Reduced-Complexity Disparity Estimation
for Efficient Multiview Imagery Encoding

Dispariteitsschatting met verlaagde complexiteit
voor de efficiënte codering van beeldmateriaal met meerdere gezichtspunten

Aykut Avci

Promotoren: prof. dr. ir. H. De Smet, dr. ir. J. De Cock
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen: Computerwetenschappen

Vakgroep Elektronica en Informatiesystemen
Voorzitter: prof. dr. ir. J. Van Campenhout
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2012 - 2013

UNIVERSITEIT
GENT

Summary

3D display technology has witnessed a rapid development in the past decades. Currently 3D displays are being widely used in different application areas such as education, broadcasting, entertainment, surgery, video conferencing, etc. However, this technology owes its success to the other complementary technologies such as image acquisition, compression and transmission. The compression is the main topic of this dissertation.

In multiview displays, the realism of the reproduced 3D scene is dependent on the number of available views that the displays can show. These view images can be captured from different viewpoints of a scene by using a camera array. A smoother transition between views can be obtained by increasing the number of cameras located in the camera array. However, this comes at the price of an increased amount of image data which needs to be encoded (compressed) to store and transmit the data efficiently.

The captured multiview videos for different views can be encoded separately by a state-of-the-art video codec like H.264/AVC, which is called simulcast coding. Although coding each video individually is an easy option to solve the problem, it is not the most efficient approach since the inter-view correlations between views are overlooked. In order to improve the efficiency of the encoder, the inter-view correspondences can be taken into account. However, in this case, the computational load of the encoder becomes very high, especially if the camera array is two dimensional, i.e. if vertically spaced views are also being captured. Limitations on processing power, memory requirement and the desirability of features like instant access to specific view frame in the multiview video may render this scheme unusable.

A periodic (2D) camera array results in a strong geometrical relationship among the captured view images. This fact forms the core of the methods I propose in this dissertation, which consequently reduces the complexity of the multiview encoder significantly.

The P and B frames are well-known frame types from the H.264/AVC video coding standard and are instrumental in the motion estimation pro-

cess that exploits the similarity between consecutive frames in the time domain. A similar process called disparity estimation can be used to exploit the similarity between views. Since the B frame offers bi-directional prediction, it shows a better coding performance than the P frame but brings high computational load to the encoder. The complexity efficient versions of these frame types, named the D_P and D_B frame respectively, are introduced in this dissertation.

The disparity estimation process is the most complex and time consuming part of the encoder. The D_P frame achieves a significant complexity reduction by skipping the disparity estimation process for some of its blocks. The skipping process is entirely based on the fact that the disparity vectors of the blocks in a D_P frame can for most blocks be derived from the previously encoded blocks in another frame due to the strong geometrical relationship between views. A derived disparity vector needs to be checked for its fidelity since the derivation process can fail in some blocks due to occlusions, anisotropic illumination effects or insufficient texture information. To do this, the rate-distortion cost value of the derived disparity vector is compared with a threshold value. The blocks whose RD cost value of the derived disparity vector is lower than the threshold value will be exempted from the disparity estimation process, which results in a net complexity gain for the encoder.

The threshold value plays a crucial role on the determination of convenience of the derived disparity vectors. Since the D_P frame is a modified standard-conforming P frame, the encoder shows the same coding performance as the P frame if the threshold value is set equal to zero. It means that none of the derived disparity vectors are used and the disparity estimation will be performed for all the blocks in the frame. As the threshold value increases, the complexity of the encoder decreases, while the quality and bit-rate of the encoded images degrade. Therefore, the threshold value is a parameter to adjust the trade-off between the complexity and the rate-distortion performance of the encoder.

I introduced five alternative prediction schemes to encode 5x3 view images taken from a 2D camera array, three of which were constructed with the D_P frame. It has been noticed that the different locations of the frames have influence on the rate-distortion performance of the encoder. The prediction scheme in which the I frame is placed in the middle gives the best performance. After investigating the impact of a wide range of threshold values on the encoding performance and the complexity, it has been realized that the complexity of the encoder can substantially be reduced without compromising the quality and bitrate. The optimum values of the

threshold should be calculated depending on the quantization parameter (QP) and the imagery content since the threshold value represents a point in the rate-distortion cost value scale.

The rate-distortion and the complexity performance of the multiview encoder are improved by applying individual threshold values for every block in a D_P frame. I propose a method which automatically calculates the optimum threshold values for blocks during the encoding, where the maximum complexity gain is achieved while maintaining the rate-distortion performance. In order to calculate the optimum threshold value of a block, the rate-distortion cost value of a previously encoded block from which the disparity vector is derived is utilized.

Basically, the B frame has a better coding efficiency than the P frame. When the B frame is employed in a prediction scheme to encode multiview images, the complexity of the encoder is much higher than the prediction schemes with only P frames. In this dissertation, the complexity efficient version of the B frame, called D_B frame, is also presented. Different prediction schemes constructed with the D_P and the D_B frames are proposed. With the help of the D_B frame, the computational load of the multiview encoder is reduced considerably. Automatic threshold values of the blocks in D_B frames are automatically generated during the encoding.

The proposed frame types allow us to encode multiview images effectively with a lower computational load while keeping the same quality and bitrate. All proposed prediction schemes are applied to different multiview image sets containing various real world objects. For this purpose, all ideas in this dissertation have been implemented in the JSVM reference software.

Samenvatting

De technologie van driedimensionale (3D) beeldschermen heeft de afgelopen decennia een snelle ontwikkeling gekend. Momenteel worden 3D-beeldschermen gebruikt in diverse toepassingen zoals onderwijs, videoconferenties, chirurgie, ontspanning en televisie-uitzendingen. Het succes van deze technologie is echter alleen maar mogelijk in combinatie met andere, complementaire, technologieën zoals beeldregistratie, datacompressie en –transmissie. Datacompressie vormt het belangrijkste onderwerp van dit proefschrift.

In de zogenaamde 'multiview' beeldschermen wordt het realisme van de gereproduceerde 3D-scène bepaald door het aantal verschillende gezichtshoeken ('views') dat het beeldscherm tegelijk kan weergeven. De beeldinformatie die bij al deze gezichtshoeken hoort, kan bijvoorbeeld opgenomen worden door middel van een rooster van camera's die op dezelfde scène gericht zijn. De overgangen tussen de verschillende gezichtshoeken worden vloeiender naarmate de camera's in het rooster minder ver uit elkaar staan en het aantal camera's dus hoger is. Dit gaat echter ten koste van een toegenomen datavolume dat moet gecodeerd, opgeslagen en getransporteerd worden.

In principe kan men de opgenomen multiview videos voor elke gezichtshoek apart coderen met een state-of-the-art video codec zoals H.264/AVC. Deze aanpak wordt simulcast-codering genoemd. Hoewel dit een eenvoudige optie is om het probleem op te lossen, is dit duidelijk niet de meest efficiënte aanpak aangezien de correlatie tussen de beelden geregistreerd vanuit verschillende gezichtshoeken over het hoofd worden gezien. Om de efficiëntie van de encoder te vergroten kunnen deze 'interview' gelijkenissen in rekening gebracht worden. Evenwel wordt in dit geval de rekenlast waarmee men de encoder opzadelt zeer hoog, in het bijzonder wanneer het camerarooster tweedimensionaal is en er dus ook verticaal gespatieerde gezichtshoeken worden geregistreerd. Beperkingen op de rekenkracht en het beschikbare geheugen alsook gewenste eigenschappen zoals het snel kunnen springen naar een willekeurig tijdstip in

de multiview video zouden deze aanpak onbruikbaar kunnen maken.

In een periodiek tweedimensionaal rooster van camera's ontstaat een sterk geometrisch verband tussen de beelden die worden geregistreerd door de individuele camera's. Deze vaststelling vormt de basis van de methodes die ik in dit proefschrift voorstel om de complexiteit van de multiview encoder aanzienlijk te reduceren.

In de H.264/AVC videocoderingsstandaard zijn bewegingsvectordetectie en 'P' en 'B'-beelden welbekende begrippen waarmee men de gelijkenissen tussen elkaar in de tijd opvolgende beelden benut om zo tot een belangrijke datareductie te komen. Dezelfde technieken kunnen gebruikt worden om de inter-view gelijkenissen te benutten. In plaats van bewegingsvectoren spreekt men hier van dispariteitsvectoren. Aangezien B beelden bi-directioneel mogelijk maken, laat dit toe een betere coderingsperformantie te bereiken dan met louter P beelden, ten koste van een hogere berekeningslast voor de encoder. In dit proefschrift introduceer ik nieuwe versies van deze beeldtypes, D_P respectievelijk D_B, die deze berekeningslast ('complexiteit') beperken.

Het proces dat de dispariteit tussen de verschillende views zoekt is het meest complexe en tijdrovende onderdeel van de multiview encoder. De introductie van het D_P-beeld realiseert een significante reductie van de rekenlast doordat het toelaat om de dispariteitsvectordetectie over te slaan voor sommige delen (pixelblokken) in ieder beeld. Dit is mogelijk doordat de dispariteitsvectoren voor de meeste pixelblokken in een D_P-beeld kunnen afgeleid worden uit de blokken die reeds in een ander beeld werden gecodeerd, door gebruik te maken van het geometrisch verband tussen de verschillende views. Dergelijke afgeleide dispariteitsvectoren moeten gecontroleerd worden op hun geschiktheid omdat het geometrisch afleidingsproces voor sommige blokken niet goed werkt als gevolg van occlusies, anisotrope belichtingseffecten of onvoldoende informatie over de textuur. Teneinde dit te doen wordt de zogenaamde 'rate-distortion' (RD) kost van de afgeleide dispariteitsvector vergeleken met een drempelwaarde. De blokken waarvoor de RD-kost lager is dan deze drempelwaarde worden uitgesloten van het dispariteitsvectordetectieproces, hetgeen leidt tot een vermindering van de complexiteit van de encoder.

De geïntroduceerde drempelwaarde speelt dus een cruciale rol bij het bepalen van de geschiktheid van de afgeleide dispariteitsvectoren. Aangezien een D_P-beeld een gemodificeerd standaard P-beeld is, vertoont de encoder net dezelfde performantie als de standaardencoder indien de drempelwaarde gelijk aan nul wordt gekozen. In dat geval wordt immers geen enkele van de geometrisch afgeleide dispariteitsvectoren gebruikt,

maar wordt het reguliere dispariteitsvectordetectieproces uitgevoerd voor alle pixelblokken. Naarmate de drempelwaarde stijgt, neemt de complexiteit van de encoder af, maar worden de kwaliteit en de bitrate slechter. Bijgevolg is de drempelwaarde een parameter die toelaat de afweging tussen complexiteit en rate-distortion van de encoder bij te sturen.

Ik heb vijf alternatieve voorspellingsschema's geïntroduceerd om een multiview registratie van 5x3 views, genomen door een tweedimensionaal camerarooster, te encoderen. Drie van deze schema's maken gebruik van D_P-beelden. Daarbij is gebleken dat de locatie van de verschillende beeldtypes in de schema's een invloed hebben op de RD performantie van de encoder. Het voorspellingsschema waarbij het I-beeld in het midden wordt geplaatst vertoont de beste performantie. Uit een grondige studie van de impact van een groot aantal drempelwaarden op de performantie en complexiteit van de encoder is gebleken dat de complexiteit van de encoder gevoelig kan gereduceerd worden zonder dat de kwaliteit en bitrate merkbaar worden aangetast. De optimum waarde voor de drempel hangt daarbij af van de kwantiseringsparameter (QP) alsook van de beeldinhoud vermits de drempelwaarde een punt vertegenwoordigt in de RD-kost schaal.

De rate-distortion en de complexiteit van de multiview encoder kunnen verder verbeterd worden door aangepaste drempelwaarden toe te passen op ieder pixelblok in een D_P-beeld. Ik heb een methode voorgesteld waarmee de optimale drempelwaarde automatisch kan berekend worden tijdens het coderingsproces. Daarmee kan de maximale winst in complexiteit behaald worden zonder dat de RD-performantie afneemt. Om de optimale drempelwaarde te berekenen wordt gebruik gemaakt van de RD kost waarde van het voordien geëncodeerde pixelblok waaruit ook de dispariteitsvector werd afgeleid.

Een B-beeld leidt tot een grotere coderingsefficiëntie. Introductie van het B-beeld in het voorspellingsschema voor multiview beelden leidt echter tot een gevoelige toename van de complexiteit van de encoder. In dit proefschrift wordt ook een variant geïntroduceerd op het B-beeld, D_B-beeld genoemd, waarbij de geometrische afleiding van dispariteitsvectoren leidt tot een reductie van de rekenlast. Verscheidene voorspellingsschema's die gebruik maken van deze D_P en D_B-beelden worden voorgesteld. Door de introductie van het D_B-beeld wordt de complexiteit van de multiview encoder aanzienlijk verminderd. Ook de drempelwaarden voor de D_B-blokken worden automatisch gegenereerd tijdens het coderingsproces.

De voorgestelde nieuwe beeldtypes laten ons toe om multiview beelden te coderen met een lagere rekenlast terwijl de bitrate en beeldkwaliteit

gelijk blijven. Alle voorgestelde voorspellingsschema's zijn toegepast op verscheidene multiview opnames van reële objecten. Daartoe werden alle in dit proefschrift geïntroduceerde concepten geïmplementeerd in de JSVM referentiesoftware.