Efficient video streaming of 360° cameras in Unmanned Aerial Vehicles: an analysis of real video sources

Stefania Colonnese, Francesca Cuomo

Department of Information Engineering,
Electronics and Telecommunications
”Sapienza” University of Rome
Italy
stefania.colonnese@uniroma1.it
francesca.cuomo@uniroma1.it

Ludovico Ferranti, Tommaso Melodia

Department of Electrical
and Computer Engineering
Northeastern University
Boston, MA, USA
ferranti@ece.neu.edu
melodia@ece.neu.edu

***Abstract*—Video streaming data acquired by Unmanned Aerial Vehicles is an innovative service that will be leveraged by several applications ranging from entertainment and surveillance to disaster recovery. 360° cameras provide unprecedented visual information and enable services to a novel level of immersive experience. However, 360° video sources are not still fully characterized, and this holds especially true for drone mounted 360° video sources. This paper presents a thorough analysis of the video traffic associated to several 360° camera sequences, acquired by a pedestrian held camera as well as by a drone mounted camera in various environments and lighting conditions. A fine-grained rate distortion analysis is presented for both video frames and video chunks, thus making this study relevant for HTTP-based video streaming services. The analysis is completed by making publicly available a dataset of 360° video traffic traces that can be used for numerical simulations of Unmanned Aerial Vehicles providing 360° video streaming services.**

***Index Terms*—360° camera, UAV, videostreaming**

I. Introduction

Video streaming data acquired by Unmanned Aerial Vehicles is a boosting service which will be exploited by several applications ranging from entertainment and surveillance to disaster recovery. 360° cameras provide unprecedented visual information and trigger the service to a novel level of immersive experience [3], [9]. Video traces characterization plays a fundamental role in system design issues [7], together with the characterization of aggregated video streaming traffic [2]. Still, while the literature reports 3D and multiview statistical characteristics of multiview video, which are significantly different from the single-view video [6], [8], 360° video sources are not still characterized, and this holds especially true for drone mounted 360° video sources [3]. The goal of this paper is to present a thorough analysis of the video traffic associated to several 360° camera sequences, acquired in different conditions, including a drone mounted camera, and encoded at different rates. The traces dataset of 360° video is made publicly available at the database web-site *www.ece.neu.edu/wineslab/360_stats/360stats.zip*.

In the following, the relevant rate-distortion features of the encoded video sequence of the acquired database are reported. Besides, the main statistics of the parsed encoded video packets (chunks) which are the data structures actually used in streaming services are summarized. Finally, the video chunk size histograms are fitted by heavy-tail Gamma distributions, and the corresponding parameters are reported. The analysis provides compact empirical models of the encoded video source ready for application in numerical simulation systems.

II. 360° Video Sequences Features

We have acquired $L = 11$ video sequences using a Ricoh Theta S Camera, whose 360° video acquisition system is build by two fish eye lenses whose field of views are slightly overlapping. Ricoh Theta S output, as most of the commercial cameras, provide a nearly lossless image output. Complex scenes can be captured through a rig of cameras, which provides a rich image quality, although its not optimal in high mobility scenes. The Ricoh Theta S was preferred to a rig of cameras as it presents a lighter solution and can easily be carried in a small Unmanned Aerial Vehicle without compromising its stability and movements. Shooting has been realized in different environments, lighting conditions, and at different camera speeds.

The first set of acquired sequences have been acquired by a pedestrian held camera, still or moving indoor and outdoor. In video A, a pedestrian is holding the 360° camera and walks around in a crowded building hall. The environment is architecturally heterogeneous and the visual content of each frame is rich. The lighting comes from both neon and daylight. The camera is subjected to small oscillation due to the walking pace of the pedestrian. In video B, the pedestrian is walking in the street with regular speed. A large portion of the frames is taken by the sky and there are few moving cars in the background. The lighting is purely daylight. In video C, the pedestrian is at a crossroad and a high number of moving objects are present. The lighting is again daylight. In video D the pedestrian is walking in an art gallery. Artificial lighting and daylight are present, although the environment is darker and more homogeneous with respect to previous footages. Video E is very similar to video D, although in the last moments, the pedestrian leaves the 360° camera in a fixed position and walks around the art gallery. The lights are purely artificial. In video F, the position of the 360° camera is fixed at the center of a large crowded room in an art gallery and a large number of moving objects is present. Daylight comes from the ceiling and artificial lighting is also present. In video G, the pedestrian is walking in a garden and leaves the 360° camera in a bench for half the duration of the video. The lighting is daylight and the environment is very colorful. We report in Tab. I the main video features of these sequences, including the camera speed, the ambience (room, open space, far field buildings), the lightening conditions (noon, evening, artificial), the number of moving people within the scene.

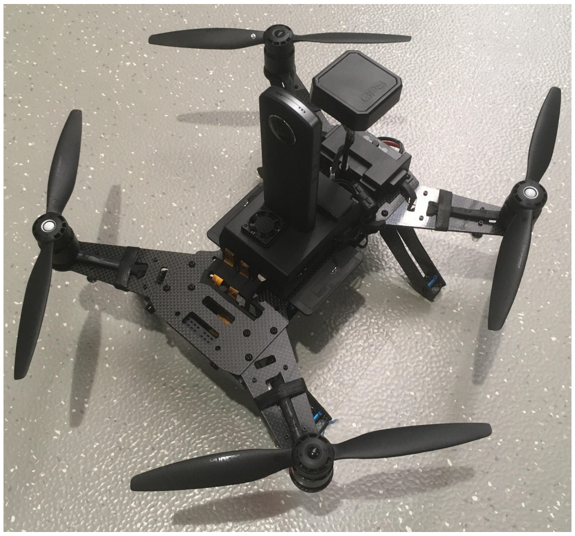

Fig. 1. Ricoh Theta S 360° camera mounted on Intel Aero drone

The second set of acquired sequences have been captured by a drone mounted camera. The Ricoh Theta S Camera was fixed on the body of an Intel Aero drone as showed in Fig. 1 and videos were acquired while flying in indoor and outdoor locations. In video H, the drone is flying at low altitude in an indoor office space. The camera is fixed and movements are linear. Light comes from neon and the environment is heterogeneous. Video I was acquired in the same building hall of video A, but during nighttime and with almost no people moving. Video J and K were acquired outdoor in a cloudy day on a hilltop. Light is neutral and equally spread from the clouds. The drone moves faster in video J while the flight style is more smooth in video K. The original video sequences characteristic are summarized in Tab.II. We notice that the overall set of sequences present a significant variety in terms of complexity and type of the framed video scene.

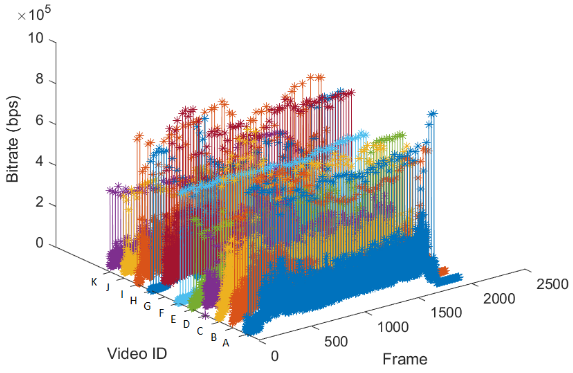

Fig. 2. Instantaneous bitrate for all footages with QP level 32

III. Statistical analysis of encoded 360° video

This section summarizes the main results of the statistical analysis conducted on the sequence database, after suitable encoding described in the followings.

A. Video encoding settings

The sequences in Tables I, II have been encoded at a fixed quality using the *ffmpeg* library [1] . Specifically, we have used the *x264* libs, since the H.264 video encoding standard has lower complexity than the H.265 one, and it is therefore more suited to be implemented on a DSP system on board of a drone. For immersive 360° rendering, sequences acquired by fish eye lens cameras are usually stitched together by implementing frame by frame processing. Due to the highly computationally demanding characteristic of the video frames stitching operation, we have assumed that it is not realized on board of the drone. Therefore, in this work we refer to encoding directly the output of the camera, which consists of two fish eye lens sequences. For each sequence, the output format of Ricoh Theta is a video sequence, pre-encoded at a rate much higher than the final targeted video quality, and straightforwardly fed to the coder. Further work will address the encoding of on-board stitched video camera frames [5]. Due to the focal structure of fish eye lenses and the subsequent need to stitch the captured videos, classical motion compensation schemes can't be applied to videos captured with such cameras.

B. Rate distortion analysis

Here, the relevant rate-distortion features observed on the $L = 11$ encoded video sequences of the database are reported. In all the referred encoding experiments, the GOP length is set to $N_{GOP} = 25$. In the following, $K = 5$ different quantization parameters (QP) values have been considered, namely $QP \in \{26, 32, 38, 44, 51\}$ covering different rate and quality conditions. We denote the frame size in bits of the l-th sequence encoded at the k-th quality level as

$$b_n^{(l,k)},\ n = 0, \cdots N-1, k = 1, \cdots K, l = 1, \cdots L$$

TABLE I
PEDESTRIAN HELD CAMERA SEQUENCES

Video ID	***A***	***B***	***C***	***D***	***E***	***F***	***G***
Setting	Indoor	Outdoor	Outdoor	Indoor	Indoor	Indoor	Outdoor
fps	30	30	30	30	30	30	30
Resolution	1080 1920	1080 1920	1080 1920	1080 1920	1080 1920	1080 1920	1080 1920
Camera speed	0-5 Km/h	0-5 Km/h	0-5 Km/h	0-5 Km/h	0-5 Km/h	0-5 Km/h	0-5 Km/h
Field of view	building hall	street	crossroad	art gallery	art gallery	crowded room	garden
Lighting conditions	daylight + neon	daylight	daylight	mixed lights	indoor lights	mixed lights	daylight
Number of MO	16	12	19	7	5	26	3

TABLE II
DRONE MOUNTED CAMERA SEQUENCES

Video ID	***H***	***I***	***J***	***K***
Setting	Indoor	Indoor	Outdoor	Outdoor
fps	30	30	30	30
Resolution	1080 1920	1080 1920	1080 1920	1080 1920
Camera speed	0-25 Km/h	0-25 Km/h	0-25 Km/h	0-25 Km/h
Field of view	Office	building hall	hilltop	hilltop
Lighting conditions	neon lights	neon lights	cloudy daylight	cloudy daylight
Number of MO	1	4	1	3

being $N = 2100$ the sequence length.

We firstly summarize the overall results in Tables III, IV where we report the average rate [kbps] and the Peak to Signal Ratio (PSNR), defined as PSNR [dB] $= 10 \log \left(\frac{MSE}{255^2} \right)$, where MSE is the Mean Square Error introduced by the coding stage. We recognize that the dataset covers several encoding rates, and can therefore be adopted for a variety of simulation scenarios.

With respect to a more generic rate distortion analysis, we present a more refined frame by frame rate distortion analysis in Figs. 3 and 4. Specifically, the Figs. 3 - 4 plot the frame size $b_n^{(l,k)}$ in bits versus the achieved frame PSNR in dB at various QPs for a PH camera sequence and a DM camera sequence, respectively. In both cases, we recognize the well known exponential trend of the rate vs quality. More in general, from these and other results not reported here for compactness' sake, some regular trends are observed. As for the PH sequence, it is clearly visible the typical less efficient encoding of Intra frames, which exhibit a less favourable rate distortion behaviour wrt predicted frame at all quality levels, and which appear as separate clusters of points in the plot. As for the DM sequence, two parameters affect the rate-distortion encoding features, namely the camera speed and the environment characteristics. Specifically, the DM camera speed is typically higher and all the frames present a high amount of innovation, making Intra and predicted frames more comparable. On the other hand, it must be noticed that in outdoor conditions with open landscapes and far field objects the rate-distortion characteristic of the sequence become more favourable.

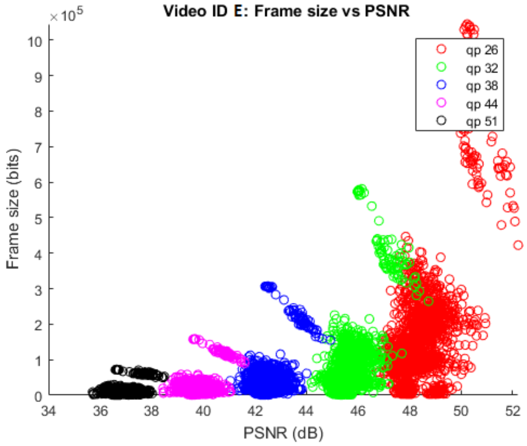

Fig. 3. Framesize vs PSNR at all different QP levels for PH video

C. Parsed encoded video features

In video streaming services, the encoded video stream is typically parsed into independently decodable packets suitable for independent transmission. In HTTP-based video streaming services, the packets, also known as chunks or segments, are built by few consecutive GOPs, covering 2-4s video, and requested by clients through suitably scheduled consecutive HTTP-Get requests. The flow of requested consecutive chunks represents the actual user plane load in streaming services. Thereby, in the following we present the main features and statistics of the parsed encoded video packets (chunks) of the database sequences. Specifically, we define the chunk size as

$$\lambda_{\nu}^{(l,k)} = \sum_{i=0}^{N_{GOP}-1} b_{i+\nu N_{GOP}}^{(l,k)}, \; \nu = 0, \cdots \lfloor N/N_{GOP} \rfloor - 1$$

TABLE III

RATE (KBPS) AND PSNR (DB) FOR PH-CAMERA SEQUENCES

Video ID	***A***	***B***	***C***	***D***	***E***	***F***	***G***
Avg bitrate (kbps) @ $QP = 26$	8455.9	7849	7562.4	6386.3	4822.7	2248.3	6938.5
PSNR @ $QP = 26$ (dB)	43.57	43.93	43.76	44.77	44.82	44.44	43.32
Avg bitrate @ $QP = 32$	4007.1	3655.7	4938.3	3124.6	2245.0	1053.7	3552.1
PSNR @ $QP = 32$	39.74	40.18	39.72	41.58	41.62	41.46	39.47
Avg bitrate @$QP = 38$	1952.2	1749.7	2396.3	1540.6	1115.3	556.4	1674.9
PSNR @ $QP = 38$	36.24	36.74	36.01	38.28	38.38	38.21	35.83
Avg bitrate @ $QP = 44$	1060.2	915.4	1268.9	809.9	601.8	301.8	815.7
PSNR @ $QP = 44$	32.86	33.46	32.60	34.92	35.00	34.85	32.52
Avg bitrate @ $QP = 51$	554.40	453.7	646.16	402.69	306.75	151.74	373.41
PSNR @ $QP = 51$	28.93	29.63	28.6	30.64	30.92	30.95	28.97

TABLE IV

RATE (KBPS) AND PSNR (DB) FOR DM-CAMERA SEQUENCES

Video ID	***H***	***I***	***J***	***K***
Avg bitrate (kbps) @ $QP = 26$	5653.27	9426.82	3833.08	4705.38
PSNR @ $QP = 26$ (dB)	43.95	43.74	45.09	45.127
Avg bitrate @ $QP = 32$	2551.27	4548.84	1642.91	2063.99
PSNR @ $QP = 32$	40.59	39.90	41.93	41.70
Avg bitrate @$QP = 38$	1243.76	2189.80	781.17	955.76
PSNR @ $QP = 38$	37.26	36.25	38.82	38.46
Avg bitrate @ $QP = 44$	649.07	1122.90	354.08	428.65
PSNR @ $QP = 44$	33.92	32.76	35.67	35.26
Avg bitrate @ $QP = 51$	322.70	555.91	128.01	155.61
PSNR @ $QP = 51$	30.06	28.79	31.69	31.33

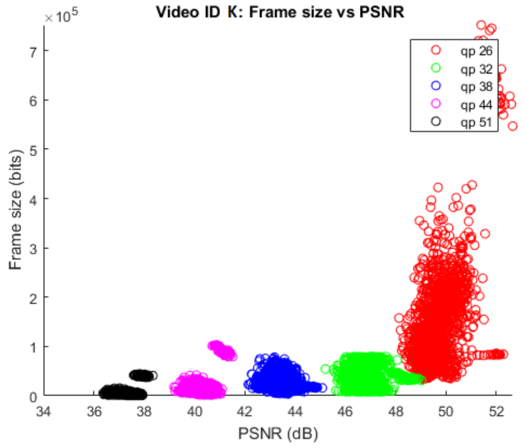

Fig. 4. Framesize vs PSNR at all different QP levels for DM video

and analyze the temporal dynamics of the chunk traces $\lambda_\nu^{(l,k)}$, $\nu = 0, \cdots \lfloor N/N_{GOP} \rfloor - 1$. Firstly, in Fig.5 we plot the temporal pattern of the chunk size $\lambda_\nu^{(l,\bar{k})}$ (bits) vs the chunk index ν for quality level $\bar{k}$ corresponding to $QP = 32$ and for all the $L = 11$ sequences. As expected, the large (wrt the average value) temporal variations observed on the frame sizes are smoothed out on the chunk traces.

Secondly, we analyzed the video chunk size histograms. It is well recognized in the video source modeling literature that video packets are well fitted by heavy-tail distributions. We have employed the Gamma probability density function (pdf):

$$p_\Lambda(\lambda; \alpha, \beta) = K \cdot (\lambda/\beta)^{\alpha-1} \cdot e^{-\lambda/\beta} \cdot u_{-1}(\lambda) \qquad (1)$$

being $K = 1/\beta\Gamma(\alpha)$. in order to fit the chunk size pdf of each considered sequence at any given quality level.

The overall Gamma fitting parameters are shown in the synoptic Tables V and VI, referring to PH and DM sequences respectively. The Gamma pdf usefully characterizes the heavy tailed, asymmetrical ($\alpha > 1$) behaviour of the observed packets distribution. This is exemplified in Figs. 6 and 7, where two histograms obtained on a PH and a DM sequence in Figs. 6 and 7, are respectively shown, together with their fitting Gamma pdfs.

Even though the Gamma pdf fitting is coarse, it provides a useful building block for characterizing the behaviour of the video source at a given average rate. Composite sources representing adaptive or interactive services may then be built by combining the fixed rate fitting with suitable models of the communication channels [4] or of the user behaviour [6].

Thereby, the above analysis provides compact empirical models of the parsed, encoded video source that come handy for application in network design simulation systems.

IV. CONCLUSION

In this paper we present a thorough analysis of the video traffic associated to several 360° camera sequences, acquired in different conditions, including a drone mounted camera. The relevant rate-distortion features of the encoded 360°

TABLE V

GAMMA α, β FOR PH-CAMERA SEQUENCES

Gamma α, β / ID	***A***	***B***	***C***	***D***	***E***	***F***	***G***
$QP = 26$	15.75, 867408	21.57, 606697	24.52, 648663	9.06, 1119615	6.29, 1277376	21.87, 165399	25.64, 4291332
$QP = 32$	12.29, 526355	14.71, 414464	18.84, 429793	7.78, 637759	6.22, 601519	21.26 , 79634	2.19 , 2566866
$QP = 38$	11.71, 269005	16.06, 181641	22.18, 177467	8.26, 296006	6.96, 266984	21.5, 41572	2.22, 1191324
$QP = 44$	12.21, 140059	20.66, 73889	26.58, 78420	9.97, 128950	6.95, 144286	20.58, 23554	2.27, 568646
$QP = 51$	12.13, 73850	25.65, 29547	26.48, 40149	11.25, 56930	6.55, 78189	19.07, 12805	2.14 , 276502

TABLE VI

GAMMA α, β FOR DM-CAMERA SEQUENCES

Gamma α, β / ID	***H***	***I***	***J***	***K***
$QP = 26$	6.47, 1479464	46.09.57 340199	20.02 320345	17.98 432167
$QP = 32$	5.86 , 738199	24.36 , 309909	16.30 , 168959	11.20 , 303144
$QP = 38$	6.20 , 339875	21.54 , 168698	183.9 , 71297	10.76 , 146274
$QP = 44$	5.81 , 189433	24.57 , 75901	18.20 , 32668	9.72 , 72869
$QP = 51$	5.42 , 101003	27.25 , 3.3902	21.54 , 9985	9.83 , 26152

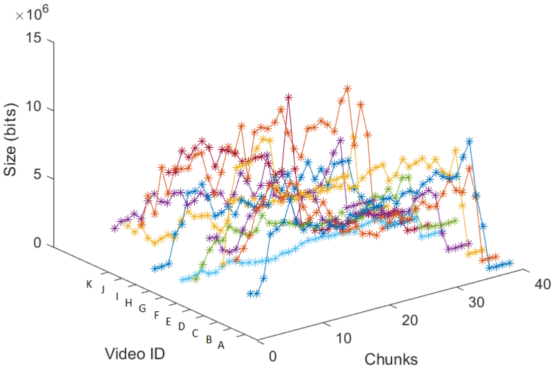

Fig. 5. Chunk size vs chunk indexes for all videos ($QP = 32$)

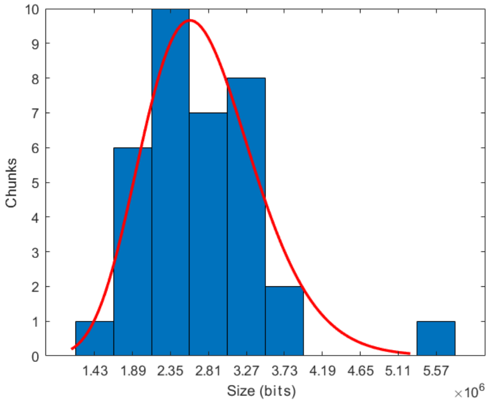

Fig. 6. Histogram of the chunk sizes for DM video J @ $QP = 32$ (blue bars) and best fitting Gamma Distribution (red line).

video sequences are reported, as well as the main statistics of the encoded video packets (chunks) which are used in streaming services. The video chunk size histograms are fitted by heavy-tail Gamma distributions, and the fitting parameters are reported. Finally, the chunk size traces of 360° video are made publicly available for numerical simulations. Future work will concentrate on acquiring more footages in different environments and comparing them with H.265 codec. Moreover the possibility to benefit from image rectification and warping prior the stitching process will be investigated. A streaming simulator is currently under development to provide and test a more refined model of video streaming using 360° cameras, which will be useful to compare with existing 360° video streaming models.

REFERENCES

[1] https://www.ffmpeg.org/.

[2] Arkadiusz Biernacki. Analysis and modelling of traffic produced by adaptive http-based video. *Multimedia Tools and Applications*, 76(10):12347–12368, 2017.

[3] Marc Van den Broeck, Fahim Kawsar, and Johannes Schöning. It's all around you: Exploring 360 video viewing experiences on mobile devices. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 762–768. ACM, 2017.

[4] S. Colonnese, P. Frossard, S. Rinauro, L. Rossi, and G. Scarano. Joint source and sending rate modeling in adaptive video streaming. *Signal Processing: Image Communication*, 28(5):403–416, 2013.

[5] C. Lyu, J. Peng, W. Zhou, S. Yang, and Y. Liu. Design of a high speed 360-degree panoramic video acquisition system based on fpga and usb 3.0. *IEEE Sensors Journal*, PP(99):1–1, 2017.

[6] Lorenzo Rossi, Jacob Chakareski, Pascal Frossard, and Stefania Colonnese. A poisson hidden markov model for multiview video traffic. *IEEE/ACM Transactions on Networking (TON)*, 23(2):547–558, 2015.

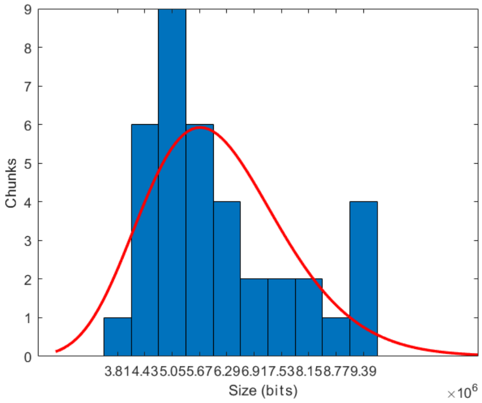

Fig. 7. Histogram of the chunk sizes for PH video B 20 @ $QP = 32$ (blue bars) and best fitting Gamma Distribution (red line).

[7] Patrick Seeling and Martin Reisslein. Video traffic characteristics of modern encoding standards: H. 264/avc with svc and mvc extensions and h. 265/hevc. *The Scientific World Journal*, 2014, 2014.

[8] Savera Tanwir, Debanjana Nayak, and Harry Perros. Modeling 3d video traffic using a markov modulated gamma process. In *Computing, Networking and Communications (ICNC), 2016 International Conference on*, pages 1–6. IEEE, 2016.

[9] Huyen TT Tran, Nam Pham Ngoc, Cuong T Pham, Yong Ju Jung, and Truong Cong Thang. A subjective study on qoe of 360 video for vr communication. In *Multimedia Signal Processing (MMSP), 2017 IEEE 19th International Workshop on*, pages 1–6. IEEE, 2017.