INTERNATIONAL ORGANISATION FOR STANDARDISATION

ORGANISATION INTERNATIONALE DE NORMALISATION

ISO/IEC JTC 1/SC 29/WG 2

MPEG TECHNICAL REQUIREMENTS

Sformatowano: Górny: 2,25 cm

ISO/IEC JTC 1/SC 29/WG 2 m 58740

Online – January 2022

Sformatowano: Włoski (Włochy)

Title: **[VCM] Recovering full CDVS description from compressed image bitstream using partial description**

Sformatowano: Wyjustowany, Wcięcie: Z lewej: 0 cm, Wysunięcie: 2,49 cm

Source:
Marek Domański, Sławomir Różek, Dominik Cywiński,
Jakub Szekiełda, Sławomir Maćkowiak, Tomasz Grajek

Poznań University of Technology, Poznań, Poland

Sformatowano: Do lewej

~~1.~~ Abstract

Sformatowano: Wcięcie: Z lewej: 0,63 cm, Bez punktorów lub numeracji

This exploratory contribution deals with recovering of the nearly full CDVS description from compressed bitstreams of images. In the decoder, the description is obtained from decoded image augmented by some differential description sent as side information together with the actual image bitstream. The experimental results are provided for VVC intra coding.

~~2.~~1. Introduction

Here, the contribution is aimed at joint coding of images and features. More precisely, this work is done in the context of joint coding of images and features where the decoded images may be viewed by humans whereas the features obtained in the decoder ~~are~~ ~~hav~~are to be consumed by computer vision software.

The decoded images suffer from some degradations due to strong compression. When the features are derived from the compressed images, these degradations yield losses in the characteristic points in the SIFT or CDVS descriptions [2-4]. In order to restore the high quality features, some side information needs to be transmitted together with the compressed image bitstream.

In the ~~document~~ paper [1] and in the documents [3], we have already proposed to extract features from both the original image and the decoded image in the encoder. ~~When~~ Assuming that the features are related to characteristic points, like in SIFT / CDVS descriptions, only the difference between two sets of characteristic points needs to be transmitted (Fig.1) as the side information together with the image bitstream. In the decoder, nearly the full set of characteristic points / features is reconstructed as a sum of the sets of transmitted features and those extracted from the decoded video. Some keypoints may be lost due to video compression, and they are sent as side information.

Therefore, the idea is to transmit two bitstreams:
- the compressed image,
- the differential description of the keypoints.

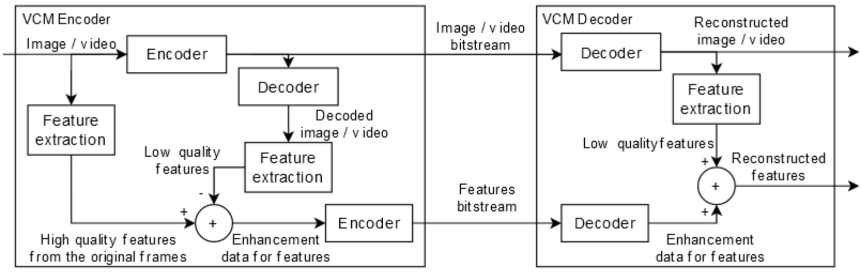


Fig. 1. The proposed VCM system.

Here, we further develop the idea from the paper [1] in the context of CDVS descriptions and VVC intra coding whereas the classic SIFT keypoints were considered in the contribution [1]. The preliminary results have been already presented also in the MPEG documents:
- the document [2] where the loss of the SIFT characteristic points was studied in the context of HEVC and VVC video coding;
- the documents [3, 4] where the number of additional SIFT characteristic points is studied for the transmission of the side information.

In the paper [5], the concept of the differential keypoint coding was proposed in another context of exploitation of the inter-frame redundancy in the descriptions of the consecutive frames of video.

Here, we study the problem for still images compressed using VVC intra encoders [6-~~8~~9].

~~3.~~2. Proposed coding system

In the proposed system for video coding for machines, we consider SIFT keypoints with their descriptions as features.

In the VCM encoder, the features are derived from both the original image and the decoded image. Selected are those features that cannot be derived from the compressed representation of the image. In other words, identified are those features (keypoints) that are derivable from the original image but are not derivable from the decoded features. Only such features (keypoints) are transmitted in the feature bitstream (Fig. 1). Due to bitstream limitations, the feature bitstream may be further reduced in such a way that some features are skipped when they are classified as the least important.

The image is decoded using some standard technique, e.g. VVC Intra [6-9~~7, 8~~].

In the decoder, the image is retrieved, and the features are derived from the decoded image. These features are denoted as “low-quality features” as some keypoints are lost or heavily degraded due to compression. Therefore, the nearly original features / keypoints are obtained by merging of the “low-quality features” and decoded features / keypoints transmitted in the feature bitstream (Fig. 1). The reconstruction is usually not perfect as some descriptors derived from the decoded image are somewhat distorted. Nevertheless, we are going to demonstrate that the feature obtained in the decoder are quite close to those obtainable form the original image.

~~CDVS mode~~	~~Maximum size of CDVS bitstream~~
~~1~~	~~512 bytes~~
~~2~~	~~1024 bytes~~
~~3~~	~~2048 bytes~~
~~4~~	~~4096 bytes~~
~~5~~	~~8192 bytes~~
~~6~~	~~16384 bytes~~

keypoints and their descriptors using standard CDVS implementation [11~~1~~], and we use CDVS syntax to transmit the differential description in the feature bitstream.

Sformatowano: Cytat

Sformatowano: Czcionka: 11 pkt, Pogrubienie

Sformatowano: Czcionka: 11 pkt, Pogrubienie

Sformatowana tabela

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

~~5.~~3. The experiment ~~ntal setup~~description and the results

In the experiment, we use the MPEG CDVS implementation. Therefore, we observe the CDVS syntax limitations upon the capacity of the keypoint descriptions (Table 1). We use this standard CDVS bitstream syntax to transmit differential description. Therefore, this side information is readable by standard CDVS decoders.

Tab~~.~~le 1. Maximum size of CDVS bitstream generated in different modes.

CDVS mode	Maximum size of CDVS bitstream
1	512 bytes
2	1024 bytes
3	2048 bytes
4	4096 bytes
5	8192 bytes
6	16384 bytes

Sformatowano: Czcionka: 12 pkt

Sformatowano: Czcionka: 12 pkt

Sformatowano: Czcionka: 12 pkt

Sformatowano: Czcionka: 12 pkt

Sformatowano: Czcionka: 12 pkt

Sformatowano: Czcionka: 12 pkt

Sformatowano: Czcionka: 12 pkt

Sformatowano: Czcionka: 12 pkt

The setup of the experiment is the following.

Firstly, from the OpenImagesV6 [12] all images with all resolutions between 1920x1088 and 1280x720 ~~has been~~are selected. From that set the biggest (in terms of jpg file size) 10 000 images ~~have been~~ are chosen as a test data for the described experiment. These images (originally in jpeg format) ~~have been~~are converted to YUV 4:2:0 10bpp format, encoded and decoded using VVC (VTM 12.0) [8,9], and converted back to jpeg format (with best quality). The images are encoded using the VVC ~~encoding have been done in~~encoder [9] in “main10” profile, “all intra” mode.

All computations related to descriptors have been done using CDVS Test ~~m~~Model in Matlab and C++ programming language. As a measure of the feature~~s~~ similarity the “local score” from the CDVS Match function have been chosen, which is the score of comparison of local descriptor sets.

3.1. Transmission of original features

Sformatowano: Nagłówek 2

The first experiment is aimed at estimation of the local score of similarity between the maximum description (mode 6) derived from the original picture and the limited description (modes 1-5) derived also from the original picture. The local score of similarity between the maximum description (mode 6) and the description (modes 6) is estimated for the reference as the maximum possible value of the local score of similarity. The values are estimated as averages over the whole set of test images used.

As a reference scenarios for the experiment two possibilities ~~have bee~~are~~n~~ considered and tested. First is the independent transmission of the images bitstream and original feature bitstream (Fig. 2.). In that case the local score is best, however the feature stream size is also biggest and can strongly affect the total bitrate, especially for stronger image compression.

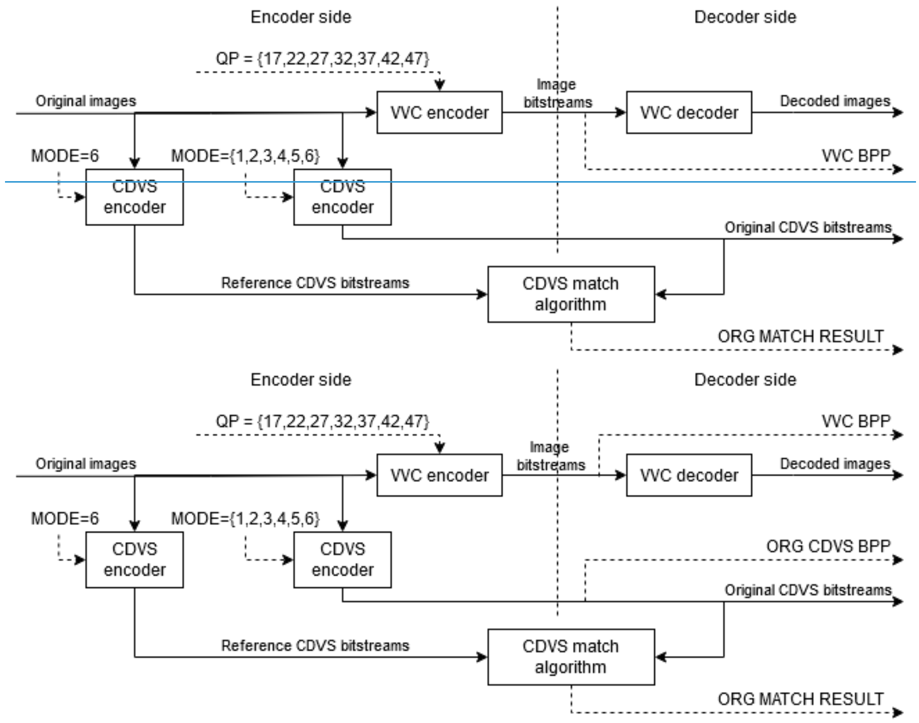


Fig. 2. Illustration of the ~~experiment~~first reference scenario – transmission of original CDVS bitstream.

Tab. 2. Local scores for original CDVS descriptors transmission

CDVS mode	Local score
1	34.97
2	118.17
3	172.99
4	207.80
5	402.52
6	556.92

Sformatowana tabela

Sformatowano: Wyrównany do środka

Sformatowano: Czcionka: 12 pkt

Sformatowano: Wyrównany do środka

Sformatowano: Czcionka: 12 pkt

Sformatowano: Wyrównany do środka

Sformatowano: Wyrównany do środka

Sformatowano: Wyrównany do środka

Sformatowano: Wyrównany do środka

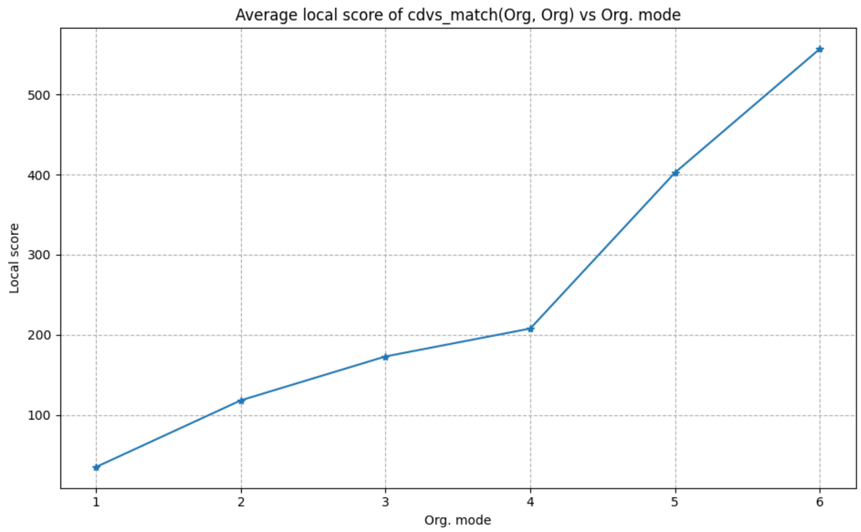


Fig. 3. The local score of comparison original (the same) descriptors for different modes.

For the transmission of the CDVS descriptions for the original images, the total bitrates are estimated as BPP, i.e. bits per pixel. For a given image, such values are calculated as the total number of bits devided by the total number of pixels in this image. The values given in Table 3 are average values over the whole set of test images used. The same method is used for BPP through the whole document.

Tab. 3. Total bitrates for original CDVS descriptors transmission

CDVS mode	**QP**						
	17	**22**	**27**	**32**	**37**	**42**	**47**
	Total BPP (VVC bitstream + Org CDVS bitstream)						
1	2.0275	1.2519	0.7429	0.4258	0.2323	0.1183	0.0559
2	2.0315	1.2559	0.7469	0.4299	0.2363	0.1223	0.0599
3	2.0395	1.2640	0.7550	0.4379	0.2444	0.1304	0.0680
4	2.0557	1.2801	0.7711	0.4541	0.2605	0.1465	0.0841
5	2.0879	1.3124	0.8034	0.4863	0.2928	0.1788	0.1163
6	2.1523	1.3768	0.8677	0.5507	0.3571	0.2432	0.1807

Sformatowano: Czcionka: Pogrubienie

Sformatowana tabela

Sformatowano: Normalny

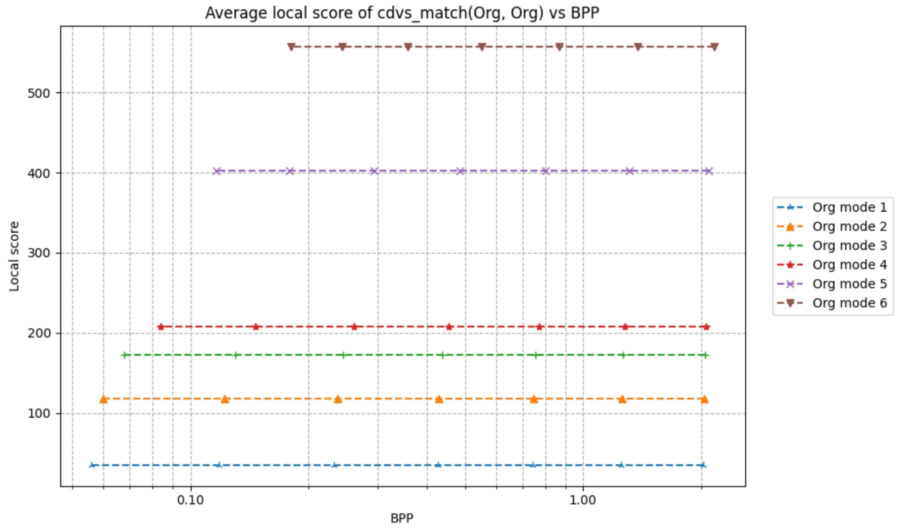


Fig. 4. The local score of ~~comparison~~similarity for -original image descriptors for different modes as a function of total bitrate (VVC bitstream ~~and~~plus original CDVS bitstream).

Sformatowano: Czcionka: 12 pkt

As can be seen above, in this scenario the local score is independent on the image compression ratio, however the original description can add significant amount of data to be transmitted.

3.2. CDVS feature extraction from decoded image

Sformatowano: Nagłówek 2

The first experiment is aimed at estimation of the local score of similarity between the maximum description (mode 6) derived from the original picture and the possible descriptions (modes 1-6) derived from the VVC-decoded picture.

In theis ~~second reference~~ scenario only image~~s~~ bitstream is transmitted (Fig. 5.). The descriptors side are extracted from decoded images only.

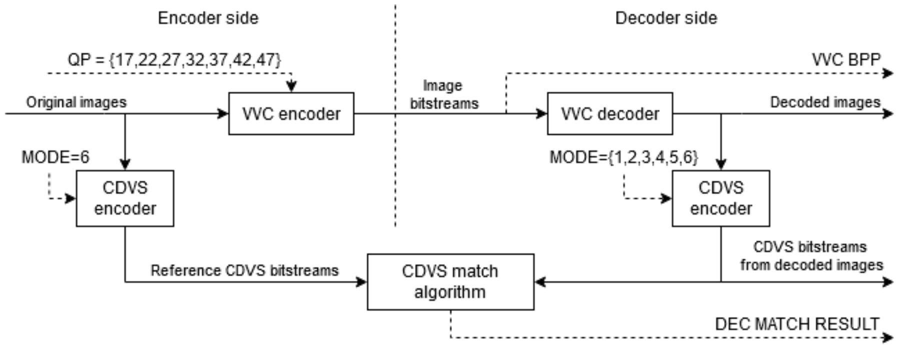


Fig. 5. Illustration of the second reference scenario – CDVS features extracted from decoded images only.

~~6. The results of the experiments~~

picture and from the decoded one. The results are averaged over the selected set of images.

Sformatowana tabela

QP~~:~~		17	22	27	32	37	42	47
BPP (VVC bitstream only)~~:~~		2.0234	1.2479	0.7389	0.4218	0.2282	0.1143	0.0518
~~MODE~~Local score	Mode 1	27.95	27.84	27.54	26.91	25.63	23.49	20.24
	Mode 2	95.10	94.36	92.64	89.02	82.45	72.42	58.74
	Mode 3	140.07	138.79	135.78	129.69	119.09	103.30	82.57
	Mode 4	168.52	166.83	162.90	155.12	141.71	122.28	96.98
	Mode 5	321.20	316.11	305.07	284.84	252.53	209.16	157.29
	Mode 6	438.43	429.59	410.96	378.49	329.61	266.96	195.39

The compression strongly affects the quality of the CDVS ~~features quality~~description. Descriptions are derived from the decoded images, ~~but~~and there is no feature~~s~~ bitstream to transmit, so the VVC bitratees remain~~s~~ ~~unchanged~~the same for a~~ny~~ll CDVS modes.

Fig. ~~3~~6. Local scores of the descriptor matching between the descriptions obtained from the original picture and from the decoded one. The results are averaged over the selected set of pictures. The local scores are depicted as a function of the quantization parameter *QP* used in VVC intra coding.

Fig. ~~4~~7. Local scores of the descriptor matching between the descriptions obtained from the original picture and from the decoded one. The results are averaged over the selected set of pictures. The local scores are depicted as a function of the average number of bits per pixel in the bitstream obtained using VVC intra coding.

3.3. Proposed solution – differential CDVS ~~bitstream~~description

Sformatowano: Nagłówek 2

The two previous experiments referred to the extreme cases:
1. The whole CDVS image description is transmitted as side information together with the compressed bitstream;
2. No CDVS description is transmitted together with VVC bitstream. The description id derived in the decoder only.
In the first scenario, the bitrate is unnecessarily high. In the second scenario, the description quality is deteriorated due to compression.

The main idea of the proposal is to the more efficient intermediate approach, i.e. to transmit only this part of the description that cannot be derived from the decoded image. We call such description as “differential description”. In this approach, we aim at the best quality of the CDVS description possible for a given image compression ratio and the minimum total bitrate, where the total bitrate is the sum of bits spent on compressed image and the differential description.

In the proposed solution, the full description on the decoder side is combined ~~using~~from the descriptors extracted from the decoded image and the additional differential ~~bitstream~~description sent a side information from the encoder (Fig. 8.). The differential descript~~or~~ion is ~~design~~obtained ~~from~~by subtraction of local descriptors sets. The global description part in the differential stream is set to void (all global parameters set to 0). Transmitted data is fully compatible with CDVS bitstream syntax.

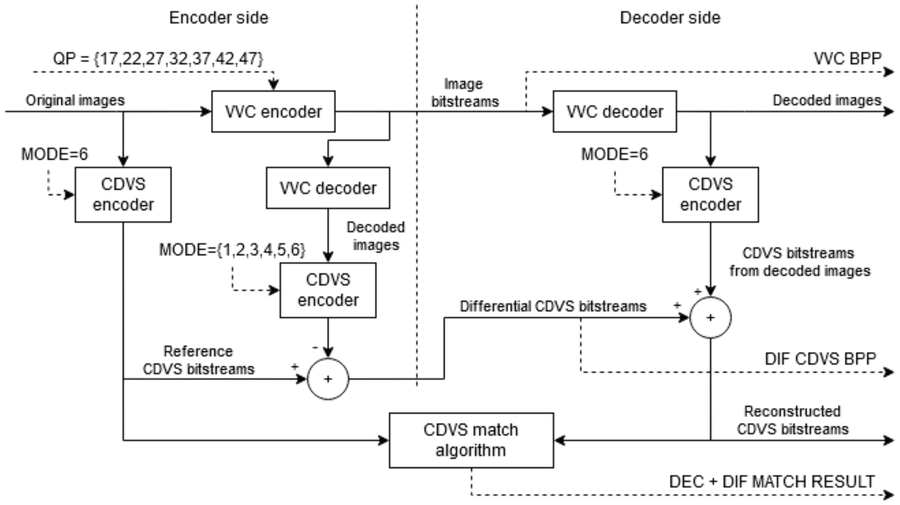


Fig. 8. Testing pipeline of the proposed system – CDVS features extracted from decoded images extended with differential description.

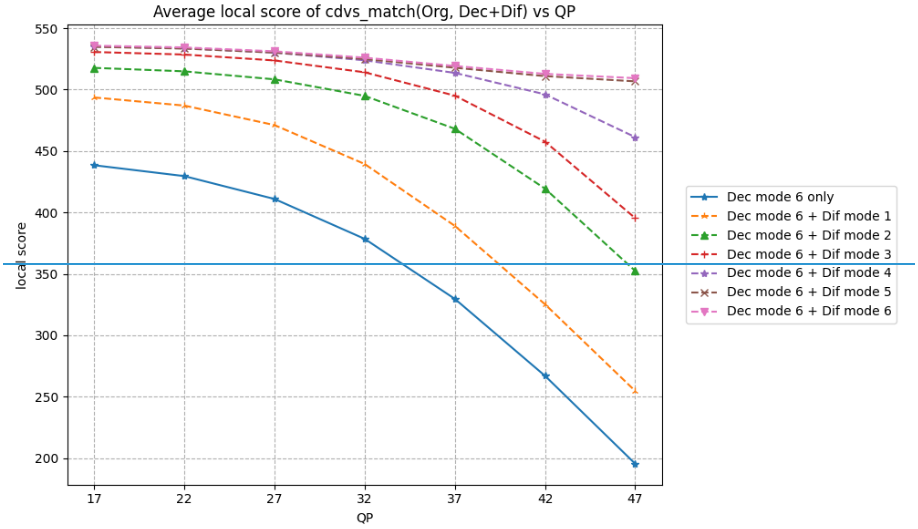
Tab. ~~2~~5. Local scores of the descriptor matching between the descriptions obtained from the original picture and descriptions obtained in the receiver using the proposed method~~from the decoded one~~. The results are averaged over the selected set of pictures.

QP:		17	22	27	32	37	42	47
Mode 1	Local score	493.62	487.02	471.13	439.43	389.05	325.20	254.67
	BPP*	2.0271	1.2517	0.7428	0.4259	0.2324	0.1184	0.0559
Mode 2	Local score	517.71	514.90	508.37	494.91	468.12	419.39	352.39
	BPP*	2.0282	1.2529	0.7444	0.4283	0.2357	0.1223	0.0600
Mode 3	Local score	530.64	528.57	523.75	514.11	495.10	457.64	395.41
	BPP*	2.0314	1.2564	0.7484	0.4331	0.2418	0.1298	0.0681
Mode 4	Local score	535.11	533.60	530.10	523.89	513.62	496.17	461.34
	BPP*	2.0353	1.2605	0.7532	0.4391	0.2499	0.1410	0.0827
Mode 5	Local score	534.81	533.41	530.20	524.88	517.81	511.00	506.75
	BPP*	2.0379	1.2634	0.7565	0.4432	0.2555	0.1493	0.0963

Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Wyrównany do środka
Sformatowana tabela
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Wyrównany do środka
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Wyrównany do środka
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Wyrównany do środka
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Wyrównany do środka
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Wyrównany do środka
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt, Pogrubienie
Sformatowano: Czcionka: 11 pkt

Mode 6	**Local score**	535.77	534.38	531.20	526.03	519.27	512.83	509.34
	BPP*	2.0455	1.2715	0.7658	0.4547	0.2702	0.1685	0.1210

*BPP – VVC bitstream and differential CDVS bitstream together.



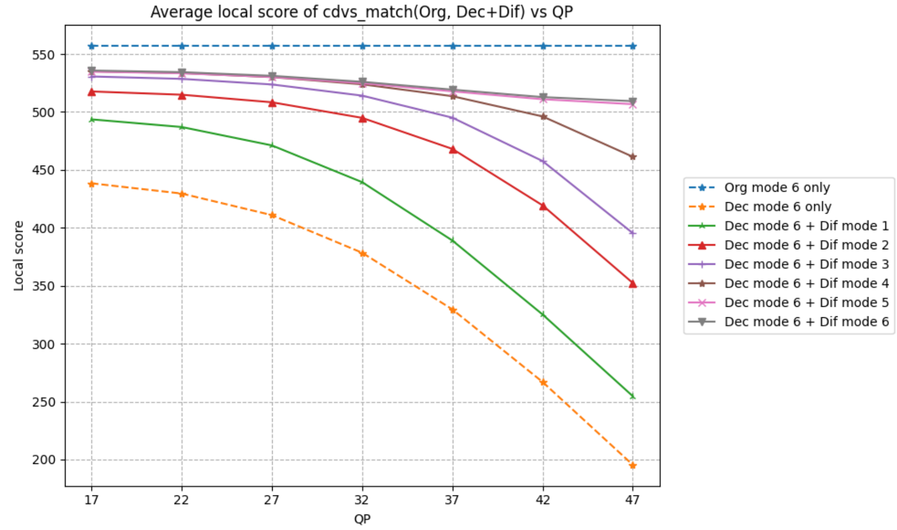


Fig. ~~5~~9. Local scores of the descriptor matching between the description obtained from the original picture and the description obtained in the receiver using the proposed method. The

Sformatowano: Czcionka: 11 pkt, Pogrubienie

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt, Pogrubienie

Sformatowano: Czcionka: 11 pkt, Pogrubienie

Sformatowano: Wyrównany do środka

Sformatowano: Czcionka: 11 pkt, Pogrubienie

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 11 pkt

Sformatowano: Czcionka: 12 pkt

Sformatowano: Wyjustowany

results are averaged over the selected set of pictures. The local scores are depicted as a function of the quantization parameter *QP* used in VVC intra coding.

Fig. ~~6~~10. Local scores of the descriptor matching between the description obtained from the original picture and the description obtained in the receiver using the proposed method. The results are averaged over the selected set of pictures. The local scores are depicted as a function of the average number of bits per pixel in the combined bitstream (obtained using VVC intra coding and differential CDVS~~.~~ coding). The line ‘Org mode 6 only” defines the maximum possible value the score.

As can be seen, the proposed solution outperforms the simple descriptor extraction from decoded images. Even for low modes, which do not add significant amounts of data to the bitstream, the local scores are noticeably higher. Although, the result description cannot reach original description score, it does not require such amount of data. For the highest QP and for the highest reasonable mode of differential CDVS (mode 5) the total bitrate is half of the bitrate in case of original CDVS transmission in highest mode.

4. Conclusions

~~...~~In this contribution, an efficient method for joint coding of images and their features proposed. The quality of the decoded image and the quality of the description can be chosen independently. Even for relatively strong image compression, the description of nearly full quality may be obtained for the bitrate that is higher than the VVC bitrate itself by a moderate number only (see Fig. 10).

The advantage of the approach is to exploit standard CDVS encoders and decoders.

~~7.~~

5. Acknowledgement

This work was ~~suppli~~supported~~ed~~ by the Project 0314/0053.

~~...~~

~~8.~~6. References

[1] Sławomir Maćkowiak, Marek Domański, Sławomir Różek, Dominik Cywiński, Jakub Szkiełda, “Video Coding for Machines: Partial transmission of SIFT features,” arXiv, Jan. 2022.

[2] Sławomir Maćkowiak, Marek Domański, Dominik Cywiński, Jakub Szekiełda, “[VCM] Influence of HEVC and VVC coding on the SIFT characteristic points extracted from the received video,” Doc. ISO/IEC JTC1/SC29/WG2/m56678, April 2021.~~.~~

[3] Sławomir Maćkowiak, Marek Domański, Dominik Cywiński, Jakub Szekiełda, “[VCM] Partial transmission of SIFT features with compressed video,” Doc. ISO/IEC JTC1/SC29/WG2/m56679, April 2021.

[4] Sławomir Różek, Marek Domański, Sławomir Maćkowiak, Olgierd Stankiewicz, Jakub Stankowski,” [VCM] New results on analysis of influence of HEVC and VVC coding on the SIFT keypoints extracted from the decoded video,” Doc. ISO/IEC JTC1/SC29/WG2/m57456, July 2021.

[5] J. Chao and E. Steinbach, “Keypoint encoding for improved feature extraction from compressed video at low bitrates,” IEEE Trans. Multimedia, vol. 18, no. 1, pp. 25–39, Jan. 2016.

~~[5]~~[6] ISO/IEC DIS 23090-3 (2020) / ITU-T Recommendation H.266 (08/2020), “Versatile video coding”.

~~[6]~~ J. Chen, Y. Ye, S. Kim, “Algorithm description for Versatile Video Coding and Test Model 3 (VTM3)”, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Doc. JVET-L1002, Macao, Oct 2018.

~~[8]~~ “Test Model ~~14~~ 12 for Versatile Video Coding (VTM ~~14~~12)”, WG 05 MPEG Joint Video Coding Team(s) with ITU-T SG 16, Doc. ~~MDS20616_WG05_N00071~~MDS20070_WG05_N00032, ~~October~~ January 2021.

~~[9]~~ https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/VTM-12.0 .~~ISO/IEC DIS 23090-3 (2020) / ITU-T Recommendation H.266 (08/2020), “Versatile video coding”.~~

[10] N15765, “Text of ISO/IEC DIS 15938-14 of CDVS Reference Software and Conformance Testing”, Geneva, CH – October 2015

[11] ISO/IEC JTC1/SC26/WG11/N15129 “Test Model 13: Compact Descriptors for Visual Search”, Geneva, 2015

~~[11] ISO/IEC 15938-15: Information Technology on Multimedia Content Description Interface, Part 15: Compact Descriptors for Video Analysis, Jul. 2019.~~

Sformatowano: Nagłówek 1, Konspekty numerowane + Poziom: 1 + Styl numeracji: 1, 2, 3, … + Rozpocznij od: 1 + Wyrównanie: Na lewo + Wyrównanie: 0 cm + Wcięcie: 0,63 cm

Kod pola został zmieniony

Sformatowano: Nie Wyróżnienie

Sformatowano: Czcionka: 12 pkt, Sprawdź pisownię i gramatykę

[13] ISO/IEC 15938-14:2018, Information technology - Multimedia content description interface - Part 14: Reference software, conformance and usage guidelines for compact descriptors for visual search.

Sformatowano: Bez podkreślenia, Kolor czcionki: Automatyczny

Sformatowano: Bez podkreślenia, Kolor czcionki: Automatyczny

Kod pola został zmieniony

Kod pola został zmieniony