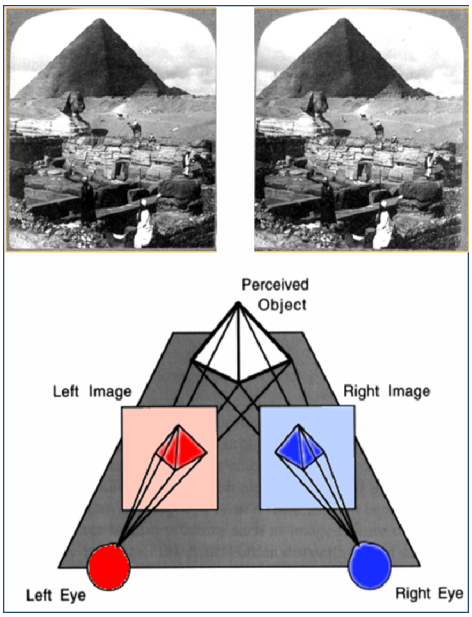
18. Stereo and Motion

Stereo Vision

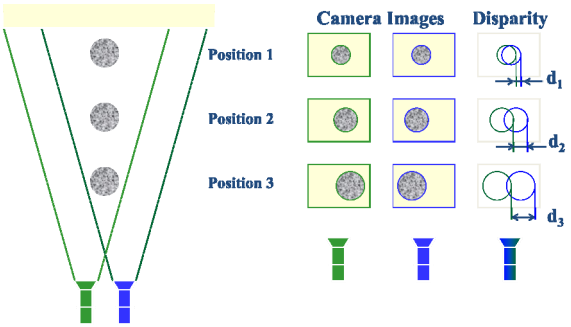


Stereo Vision ist ein zusammengesetztes Wort aus den beiden Begriffen ‚stereos', welches im Griechischem "räumlich, fest" bedeutet und ‚videre', welches aus dem Lateinischem stammt und für „sehen" steht. Dieses zusammengesetzte Wort soll somit das räumliche Sehen bezeichnen, welches auch als das menschliche bzw. binokulare Sehen bezeichnet wird. Dabei steht der Begriff für die Erstellung eines Tiefenbildes aus zwei Bildern einer Szene. Bilder kann man als zweidimensionale Projektionen des dreidimensionalen Raumes verstehen, bei denen eine Koordinate verloren geht. Dies ist der Informationsverlust der Tiefe. Der Mensch ist jedoch in der Lage, Tiefeninformationen aus seiner Umgebung zu ermitteln. Dieses Prinzip wird auch bei der Stereo Vision umgesetzt und auf zwei Kameras übertragen. Durch die bekannte Lage der Kameras zueinander und unter Verwendung der geometrischen Prinzipien der Triangulation und der Epipolargeometrie können korrespondierende Bildpunkte und die dazugehörigen Tiefenwerte rekonstruiert werden.
Durch den leichten Versatz eines Stereokamerasystems entstehen unterschiedliche Bilder. Dabei ist der Versatz von weiter entfernten Objekten geringer als der von näher befindlichen Objekten. Dieser Versatz wird auch als **Disparität** bezeichnet. Durch Zuweisung der unterschiedlichen Disparitäten zu jedem Bildpunkt wird eine Disparitätsmatrix erstellt. Aus dieser lässt sich für jeden Bildpunkt die Tiefe ableiten.

Stereoskopie

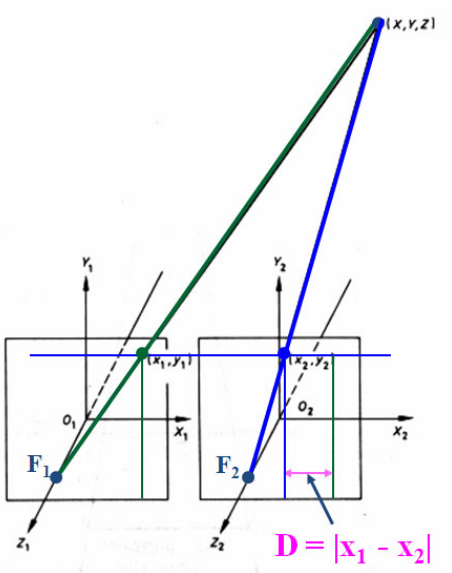
Die Stereoskopie (griechisch ‚skopeo' = betrachten) ist die Wiedergabe von Bildern mit einem räumlichen Eindruck von Tiefe, der physikalisch nicht vorhanden ist. Umgangssprachlich wird Stereoskopie fälschlich als "3D" bezeichnet, obwohl es sich nur um zweidimensionale Abbildungen handelt, die einen räumlichen Eindruck vermitteln. Das Prinzip des räumlichen Sehens war nachweislich schon dem griechischen Mathematiker *Euclid* im 3. Jh. v. Chr. bekannt. Im Laufe der Jahrhunderte haben sich viele Wissenschaftler, unter Ihnen auch *Leonardo da Vinci*, mit diesem Phänomen beschäftigt. Erst im 19. Jh. entdeckte *Charles Wheatstone* die Stereoskopie. Seinen bahnbrechenden Vortrag „Über einige bemerkenswerte und bisher nicht beobachtete Erscheinungen beim beidäugigen Sehen" hielt er vor der Royal Society im Jahre 1838 (also noch vor Erfindung der Fotografie). Er berechnete und zeichnete Bildpaare und konstruierte einen Apparat, das Stereoskop, um diese Bildpaare räumlich betrachten zu können. 1849 stellte *David Brewster* die erste Stereokamera vor. Mit diesem Apparat mit zwei Objektiven war es nun erstmals möglich, ein bewegtes Motiv aufzunehmen. Der Durchbruch kam 1851 zur Weltausstellung in London.
Wie funktionieren nun Stereoskope? Eine einfache Methode besteht darin, zwei stereoskopische Halbbilder nebeneinander abzubilden. Mit einer speziellen Blicktechnik (Parallelblick oder Kreuzblick) können sie dann ohne weitere Hilfsmittel als räumliches Bild wahrgenommen werden. Zur Vereinfachung gibt es jedoch spezielle Prismenbrillen. Bei den Anaglyphenbildern werden die beiden Halbbilder übereinander gedruckt, wobei die Halbbilder in Komplementärfarben eingefärbt werden. Zur Trennung der beiden Einzelbilder werden verschiedene Farbfilter in 3D-Brillen verwendet, ursprünglich Rot vor dem rechten Auge und Grün vor dem linken. Dies reduziert jedoch den Farbeindruck, daher arbeitet man heute mit polarisiertem Licht. Bei dieser Methode werden die Bilder des Stereobildpaars jeweils in entgegengesetzt polarisiertem Licht ausgestrahlt. Es befinden sich dazu jeweils entsprechend versetzte Polarisationsfilter vor den Projektionsobjektiven und in den 3D-Brillen der Betrachter. Alternativ werden Shutter-3D-Systeme verwendet, die bei der Wiedergabe von 3D-Bildern sogenannte Shutterbrillen einsetzen. Diese Spezialbrillen haben Gläser, die aus zwei Flüssigkristallflächen bestehen (je eine für das linke und rechte Auge), die elektronisch zwischen durchlässig und undurchlässig umgeschaltet werden können. Damit lässt sich wahlweise das linke oder das rechte Auge abdunkeln. Die Brillen werden mit dem Anzeigegerät synchronisiert, sodass abwechselnd das linke und rechte Bild gezeigt wird und die Brille das jeweils richtige Auge dafür auswählt.

Disparität



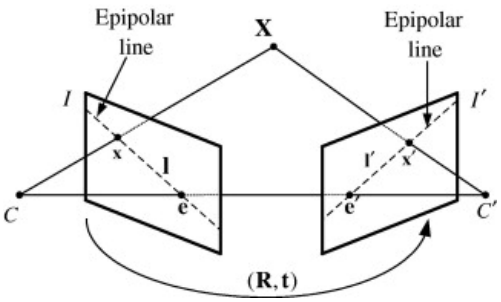
Ein einzelner Punkt wird in beiden Bildern des Stereosystems auf unterschiedliche Bildkoordinaten abgebildet. Die horizontale Differenz zwischen diesen nennt man Disparität. Wenn zum Beispiel zwei Kameras parallel angeordnet sind, wie in nebenstehender Abbildung gezeigt, und eine Kugel an drei verschiedenen Positionen (Abständen zum Kamerapaar) aufgenommen wird, so kann man die Disparität in der Überlagerung der Bilder beobachten. An Position 1 ist der Unterschied der Position der Kugel im überlagerten Bild d_1 klein. Wenn der Abstand zu den Kameras geringer wird (Position 2 und 3), werden die Bilder der Kugeln einerseits größer, da sie näher sind, andererseits wird auch die Differenz der Punkte in der überlagerten Bildern (d_2, d_3) größer. Daher wird die Disparität größer, je näher das Objekt zur Kamera ist, und ist 0 im Unendlichen. Die Disparität ist umgekehrt proportional zur Tiefe.

Normalfall



Das achsparallele Stereosystem zeichnet sich durch zwei Kameras aus, die nur horizontal verschoben und deren Koordinatensysteme nicht gegeneinander verdreht sind. Der Abstand zwischen den beiden optischen Zentren wird Basislänge *B* genannt und die Verbindungsgerade Basislinie (engl. *baseline*). Die Brennweite *f* legt den Abstand der beiden Brennpunkte zu ihren Bildebenen fest und wird für beide Kameras als identisch vorausgesetzt. Ein 3D-Punkt *X* wird somit über die beiden optischen Zentren in die Abbildungen *x* und *x'* projiziert. Da bei einem achsparallelen Stereosystem die Bildzeilen identisch sind, führt die unterschiedliche Perspektive der Kameras hinsichtlich des 3D-Punktes *X* nur zu einer horizontalen Disparität D in der Abbildung. Die Disparität wird im Allgemeinen in Bildkoordinaten berechnet, so dass die Einheit Pixel ist: $D = |x_1 - x_2|$.
Der Abstand *Z* eines Punktes *X* von der Kamera lässt sich aus den bekannten konstanten Kameraparametern *f, B* sowie der Disparität D berechnen: $Z = (B \cdot f)/D$. Damit stellt die Disparität ein Maß für die Raumtiefe des 3D-Punktes *X* dar und verhält sich umgekehrt proportional zu ihr. Für Punkte im Unendlichen muss daher die Disparität gegen Null konvergieren.

Epipolargeometrie



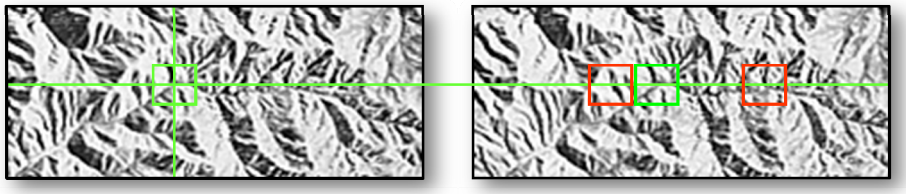
Kameraanordnungen mit zwei Kameras werden Stereosysteme genannt und lassen sich bezüglich ihrer räumlichen Anordnung in zwei grundsätzliche Klassen einteilen. Dies ist zum einen das bereits vorgestellte achsparallele Stereosystem (Normalfall). Zum anderen ist die konvergente Anordnung zu nennen, die sich i. A. durch eine Ausrichtung der optischen Achsen auf einen Konvergenzpunkt auszeichnet. Bei der allgemeinen Stereogeometrie, auch Epipolargeometrie genannt, sind die beiden Kameras nicht nur verschoben, sondern auch noch zueinander gedreht (siehe Abbildung rechts). Die Verbindungsgerade zwischen den beiden optischen Zentren der Kameras wird Basislinie genannt. Aufgrund der gedrehten Bildebenen schneidet diese Gerade beide Bildebenen. Diese Schnittpunkte werden **Epipole** (e und e´) genannt und ihre Lage in den Bildebenen ist nur durch die Anordnung der Kameras zueinander bestimmt. Die Epipole können auch als Projektion der optischen Zentren in die jeweils andere Bildebene aufgefasst werden. Ein 3D-Punkt *X* und die beiden Brennpunkte *C* und *C´* spannen eine Ebene auf, die als **Epipolarebene** bezeichnet wird. Die Epipolarebene schneidet nun die beiden Bildebenen in zwei Schnittgeraden, die **Epipolarlinien** (*l* und *l´*) genannt werden. Stellt man sich nun die beiden Sehstrahlen des 3D-Punktes in die beiden Kameras als Gummiband vor und bewegt man diesen Punkt innerhalb der Epipolarebene, so führt dies zu unterschiedlichen Abbildungen in den beiden Bildebenen, wobei diese jedoch immer auf den Epipolarlinien liegen. Lässt man den 3D-Punkt X entlang eines Sehstrahles z. B. in Richtung Kamera 1 laufen, so ergibt sich immer die gleiche Abbildung *x* in Kamera 1, während in Kamera 2 die Abbildung *x´* entlang der Epipolarlinie *l´* in Richtung des

Epipols *e´* wandert. Der Sehstrahl von jedem 3D-Punkt in eine Kamera liefert somit als Projektion in der anderen Kamera die entsprechende Epipolarlinie. Folglich muss auch für jeden Bildpunkt in der einen Kamera in der anderen Kamera eine Epipolarlinie existieren, auf der der Korrespondenzpunkt liegt.

Korrespondenzproblem

Das Korrespondenzproblem stellt das zentrale Problem der Stereo Vision dar. Es bezeichnet die Aufgabe, für jeden Bildpunkt im linken Bild jenen Punkt im rechten Bild zu finden, der denselben Objektpunkt abbildet. Das entsprechende Suchverfahren bezeichnet man als Korrespondenzanalyse oder auch als Stereo Matching.
Aufgabe der Epipolarrektifizierung ist es, Stereobildpaare anhand der Epipolargeometrie so zu transformieren, dass zusammengehörende Bildpunkte auf derselben horizontalen Linie liegen. Statt in zwei Dimensionen muss ein korrespondierender Punkt hier nur mehr entlang einer einzigen Scanline gesucht werden. Unter dieser Voraussetzung wird das Korrespondenzproblem wesentlich vereinfacht und die Korrespondenzanalyse somit beschleunigt. Gängige Stereo Matching Verfahren gehen in der Regel davon aus, dass die Bildpaare in rektifizierter Form vorhanden sind.

Regionenbasiertes Matching

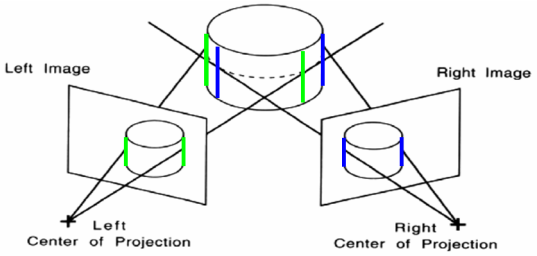


Regionenbasiertes Matching (engl. *Area-Based Matching* - ABM) vergleicht kleine Ausschnitte zwischen den rektifizierten Bildern. Dabei wird für jede Position eines Beobachtungsfenster im linken Bild das entsprechende Beobachtungsfenster im rechten Bild entlang der Epipolarlinie (Scanline) bewegt. Für jede Position wird geprüft, wie gut die Grauwerte mit den zu vergleichenden Grauwerten im linken Bild übereinstimmen, indem eine Ähnlichkeitsmessung durchgeführt wird. Die Wahl der Fenstergröße wirkt sich dabei auf die Robustheit und Geschwindigkeit des Matchings aus: Zu kleine Fenster beinhalten zu wenig Information für eine Korrespondenzzuordnung, zu große führen zu erhöhten Rechenzeiten. Zur Ähnlichkeitsmessung werden meistens die Summe der absoluten Differenzen (*Sum of Absolute Differences* - SAD), Summe der quadrierten Differenzen (*Sum of Squared Differences* – SSD) oder die Normalisierte Kreuzkorrelation (*Normalized Cross Correlation* - NCC) eingesetzt. Rechts sind SSD und SAD angegeben. Prinzipiell wird bei allen diesen Verfahren die Ähnlichkeit durch einen Vergleich von Pixeln innerhalb einer quadratischen Nachbarschaft zwischen dem linken und dem rechten Bild berechnet. Wenn das linke und rechte Bild exakt aufeinander passen, erhält man als Resultat ein Maximum (bei NCC) oder Minimum (bei SAD und SSD) in der Ähnlichkeitsfunktion. Mit Hilfe der Position des Maximums oder Minimums der Ähnlichkeitsfunktion wird die Position des korrespondierenden Punktes bestimmt (Disparität). Es ist möglich, dass kein eindeutiges Minimum oder Maximum gefunden werden kann. Das bedeutet, dass die Intensitäten in den beiden Bildern nicht zueinander passen, der korrespondierende Punkt ist nicht vorhanden. Dies kann zum Beispiel durch Verdeckungen passieren, wenn ein Punkt von einer Kamera aus sichtbar ist und von der anderen nicht. Der Hauptvorteil der regionenbasierten Matching-Methode gegenüber den merkmalbasierten Verfahren ist eine dichteres Tiefenbild. Hier werden die Tiefenwerte für alle Pixel direkt ausgerechnet, nicht nur für einige ausgewählte Merkmalspunkte. Nachteilig sind eine höhere Komplexität und ein entsprechend höherer Rechenaufwand.

$$\sum_{(i,j)\in W} \left(I_1(i,j) - I_2(x+i, y+j)\right)^2$$

$$\sum_{(i,j)\in W} |I_1(i,j) - I_2(x+i, y+j)|$$

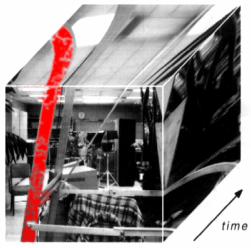
Merkmalbasiertes Matching



Ein Problem der regionenbasierten Matching-Methode sind homogene Bildbereiche. Sie beinhalten sehr wenig Information, werden aber in die Berechnung miteinbezogen und führen zu groben Fehlern. Eine Möglichkeit der Vermeidung ist die Verwendung einzelner, ausgewählter Pixel, die sich gut zueinander zuordnen lassen. Bei der merkmalbasierten Methode werden Merkmale aus jedem Bild individuell extrahiert, bevor sie verglichen werden. Lokale Merkmale können Ecken, Kanten oder andere Interest Points sein (siehe Kap. 17). Die Interest Points können durch lokale Operatoren extrahiert werden wie z.B. Moravec oder SIFT. Der eigentliche Korrespondenzvergleich kann schneller durchgeführt werden, da bei der Merkmalsextraktion eine wesentliche Datenreduktion stattfindet. Der Hauptnachteil von diesen Verfahren liegt aber darin, dass man hier zuverlässige Tiefeninformationen nur für diese ausgewählten Merkmale erhalten kann (keine dichten Tiefenbilder wie

bei ABM). Die Bereiche zwischen den Interest Points bleiben zuerst unberücksichtigt und müssen ggf. weiteren Verarbeitungen unterzogen werden (z.B. Interpolation).

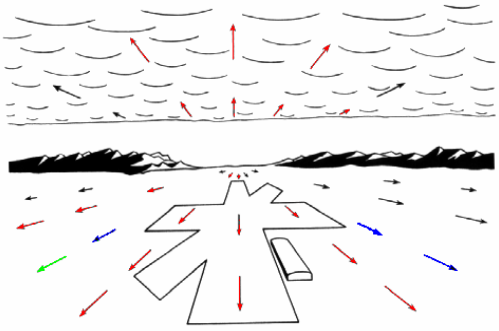
Structure-from-Motion



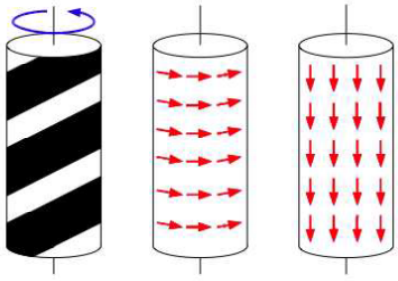
Structure from Motion (SfM) bezeichnet den Prozess der Gewinnung dreidimensionaler Information von Objekten oder einer ganzen Szene durch die Auswertung einer zeitlichen Folge von Bildern. Das zentrale Problem ist hier ebenfalls das Finden von Korrespondenzen zwischen den Bildern, im Gegensatz zur Stereo Vision ist die Kamerageometrie zw. aufeinander folgenden Bildern aber zunächst nicht bekannt. Die anhand der Korrespondenzen geschätzten Bewegungsfelder können dazu verwendet werden, bestimmte Kamerabewegungen zu bestimmen.

Bewegungsfeld

Fixiert ein Beobachter den Horizont, so "sind der Mond, die Sterne und das ganze obere Gesichtsfeld ohne Bewegung, die Welt und der Erdboden aber fließen in einem kontinuierlichen Strom vorbei. Projiziert man die Umgebung auf eine Abbildungsebene vor dem Betrachter, so ist die relative Geschwindigkeit eines Objektes umgekehrt proportional zu seiner Entfernung vom Betrachter" (*Helmholtz*, 1950). Je weiter ein Objekt vom Betrachter entfernt ist, desto geringer ist dessen Bewegung: Die Sterne und der Mond bewegen sich nicht, die Straße direkt neben dem Beobachter bewegt sich sehr schnell. Die Richtung des Bewegungsvektors eines Ortspunktes ist dabei abhängig von der Lage dieses Ortspunktes zum Betrachter. In einer Umgebung, in der sich die Objekte zueinander nicht bewegen, kann so die Eigenbewegung berechnet und eine relative Tiefenkarte der Umgebung erstellt werden. Das Bewegungsfeld besteht aus den einzelnen Bewegungsvektoren (siehe Abbildung). Bewegt sich die Kamera ohne zu rotieren, bewirkt dies ein „nach außen“ oder „nach innen“ Zeigen aller Vektoren zu einem einzigen Punkt, auch *Focus of Expansion* (FoE) oder *Focus of Contraction* (FoC) genannt. Dieser Punkt befindet sich dort, wo sich der Verschiebungsvektor der Kamera mit der Bildebene schneidet (= Epipol). Bei der Kameraverschiebung hängt die Größe der Bildbewegung eines Szenenpunktes umgekehrt proportional von der Entfernung eines Punktes zur Kamera und direkt proportional von dem Sinus des Winkels zwischen der Richtung, in der dieser Szenepunkt liegt, zu der Richtung, in welcher die Kamera verschoben wird, ab.



Bewegungsfeldbestimmung (Optischer Fluss)



Der optische Fluss (engl. *optical flow*) ist definiert als ein Vektorfeld, das die wahrgenommene Bewegung innerhalb einer Bildsequenz beschreibt. Diese Bewegung entspricht der Bewegung der einzelnen Pixel zwischen den Bildern einer solchen Sequenz. Die wahrgenommene Bewegung muss dabei nicht unbedingt mit der tatsächlichen Bewegung des betrachteten Objektes übereinstimmen. Nimmt man zum Beispiel einen Zylinder und zeichnet auf dessen Mantelfläche eine gewindeartige Linie, so entsteht beim Drehen dieses Zylinders entlang der Längsachse der optische Eindruck einer vertikalen Bewegung, wohingegen die tatsächliche Bewegung des Zylinders horizontal verläuft (Barber-Pole-Illusion). Es ist daher nicht immer möglich, vom optischen Fluss eindeutig auf die physikalische Bewegung eines Objektes zu schließen. Auch der bekannte Apertur-Effekt steht damit in Verbindung: Wenn sich ein Balken hinter einem Fenster bewegt, kann nur jener Bewegungsanteil bestimmt werden, der normal zu dem Balken verläuft.

Durch den Vergleich von 2D-Merkmalen kann man einen merkmalsbasierten optischen Fluss finden. Ähnlich wie im Stereo-Fall findet der merkmalbasierte Ansatz Merkmale und verfolgt diese während sie sich von Frame zu Frame bewegen. Dieser Vorgang erfolgt in zwei Schritten. Im ersten Schritt werden die Merkmale in zwei oder mehreren aufeinander folgenden Bildern extrahiert. Im zweiten Schritt werden diese Merkmale zwischen den Frames einander zugeordnet. Methoden des optischen Flusses, die auf der Erkennung von 2D-Merkmalen basieren, haben den Vorteil, dass der optische Fluss an jeder Position der Messung meist zuverlässig bestimmt werden kann. Der Nachteil ist, dass nur eine geringe Menge an Messungen zur Verfügung steht, im Gegensatz zu den Methoden, die Messungen an jeder Bildposition zur Verfügung stellen. Gradientenbasierte Verfahren benutzen räumliche und zeitliche Ableitungen, um den optischen Fluss für jeden Bildpunkt zu berechnen.