Fortschritt-Berichte VDI

VDI

Reihe 10

Informatik/
Kommunikation

Dipl.-Ing. Matthias Reso,
San Francisco

Nr. 861

Temporally Consistent Superpixels

tnt

Institut für Informationsverarbeitung
www.tnt.uni-hannover.de

Temporally Consistent Superpixels

Von der Fakultät für Elektrotechnik und Informatik

der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des akademischen Grades

Doktor-Ingenieur

(abgekürzt: Dr.-Ing.)

genehmigte

Dissertation

von

Dipl.-Ing. Matthias Reso

geboren am 9. Januar 1985 in Hannover.

2017

Referent:	Prof. Dr.-Ing. J. Ostermann
Korreferent:	Prof. Dr.-Ing. R. Stiefelhagen
Vorsitzender:	Prof. Dr.-Ing. B. Rosenhahn
Tag der Promotion:	09.08.2017

Fortschritt-Berichte VDI

Reihe 10

Informatik/
Kommunikation

Dipl.-Ing. Matthias Reso,
San Francisco

Nr. 861

Temporally Consistent Superpixels

Institut für Informationsverarbeitung
www.tnt.uni-hannover.de

Reso, Matthias
Temporally Consistent Superpixels
Fortschr.-Ber. VDI Reihe 10 Nr. 861. Düsseldorf: VDI Verlag 2018.
124 Seiten, 54 Bilder, 3 Tabellen.
ISBN 978-3-18-386110-1, ISSN 0178-9627,
€ 48,00/VDI-Mitgliederpreis € 43,20.

Keywords: Superpixels – Temporal Consistency – Supervoxels – Oversegmentation – Occlusion – Interactive Video Segmentation

This thesis addresses the field of early stage video preprocessing in order to improve and accelerate subsequent processing steps like semantic video segmentation or video-based object tracking. A framework is proposed to segment video streams into temporally consistent superpixels in order to create a representation of the video with far less image primitives than the voxel-grid. The proposed energy-minimization-based approach utilizes a novel hybrid clustering strategy for a multidimensional feature space. Techniques are presented to ensure the consistency of the superpixel flow with the image movement while considering visual occlusion and disocclusion effects. The effectiveness of the proposed method is shown by a comparison to state-of-the-art spatio-temporal oversegmentation algorithms using established benchmark metrics. Additionally, its effectiveness is further demonstrated by showing its application on the real-world scenario of interactive video segmentation.

Bibliographische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter www.dnb.de abrufbar.

Bibliographic information published by the Deutsche Bibliothek
(German National Library)
The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at www.dnb.de.

Als Manuskript gedruckt. Printed in Germany.

ISSN 0178-9627

ISBN 978-3-18-386110-1

Acknowledgement

The present thesis was written during my time at the Institut für Informationsverarbeitung (TNT) of the Gottfried Wilhelm Leibniz Universität Hannover.

I owe my deepest gratitude to Prof. Dr.-Ing. Jörn Ostermann for giving me the opportunity to work in his group. His excellent supervision, encouragement and patience during my time at the institute helped to make this thesis a success.

I also would like to thank Prof. Dr.-Ing. Bodo Rosenhahn for his guidance and helpful comments on my work done at the institute as well as for being the chair of the defense committee.

Thanks is also due to Prof. Dr.-Ing. Rainer Stiefelhagen for being the second examiner of my thesis.

Special thanks goes to Jörn Jachalsky from whom I learned a lot during our fruitful discussions and our collaboration on many research papers.

I am very grateful to my dear colleagues and friends at the TNT like Arne Ehlers, Benjamin Jungbluth, Florian Baumann, Björn Scheuermann, Hanno Ackermann, Kai Cordes and Stephan Preihs who contributed their support and made my stay at the institute to one of the greatest times in my life.

I also thank Matthias Schuh, Martin Pahl, Thomas Wehberg and the secretaries, Mrs. Brodersen, Mrs. Bank and Mrs. Jaspers-Göring for their commitment and assistance.

A very special thanks goes to my family and particularly my parents Andrea and Henning Reso for enabling my studies and always supporting me in my endeavours. And to my beloved wife Julia who stood by my side even in the darkest hours.

This work has been partially funded by the ERC within the starting grant Dynamic MinVIP. The author gratefully acknowledges the support.

Dedicated to my family

Contents

Abbreviations

2D	two-dimensional
3D	three-dimensional
5D	five-dimensional
6D	six-dimensional
CCD	charge coupled device
CMOS	complementary metal oxide semiconductor
CPU	central processing unit
CRF	conditional random field
FIFO	first-in, first-out
GPU	graphics processing unit
HD	high-definition
KLT	Kanade-Lucas-Tomasi feature tracker
MAP	maximum a posteriori
MRF	Markov random field
OVS	online video seeds
SA	3D segmentation accuracy
sGBH	streaming hierarchical video segmentation
SLIC	simple linear iterative clustering
SOR	successive over-relaxation
TCS	temporally consistent superpixels
TEX	temporal extent
TSP	temporal superpixels
UE	3D undersegmentation error

Symbols and Notation

α	spatial weighting term with an interval of [0,1]
$\widetilde{\alpha}$	alternative spatial weight with an interval of $[0,\infty)$
B	number of labels i.e. cluster, superpixels, segments
$\boldsymbol{b}$	labeling of a set of pixels
b	label variable
$\hat{\boldsymbol{b}}$	optimal labeling
$\hat{b}$	optimal label
β	weighting term including the feature variance
$\bar{C}$	superpixel compactness
$\boldsymbol{c}$	set of pixel pairs
$D_i(b)$	unary term for assigning label b to pixel i
$E[\cdot]$	expectation value
$E_{\text{asm}}(i,b)$	assignment energy function
$E_{\text{clg}}(\underline{\vec{\omega}})$	combined local-global optical flow energy function
$E_{\text{clg,nq}}(\underline{\vec{\omega}})$	non-quadratic combined local-global energy function
$E_{\text{crf}}(\boldsymbol{b})$	energy function of a conditional random field
$E_{\text{hs}}(\underline{\vec{\omega}})$	Horn-Schunck optical flow energy functional
$E_{\text{ivs}}(\boldsymbol{b})$	energy function for interactive video segmentation
$E_{\text{kms}}(\boldsymbol{b})$	total energy of general k-means
$E_{\text{lk}}(\underline{\vec{\omega}})$	Lucas-Kanade optical flow energy function
$E_{\text{mrf}}(\boldsymbol{b})$	energy function of a Markov random field
$E_{\text{tcs}}(\boldsymbol{b})$	total energy of a temporally consistent superpixel labeling
$E^{\text{tcs}}_{\text{col}}(i,b)$	color part of the TCS assignment energy
$E^{\text{tcs}}_{\text{pos}}(i,t,b)$	spatial part of the TCS assignment energy
ϵ	misclassification rate
EV	explained variation
F	number of future frames
$G(\vec{y})$	window around a vector $\vec{y}$
$\boldsymbol{g}$	set of pixels contained in a ground truth segment
$\Gamma(\cdot,\cdot)$	indicator function
γ	scale and growth parameter of the penalty function Ψ
$I(x,y,t)$	image functional
i,j,k	pixel indices
$K(\cdot)$	kernel function

$\boldsymbol{k}$	set of possible labels
L_b	perimeter of superpixel b
l	triangle index
$\lambda_{\mathcal{F}}$	smoothness regularization constant for the optical flow field
$\lambda_{\mathcal{G}}$	label consistency weight
$\vec{\mu}_b^{\mathcal{C}}$	superpixel color center
$\vec{\mu}_b^{\mathcal{S}}$	superpixel spatial center
$\boldsymbol{\mu}$	set of superpixel centers
$\vec{\mu}_{b,t}^{\mathcal{S}}$	spatial center of a temporally consistent superpixel
$\vec{\mu}_{b,t}^{\mathcal{F}}$	weighted average flow of superpixel b in frame t
N	number of pixels in an image
$\boldsymbol{n}$	set of pixels or sites
$\boldsymbol{n}_b$	set of pixels in superpixel with label b
$\boldsymbol{n}_{b,t}$	set of pixel in superpixel with label b at time t
$\boldsymbol{n}_{b,t}^{\mathcal{H}}$	hidden fraction of superpixel b in frame t
$\boldsymbol{o}$	set of observations
o	observation
$\vec{\omega}$	optical flow vector
$\underline{\vec{\omega}}$	optical flow vector in homogeneous coordinates
P	number of past frames
$P(\cdot)$	probability function
$\Phi(\cdot)$	potential function
$\Psi_i(\cdot)$	non-quadratic penalty function
Q_b	iso-perimetric quotient of superpixel b
$\bar{S}$	average superpixel edge length
SA	3D segmentation accuracy
$\sigma_{\mathcal{F}}^2$	variance of the average flow weighting function
T	total number of frames
t	time index
τ	local frame index
UE	3D undersegmentation error
$V_{i,j}(b,d)$	pairwise energy for assigning labels b, d to the pixels i, j
VoA	variance of area
W	length of the sliding window
$\vec{x}_i^{\mathcal{C}}$	color vector of a pixel
$\vec{x}_i^{\mathcal{S}}$	spatial vector of a pixel
Z	number of iterations
z	iteration index
$\bar{\zeta}_m$	averaged benchmark result m

Abstract

A wide variety of computer vision applications rely on superpixel or supervoxel algorithms as a preprocessing step. This underlines the overall importance that these approaches have gained in recent years. However, most methods show a lack of temporal consistency or fail in producing temporally stable segmentations.

In this regard, this thesis presents a highly competitive approach for temporally consistent superpixels for video content. The approach is based on energy-minimizing clustering utilizing a novel hybrid clustering strategy for a multidimensional feature space. By working in a joint global color space, but keeping the positions of the superpixels localized to the frame level, the framework allows for arbitrary large displacements of the superpixels along the image plane over time. By applying a contour-based optimization the spatial coherency of the pixels of each superpixel is ensured while obeying the optimization target at all times. A sliding window technique enables the approach to process videos of arbitrary length in a streaming fashion. To propagate the superpixel segmentation while shifting the sliding window over the video volume, this thesis proposed two novel propagation methods. While the first approach is trimmed for efficiency and utilizes sparse optical flow vectors in combination with a Delaunay triangulation, the second approach individually propagates the shape of each superpixel. The individual propagation enables the detection of occluded and disoccluded image regions. In order to provide equally sized superpixels, this thesis further proposes a novel approach to handle structural changes in the video volume by utilizing the collected dis-/occlusion information.

For a thorough evaluation, the proposed approach is compared to state-of-the-art spatio-temporal oversegmentation algorithms using established benchmark metrics. The benchmark results show that the proposed framework produces the lowest spatio-temporal segmentation error of all approaches. Thereby, creating longer temporal superpixel trajectories than approaches with a comparable segmentation error. This shows that the proposed method extracts the temporal connections of the image regions inherent in the video volume to a higher extent than previous methods. Simultaneously, its run time scales better than approaches of comparable quality, as it only depends linearly on the number of pixels as well as the number of superpixels.

The effectiveness of the proposed method is further evaluated by showing its application to the task of interactive video segmentation using graph cut techniques. When compared to a voxel level processing of the video material the proposed oversegmentation method decreases the initial segmentation error by over 47 %. Additionally, its application reduces the average run time of the performed graph cut from 31 minutes to under 7 ms per sequence.

Keywords: superpixels, temporal consistency, supervoxels, oversegmentation, occlusion, interactive video segmentation

Kurzfassung

Eine große Anzahl Computer Vision Applikationen basiert auf der Verwendung von Superpixeln oder Supervoxeln als Vorverarbeitungsschritt. Dies unterstreicht die Wichtigkeit, welche diese Ansätze in den letzten Jahren erlangt haben. Viele dieser Methoden erzeugen allerdings zeitlich inkonsistente oder instabile Segmentierungen.

Ziel dieser Arbeit ist die Beschreibung eines Systems zur Erzeugung zeitkonsistenter Superpixelsegmentierungen für Videos. Der Ansatz basiert auf einem energieminimierenden Verfahren zur Cluster Analyse und nutzt einen neuen, hybriden Ansatz für den multidimensionalen Merkmalsraum. Dabei kommt ein globaler, zusammengefasster Farbraum zur Anwendung, während die räumlichen Positionen der Superpixel auf den Einzelbildern betrachtet werden. Somit lassen sich beliebig große Bewegungen von Bildregionen entlang der Bildebene durch die Superpixel abbilden. Indem eine konturbasierte Optimierung Anwendung findet, wird der räumliche Zusammenhalt der Pixel jedes Superpixels garantiert, während das Optimierungskriterium zu jedem Zeitpunkt Berücksichtigung findet. Durch den Einsatz einer Fensterungstechnik lassen sich dabei beliebig lange Videosequenzen sukzessiv verarbeiten. Um die Segmentierung während der sukzessiven Verarbeitung zu propagieren, werden in dieser Arbeit zwei neue Ansätze hierfür vorgestellt. Während beim Ersten großes Augenmerk auf die Effektivität gelegt wird und eine Delaunay Triangulation in Kombination mit einzelnen, verfolgten Merkmalspunkten Anwendung findet, propagiert der Zweite jeden Superpixel einzeln. Hierbei lassen sich Rückschlüsse auf verdeckte und aufgedeckte Bildregionen ziehen. Diese Informationen werden im weiteren Verlauf dazu genutzt, um auf strukturelle Änderungen im Videovolumen zu reagieren und hierdurch möglichst gleichgroße Superpixel zu generieren.

In einer umrangreichen Evaluierung mit etablierten Testverfahren wird das vorgstellte System mit aktuellen Verfahren zur Videoübersegmentierung verglichen. Die Ergebnisse zeigen, dass das vorgeschlagene Verfahren den geringsten Segmentierungsfehler aufweist. Gleichzeitig werden zeitlich längere Superpixeltrajektorien erzeugt als von Verfahren vergleichbarer Segmentierungsqualität. Dies zeigt, dass das vorgestellte Verfahren die im Video enthaltenen zeitlichen Verbindungen der Bildregionen besser extrahiert als frühere Ansätze. Gleichzeit skaliert die Laufzeit des Verfahrens besser, da sie nur linear mit der Anzahl der Pixel und Superpixel ansteigt.

Darüber hinaus wid die Leistungsfähigkeit des Verfahren am Beispiel der interaktiven Videosegmentierung mittels des Graph-Cut Algorithmus demonstriert. Verglichen mit einer pixelweisen Verarbeitung des Videomaterials veringert sich der initiale Segmentierungsfehler bei Anwendung des vorgestellten Verfahrens um über 47 %. Zusätzlich verkürzt sich die durchschnittliche Ausführungszeit des Graph-Cut Algorithmus von 31 Minuten auf unter 7 ms pro Sequenz.

Stichworte: Superpixel, Zeitkonsistenz, Supervoxel, Übersegmentierung, Verdeckung, interaktive Videosegmentierung