Depth-supported real-time video segmentation with the Kinect

Alexey Abramov[1], Karl Pauwels[2], Jeremie Papon[1], Florentin Wörgötter[1], Babette Dellen[3]

[1]BCCN, Georg-August-University Göttingen, III Physikalisches Institut, Friedrich-Hund Platz 1, 37077 Göttingen, Germany

[2]Computer Architecture and Technology Department, University of Granada, Granada, Spain; [3]Institut de Robótica i Informática Industrial (CSIC-UPC), Barcelona, Spain

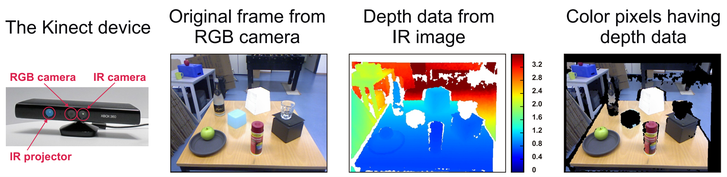
{abramov, jpapon, worgott}@physik3.gwdg.de, kpauwels@atc.ugr.es, bdellen@iri.upc.edu

WACV & MMBIA 2012

1. Introduction

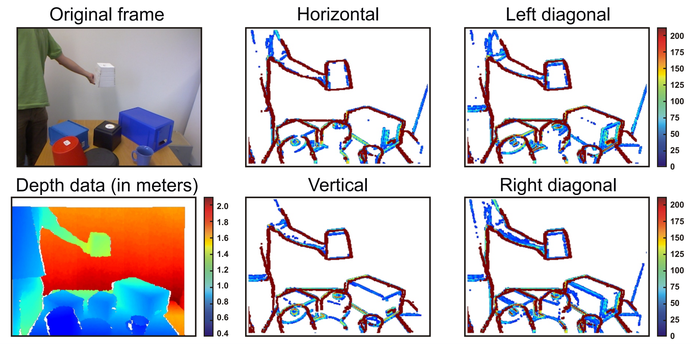
Video segmentation aims at representing image sequences through homogeneous regions (segments), where the same object part should carry the same unique label along the whole movie. The segmented visual data can be used for higher-level vision tasks which require temporal relations between objects to be established, including object tracking, action recognition, and content-based image retrieval. In the past, joint segmentation and tracking of segments in videos have been addressed in various works.However, all these methods are based on color cues alone.

In this work, we show that the inclusion of depth information can greatly improve video segmentation. We extended a recently developed framework for parallel color video segmentation (Abramov et al., 2010) by including depth information to the segmentation kernel. For depth acquisition, the Microsoft Kinect device, which was first released in the fall of 2010 for the Xbox video game platform, is used. The OpenNI toolbox is used for the Kinect calibration and mapping of color pixels with range values.



2. Image segmentation kernel

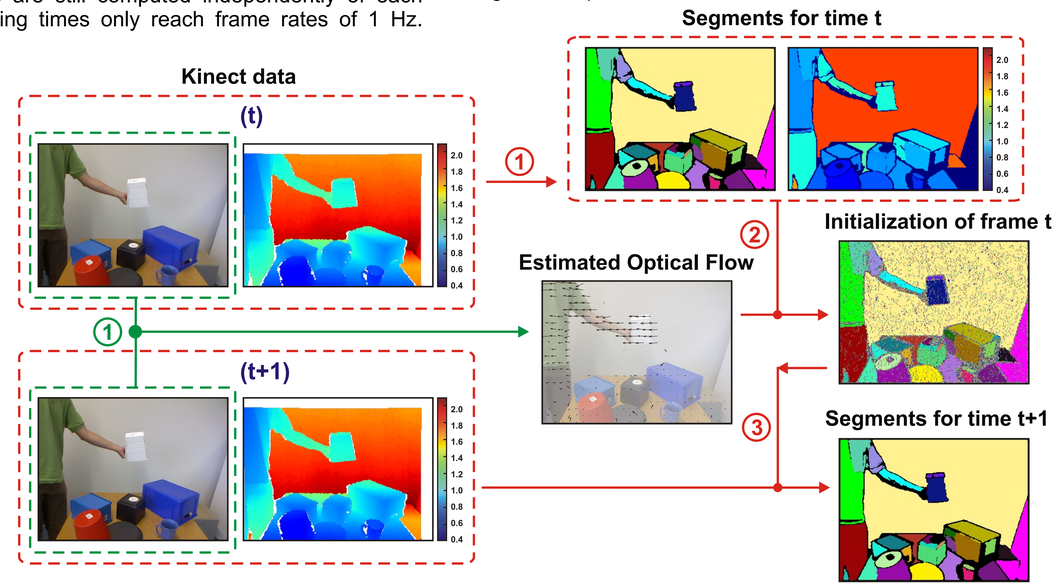
The segmentation corresponds to the equilibrium state of a Potts model, which is computed using a parallel Metropolis algorithm on the GPU. Since the method uses 8-connectivity of pixels, interaction strengths for one pixel need to be computed in four different directions: vertical, horizontal, left diagonal, right diagonal. The depth data acquired along with the color image is used to prevent interactions between pixels having a large range difference. This is done by replacing all color differences having a displacement larger than a predefined threshold with the maximum possible value.



3. Depth-supported video segmentation

Many methods for video segmentation are usually performing independent segmentations of each frame and then matching segments for tracking, having the disadvantage that segmentations need to be computed from scratch for each frame, affecting the efficiency of the method. Another problem is that the partition of the image may have changed in the new frame, leading to temporal consistency problems between segmentations. To resolve these problems, a graph-based model can be used to construct a consistent video segmentation from the over-segmented frames (Grundmann et al., 2010). Hovewer, the over-segmentations are still computed independently of each other, and processing times only reach frame rates of 1 Hz.

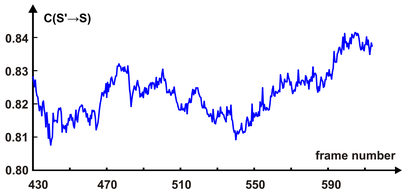
In the proposed method consistent video segmentation and segment tracking is achieved by transferring the solutions obtained at the precedent frame to the current frame. Only the first frame of the video stream is segmented completely from scratch, i.e., all spin variables are initialized by random values. Consecutive frames are initialized by spin state configurations taken from previous frames considering spatial shifts due to motion. To estimate the motion we use the real-time dense optical flow algorithm (Pauwels et al., 2010). The algorithm provides a vector field at each pixel indicating its motion. Having segments with correspondent average range values for a time step t, estimated optical flow vector field, labels of segments at time t are transferred to frame t+1, excluding transfers between pixels having a range difference larger than a pre-defined threshold.

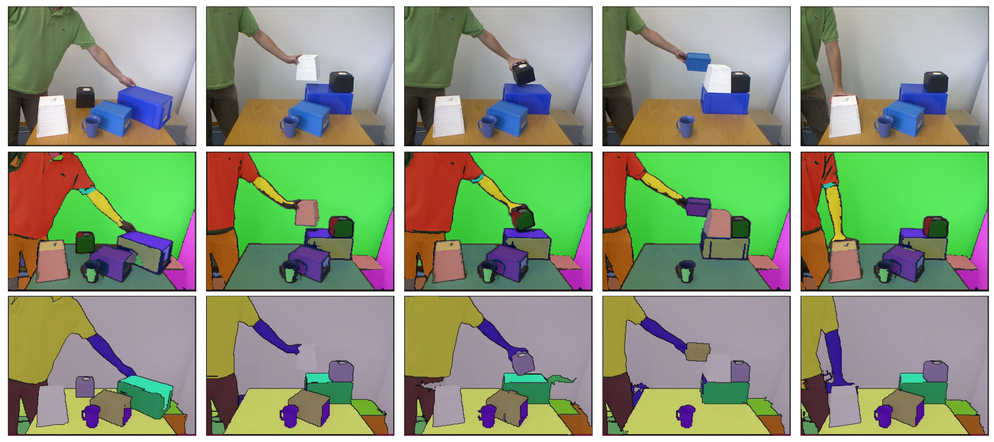


4. Results

Segmentation results for a 2 min frame sequence of the sample action “Building a pyramid” are presented. Our method provides a temporally coherent video segmentation, in which all segments carry their initially assigned labels along the whole movie. For comparison, we show segmentation results for the same sequence obtained with a recent graph-based video segmentation method (Grundmann et al., 2010). Depending on the hierarchy level of the graph-based method, a coarser or finer segmentation is obtained. At coarse levels, merging problems leading to under-segmentation are observed, while at finer levels, more segments are formed, leading however to some temporal coherency problems. The proposed method runs in real-time for medium image resolutions and can process video sequences of arbitrary length, while the graph-based video segmentation needs about 20 min to process a 40 sec video and only sequences that are not longer than 40 sec (with 25 fps) can be processed in the hierarchical mode.

To measure the quality of video segmentations we use the segmentation covering metric (Arbelarz et al., 2009). The idea of the metric is to evaluate the covering of a human segmentation, called also ground truth segmentation, by a machine segmentation. A human segmentation, is a manual annotation of a video sequence showing how humans perceive the scene, whereas a machine segmentation is an output result of the considered video segmentation algorithm.





5. Conclusions

In the current study we presented a novel real-time technique for the spatiotemporal segmentation of depth/color videos. The proposed method performs a homogeneous video segmentation, i.e. all objects visible in the scene carry the same unique labels along the whole video sequence. A Kinect device was used as a hardware setup for simultaneous real-time acquisition of color images and correspondent range data. Usage of depth data makes it possible to track relatively fast moving objects by preventing interactions between pixels having significant range differences.

6. Acknowledgments

This research has received funding by the EU GARNICS project FP7-247947 and the EU IntellAct project FP7-269959. B.Dellen acknowledges support from the Spanish Ministry for Science and Innovation via a Ramon y Cajal fellowship. K. Pauwels acknowledges support from CEI BioTIC GENIL (CEB09-0010) of the MICINN CEI program.

References

Abramov A. et al.: 3D Semantic Representation of Actions from Stereo-image-sequence segmentation on GPUs, 3DPVT2010.
Grundmann M. et al.: Efficient hierarchical graph-based video segmentation, CVPR, 2010.
Pauwels K. et al.: A cortical architecture on parallel hardware for motion processing in real time, Journal of Vision, 2010.
Arbelaez et al.: From contours to regions: An empirical evaluation, CVPR 2009.
OpenNI, http://www.openni.org .