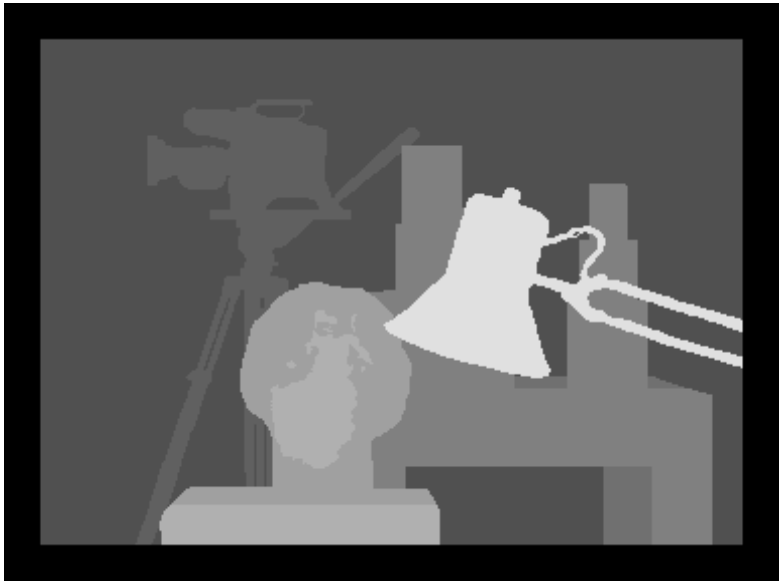
Stereo Analysis (1)

- **Stereo disparity**
  - **The stereo impression perceived by humans relies on the fact that the 2 images captured by the two eyes are not completely identical. The two images show certain geometric differences (*stereo disparities*).**
  - **Further processing of these disparities provides information about the depth of a scene point, i.e., its distance from the eye (or the camera).**

Left and right image of a stereo pair with disparity vectors (Source: Yao Wang, *Video Processing and Communications*)



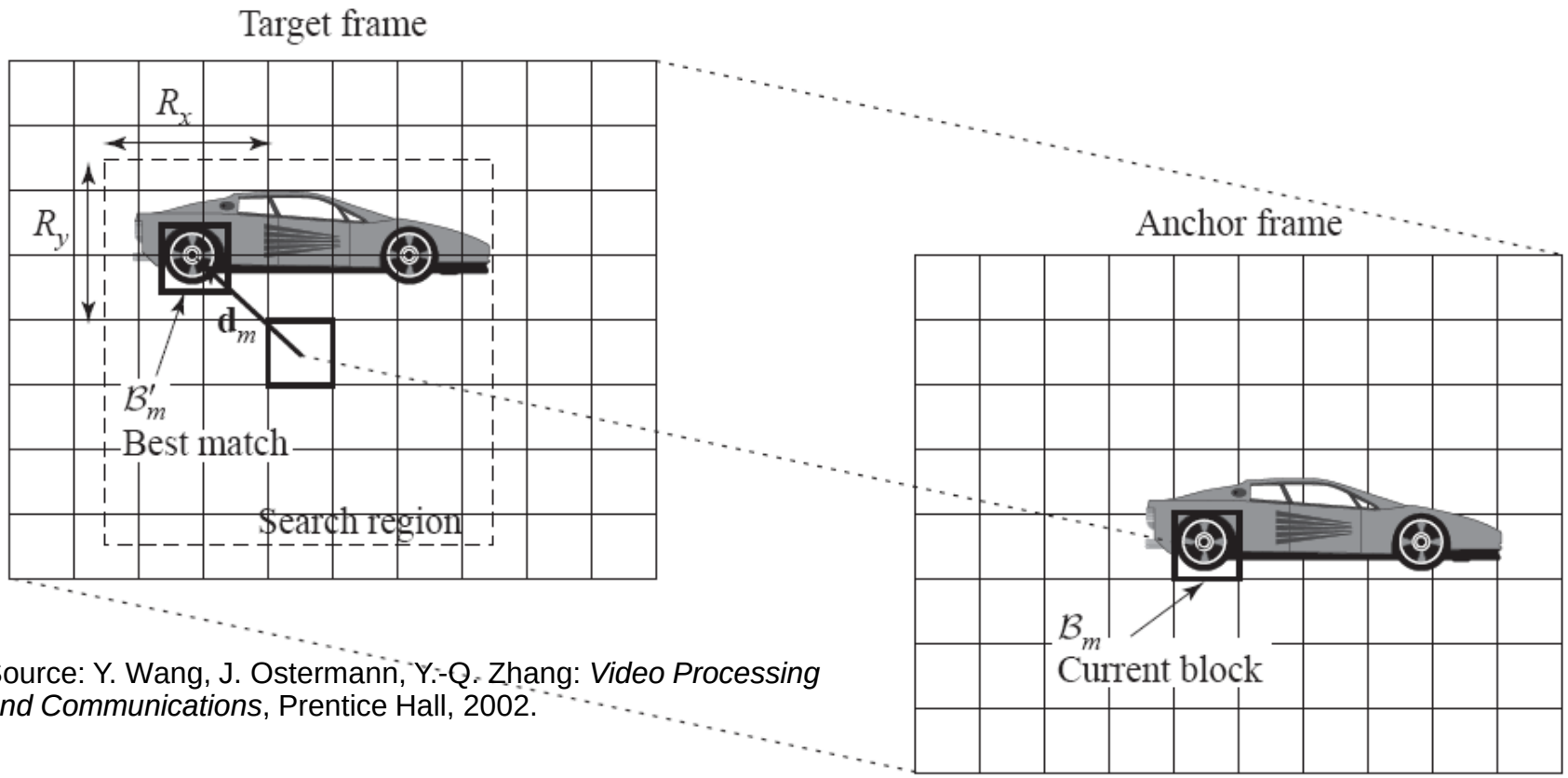
Stereo Analysis (2)



- **Research area: stereo matching**
  - **Automated identification of corresponding points in the left and right image.**
- **Results**
  - **Depth map (~ disparity map)**
  - **Further processing -> novel views**

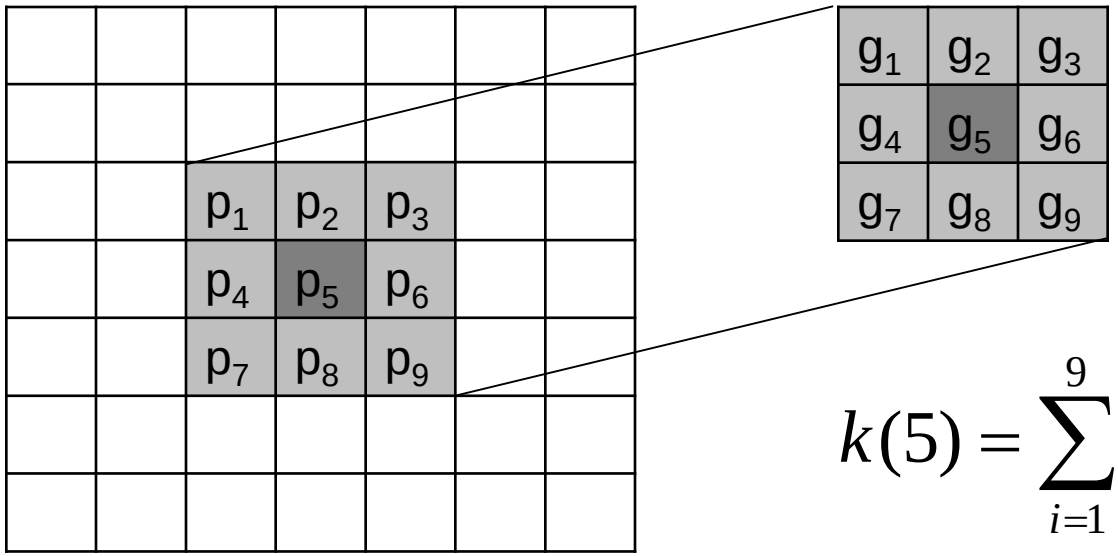
Basic Matching - Image Correlation (1)

- **Task**
  - **Identification of corresponding points in consecutive frames of a video sequence or a stereo image pair.**



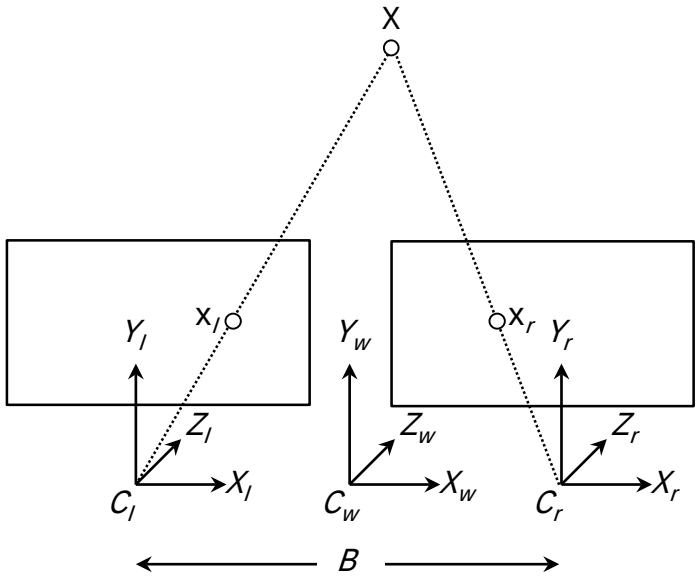
Source: Y. Wang, J. Ostermann, Y.-Q. Zhang: *Video Processing and Communications*, Prentice Hall, 2002.

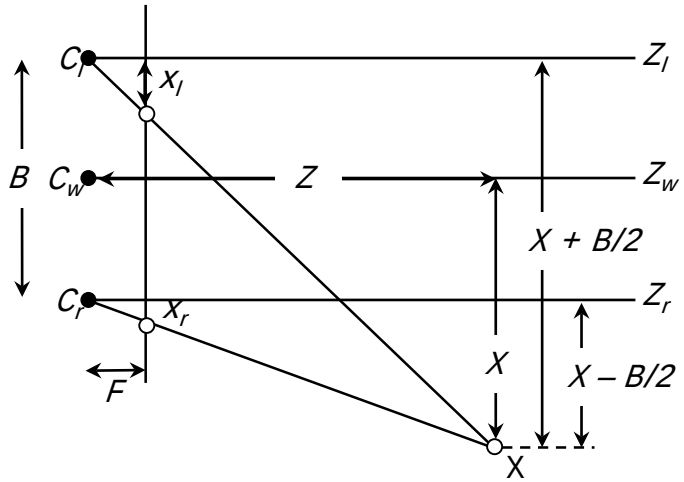
- **Basic technique**
  - **The image content in the regarded section of the master image should coincide as far as possible with the image content at the corresponding position in the search image.**
  - **The differences between gray values – e.g., sum of absolute values – should be a minimum.**



$$k(5) = \sum_{i=1}^{9} \left| p(i) - g(i) \right| \rightarrow \min$$

- **Stereo geometry**





$$X_l = X + \frac{B}{2}, \quad X_r = X - \frac{B}{2}, \quad Y_l = Y_r = Y, \quad Z_l = Z_r = Z;$$

$$x_l = F\frac{X + B/2}{Z}, \quad x_r = F\frac{X - B/2}{Z}, \quad y_l = y_r = y = F\frac{Y}{Z}.$$

$$\boxed{d_x = x_l - x_r = \frac{FB}{Z}}$$

i) **Disparity is inversely proportional to Z**
ii) **Range of disparity increases with B**

Stereo Analysis (4)

- **Some notations on the previous slide:**
  - ***B*** **… base line (distance between the 2 cameras)**
  - ***F*** **… focal length**
  - ***Z*** **… depth (3rd dimension)**
  - ***d*** **… disparity (also called stereo parallax):**
    Specifies the shift (difference in position) between corresponding points in the left and right image.
  - **Capital letters refer to the 3D coordinate systems, lower case letters to the image coordinate system.**
  - **Indices *l, r* and *w* refer to the *left* and *right* camera (or image) and the *world coordinate system,* respectively.**
- **The lowermost formula is important: It shows the relation between the disparity *d* and the depth *Z*.**

Stereo Analysis (5)

- **The relationship between the disparity *d* and the depth *Z* is inversely proportional. The base line *B* and the focal length *F* act as scaling factors.**
- **The (automatic) determination of the disparity *d* for the individual pixels of the image pair is a central task of stereo processing called stereo matching.**
- **If the stereo matching problem is solved and if the positions and other parameters of the cameras (e.g. lens distortion) are known, the depth *Z* and, consequently, the 3D position of the imaged scene points can be determined using the given formulae.**

- **Epipolar geometry**
  - **If the camera position and other camera parameters are known, the stereo image pair can be transformed into epipolar geometry. In epipolar geometry, corresponding points are located along corresponding horizontal lines. The determination of the disparity is thus reduced from a 2D to a 1D search problem (all disparity vectors are horizontal), which facilitates the matching task significantly.**

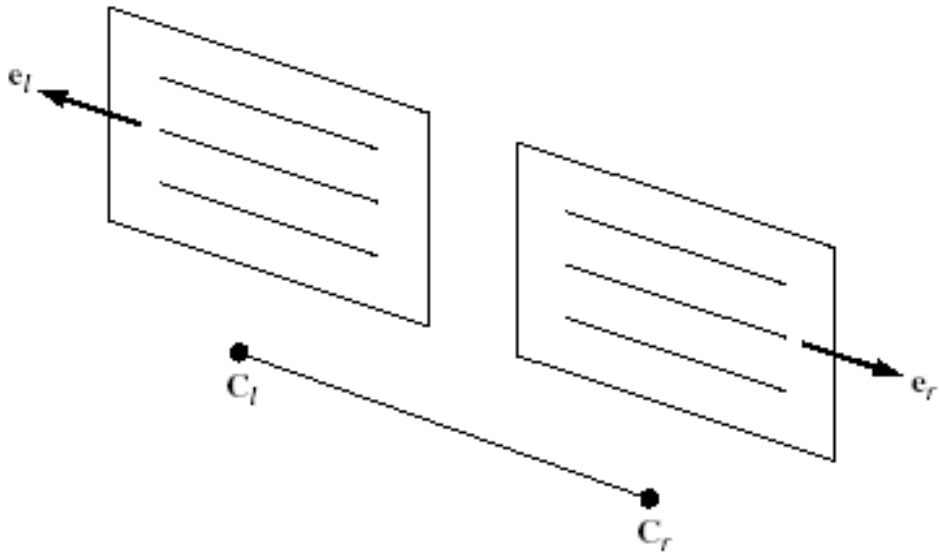


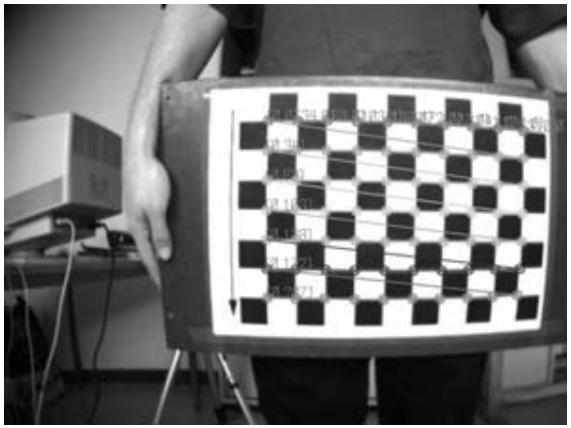
Figure 12.8 Epipolar geometry for a parallel camera configuration: epipoles are at infinity, and epipolar lines are parallel. Adapted from O. Faugeras, *Three-Dimensional Computer Vision—A Geometric Viewpoint*, Cambridge, MA: MIT Press, 1993. Copyright 1993 MIT Press.

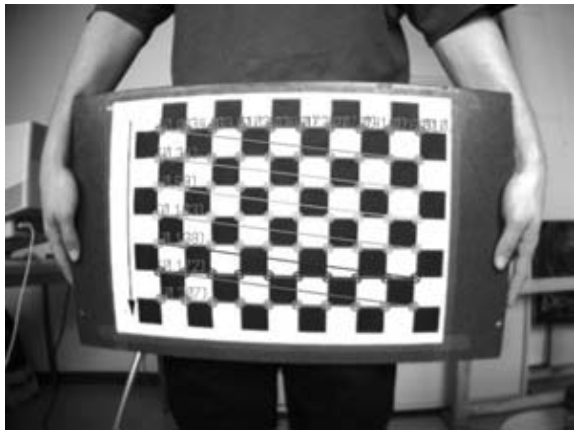
Stereo Analysis (7)

- **Camera calibration**
  - **A defined reference pattern is used to calibrate the cameras (see chessboard on the next slide).**
  - **The reference pattern must be visible for every camera in the system to be calibrated.**
  - **For a robust calibration, a variety of different positions of the calibration pattern must be recorded.**
  - **Corresponding points are identified in the set of images (preferably automatically). The positions and intrinsic parameters of the individual cameras can be calculated from these correspondences.**

Stereo Analysis (8)

- **Result of the calibration**
  - **The position of the cameras (also relative to each other) as well as individual ("intrinsic") parameters of the involved cameras (lens distortion, etc.) are determined.**





Stereo Analysis (9)

- **Usage of the calibration**
  - **The absolute position of the imaged scene points in 3D space can be calculated from the disparity vectors measured in the image and the calibration parameters.**
  - **Also, the calibration information allows the projection of the original images into epipolar geometry. This facilitates the computation of the disparity vectors (see before).**

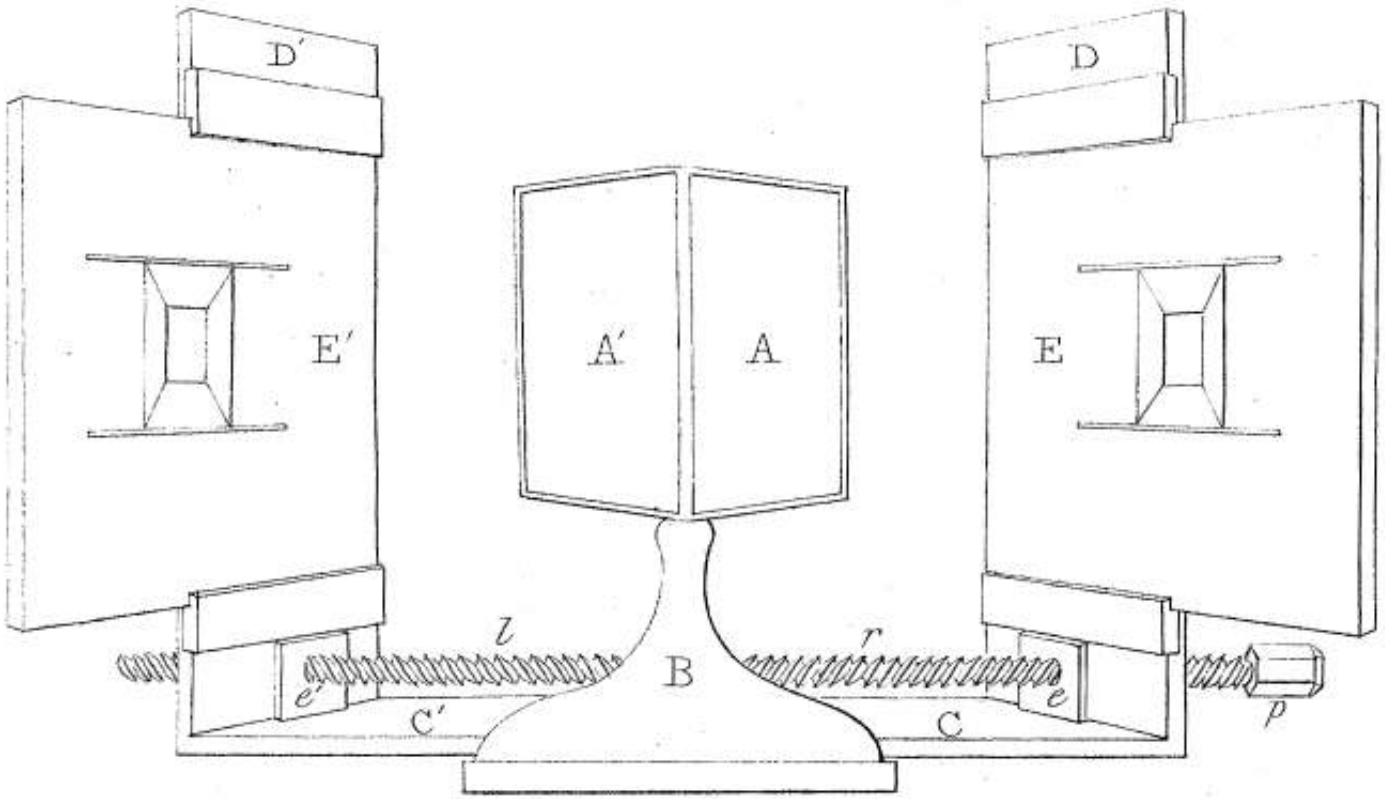
- **Stereo cameras**
  - **Lab equipment: synchronized individual cameras (below)**
  - **Mobile systems (right)**
  - **Cameras: Pointgrey, Fuji**

Stereo Display

- **Principle**
  - **The left and the right eye must perceive the left and the right image separately.**
- **Separation mechanisms with standard stereo displays (with glasses!)**
  - **Color filter (only for grey scale images – not suitable for display of color stereo images)**
  - **Polarization filter (polarization glasses)**
  - **The left and right image are displayed alternatively, with respective synchronization of the glasses (shutter glasses).**
- **Autostereoscopic displays (no glasses necessary!)**
  - **The screen displays 2 (or more) views simultaneously.**
  - **The user may see different views depending on the viewing direction -> multi-view displays.**

Stereo Displays – Historic Techniques (1)

- **Mirror Stereoscope (Charles Wheatstone, 19th century)**



Stereo Displays – Historic Techniques (2)

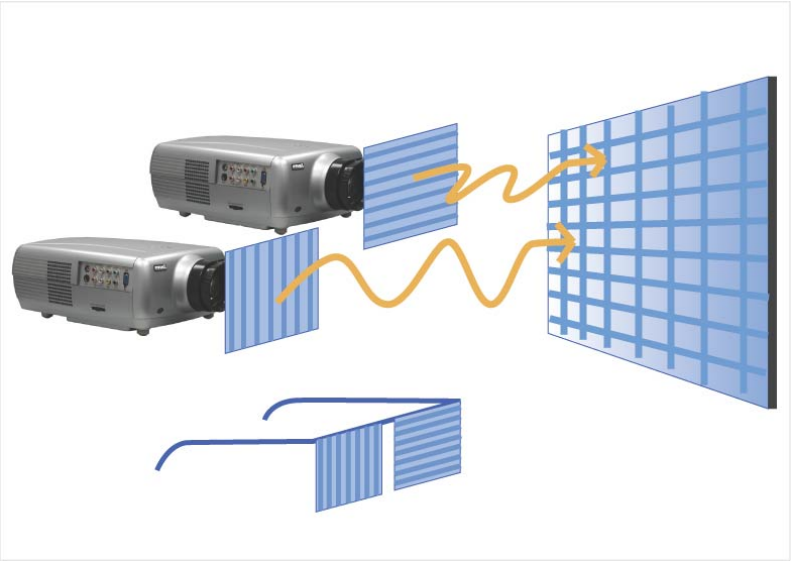
- **Lenticular (left) and portable (right) stereoscope**

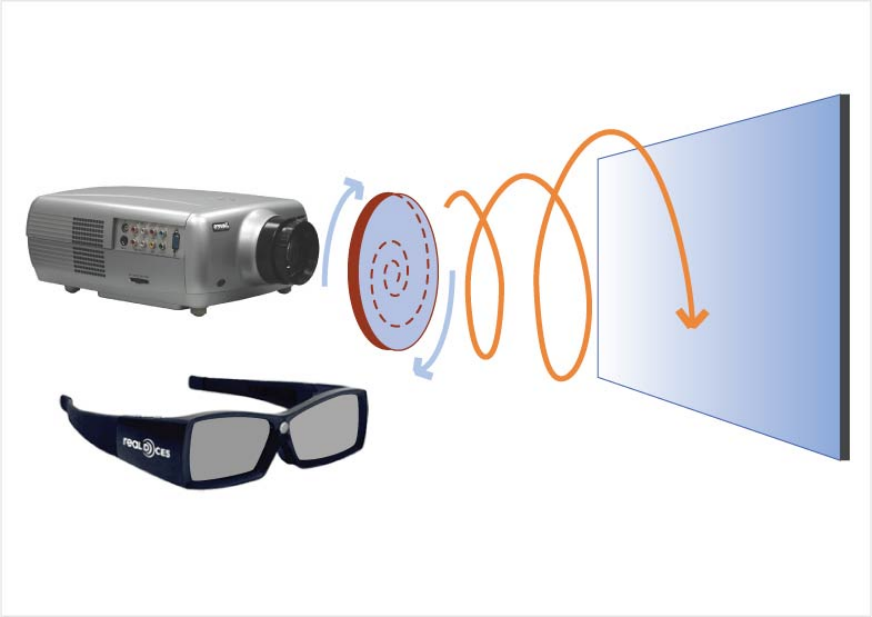
Stereo Displays – Current Techniques (1)

- **Displays with glasses based on different techniques**
  - **Red/green**
  - **Shutter glasses**
  - **Polarization glasses**

Stereo Displays – Current Techniques (2)

- **Polarization principle**
  - **Vertical/horizontal polarization (left)**
  - **Circular polarization (right)**





Source: http://www.visionnew.net/

Stereo Displays – Current Techniques (3)

- **Principle of shutter glasses**

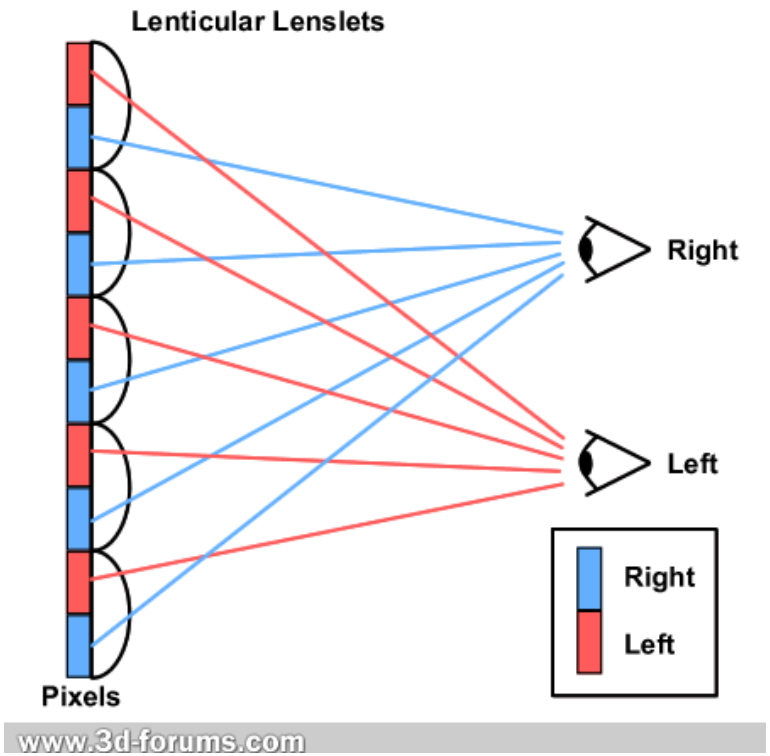
Source: http://www.visionnew.net/

Stereo Displays – Current Techniques (4)

- **Autostereoscopic displays**
  - **3D impression without additional aids.**
  - **examples: Philips WOWvx (left), Tridelity MV4200 (right)**

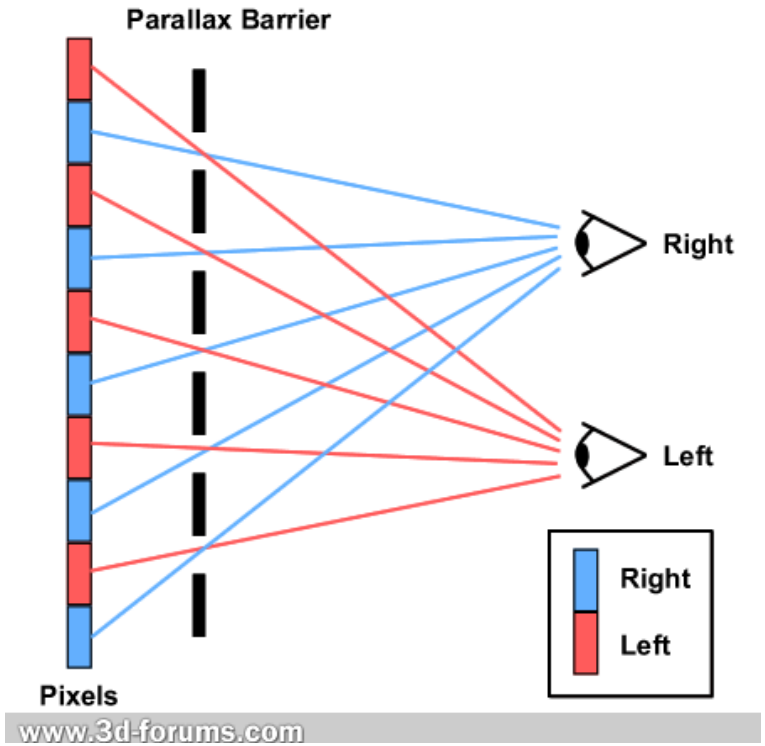
Principle of Autostereoscopic Displays (1)

- **Lenticular technique**
  - **The columns of the individual images are displayed in an alternative order.**
  - **The lenslets direct the columns that belong to the left image to the left eye, and the columns that belong to the right image to the right eye.**
  - **The use of multiple images allows to generate multiple views („multiview display“) with multiple „sweet spots“.**



Principle of Autostereoscopic Displays (2)

- **Parallax barrier (another technique)**



Format: Video (2D) + Depth

- **The calculated depth map can be used for**
  - **Multi-user autostereoscopic displays**
  - **Adaptation of the original film/video material to different types of displays and viewing environments.**
- **The depth map allows the creation of novel virtual views.**
  - **Format: 2D + Depth (see below).**

Temporal Smoothing of Depth Videos

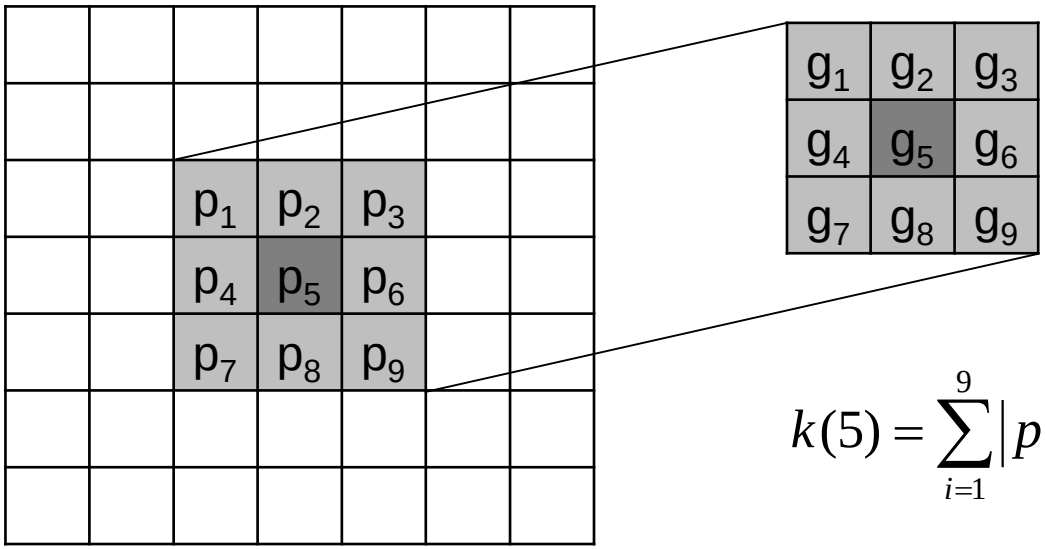
- **Automatic generation of depth videos from stereo videos (input)**

Temporal Smoothing of Depth Videos

- **Automatic generation of depth videos from stereo videos (output) – result before and after smoothing**

Stereo Research

- **Principle: window-based correlation**



match window

$$k(5) = \sum_{i=1}^{9} \left| p(i) - g(i) \right| \rightarrow \min$$

- **A major challenge of local stereo methods is to find a suitable size (and shape) of the match window.**
- **Contradictory requirements: large enough to capture sufficient texture, but small enough not to overlap object boundaries -> adaptive support weight approaches.**

- **„Local Stereo Matching Using Geodesic Support Weights“ (ICIP 2009)**
  - **The main idea is to use geodesic distances to compute the local window size, shape and support weights.**
  - **We include only those pixels in the support window that can be connected to the center pixel through a path along which the color does not vary significantly.**

➢ The geodesic distance $D(p,c)$ between a pixel *p* of the support window and the window's center *c* is the shortest path that connects *p* with *c* in the color volume:

$$D(p,c) = \min_{P \in P_{p,c}} d(P)$$

$P_{p,c}$ represents the set of all paths between pixels *p* and *c*.

- **„Local Stereo Matching Using Geodesic Support Weights“ (ICIP 2009)**

➢ A path P = <p_1,..,p_n> is a sequence of spatially neighboring points in 8-connectivity. The costs $d(P)$ are computed by:

$$d(P) = \sum_{i=2}^{i=n} d_C(p_i, p_{i-1})$$

where $d_C()$ determines the color difference between two pixels.

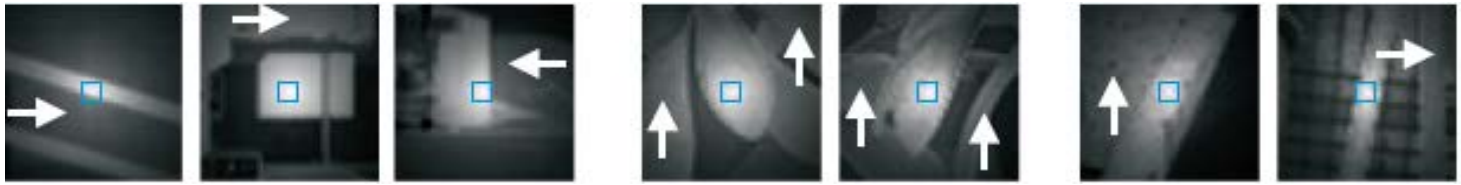
➢ Support weights are inversely proportional to the geodesic distance:

$$w(p,c) = \exp\left(-\frac{D(p,c)}{\gamma}\right)$$

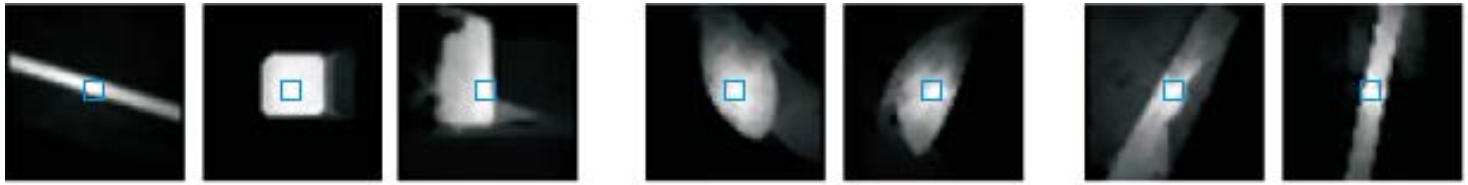
γ is a parameter defining the strength of segmentation (the only parameter).

- **Advantage of geodesic support weights**

Image crops



Color-based support weights (earlier approach)



Geodesic support weights (our approach)

- **Results**
  - **At the time of submission, our approach (GeoSup) was the best-performing local method in the Middlebury stereo evaluation benchmark (http://vision.middlebury.edu/stereo).**

Algorithm	Rank	Avg. Error [%]	Error non-occluded pixels [%]			
			Tsukuba	Venus	Teddy	Cones
GeoSup	**10**	**5.80**	**1.45**	**0.14**	**6.88**	**2.94**
AdaptDispCalib	13	6.10	1.19	0.23	7.80	3.62
DistinctSM	18	6.14	1.21	0.35	7.45	3.91
CostAggrOcc	22	6.20	1.38	0.44	6.80	3.60
SegmentSup	24	6.44	1.25	0.25	8.43	3.77
AdaptWeight	29	6.67	1.38	0.71	7.88	3.97
SSD+MF	70	15.7	5.23	3.74	16.5	10.6

- **„Near Real-time Stereo with Adaptive Support Weight Approaches“ (3DPVT 2010)**
  - **Modification of the *GeoSup* Algorithm to achieve significant reduction in computation time.**
  - **Approximation of the original algorithm allows efficient computation by application of a „sliding window“ technique.**

Image	Resolution	Time(sec.)	
		GeoSup	**Our Algorithm**
Tsukuba	384 x 288	10.874	**0.176**
Venus	434 x 383	15.97	**0.406**
Teddy	450 x 375	18.532	**0.704**
Cones	450 x 375	18.56	**0.688**

Performance on video stream: 10 frames/sec (320 x 240 resolution, disparity range of 26 pixels).

Recent Results

- **Hosni, A., C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, Fast Cost-Volume Filtering for Visual Correspondence and Beyond, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013**

Some results … →

Stereo Results

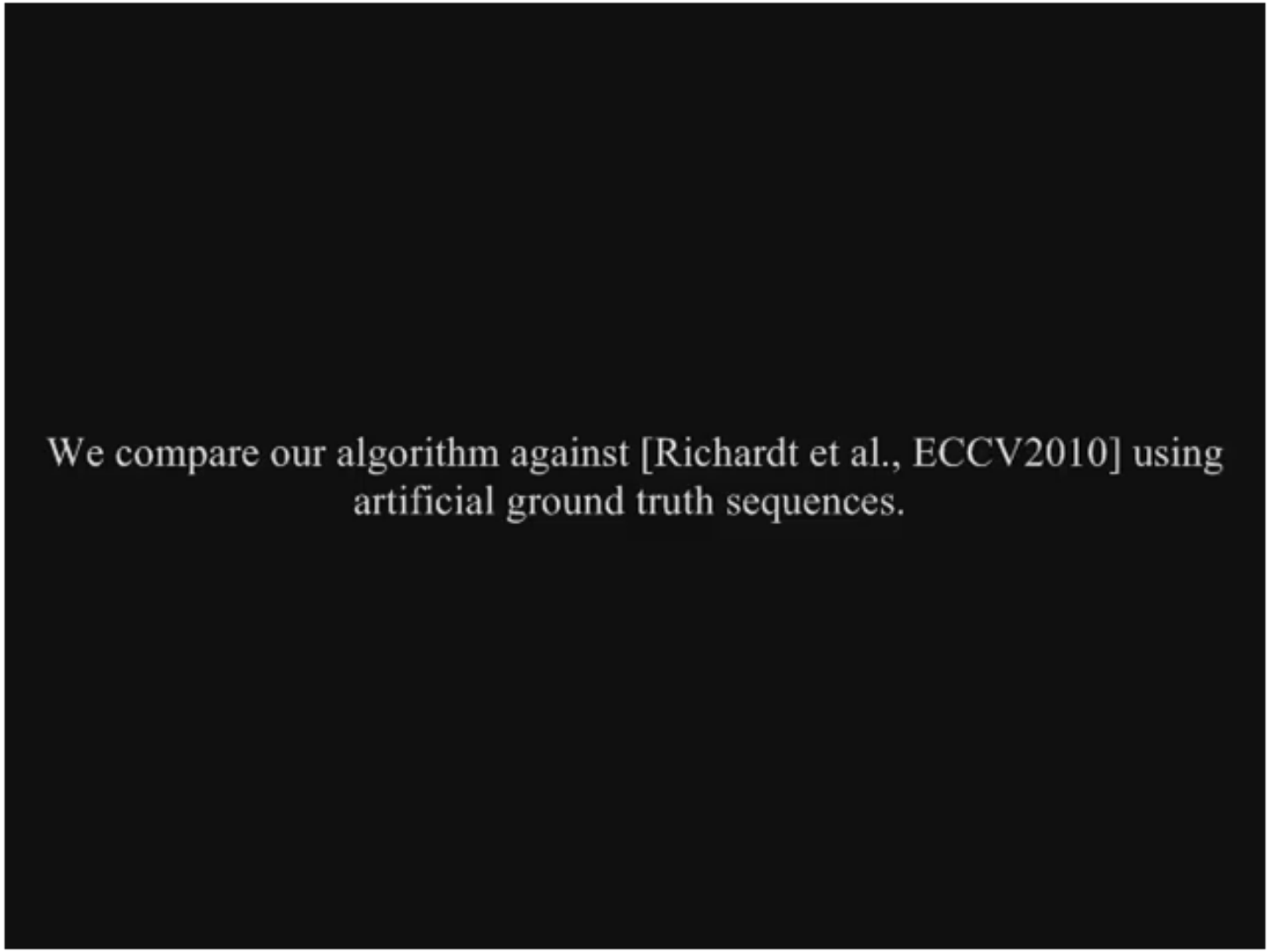
Left input frames | **Results of “Temporal DCBGrid”** | **Our spatio-temporal results**

Visual comparison to “Temporal DCBGrid”

1st row: Sample result from “Avatar” movie. Our temporal regularization could remove artifacts better than the temporal approach of “DCBGrid”.

2nd row: Sample result from “Skydiving” movie. Our spatio-temporal method can preserve some object outlines better than the temporal approach of “DCBGrid”. Red boxes mark major differences.

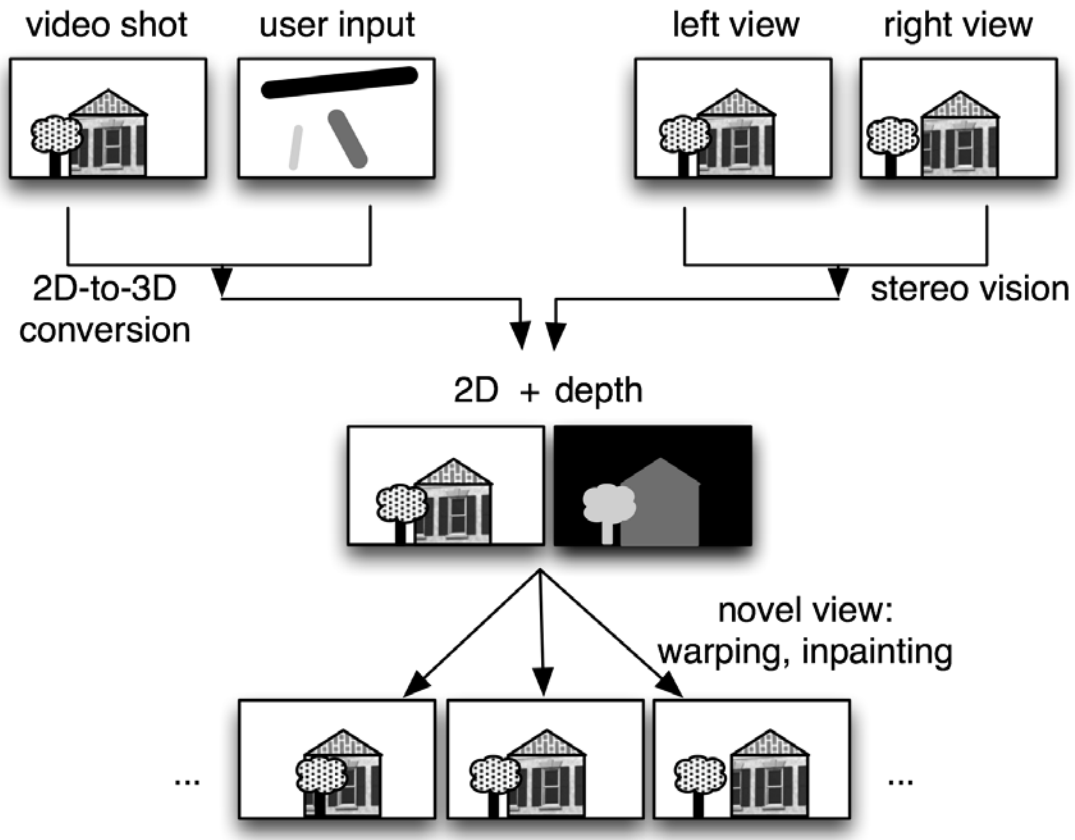
Stereo Results



3D Content Generation:

Novel Views & Inpainting

- **A processing chain for generating stereo content**



Stereo Format: 2D + Depth

- **Stereo content can be transmitted in different formats, including**
  - **2D + Depth**
    - **2D image (left)**
    - **Depth image (right) - brighter pixels are closer to the camera/viewer than darker pixels.**

Stereo Images – Novel View Generation

- **The proper handling of occlusions poses special challenges for novel view generation.**

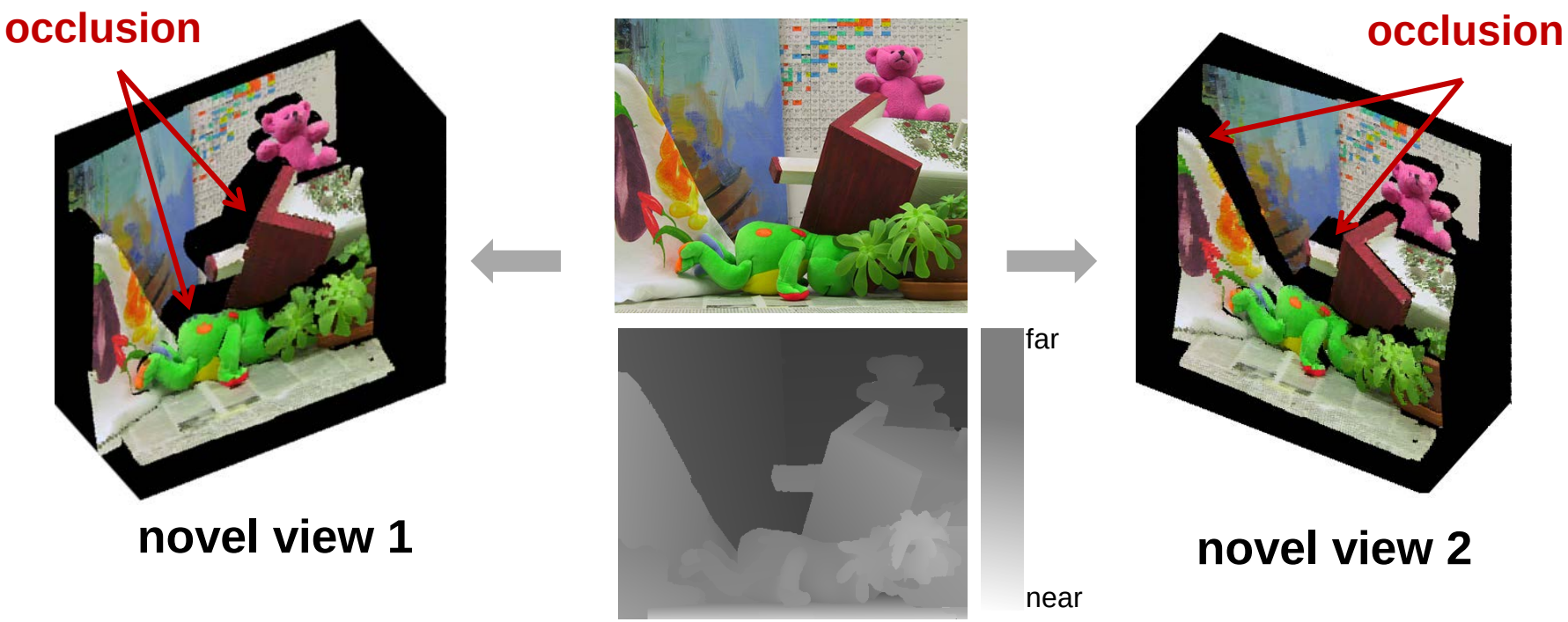
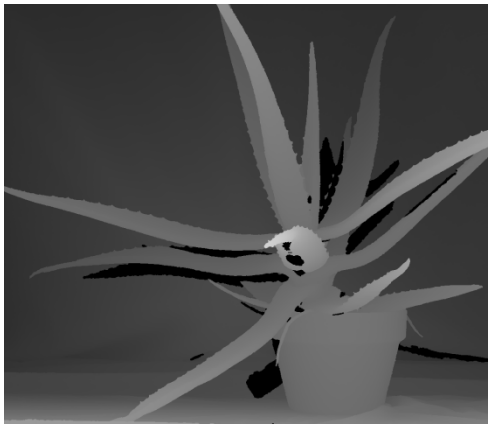


Image Inpainting

Left image



Depth map

Novel view (without inpainting)

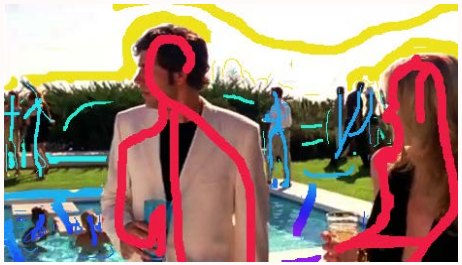
Right image

Inpainting result

2D-to-3D Conversion

- **Input: 2D Video & User Input**

First frame



Last frame

near

far

- **Output: Depth Video**

near

far

Copyright of original video: Warner Bros.

Image Matting

Image Matting and Compositing

- **Precise extraction of foreground object (matting).**
- **Object is pasted onto new background (compositing).**

- **The alpha matte includes semi-transparent pixels.**

Compositing Equation

- **Splitting up a natural image**

$$\mathbf{C} = \boxed{\alpha} \cdot \boxed{\mathbf{F}} + (1 - \alpha) \cdot \boxed{\mathbf{B}}$$

Matting = inverse process of compositing:

Determine: **F**(oreground), **B**(ackground), **α** (matte)
Given: **C**(omposite image)

Compositing Equation

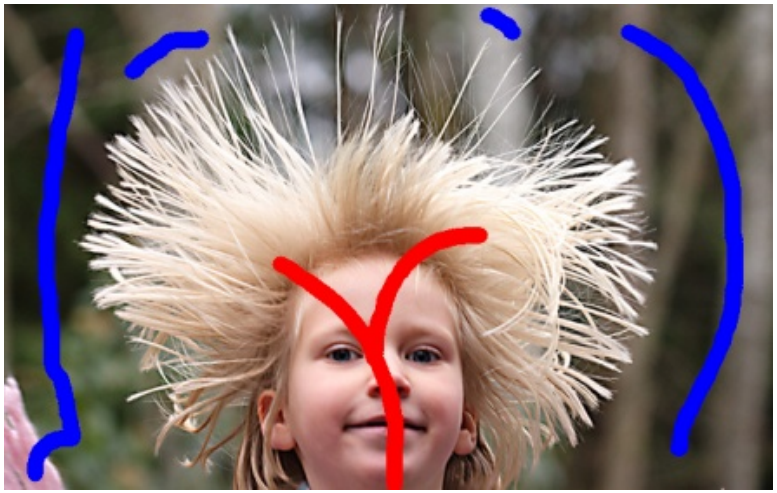
- **The problem is ill-posed**

$$C = \alpha \cdot F + (1 - \alpha) \cdot B$$

$$C^r = \alpha \cdot F^r + (1 - \alpha) \cdot B^r$$
$$C^g = \alpha \cdot F^g + (1 - \alpha) \cdot B^g$$
$$C^b = \alpha \cdot F^b + (1 - \alpha) \cdot B^b$$

7 unknown variables in only 3 equations -> no unique solution!

Image Matting Steps



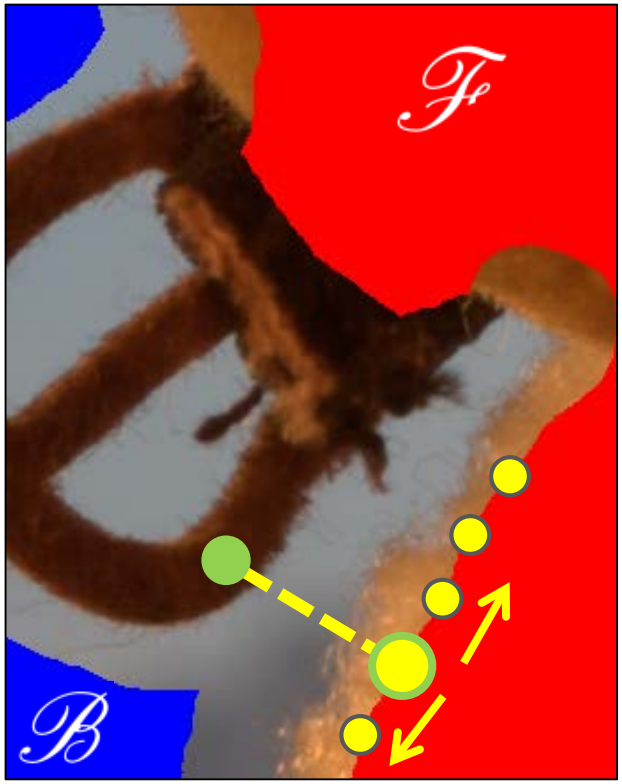
Scribble-based interface

Trimap

Alpha matte – result of our matting algorithm (BMVC08)

Image Matting Algorithm - Detail

- **The goal is to estimate the (mixed) colors in the unknown image regions from the colors in the known fore- and background regions.**



Foreground color (true)

Foreground color (assumed)

Image Matting Algorithm - Detail

- **The goal is to estimate the (mixed) colors in the unknown image regions from the colors in the known fore- and background regions.**

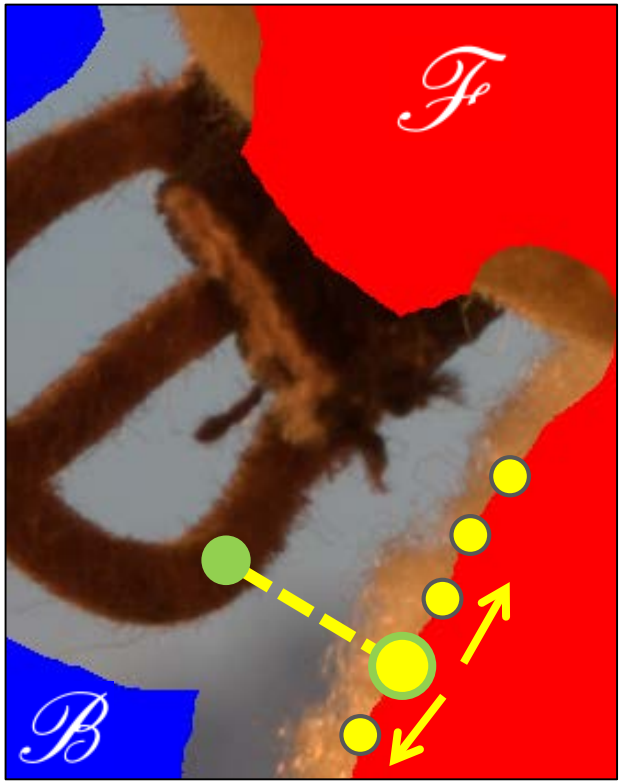
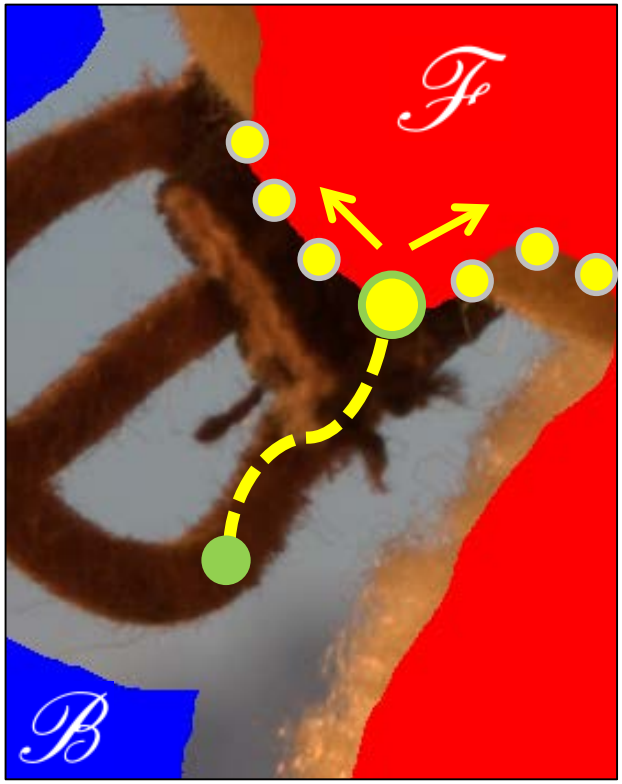


Image Matting Algorithm - Detail

- **We improve the results significantly by incorporating *geodesic distances* (exploiting the spatial connectivity of the foreground object).**

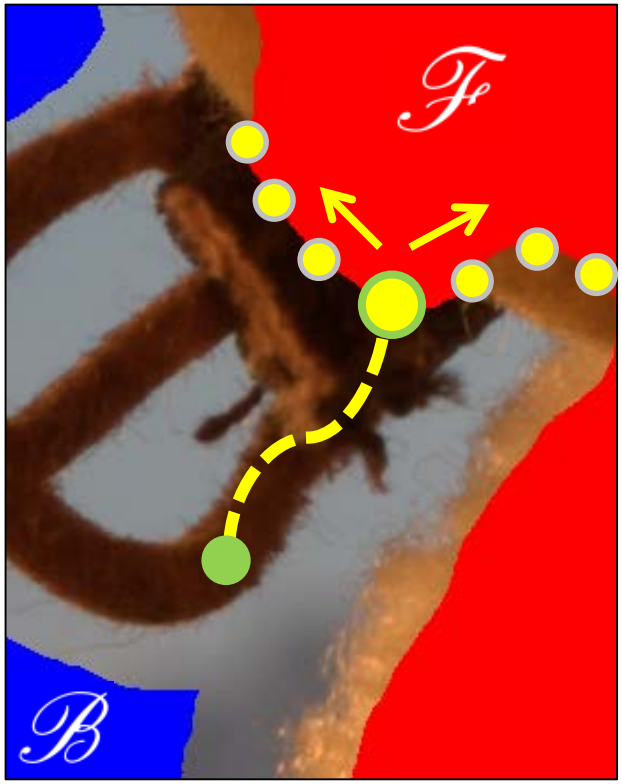


Foreground color (true)

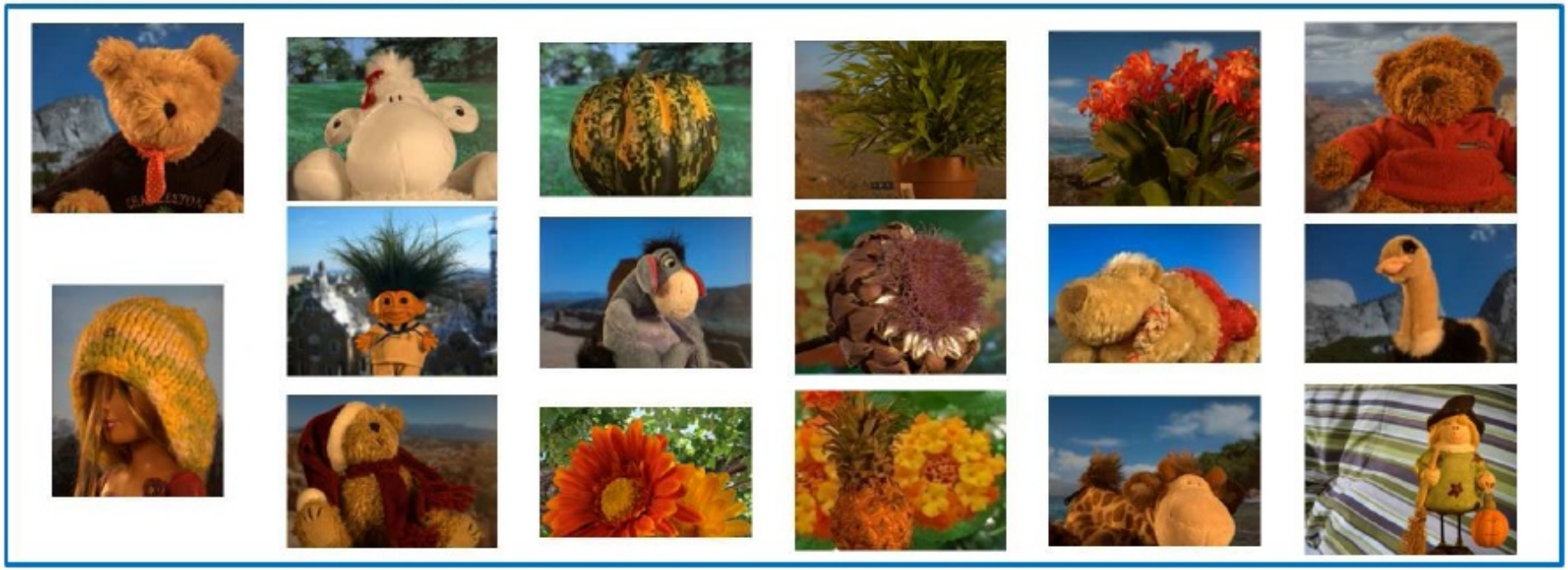
Foreground color (assumed)

Image Matting Algorithm - Detail

- **We improve the results significantly by incorporating *geodesic distances* (exploiting the spatial connectivity of the foreground object).**



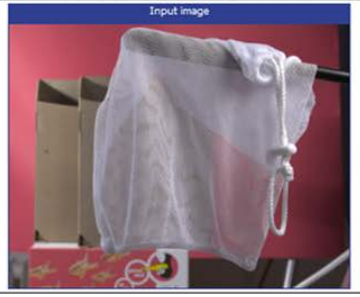
- **Ground truth dataset**

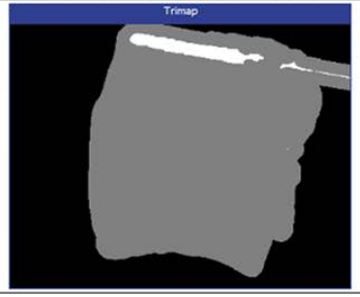


- **Online evaluation: www.alphamatting.com (CVPR 2009)**

Image matting evaluation results Competition: Low resolution High resolution
Error type: SAD MSE Gradient Connectivity

Sum of Absolute Differences	overall rank	avg. small rank	avg. large rank	avg. user rank	Troll (Strongly Transparent) Input small	Troll large	Troll user	Doll (Strongly Transparent) Input small	Doll large	Doll user	Donkey (Medium Transparent) Input small	Donkey large	Donkey user	Elephant (Medium Transparent) Input small	Elephant large	Elephant user	Plant (Little Transparent) Input small	Plant large	Plant user	Pineapple (Little Transparent) Input small	Pineapple large	Pineapple user	Plastic bag (Highly Transparent) Input small	Plastic bag large	Plastic bag user	Net (Highly Transparent) Input small	Net large	Net user
Closed-Form Matting	1.3	1.4	1.4	1.3	12.7 1	21.9 2	17.2 1	5.9 1	8.5 1	8.6 1	4.7 1	6 1	4.3 1	2.2 1	4.6 1	3.3 1	9.3 2	12.1 1	19.3 2	8.3 2	14.9 2	13.4 2	34.2 2	32.4 2	27.4 1	26.5 1	25.7 1	27.1 1
Robust Matting	1.9	1.6	2.1	1.9	17.3 2	28.4 3	21.1 3	10.1 2	16.9 3	11.4 2	4.8 2	6.5 2	5 2	2.8 2	7.3 3	4.4 2	7.3 1	14 2	18.1 1	6.8 1	14.6 1	10.6 1	22.7 1	26.1 1	32.1 2	34.4 2	37 2	38 2
Random Walk Matting	3.3	3.5	3	3.5	17.9 3	20.3 1	19.4 2	11.3 3	15.6 2	11.8 3	5.8 3	7 3	6.3 4	3.4 3	6.7 2	4.6 3	13.1 4	22.1 4	27.4 4	12.3 4	18 4	15.7 4	44.1 4	43.5 4	41 4	75.1 4	81.8 4	72.2 4
Easy Matting	4	4	4.1	4	23.9 4	32.6 4	30 4	17.1 4	21.8 4	19.4 5	6.3 4	7.5 4	5.8 3	4.7 4	10.5 4	5.6 4	12.1 3	15.7 3	22.9 3	11.2 3	17 3	14.8 3	49.5 5	49.6 5	46.2 5	77.8 5	108.6 6	109.2 5
Bayesian Matting	4.5	4.5	4.6	4.5	30.3 5	42.4 5	33.4 5	19.2 5	25.8 5	18.4 4	10.8 5	12.4 5	10.8 6	6.6 5	18.5 6	6.2 5	14.2 5	29.8 5	33.2 5	15.4 5	30.6 5	19.7 5	35.8 3	40.6 3	39.6 3	45.3 3	76.8 3	43.6 3
Poisson Matting	5.9	6	5.8	5.9	51.8 6	56.2 6	52 6	28.3 6	43.5 6	30.7 6	12.1 6	13.7 6	9.2 5	11.7 6	18.4 5	11.2 6	22.4 6	36.8 6	55.5 6	21.4 6	32.2 6	22.7 6	53.6 6	72.9 6	58.4 6	125.5 6	84.8 5	139.7 6

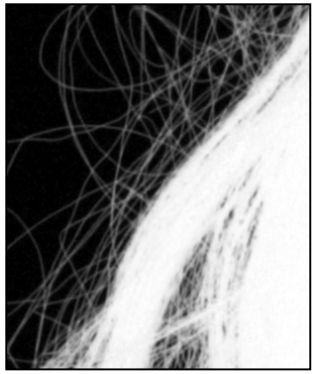




Joint work with Microsoft Research Cambridge

Input image

Ground truth



Zoom in