INFRARED-AIDED SUPERPIXEL SEGMENTATION

Christopher Haccius[1], Harini Priyadarshini Hariharan[1], Thorsten Herfet[1], Jörn Jachalsky[2], Wolfram Putzke-Röming[2], Thomas Hach[3]

1 Intel Visual Computing Institute, Saarland University, 66123 Saarbrücken, Germany
2 Technicolor Research & Innovation, 30625 Hannover, Germany
3 Arnold & Richter Cine Technik GmbH & Co, 80799 München, Germany

ABSTRACT

Image segmentation is a fundamental preprocessing step in multiple tasks for the recognition and detection of semantically meaningful objects. In the past decades numerous image segmentation algorithms have been proposed. However, complexities at and above $O(n^2)$ make many of these computationally very expensive, several approaches require human input or have poor boundary recall. Among the state-of-the-art algorithms are superpixel segmentations, for which fast and fully automatic approaches exist. However, often scenes have content which cannot be segmented precisely based purely on color information. Novel image and video acquisition hardware can capture not only color, but also depth and infrared information. This additional information can be used to enhance existing segmentation algorithms. We present a novel multi-channel extension to existing superpixel segmentations which makes use of this additional information in order to improve the boundary recall by more than 11% while maintaining the same oversegmentation factor compared to a purely color based segmentation.

1 INTRODUCTION

Detecting, tracking and recognizing objects is a core problem of today's computer vision research. Fundamental to this research is the segmentation of images into meaningful parts. Numerous image segmentation algorithms have been proposed in the past decade. Among these, the SLIC superpixel algorithm [1] has outperformed other approaches due to its run-time performance, nearly unsupervised processing and finally the good achievable boundary recall.

However, often images contain content which cannot be segmented using only color information. This might be overlapping objects of the same color, mist hiding distant objects, or complex patterns obscuring object boundaries. Information exceeding the pure color information is necessary in such cases to still obtain meaningful and precise segmentations. Novel camera technologies capture such additional information, which can be used for image segmentation. The Motion Scene Camera [2], which we used for the data acquisition, captures not only color, but also depth and infrared information.

This paper focuses on using the available infrared information for enhanced superpixel segmentation. Using additional infrared intensity values we can increase the boundary recall by over 11%, thus significantly increasing the boundary detection of the superpixel segmentation.

In the following sections we will introduce related previous work in Section 2. The algorithmic extension of the superpixel algorithm in order to utilize such infrared information is described in Section 3. Our dataset, which is acquired by a novel camera for color and infrared capture, is introduced in Section 4. Experiments and experimental results are explained in Sections 5. Conclusions are drawn and possible future work is outlined in Section 6.

2 RELATED WORK

The importance of meaningful image segments to the human understanding of image content has been known since almost a century [3]. However, algorithmic development for image segmentation has started in the 1980's with the advance of digital image processing techniques. Among the first segmentation approaches were edge-based methods [4] or eigenvector-based approaches [5]. However, those methods were computationally expensive. In 2003 Felzenswalb and Huttenlocher have presented a graph-based image segmentation algorithm with the two major goals to "capture perceptually important groupings or regions" and at the same time "be highly efficient" [3]. Both goals are characteristic for superpixel algorithms. Next to graph based approaches like the aforementioned and geometric flow based algorithms [6] Achanta et al. have presented a Simple Linear Iterative Clustering (SLIC) algorithm for superpixel clustering in 2012 [1]. Due to its locally restrained search region SLIC superpixels are computationally very efficient and define the state-of-the-art with respect to boundary recall.

Employing infrared information for image segmentation is not a new idea. Texture based image segmentation approaches using near infrared (NIR) [7] as well as content classification using both, color and NIR [8] have been presented in recent years. Two core aspects in our research differ from already published approaches. First, SLIC superpixels currently present the state of the art, but superpixel extensions to NIR are currently unknown to the authors. Second, a SLIC extension is not only interesting, as this approach has superior boundary recall at a low computational complexity, but we will show that human input can further be minimized.

A problem for the approach of using NIR for computer vision is the availability of such information. Traditionally, cameras are restricted to the human visual perception and therefore capture only RGB information. While most sensors of digital cameras are sensible in NIR, it requires a camera modification to acquire this data as well. Even after such a modification, traditional sensors can capture either RGB or NIR information, as the sensor architecture is designed to capture only three channels.

In 2011 the SCENE project funded by the European Commission set out to capture and use additional environment information for computational videography. A core development in this project is the Motion Scene Camera (Figure 1), which captures color, depth and infrared information through the same optical system at the same time **[2]**. This novel hardware provides color and infrared information without spatial or temporal offset; a characteristic previously unavailable. Considering the infrared information utilized in this work, we decided to exploit the NIR light source which is attached to the camera and the second sensor's capability to natively generate an NIR image. This means, we record the NIR image of a scene which represents only the amount of the reflected NIR light sent out by the light source. Hence the characteristics of this NIR image are typical for images taken under artificial LED lighting. First, there is quadratic light decrease for objects at increasing distances from the camera retaining constant reflectivity. Second, the spectrum of the NIR image is discrete and centered at 850nm. This distinguishes our approach from using ambient NIR light sources. Last, the NIR image has a resolution of 160x90 pixels compared to the 1920x1080 pixels available for the RGB image. This mismatch between the spatial resolutions of both sources is posing a problem for postproduction. Typically, RGB-guided upscaling filters are applied to deal with this kind of problem. However, we do not want to introduce RGB information into the NIR image which is inherently done using RGB-guided upscaling. Hence we decided to use bicubic upscaling only. We want to outline at this point that the improvements using our method are achieved with this kind of low information NIR images.

Figure 1: Motion Scene Camera which captures Color, Infrared and Depth information through the same optical system

In recent years, the research focus for superpixel shifted from still images towards video content in order to generate temporally and spatially consistent superpixels for video sequences, as e.g. presented in [9]. The idea to employ NIR information also for temporally consistent superpixel segmentation is therefore highly interesting; an issue we have approached in this paper as well.

3 ALGORITHMIC EXTENSION

Current superpixel clustering algorithms employ color and spatial information to generate superpixels. For each channel, the distance to the superpixel center is computed, where Δ is the difference. Color information is converted into CIELab color space such that distances between colors closely correspond to subjectively perceived differences. In detail, SLIC superpixels presented by Achanta et al. [1] calculate the color distance

$$D_C = \sqrt{\Delta L^2 + \Delta a^2 + \Delta b^2} \quad (1)$$

where L, a and b represent the Lab color channels. Furthermore, the spatial distance is calculated

$$D_S = \sqrt{\Delta x^2 + \Delta y^2} \quad (2)$$

where x, y are the spatial parameters. Both distances, D_C and D_S, are combined to one single distance measure

$$D = \sqrt{D_C^2 + m \cdot D_S^2} \quad (3)$$

with m as a weighting factor between color and spatial distances. The authors of [1] suggest $1 \leq m \leq 40$, but use $m = 10$ as an optimized compromise between superpixel compactness and boundary recall.

We include the infrared information as a further channel into the distance measure, and set

$$D_I = \sqrt{\Delta IR^2} \quad (4)$$

where IR is the infrared information. We include this additional distance into distance D by providing new weights between the different parts

$$D = \sqrt{m \cdot D_S^2 + n \cdot D_C^2 + o \cdot D_I^2} \quad (5)$$

Choosing fixed values for m, n and o does, however, not work satisfactorily. The parameters are set as: $m = 1$, $n = o = 0.1$. This results in the same compactness and boundary recall as the original paper if either color or

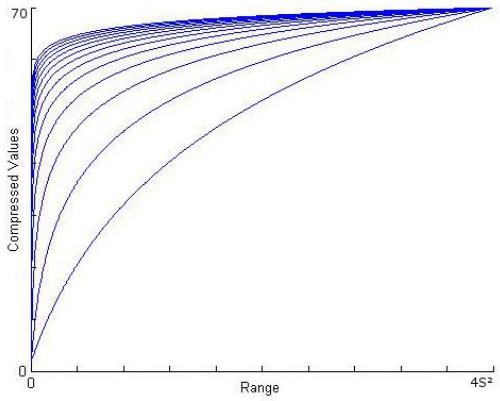

Figure 2: Functions assigning parameter values according to edge pixels found in subimages

infrared distance is 0, but becomes a lot more fuzzy if both, color and IR distance are large. Therefore, a dynamic allocation of weights is required.

Thereby, we take the following basic approach. We measure the fidelity of a search space surrounding a superpixel, and assign weights according to the fidelity of the individual channels. First, we calculate the Laplacian, as given in Equation (6), on each color and IR channel.

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \qquad (6)$$

The Laplacian is a good detector of edges in an image. For the color information we combine the results for the three Lab channels using a logical OR operation, resulting in a single channel depicting edges in either of the L, a or b channel.

The number g of edge pixels in a subimage of size $2S \times 2S$ is bound by $0 \leq g \leq 4S^2$. Counting edge pixels is threshold based, and we found through subjective tests that a threshold $t = 0.05$ returns good results for input values v scaled to $0 \leq v \leq 1$.

In further subjective tests, described in Subsection 3.1, we derived a function which assigns adequate parameters m, n and o based on the different fidelity values of color and infrared information.

3.1 Parameter Decision

With the dynamic approach for the parameter settings described above only few decisions need to be taken in advance. First, it was observed that many structures are contained in both, color and infrared images. We therefore asked a group of people to identify the subjectively best tradeoff between compactness and boundary recall of superpixels, when distances were weighted with $m = 1$, $n = o = x$ where $1 \leq x \leq 15$. This subjective test returned an optimal value of $x = 7$. Further, with $m = 10$ from [1], we set $1 \leq n, o \leq 70$ based on the number of edge pixels detected in a subimage surrounding a superpixel center.

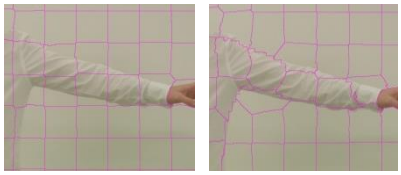

Figure 3: Effect of different compression functions, most linear (left) and most curved (right) from Figure 2

To find a suitable function we derived a set of curves assigning output values between $1 \leq n, o \leq 70$ to input edge values of $0 \leq g \leq 4S^2$, as shown in Figure 2. A set of sample images segmented with the different curves was presented to several people, and their subjective decision for the best segmentation result was evaluated. The function given in Equation (7) was the result of this process.

$$f(g) = \frac{69\left(2.25ln\left(\frac{2200g}{S^2}+1\right)-2.25ln\left(\frac{2200}{S^2}+1\right)\right)}{\left(2.25ln\left(\frac{2200\cdot 10^6}{S^2}+1\right)-2.25ln\left(\frac{2200}{S^2}+1\right)\right)} + 1 \qquad (7)$$

This equation takes the number of edge pixels as input g and requires the additional parameter S, which corresponds to the subimage size and is fixed by the number of superpixels per image. The output $f(g)$ corresponds to the missing parameters n and o for RGB and IR respectively. The function f is a modification of the μ-law [10], with $\mu = 86.7$ and $V = 10$ fitted to the required range. Figure 3 shows the effect of choosing different functions. If the assignment is rather linear, the spatial weighting tends to be too high, resulting in too compact superpixels with little boundary recall. On the other hand, if the function assigns too high values for few edge pixels already, superpixels tend to become too irregularly shaped.

4 DATASET

A video sequence was captured by the Motion Scene Camera. This sequence was directed in a way to produce content that can show the benefit of using an additional IR channel. One scenario where additional infrared information can help is when objects of the same color overlap. In this specific video sequence a talent wearing a white shirt moves from a green background to a white background. Figure 4 shows two frames of this data set, one with the talent in front of the green wall, one with the talent in front of the white wall. As the video sequence was captured with the Motion Scene Camera through the same optical system, the color and infrared information are aligned perfectly, temporally as well as spatially.

For this dataset a ground truth segmentation was created. The ground truth contains 10 different semantic objects; the green wall, the white wall, a wooden line connecting both wall pieces, the floor, the talent's face and two hands, hair, shirt and pants. The ground truth corresponding to the frames in Figure 4 is shown in Figure 5.

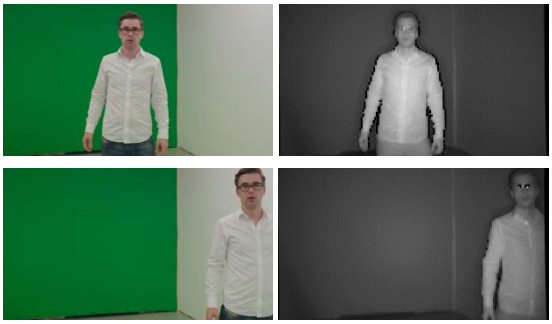

Figure 4: data set sample frames RGB and IR

Figure 5: ground truth for frames in Figure 4

5 EXPERIMENTS AND RESULTS

Boundary recall is a value reflecting the amount of semantic image boundaries that an algorithm finds. We define boundary recall br of a segmentation S compared to a ground truth T as the percentage of border pixels picked up by the segmentation algorithm, calculated as

$$br(S,T) = \frac{bp(S,T)}{bp(T)} \cdot 100\% \quad (8)$$

where $bp(S,T)$ are the boundary pixels of both, ground truth and segmentation, and $bp(T)$ are the boundary pixels of only the ground truth.

We have extended the implementation from [9] by our novel distance metric and created a temporally consistent segmentation of the input sequence. Figure 6 (top) depicts two segmentations of an input frame showing the talent in front of the white background without NIR information (left) and using available NIR information (right). In Figure 6 (bottom) superpixels belonging to the actor are colored black, those belonging to the background are colored white, and superpixels crossing the object boundaries are colored red.

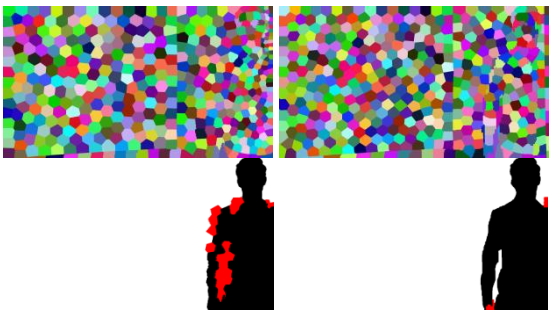

Figure 6: superpixel border recall without and with IR information using [11] extended by our distance metric

Table 1: boundary recall without and with IR in front of green and white background for 400 superpixels using [11]

	w/o IR	w/ IR
green	95%	94%
white	89%	95%

The oversegmentation factor relates the number of superpixels in a segmentation S to the number of segments in the corresponding ground truth$T$. The oversegmentation factor is calculated as

$$os(S,T) = \frac{1}{k}\sum_{i=1}^{k} r(S,T_i) \quad (9)$$

where k is the number of image segments in the ground truth and $r(S,T_i)$ is the number of segments in S overlapping with segment i from ground truth T.

Due to their locally restricted search regions SLIC superpixels are – by default – oversegmenting image content. Oversegmentation increases the probability of finding the correct boundary, but comes at the cost of many insignificant boundaries. This is illustrated when looking at the two edge cases and the desired case.

Case 1: every pixel is a superpixel

In this case every pixel in S is also a region boundary pixel. By default, each boundary pixel in T has a corresponding boundary pixel in S, which results in a boundary recall $br(S,T) = 100\%$. As the number of pixels is usually significantly larger than the number of regions in T, it is $os(S,T) \gg 1$.

Case 2: S is only one region

In this case every region T_i overlaps with only one region in S. Therefore we have $\sum_{i=1}^{k} r(S,T_i) = k$ and consequently $os(S,T) = 1$. On the other hand there are no region boundaries in S, therefore $br(S,T) = 0\%$.

Case 3: every region in S matches a region in T

Here all boundary pixels match, therefore $bp(S,T) = bp(T)$ and $br(S,T) = 100\%$. At the same time each region T_i overlaps with exactly one region in S, therefore $\sum_{i=1}^{k} r(S,T_i) = k$ and hence $os(S,T) = 1$.

A boundary recall value of 100% and at the same time an oversegmentation of 1 therefore is desirable.

Finally, we employ a measure of compactness for superpixels, which was introduced by Schick et al. [12]. They introduce compactness of a segmentation $co(S)$ as

$$co(S) = \frac{1}{k} \cdot \sum_{i=1}^{k} Q_i \quad (10)$$

with the isoperimetric coefficient of the i-th superpixel of a segmentation S defined as

$$Q_i = \frac{4\pi A_i}{L_i^2} \quad (11)$$

where A_i and L_i are the area and the perimeter of the i-th superpixel respectively.

Table 1 gives the boundary recall comparing a segmentation of the white shirt in front of white background and in front of green background. While the

segmentation based purely on RGB works well when the white shirt is in front of the green background, the benefit or using additional IR information becomes clear when the white shirt is in front of the white background.

Figure 7 shows the oversegmentation factor in relation to the boundary recall for segmentations without (green curve) and with (blue curve) infrared information. As expected, a small oversegmentation factor corresponds to a low boundary recall value, and the higher the oversegmentation the better the boundary recall. More interesting, however, is the gradient of the curve belonging to the segmentation process with infrared information which is continuously higher than the curve belonging to the segmentation process without additional infrared information. At an oversegmentation factor of roughly 4 we can enhance the boundary recall of only color based segmentation by more than **11**% through the additional use of NIR information.

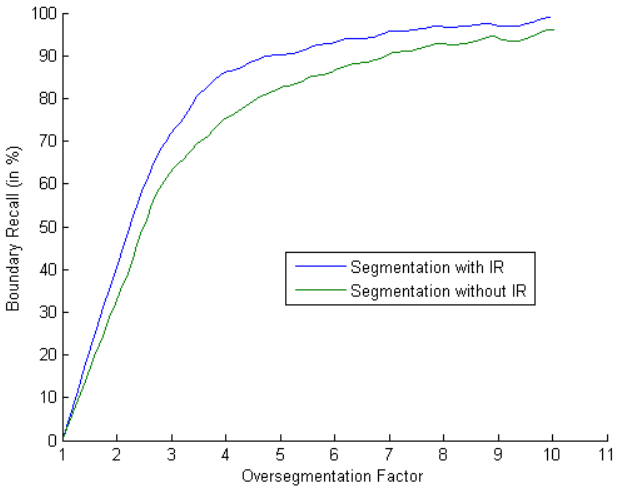

Figure 7: boundary recall vs. oversegmentation comparing segmentations with IR (blue line) to segmentations without IR information (green line)

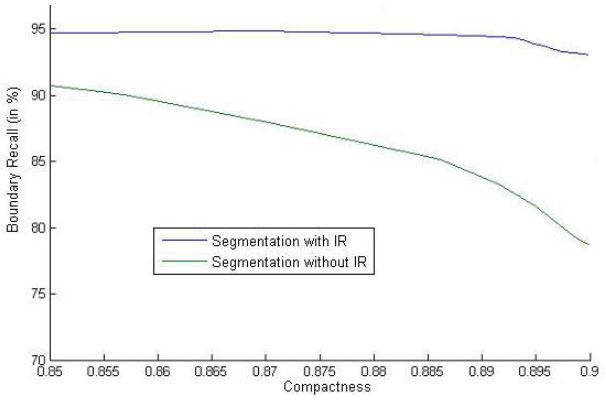

Figure 8: boundary recall vs. compactness comparing segmentations with IR (blue line) to segmentations without IR information (green line)

Table 2: boundary recall without and with IR with brightness change and salt-and-pepper noise

	w/o IR	w/ IR
brightness	62%	80%
noise	87%	95%

Figure 9: brightness change (left) and salt-and-pepper noise (right)

5.1 Robustness

Often images are affected by different kinds of noise. We have evaluated the robustness of our approach with respect to Gaussian distributed salt and pepper noise as well as brightness changes. Figure 9 shows sample frames with both forms of image distortion applied. In Table 2 the boundary recall achievements for noisy and images and images affected by lighting changes are shown, all other parameters remain unchanged with respect to the results achieved in Table 1. While the segmentation is affected by the noise, it remains obvious that results achieved with the additional IR information are superior to segmentations based purely on RGB information.

6 CONCLUSION AND FUTURE WORK

The experiments described above show two major benefits. First and most important is that the use of additional information not visible to the human eye is useful for the segmentation of semantically meaningful objects. Object boundaries can be invisible in color space but prominent in infrared.

Second, use of additional information channels enforces the use of dynamic parameter allocation. Using boundaries known from experiments on only color based segmentation, such dynamic parameter decisions reduce the required human input even compared to only color based segmentations. While the SLIC algorithm proposed in [1] requires an input image, the number of desired superpixels and the weighting factor m to weight the color information, our implementation requires only the input image and the desired number of superpixels.

Our experiments with respect to robustness allow the conclusion that in presence of noise or changing lighting conditions boundary recall may vary. However, independent of the amount of initial RGB information (clear, noisy, change of brightness) employing additional infrared information yields superior segmentations.

Looking at the proposed way of adding the additional infrared channel to the information the idea to add further information channels suggests itself. In addition to color

and IR information the Motion Scene Camera provides depth information per pixel. Depth information was already applied for enhanced superpixel segmentation [13] [14], and the addition of this information to our algorithm and comparison to existing implementations will yield new insights into the use of additional information.

Traditional cameras can also capture infrared information with their native RGB sensors. In order to do that, a filter usually blocking NIR needs to be replaced by an NIR-pass filter. As shown in Figure 10 the NIR area does not necessarily be considered as one additional channel, but taking the blue channel and the red channel independently it is possible to separate NIR at $\sim 7000nm$ and $\sim 8000nm$, thus yielding two additional channels. Furthermore, using thermal imaging systems, which capture infrared between $1\mu m$ and $1mm$, can offer a whole new range of information invisible to the human eye, but interesting for object segmentation.

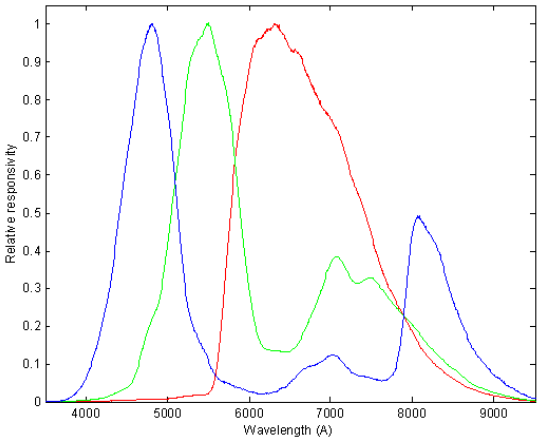

Figure 10: sensibility of color channels of traditional camera without a hot mirror filter [15]

ACKNOWLEDGEMENT

This work bases on research conducted in the SCENE project [16]. It has been supported by the EC within the 7th framework programme under grant agreement no. FP7-IST-287639.

7 REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," in *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34(11), 2274-2282*, 2012.

[2] T. Hach and J. Steurer, "A novel RGB-Z camera for high-quality motion picture applications," in *Proceedings of the 10th European Conference on Visual Media Production, CVMP*, London, 2013.

[3] P. Felzenswab and D. Huttenlocher, "Efficient Graph-Based Image Segmentation," *International Journal of Computer Vision*, vol. 2, no. 59, pp. 167-181, 2004.

[4] R. Urquhart, "Graph Theoretical Clustering Based on limited neighbourhood sets," *Pattern Recognition*, vol. 3, no. 15, pp. 173-197, 1982.

[5] J. Shi and J. Malik, "Normalized cuts and Image Segmentation," *Pattern Analysis and Machine Intelligence*, vol. 8, no. 22, pp. 888-905, 2000.

[6] A. Levinshtein, A. Stere, K. Kutulakos, D. Fleet, S. Dickinson and K. Siddiqi, "Turbopixels: Fast Superpixels using geometric flow," *Pattern Analysis and Machine Intelligence*, vol. 12, no. 31, pp. 2290-2297, 2009.

[7] W. Blanton and K. Barner, "Texture-Based Infrared Image Segmentation by Combined Merging and Partitioning," in *IEEE International Conference on Image Processing, 2007. ICIP 2007.*, San Antonio, TX, 2007.

[8] N. Salamati, D. Larlus, G. Csurka and S. Süsstrunk, "Semantic Image Segmentation," in *4th Workshop on Color and Photometry in Computer Vision at ECCV12*, Florence, Italy, 2012.

[9] M. Reso, J. Jachalsky, B. Rosenhahn and J. Ostermann, "Temporally Consistent Superpixels," in *IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013.

[10] B. Sklar, "Companding Characteristics," in *Digital Communications: Fundamentals and Applications*, NJ, Prentice-Hall, 1988, pp. 84-85.

[11] M. Reso, J. Jachalsky, B. Rosenhahn und J. Ostermann, „Temporally Consistent Superpixels," in *IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013.

[12] A. Schick, M. Fischer and R. Stiefelhagen, "Measuring and Evaluating the Compactness of Superpixels," in *International Conference on Pattern Recognition*, Tsukuba Science City, Japan, 2012.

[13] D. Weikersdorfer, D. Gossow and M. Beetz, "Depth-Adaptive Superpixels," in *21st International Conference on Pattern Recognition*, 2012.

[14] I. Jebari and D. Filliata, "Color and Depth-Based Superpixels for Background and Object Segmentation," in *International Symposium on Robotics and Intelligent Sensors*, 2012.

[15] Astrosurf.com, "Canon EOS 350d - Filter Removal Operation and Performance," 2014. [Online]. Available: http://www.astrosurf.com/buil/350d/350d.htm. [Accessed 20 October 2014].

[16] V. López, E. Fuenmayor and A. Hilton, "Novel Scene Representations for Richer Networked Media," October 2014. [Online]. Available: http://3d-scene.eu. [Accessed 30 October 2014].

[17] T. Hach und T. Steurer, „A novel RGB-Z camera for high-quality motion picture applications," in *Proceedings of the 10th European Conference on Visual Media Production, CVMP*, London, 2013.