

Performance of multi camera views' detection using MPEG-7 visual signature tools

K. WNUKOWICZ* and G. GALIŃSKI

Institute of Radioelectronics, Warsaw University of Technology, 15/19 Nowowiejska Str.,
00–665 Warszawa, Poland

The paper presents experimental results for the detection of related video streams (views) of multi camera system using MPEG-7 visual signature tools. This detection can be used for identification of video streams stored in video repositories which constitute the same multi camera recording set. Two signature tools have been investigated: Image Signature descriptor and Video Signature descriptor. The motivation for the usage of the tools designed primarily for replica detection is the fact that the video material produced by multi camera system typically consists of a few related video streams which are similar to each other and, thus, the video streams produced by multi camera system can be regarded as semi-replicas as they contain predominantly the same scenes or objects recorded at the same time. Experiments show that only in some restricted scenarios of a multi camera system setup the MPEG-7 visual signature tools can be used for reliable detection of multi camera views. These reliable scenarios include systems having cameras with perpendicular optical axes and “far” scene.

Keywords: multi view video detection, multi camera systems, MPEG-7 visual signature tools.

1. Introduction

The material produced by a multi camera video system typically consists of a few related video streams which are similar to each other and, thus, video streams produced by a multi camera system can be regarded as semi-replicas because they contain predominantly the same scenes or objects recorded at the same time. The paper presents the results of experiments for the detection of video streams (views) recorded at the same time in multi camera systems by using MPEG-7 visual signature tools. This detection can be used in applications such as identification of video streams constituting one multi camera set in a video repository, or in an application for verification that the given sets of video streams are multi view video recordings.

MPEG-7, an international standard formally called Multimedia Content Description Interface (ISO/IEC 15938) [1], specifies, among others, visual signature tools [2] designed to be used primarily in applications of image and video replica detection. The goal of visual signature tools is to allow the detection of all variants of the same multimedia material, including the versions modified by various image/video processing techniques, e.g.: resizing, smoothing, rotation, colour conversion, translation, cropping, and lossy compression. Two signature tools have been investigated [3]: Image Signature descriptor and Video Signature descriptor. Image Signature specifies a tool for applications of a content based image identification and a replica detection. This

tool uses a concise signature calculated from an image content which is invariant to many image processing techniques with fast matching function. The signature allows for an identification of a different version of the same image but discriminates images which are not replica of each other with high reliability. In the paper this tool was used to compare video frames obtained from different cameras in a multi camera system. Video Signature, another visual signature tool, is designed to detect duplicated video segments in video streams. Again, this tool was used to compare views obtained from different cameras in a multi camera system.

The paper's outline is as follows: Sect. 2 contains the description of Image Signature extraction and matching, Sect. 3 contains the description of Video Signature extraction and matching, Sect. 4 presents the experimental results of a multi view detection using MPEG-7 visual signature tools, and Sect. 5 concludes the paper.

2. Image Signature

Image Signature defined in a MPEG-7 standard is a concise image identifier designed for various applications of an image replica detection [4]. This signature is designed to allow for a fast and reliable identification of all variants of an original image, also those which were modified by many common processing techniques, like resizing, smoothing, noising, rotation, lossy compression, etc. Two versions of signature have been proposed: global and local features. Global signature allows for very fast searching (matching of

* e-mail: K.Wnukowicz@ire.pw.edu.pl

up to 100 million signature pairs per second on a contemporary PC), but is only robust to simple “global” modifications such as resizing, colour conversion, lossy compression. The size of a global signature is 64 bytes and it represents statistics of global edge information computed with Trace Transform [5] which is a modified version of Radon Transform. The “local” version of Image Signature uses local feature point descriptors extracted in a scale-space representation of image, and is designed to be robust to more complex modifications, such as cropping, translation or perspective, but it is more complicated and much slower than the global signature with the matching speed about 100 000 signature pairs per second.

It appeared that the “global” image identifier is not robust to modifications usually present in multi camera systems, such as view translation, and does not detect related frames obtained from different views nearly at all, so only a “local” version of Image Signature was used in the following experiments.

2.1. Image Signature extraction

The extraction algorithm of “local” part of Image Signature consists of image pre-processing, feature point (key point) detection in image scale-space, local feature extraction, and creation of signature bit-stream [4,5].

In the pre-processing step an image is converted into a grayscale, and resized to the resolution $N \times 256$ or $256 \times N$ preserving aspect ratio, where N is the size of the longer edge after resizing. After that, a low-pass Gaussian filter with 3×3 kernel is applied to smooth the image. This pre-processed image is an input to the next stage: feature point detection.

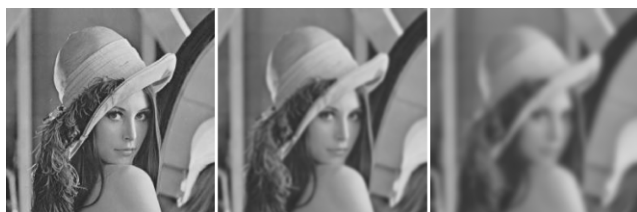


Fig. 1. Original “Lena” image and smoothed representations at scale levels 5 and 11.

The feature points are detected in a scale-space representation of an image. The scale-space representation is obtained by repeatedly smoothing the image with an increased standard deviation of Gaussian filter. The detection is performed in 12-scale levels. Figure 1 shows the example of an original image and its two smoothed versions at scales 5 and 11.

In order to detect a feature point location and scale, two operators are applied on the smoothed images on each scale: Harris detector and Scharr operator. The points for which the response of Harris detector is above a predefined threshold are added to the candidate set of feature points. The out-

put of Scharr operator is used to obtain gradient magnitudes and directions for each detected point in the candidate set. Each feature point is represented by its horizontal and vertical coordinates in image (x , y), scale number, gradient magnitude normalized by scale, and gradient direction. In the next step, a list of up to 80 feature points with the highest magnitudes is selected from the candidate feature points. The selection procedure for a feature point is the following: select the feature point with the strongest magnitude value from unselected feature points and reject the point if it is closer to image edge than a radius r_1 computed from the scale, or it is closer to any previously selected feature point than a radius r_2 (exclusion zone of selected feature point, in MPEG-7 the radius r_2 was set to 12 pixels). The selection procedure continues until number of selected features is 80 or no more features in the candidate set are left.

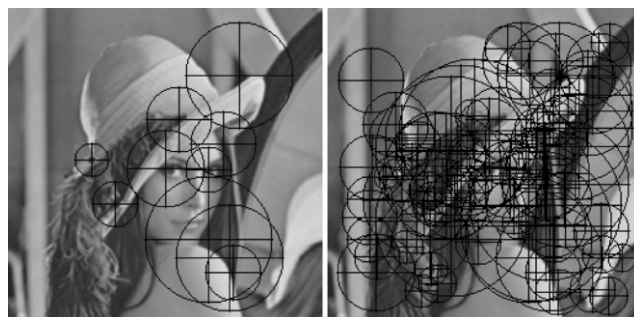


Fig. 2. Feature point locations and circular region sizes: 10 feature points with the highest magnitudes (left image), all 80 feature points (right image).

For each selected feature point a local descriptor is calculated. First, a circular region in image centred at the feature point location with radius dependent on scale is extracted – the higher is the scale the higher is the region radius. If the feature point is detected at various scales, the scale with the highest magnitude is taken for further analysis. Figure 2 shows circular regions in an example image centred at feature point locations; their radii depend on the scale representations. An image with all 80 points is presented, and a second image containing only 10 points with the highest magnitudes is shown for better visibility. The circular region for each feature point is extracted with sub-pixel accuracy and normalized to predefined size by using cubic interpolation. The local descriptor is computed by using Trace Transform, which is a modified version of the method used for computation of global signature. The local descriptor is computed by using 14 functions applied to the output of Trace Transform giving 14 scalar components. Each component is quantized to 8 bits which gives 112 total bits from which 60 bits is selected to form the feature descriptor. Image Signature consists of up to 80 feature point descriptors, each descriptor consists of 80 bits: x position in image (8 bits), y position in image (8 bits), direction (4 bits), and 60 bits of feature descriptor.

2.2. Image Signature matching

Matching of two “local” Image Signatures is carried out to decide if the two images are replicas (or semi-replicas in our case) of each other or not. The matching algorithm takes into account similarity of individual feature point descriptors and their geometric relationships in two matched images. Two images are said to be replica of each other if there exists a set of feature descriptor pairs with distances below a predefined threshold T_{dist} and their geometrical relationships are similar in the two images.

The distance between two individual feature descriptor pairs in two images is calculated as a Hamming distance of 60-bit descriptors. The distances of all unique descriptor pairs from two images are calculated and the descriptor pairs with distance below the threshold T_{dist} are added to the candidate set used in geometrical matching. The geometrical matching is performed if at least 3 descriptor pairs exist in the candidate set. For geometrical matching all combinations of 3 unique descriptor pairs from the candidate set are tested. If the geometrical relationships of 3 feature point pairs within two images meet predefined constraints the images are regarded as replicas.

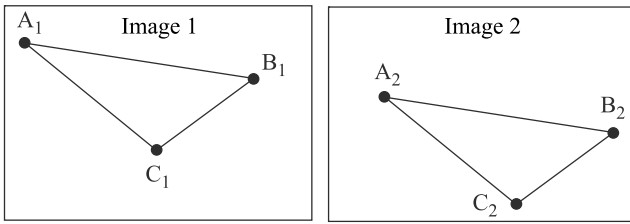


Fig. 3. Geometrical relationships of feature points in two images.

Figure 3 shows three matched image pairs in image 1 and 2. If matched feature points are (A_1, A_2) , (B_1, B_2) and (C_1, C_2) , the geometrical constraints G_a , G_b are calculated from line length ratios in the two images

$$L_i^a = \frac{A_i B_i}{A_i C_i} \quad L_i^b = \frac{A_i B_i}{B_i C_i} \quad i=1,2, \quad (1)$$

$$G_a = \frac{|L_1^a - L_2^a|}{L_1^a + L_2^a} \quad G_b = \frac{|L_1^b - L_2^b|}{L_1^b + L_2^b}. \quad (2)$$

Two thresholds are used for geometrical constraints to decide if the images are replicas, one threshold $T1_{geom}$ for each measure G_a and G_b , and the threshold $T2_{geom}$ for the sum $(G_a + G_b)$.

The calculation of distance is implemented in a 4-stage matching algorithm, designed to increase the matching speed, where each stage has higher computational complexity and the precision for each stage is higher. At each stage some of the images can be discarded as non-replicas avoiding some costly computations for them.

3. Video Signature

The purpose of Video Signature descriptor [3,6] of MPEG-7 is the detection of replicated video content. It was designed to be fast and robust to many video editing operations, both in spatial and temporal domain. The signature supports the detection and localization of duplicated temporal video segments which is present in two video sequences with the minimum duration of about 2 seconds. Moreover, the signature design allows for matching video segments with different frame rates. The signature is robust to video editing techniques such as: text or logo overlay, lossy compression, colour conversion, frame rate modification, analogue recording and recapturing, capturing on camera with additional background, change of video resolution.

The signature consists of 3 elements:

- Frame Signatures (608 bits for each frame);
- Frame Words – a set of 5 compact frame summaries extracted from Frame Signatures, where each word is encoded with 8 bits;
- Bag-Of-Words – representation of a group of frames extracted from Frame Words; the typical Bag-Of-Words represents a segment of 90 frames and it is encoded with 243 bits.

The basic frame descriptor is Frame Signature which describes single video frame in a similar way as Image Signature, but it is designed to be robust to modifications specific to video such as text/logo overlay. Another difference is the extraction time of Frame Signature which is much faster than Image Signature. Fast extraction time is crucial for video content that may contain many thousands of frames in a single video sequence.

3.1. Video Signature extraction

In the first step of Frame Signature extraction the video frame is divided into 32×32 blocks of equal size. The Frame Signature is computed from the relationships between block luminance of predefined regions in video frame. 380 block relationships (called dimensions) within video frame have been defined experimentally and each dimension is quantized to 3 values (0, 1, 2). The elements are encoded with an encoding scheme where five elements are packed to 8 bits, giving a 76-byte signature for 380 elements.

Frame Signature forms a complete description of video content but the matching speed using only these descriptions is low. The matched video sequences can contain thousands of frames and to find matched segments all frames from one video sequence should be matched to all frames of the second video sequence. To speed up the matching process two additional elements are included in the Video Signature: Frame Words and Bag-of-Words. Frame Word is a compact representation of a complete Frame Signature. The formation of Words is based on a projection from 380-dimensional space of Frame Signature to 5-dimensional space of Frame Words. For two frames the distance between two corresponding Words is an approxima-

tion of the distance between two full Frame Signatures. Bag-of-Words is a representation of a video segment containing 90 frames. Bag-of-Words represent words which are present in a video segment and it is extracted as histogram of word occurrences in the temporal segment.

3.2. Video Signature matching

Video Signature matching process returns the matched segments (e.g., segments that are replicas of each other) in two video sequences which may be the whole sequences or any temporal parts of them. The matched segments should have a duration of at least 2 seconds or more to get reliable results. The matched segments may have different frame rates, both known or unknown to the user.

Video Signature matching algorithm consists of 3 stages. The first and second stages operate at segment level of the size corresponding to Bag-of-Words (BoW segment, usually 90 frames). At these stages the matching process may end resulting in a "non-replica" decision, or the matching process goes to the next stage for a more precise matching. The matching algorithm is the following:

- stage 1: BoW segment matching: Bag-of-Words (BoW) from one video sequence are matched to BoW of the second video sequence, if the matching score for a BoW segment pair is above a predefined threshold add the BoW segment pair to the candidate list for stage 2. All BoW segment pairs from the two video sequences are matched to each other. If the candidate list of BoW segment pairs is empty, returns a "non-replica" decision.
- stage 2: frame to frame matching for BoW segment pairs detected in stage 1. In this stage all frames in the detected BoW segment pair are matched to each other by using Frame Signature descriptors. The matrix of distances which are below the predefined threshold are added to the distance matrix of the size 90x90 corresponding to the two BoW segment sizes. The best matched frame pair and corresponding frame rate ratio are detected by using Hough Transform of the distance matrix. If all elements of the distance matrix are zeros, remove the BoW pair from the candidate list; otherwise add the best matched frame pair and corresponding frame rate ratio to the candidate list for stage 3.
- stage 3: frame to frame matching by using best matched frame pair and frame rate ratio obtained in stage 2. In this stage, the frames located just before the best matched frame pair and located just after the best matched frame pair are matched on a frame-to-frame basis to get the precise location of full matched video segments in the two video sequences. The frame pairs from the two video sequences with distances below a predefined threshold are appended to the best matched frame pairs. If the frame rate ratio is not one, the temporal frame positions are interpolated according to the frame rate ratio. The locations of matched segments in the two video sequences are returned as the result of a replica detection.

4. Experiments and results

The aim of the experiments was to check the possibility of using visual signature tools, defined in MPEG-7, for the detection of video streams recorded simultaneously by different cameras in a multi camera system. The test set used in the experiments contained 7 multi camera recording sets [7–9], where each recording set consist of 8 video streams (views) recorded at the same time from different cameras in the system. In each case the cameras were set in a row, with a 20-cm spacing. In most recording sets the cameras were set perpendicularly to each other, but in two sets the optical axes of the cameras were convergent, causing more complex view differences than only horizontal shift. Figure 4 shows example recording sets. The experiments for Image Signature using local feature points and Video Signature have been carried out which are presented in the following subsections.

4.1. Views detection using Image Signature

The experiments of views' detection in a multi camera system were first carried out using the Image Signature description. In the experiments, MPEG-7 reference software has been used for Image Signature extraction and matching. First, Image Signature descriptors of all video frames have been extracted and the matching of corresponding frames was performed to detect semi-replicas. The results contain the number of frames detected as semi-replica where all frame pairs of two video streams from different views (but from one recording set) were compared frame by frame, where only frames with the same position in videos were compared (the first frame to the first frame, the second frame to the second frame, etc.). The tables below present the experiments' results. Tables 1 to 6 contain the number of detected semi-replicas in a view to view test: the first row and the first column contain the view number; other cells contain the number of frame pairs detected as semi-replica for corresponding video views. The cross-sequence verification was also performed in order to check if the images from various multi view sequences are not detected as semi-replicas of each other and there were no false detection.

Table 1. Number of semi-replicas in view to view test – "ballroom" (250 frames).

View	1	2	3	4	5	6	7
0	236	167	123	96	40	43	18
1		233	152	156	100	43	37
2			220	149	123	78	66
3				220	167	150	88
4					224	159	110
5						223	184
6							224



Fig. 4. Example multi-view sequences: “ballroom” [7], “exit” [7], “vassar” [7], “race”, “pozstreet” [8], “jungle” [9].

Table 2. Number of semi-replicas in view to view test – “exit” (250 frames).

View	1	2	3	4	5	6	7
0	207	130	50	1	0	0	0
1		189	124	5	2	0	0
2			208	50	11	0	0
3				179	46	1	0
4					133	4	0
5						43	1
6							20

Table 3. Number of semi-replicas in view to view test – “vassar” (250 frames).

View	1	2	3	4	5	6	7
0	250	250	239	244	185	43	118
1		250	248	208	80	84	64
2			250	240	156	154	85
3				250	227	174	194
4					250	189	226
5						250	240
6							250

Table 4. Number of semi-replicas in view to view test – “race” (300 frames).

View	1	2	3	4	5	6	7
0	300	297	291	273	253	235	161
1		300	299	292	282	269	165
2			300	298	294	272	240
3				300	300	292	287
4					300	299	298
5						300	299
6							300

Table 5. Number of semi-replicas in view to view test – “pozstreet” (450 frames).

View	1	2	3	4	5	6	7
0	450	450	448	448	436	433	431
1		450	450	450	450	447	401
2			450	450	450	449	447
3				450	450	450	450
4					450	450	450
5						450	450
6							450

Table 6. Number of semi-replicas in view to view test – “jungle” (250 frames).

View	1	2	3	4	5	6	7
0	1	0	0	0	0	0	0
1		0	0	0	0	0	0
2			0	0	0	0	0
3				0	0	0	0
4					0	0	0
5						0	0
6							0

The presented results show that the performance of Image Signature for a view detection strongly depends on the setting of the multi camera environment and the characteristics of a recorded scene. The Image Signature descriptor was designed to detect modified versions of the same source image. However, in case of a multi view camera system there is no one source image that is modified, but different cameras give images with similar content. When the scene is “far” (e.g., objects are far away from the cameras) and the cameras are set perpendicularly, the images from different

views are just shifted version of each other. In such cases an image signature tool is quite good (e.g., “vassar”, Table 3, or “race”, Table 4) and even very good (“pozstreet”, Table 5). However, the detection rate decreases as the distance between cameras increases because the offset between images increases. Also objects close to the camera cause greater differences between images taken from different views – object offset is greater. Since there is the content that is visible only in one of the views, so different feature points can be found. In such cases, the performance of the detection even for neighbouring views is significantly worse (e.g., last views of exit sequence, Table 2). Convergent optical axes of cameras cause different types of modifications, including changing of geometrical relations between feature points, so in such cases (“jungle”, Table 6, the result for other such recording set was similarly poor) related semi-replica images were not detected nearly at all.

4.2. Views detection using Video Signature

Next, the Video Signature descriptor was used for views' detection in a multi camera system. MPEG-7 reference software has been used for the video signature extraction and matching. First, Video Signature descriptors of all views have been extracted and the matching of the descriptors was performed to detect semi-replicas. The tables below present the results of experiments. For each view pair within multi camera video sets the detection result is presented (“semi-replica” decision is marked with “+” sign, “non-replica” with “–”). The cross-sequence verification was also performed, in order to check if the views from various multi view sequences are not detected as semi-replicas of each other and there were no false detection.

Table 7. Result of video semi-replica detection in view to view test – “ballroom”.

View	1	2	3	4	5	6	7
0	+	+	–	–	–	–	–
1		+	+	–	–	–	–
2			+	–	–	–	–
3				+	–	–	–
4					+	–	–
5						+	+
6							+

Table 8. Result of video semi-replica detection in view to view test – “exit”.

View	1	2	3	4	5	6	7
0	+	–	–	–	–	–	–
1		+	–	–	–	–	–
2			–	–	–	+	–
3				+	–	–	–
4					+	–	–
5						–	–
6							–

Table 9. Result of video semi-replica detection in view to view test – “vassar”.

View	1	2	3	4	5	6	7
0	–	–	–	–	–	–	–
1		+	–	–	–	–	–
2			+	+	–	–	–
3				+	–	–	–
4					+	+	–
5						+	+
6							+

Table 10. Result of video semi-replica detection in view to view test – “race”.

View	1	2	3	4	5	6	7
0	+	+	+	–	+	–	+
1		+	+	–	+	+	+
2			+	+	+	+	+
3				+	+	+	+
4					+	+	+
5						+	+
6							+

Table 11. Result of video semi-replica detection in view to view test – “pozstreet”.

View	1	2	3	4	5	6	7
0	+	+	+	+	+	+	+
1		+	+	+	+	+	+
2			+	+	+	+	+
3				+	+	+	+
4					+	+	+
5						+	+
6							+

Table 12. Result of video semi-replica detection in view to view test – “jungle”.

View	1	2	3	4	5	6	7
0	+	–	–	–	–	–	–
1		–	–	–	–	–	–
2			–	–	–	–	–
3				–	–	–	–
4					+	–	–
5						–	–
6							+

As for Image Signature, the results for video semi-replica detection by using Video Signature strongly depend on the setting of a multi camera environment. Again, in the multi camera system there is no one modified video sequence, but multiple views with similar content. Although, for recording sets, where the cameras were set perpendicularly, the neighbouring views were mostly detected as semi-replicas. Especially good performance was obtained for “race” and “pozstreet” sequences. In this sequences the scene is really “far”, there are no object close to the cameras, so the scene offset between views (and hence the content presented only in one of the views) is relatively small, as

a result almost all views are detected as semi-replicas of each other (within recording set, cross-set views are not detected as replicas, of course), see Tables 10 and 11. On the contrary, object close to the camera, as in the last views of “exit” sequence, could cause that even neighbouring views are not detected (Table 8). Similarly, as for the result of Image Signature, the view detection does not work nearly at all in systems with convergent cameras (Table 12, similarly poor results were obtained for other such recording set) because of the same reasons.

It is worth noticing, that the detection rate could be increased by modifying the thresholds used in the descriptor matching process. However, this also increases a risk of false detection, e.g., taking as a semi-replica two video sequences coming from various multi camera systems.

5. Conclusions

In this paper the results of experiments for semi-replica detection in multi camera systems using MPEG-7 visual signature tools have been presented. The aim was to detect video recordings from different views recorded simultaneously in a multi camera system. Two signature tools have been used: Image Signature descriptor and Video Signature descriptor. The detection results strongly depend on the multi camera system settings and the characteristics of recorded scene. These tools can be used for the reliable detection of multi camera views only in some restricted scenarios of a multi camera system setup, such as systems having cameras with perpendicular optical axes and scene distant from the cameras. These tools cannot be used in generic application of detecting multi camera views. Such application would require a new description designed especially for

such search scenario, taking into account not only popular image and video processing techniques, but also various types of image modifications that occur between views in multi camera systems.

References

1. B.S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, edited by John Wiley & Sons Inc., New York, 2002.
2. M. Bober and P. Brasnett, “MPEG-7 visual signature tools”, *IEEE I. Conf. Multimedia and Expo*, pp. 1540–1543, New York, 2009.
3. P. Brasnett, S. Paschalakis, and M. Bober, “Recent developments on standardisation of MPEG-7 visual signature tools”, *IEEE I. Conf. Multimedia and Expo*, pp. 1347–1352, Singapore, 2010.
4. P. Brasnett and M. Bober, “Fast and robust image identification”, *19th Int. C. Patt. Recog.*, pp. 871–876, Tampa, 2008.
5. P. Brasnett and M. Bober, “A Robust visual identifier using the trace transform”, *Proc. Visual Information Engineering Conference*, pp. 25–27, London, 2007.
6. M. Bober and S. Paschalakis, “MPEG image and video signature”, in *MPEG Representation of Digital Media*, pp. 81–95, edited by L. Chiariglione, Springer, New York, 2012.
7. A. Vetro, M. McGuire, W. Matusik, A. Behrens, J. Lee, and H. Pfister, “Multiview video test sequences from MERL”, *ISO/IEC JTC1/SC29/WG11 doc. M12077*, Busan, Korea, 2005.
8. M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner, “Poznań multi-view video test sequences and camera parameters”, *ISO/IEC JTC1/SC29/WG11 doc. M17050*, Xian, China, 2009.
9. A. Smolic, “MPEG-3DAV FhG-HHI Test Data”, available at <https://www.3dtv-research.org/downloadlist.php?dat&s=61>