Improving the Prediction Efficiency for Multi-View Video Coding Using Histogram Matching

Ulrich Fecker, Marcus Barkowsky, and André Kaup

Chair of Multimedia Communications and Signal Processing,
University of Erlangen-Nuremberg,
Cauerstr. 7, 91058 Erlangen, Germany

Abstract. Applications for video data recorded with a setup of several cameras are currently attracting increasing interest. For such multi-view sequences, efficient coding is crucial to handle the enormous amount of data. However, significant luminance and chrominance variations between the different views, which often originate from imperfect camera calibration, are able to reduce the coding efficiency and the rendering quality. In this paper, we suggest the usage of histogram matching to compensate these differences in a pre-filtering step. After a description of the proposed algorithm, it is explained how histogram matching can be applied to multi-view video data. The effect of histogram matching on the coding performance is evaluated by statistically analysing prediction from temporal as well as from spatial references. For several test sequences, results are shown which indicate that the amount of spatial prediction across different camera views can be increased by applying histogram matching.

Index Terms – multi-view video, video coding, image filtering

1 Introduction

Image-based rendering is an approach to visualise an object or a scene for any desired viewpoint and viewing angle. For that, the scene is recorded using a setup of multiple cameras. If the scene is dynamic, this means that a video sequence needs to be acquired for each camera position. These video streams can then be used to render photorealistic views for arbitrary viewpoints and arbitrary viewing angles [1–3].

Applications for multi-view video include three-dimensional television (3D TV) or free-viewpoint television (FTV) [4, 5], as well as the visualisation of medical data, e. g. inner organs of the human [6].

The amount of data involved in image-based rendering is huge, and therefore, efficient compression techniques are required to store or transmit multi-view video streams. Different coding schemes have been proposed which exploit not only the temporal correlation between subsequent frames, but also the spatial correlation between neighbouring camera views (see e. g. [7, 5]).

However, significant discrepancies between the luminance components as well as the chrominance components of the different camera views can often be observed. These variations may well impair the performance of a multi-view coder or a renderer. In many cases, it is therefore desirable to compensate these differences in a pre-filtering step.

In this paper, we propose the use of histogram matching, as outlined by Hekstra et. al. in [8], for the compensation of luminance and chrominance variations in multi-view sequences. This method assumes that a good fit of a distorted image to a reference image may be obtained by adapting the cumulative histogram of the distorted image to the cumulative reference histogram. The advantage of this procedure is that no assumptions on the type of distortion like brightness or contrast variations are made and also non-linear operations may be considered.

First, the proposed algorithm is described, and it is explained how this method can be applied to multi-view sequences. To analyse to which extent histogram matching is able to improve the prediction efficiency in multi-view video coding, a statistical analysis of block matching with temporal as well as spatial references is performed. Results are shown for several multi-view test data sets.

2 Histogram Matching

2.1 Description of the Algorithm

Based on the idea described in [8], we suggest the following method for histogram matching. The algorithm is exemplarily shown for the luminance component. It can be done for the chrominance components in an analogous manner.

The first step is to compute the histograms of the reference and the distorted image. Let $y_R[m, n]$ denote the amplitude of the luminance signal of the

reference image. Then, the histogram is calculated as follows:

$$h_R[v] = \frac{1}{w \cdot h} \sum_{m=0}^{h-1} \sum_{n=0}^{w-1} \delta\big[v, y_R[m,n]\big]$$
$$\text{with} \quad \delta[a,b] = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{else} \end{cases} \tag{1}$$

In this equation, w denotes the width and h the height of the image. The calibration is done by mapping each value of the distorted image to a corrected value. Two steps are necessary to generate the mapping function M. First, the cumulative histogram $c_R[v]$ of the reference image is created:

$$c_R[v] = \sum_{i=0}^{v} h_R[i] \tag{2}$$

The cumulative histogram $c_D[v]$ of the distorted image is calculated in the same manner. An example for a reference image and a distorted image together with their histograms and cumulative histograms is shown in Fig. 1. The distortion that was used on the distorted image included a gamma correction and a decrease in brightness. Therefore, its histogram is stretched and moved to the left.

Based on the cumulative histograms, we find the mapping by matching the number of occurrences in the distorted image to the number of occurrences in the reference image:

$$M[v] = u \quad \text{with} \quad c_R[u] < c_D[v] \leq c_R[u+1] \tag{3}$$

This process is illustrated in Fig. 2(a). The mapping may then be applied to the distorted image $y_D[m,n]$, resulting in the corrected image $y_C[m,n]$:

$$y_C[m,n] = M\big[y_D[m,n]\big] \tag{4}$$

2.2 Correction of the First and Last Active Bin

When the described algorithm is applied, all values of the distorted image below a certain threshold are clipped and are now remapped to the first active bin in the histogram. This induces a brightness offset in the corrected image for those dark values in the luminance image. In order to avoid this effect and its counterpart, when clipping at the upper boundary occurs, the first and the last active bin values are modified. The following additional step is only applied to the luminance component, because usually the colour components do not suffer from clipping artifacts.

The algorithm described so far implies that the highest value of the reference image is used for the values in the clipped interval. The quality of the mapping is improved by using the centre of mass of the values for this interval in the reference image. The lower interval is $\big[0 \ldots M[0]\big]$ and the upper interval is $\big[M[255] \ldots 255\big]$. Now, the centres of mass s_l and s_u for these intervals are calculated:

$$s_l = \frac{\sum_{i=0}^{M[0]} i \cdot h_R[i]}{\sum_{i=0}^{M[0]} h_R[i]} \tag{5}$$

and

$$s_u = \frac{\sum_{i=M[255]}^{255} i \cdot h_R[i]}{\sum_{i=M[255]}^{255} h_R[i]} \tag{6}$$

These values are then applied in the mapping:

$$M[0] = s_l \quad \text{and} \quad M[255] = s_u \tag{7}$$

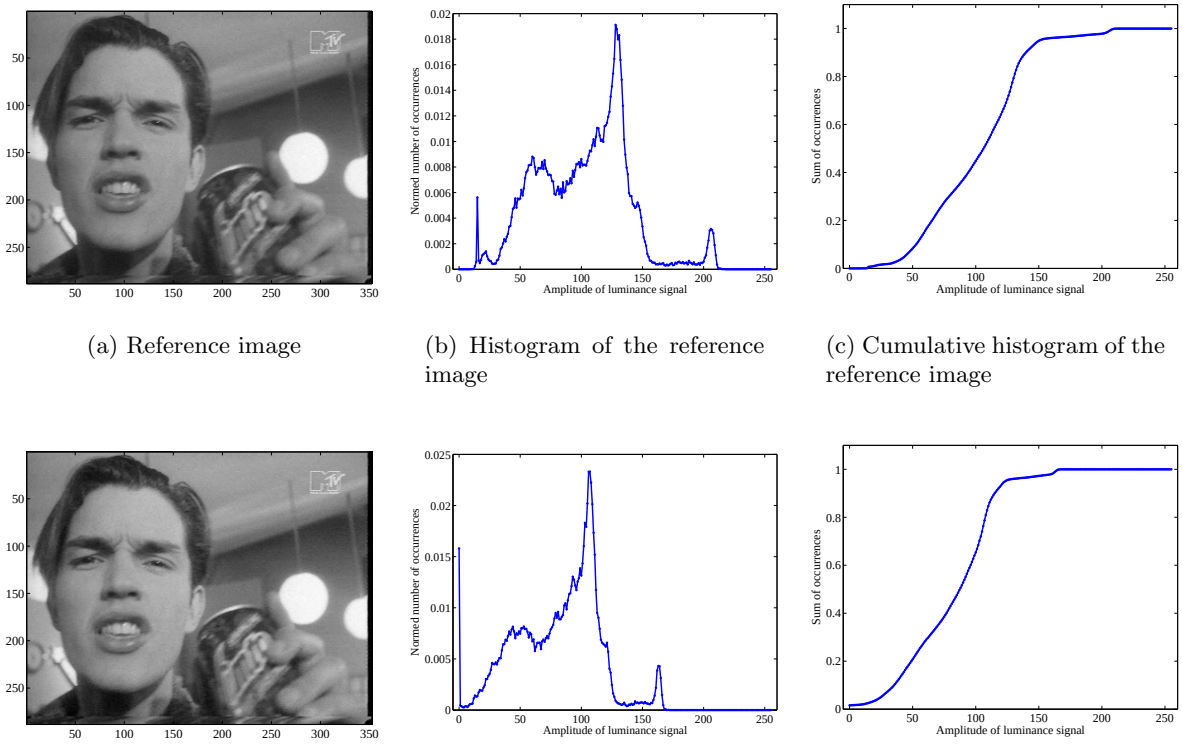
The resulting histogram of the corrected image with this mapping applied is shown in Fig. 2(b). The histogram of the corrected image is more similar to the reference histogram, and the smallest values are now mapped to a darker brightness value. It can be seen that there may be luminance levels which do not occur in the corrected image. This effect is however not visible in the image.

2.3 Application to Multi-View Sequences

The described algorithm can be applied to multi-view sequences in the following way: One camera view, which is close to the centre of the camera setup, is chosen as the reference view. All other camera views are corrected so that their histograms fit the histogram of the chosen reference view. This is done frame by frame for the whole sequence.

3 Evaluation Method

To evaluate the effect of histogram matching on multi-view sequences, the method described in [9] was used. It was assumed that an encoder would hold frames from a certain number of time steps for



(a) Reference image

(b) Histogram of the reference image

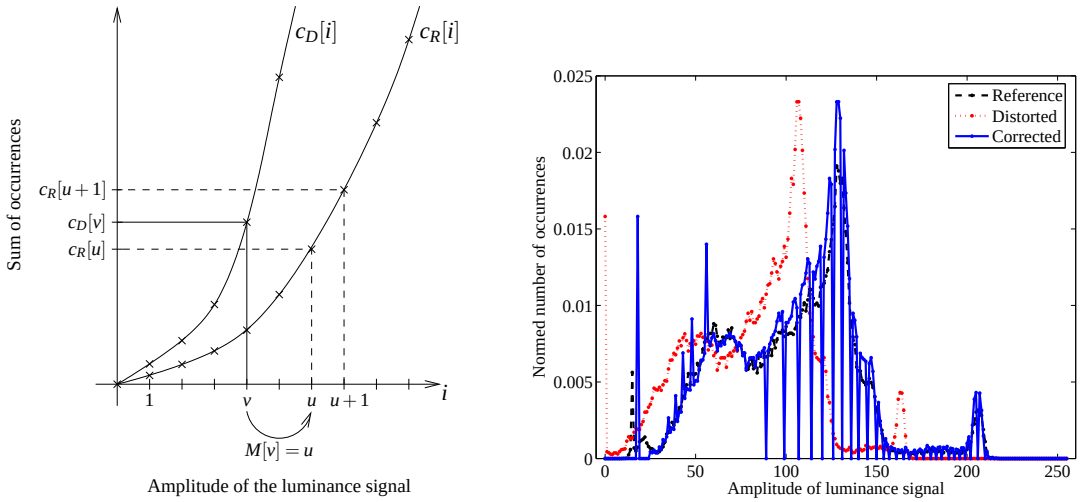
(c) Cumulative histogram of the reference image

(d) Distorted image

(e) Histogram of the distorted image

(f) Cumulative histogram of the distorted image

Fig. 1. Example histograms of the luminance component for a reference image and a distorted image



(a) Details of the mapping algorithm shown in a section of the cumulative histogram

(b) Histogram of the corrected image compared to the reference image and the distorted image

Fig. 2. Mapping of the cumulative histograms

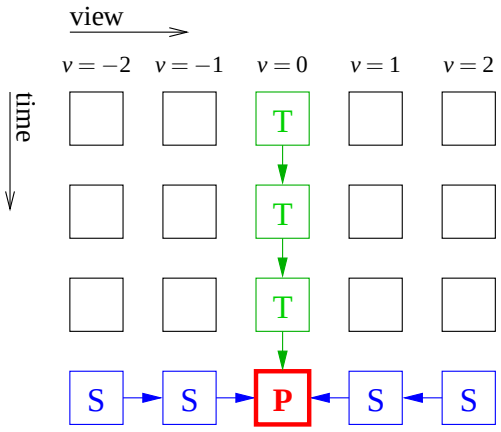


Fig. 3. Prediction scheme assumed for the analysis

all camera views in its memory (see Fig. 3). The current frame to be predicted is marked as "P"-frame. For each block in the frame, block matching is performed to find the reference block with the minimum mean square error (MSE) compared to the current block. For that, all temporally preceding framed ("T"-frames) as well as frames from the same time step but from different camera views ("S"-frames) are searched. The frame delivering the minimum MSE is then chosen as the best reference for the current block.

The number of temporal (T) and spatial (S) references is counted for all frames in all views of the multi-view sequence and converted into percentages. These probabilities give a rough feeling of the gain such a scheme could achieve compared to simulcast coding, where only temporal prediction is possible.

In [9], the search was extended to the remaining frames in the buffer (marked by empty rectangles in Fig. 3). These references were called "mixed modes". However, it turned out that for a practical coder, it might be suitable to skip these modes and search only temporal and spatial references because the probability for each single mixed mode is rather small. In this paper, we therefore assume that mixed mode prediction is not used. The results are however similar for the case of mixed mode prediction.

4 Results

For the evaluation, several multi-view test sequences were used. The described block matching scheme was applied to the original sequences as well as to the sequences after histogram matching had been performed. The results are shown in Table 1.

The results show that the percentage of spatial prediction increases when the sequences have been compensated using histogram matching. For all tested sequences — except for the sequence *Xmas* — a gain could be observed. As the sequence *Xmas* was recorded using a single camera only, the luminance and chrominance components are already well-calibrated and therefore the sequence does not benefit from histogram matching. For the other sequences, the gain in spatial prediction varies between 1.6 % and 12.9 %, depending on their specific characteristics. The results indicate that a real coding gain could be possible in a practical multi-view coder when luminance and chrominance compensation is applied before the encoding step.

After luminance and chrominance compensation, the input sequences are of course changed compared to their original versions. For the sequence *Ballet,* for example, the PSNR between the compensated sequence and the original sequence is 39.2 dB (averaged over all changed views). For the sequence *Exit,* the PSNR is 32.5 dB. The changes to the sequences do however hardly affect the visual quality and may even be helpful for the renderer when interpolation of views is necessary.

5 Summary

Histogram matching was proposed for the compensation of luminance and chrominance variations between the different camera views of multi-view sequences. A description of the histogram matching algorithm was given. All camera views of a multi-view sequence were matched to a reference view, which was chosen to be a view close to the centre of the camera arrangement.

To evaluate the effect of histogram matching, a statistical analysis of the block matching step of a multi-view video coder with temporal as well as spatial references was performed. The results with and without histogram matching were compared for several test data sets. A gain in spatial prediction of 1.6 % to 12.9 % could be observed, meaning that histogram matching is able to improve the efficiency of prediction between the camera views. This leads to the assumption that a coding gain could be achieved in a practical multi-view video coder.

Table 1. Results of the statistical analysis if histogram matching is used compared to the results without histogram matching

Sequence	Number of views	Number of frames	*Without Histogram Matching* Temporal	*Without Histogram Matching* Spatial	*With Histogram Matching* Temporal	*With Histogram Matching* Spatial	Gain in Spatial Percentage
Crowd	5	1 002	86.20 %	13.80 %	84.61 %	15.39 %	1.59 %
Flamenco1	8	624	77.43 %	22.57 %	72.42 %	27.58 %	5.01 %
Ballet	8	100	87.20 %	12.80 %	80.29 %	19.71 %	6.91 %
Breakdancers	8	100	63.70 %	36.30 %	50.76 %	49.24 %	12.94 %
Ballroom	8	250	84.91 %	15.09 %	83.26 %	16.74 %	1.65 %
Exit	8	250	87.26 %	12.74 %	83.96 %	16.04 %	3.30 %
Jungle	8	250	96.13 %	3.87 %	93.66 %	6.34 %	2.47 %
Uli	8	250	91.67 %	8.33 %	85.49 %	14.51 %	6.18 %
Xmas	10	101	19.17 %	80.83 %	19.91 %	80.09 %	-0.74 %

Acknowledgemenents

This work was funded by the German Research Foundation (DFG) within the Collaborative Research Centre "Model-based analysis and visualisation of complex scenes and sensor data" under grant SFB 603/TP C8. Only the authors are responsible for the content.

The authors express their thanks for providing test sequences to KDDI Corporation, the Interactive Visual Media Group at Microsoft Research, Mitsubishi Electric Research Laboratories (MERL), Fraunhofer HHI and Tanimoto Laboratory at Nagoya University.

References

1. Levoy, M., Hanrahan, P.: Light field rendering. In: Proceedings SIGGRAPH 96, New Orleans, Louisiana, USA (1996) 31–42
2. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: Proceedings SIGGRAPH 96, New Orleans, Louisiana, USA (1996) 43–54
3. Smolic, A., Kauff, P.: Interactive 3-D video representation and coding techniques. Proceedings of the IEEE **93** (2005) 98–110
4. Vetro, A., Matusik, W., Pfister, H., Xin, J.: Coding approaches for end-to-end 3D TV systems. In: Picture Coding Symposium (PCS 2004), San Francisco, CA, USA (2004)
5. Tanimoto, M.: Free viewpoint television — FTV. In: Picture Coding Symposium (PCS 2004), San Francisco, CA, USA (2004)
6. Vogt, F., Krüger, S., Schmidt, J., Paulus, D., Niemann, H., Hohenberger, W., Schick, C.H.: Light fields for minimal invasive surgery using an endoscope positioning robot. Methods of Information in Medicine **43** (2004) 403–408
7. Kimata, H., Kitahara, M., Kamikura, K., Yashima, Y.: Multi-view video coding using reference picture selection for free-viewpoint video communication. In: Picture Coding Symposium (PCS 2004), San Francisco, CA, USA (2004)
8. Hekstra, A.P., Beerends, J.G., Ledermann, D., de Caluwe, F.E., Kohler, S., Koenen, R.H., Rihs, S., Ehrsam, M., Schlauss, D.: PVQM — a perceptual video quality measure. Signal Processing: Image Communication **17** (2002) 781–798
9. Fecker, U., Kaup, A.: Statistical analysis of multireference block matching for dynamic light field coding. In: Proc. 10th International Fall Workshop Vision, Modeling, and Visualization (VMV 2005), Erlangen, Germany (2005) 445–452