Information Sciences 506 (2020) 395–423

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Machine learning based video coding optimizations: A survey

Yun Zhang [a], Sam Kwong [b,*], Shiqi Wang [b]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[b] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, China

ARTICLE INFO

Article history:
Received 16 November 2018
Revised 28 July 2019
Accepted 29 July 2019
Available online 29 July 2019

Index terms:
Video coding
High efficiency video coding
Machine learning
Mode decision
Visual quality assessment
Convolutional neural network
Deep learning
Versatile video coding

ABSTRACT

Video data has become the largest source of data consumed globally. Due to the rapid growth of video applications and boosting demands for higher quality video services, video data volume has been increasing explosively worldwide, which has been the most severe challenge for multimedia computing, transmission and storage. Video coding by compressing videos into a much smaller size is one of the key solutions; however, its development has become saturated to some extent while the compression ratio continuously grows in the last three decades. Machine leaning algorithms, especially those employing deep learning, which are capable of discovering knowledge from unstructured massive data and providing data-driven predictions, provide new opportunities for further upgrading video coding technologies. In this article, we present a review on machine learning based video encoding optimization, aiming to provide researchers with a strong foundation and inspire future developments for data-driven video coding. Firstly, we analyze the representations and redundancies of video data. Secondly, we review the development of video coding standards and key requirements. Subsequently, we present a systemic survey on the recent advances and challenges associated with the machine learning based video coding optimizations from three key aspects, including high efficiency, low complexity and high visual quality. Their workflows, representative schemes, performances, advantages and disadvantages are analyzed in detail. Finally, the challenges and opportunities are identified, which may provide the academic and industrial communities with groundwork and potential directions for future research.

© 2019 Published by Elsevier Inc.

1. Introduction

With the development of multimedia computing, communication and display technologies, many video applications have emerged, such as TV broadcasting, movies, video-on-demand, video conference, mobile video, video surveillance, remote control, robotic, 3D videos and free viewpoint TV, Virtual Reality (VR), as shown in Fig. 1, which can provide immersive telepresence and realistic visual perception experience. These video applications have been widely employed for multiple roles in human daily life, such as manufacturing, communication, national security, military, education, medication, and entertainment. Nowadays, video data has been the majority of the data traffic over the internet and its volume grows explosively each year. In 2016, global IP video traffic was or 70 exabytes [EB] (one billion gigabytes [GB]) per month, which accounted for 73% of all consumer internet traffic [13]. Cisco Visual Networking Index (VNI) forecasts the video traffic will be increased to 82% of all consumer internet traffic by 2021 [13]. On that occasion, million minutes of video contents will

* Corresponding author.
E-mail addresses: yun.zhang@siat.ac.cn (Y. Zhang), cssamk@cityu.edu.hk (S. Kwong), shiqwang@cityu.edu.hk (S. Wang).

https://doi.org/10.1016/j.ins.2019.07.096
0020-0255/© 2019 Published by Elsevier Inc.

Fig. 1. Examples of typical video applications.

be delivered through the network in every second. Regarding Internet video, for example, 400 h of videos were uploaded to YouTube every minute (*i.e.,* 65 years video a day) and one billion hours of YouTube videos were watched every day at the end of 2017 [121]. Besides, mobile video was expected to account for a staggering 78% of total mobile data traffic by the end of 2021 [14]. IHS Markit reported in China that there were about 176 million surveillance cameras in 2017, which generated 104 EB every month. To further enhance the immersive and realistic visual experiences, more high-end video applications emerge, such as High Definition (HD)/Ultra HD (UHD), holograph 3D and VR, High Dynamic Range (HDR) and Wide Color Gamut (WCG) videos, which require larger data volume to represent higher fidelity and more details. Meanwhile, the number of video clients and cameras in use grows rapidly as the video demands keep boost in recent years, such as HDTV, surveillance camera, laptop and smart phones. The total amount of global video data doubles every two years, which has been for the bottleneck for data processing, storage and transmission.

Video coding is one of the core technologies in video applications that enables to structure and compress the video data in a more effective manner for computing, transmission and storage. It has been developed over three decades with four generations and the coding efficiency doubles every ten years. But there is a big gap as compared with the rapid growth of global video data doubling every two years. Achieving much higher compression efficiency and narrowing the gap in an effective way become urgent missions for video coding. Machine learning is a field of study that can learn from data, discover hidden patterns and make data-driven decisions. Due to its superior performance in learning from data, many emerging works have applied machine learning algorithms to video coding to further promote the coding performances, which becomes one of the most promising directions in both academic and industrial communities.

In this paper, we aim to provide a comprehensive overview on machine learning based video coding optimization. The main contributions of this work are: 1) We summarize the representations and redundancies of video and figure out three key challenging issues in video coding; 2) Subsequently, we overview the recent advances on learning based low complexity video coding optimization, which are categorized into statistical, machine learning based and end-to-end learning based schemes. Their decision problems, representative features, workflows, advantages and disadvantages are analyzed. 3) We review the learning based high efficiency video coding with four key problems, including predictive coding, transform coding, entropy coding and enhancement. Their problem formulation, representative schemes and coding performances are presented. 4) We conduct comprehensive survey on the subjective visual quality assessment and learning based visual quality prediction, which is the key to perceptual video coding. The quality prediction is summarized and reviewed in four categories based on the functionalities of learning models in feature extraction and fusion. 5) The challenging issues and potential research opportunities in learning based video coding optimizations are identified.

The paper is organized as follows. In Section 2, the representations and redundancies of videos are first analyzed. Subsequently, the milestones of video coding standards and challenging issues are presented in Section 3. In Sections 4–6, the recent technical advances on machine learning based coding optimization are further analyzed from three key aspects, including low complexity optimization, high efficiency coding tools design and perceptual encoding optimization. Meanwhile, their workflows, advantages and disadvantages are analyzed in detail. Finally, we draw the conclusions and identify future research opportunities in Section 7.

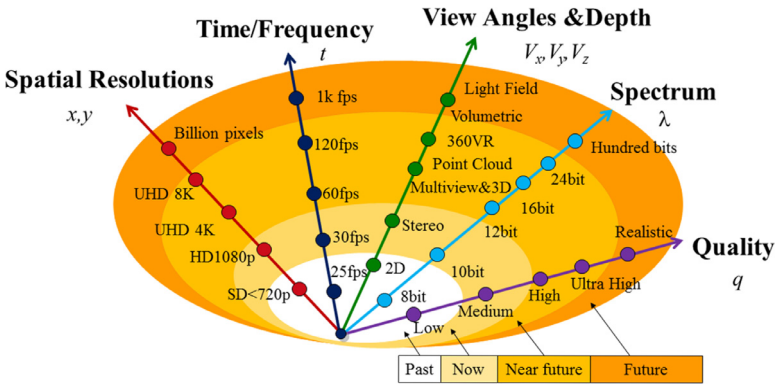


Fig. 2. Trends of video data representation from user end [16].

2. Representations and redundancies of video data

2.1. Representations of video data

The 3D world scene (**P**) can be modelled as a plenoptic function [5] with 7 parameters,

$$\mathbf{P} = F_7(\varphi,\ \theta,\ \lambda,\ t,\ V_x,\ V_y,\ V_z), \tag{1}$$

where V_x,V_y,V_z indicate the horizontal, vertical and depth viewing position in the 3D world coordinates, φ and θ represent viewing directions, λ is the spectrum wave length and t is the time sampling for dynamic scene. It can also be presented in Cartesian coordinates as [5]

$$\mathbf{P} = G_7(x,\ y,\ \lambda,\ t,\ V_x,\ V_y,\ V_z), \tag{2}$$

where x and y are coordinates on an image plane. With the development of video technologies, the video representations are extended with the following five trends, as shown in Fig. 2. 1) spatial resolution (x,y): the spatial resolution of video (x,y) grows continuously to enhance the video clarity. It is from the Common Intermediate Format (CIF) (320×240) to Standard Definition (SD) (720p), HD (1080p) and now 4 K (3840×2160)/8 K (7680×4320), which may be further extended to billions of pixels beyond the fidelity of human vision. 2) Viewing angle and depth (V_x,V_y,V_z): the video formats are developed from 2D to stereo (2-views), multiview, free viewpoint video, 360° VR [10], light field and volumetric, towards providing 3D, immersive and six Degree of Freedoms (DoFs) vision. 3) Spectrum (λ) indicating color fidelity and amplitude resolution: Video develops from black/white, color with RGB, and now targets to the WCG and HDR for more colorful and higher dynamic presentations. It will even be upgraded to high spectrums with 16 to 24-bit per channel for some specific applications. 4) Time sampling (t): with the development of capturing and computational photography technologies [16], the video frame rate increases from 25/30 frames per second (fps) for SD video to 60 fps for HD video, and will probably be 120 fps or even higher frequency.

Since video will be distorted during acquisition, compression, transmission, processing or display, the video presented to users is no longer the original representation complying with the plenoptic function, but has quality degradations. Therefore, in addition to the above four dimensions of video representation, there is another important dimension, the quality (q), from the user perspective. With the development of communication and display technologies, the user requirements on the video Quality-of-Experience (QoE) increases continuously. It is worth noting that the QoE of videos is not only limited to the picture quality or clarity, but also the visual comfort, depth quality, fatigue, immersion, DoF, delay and other visual factors relevant to the visual experience. It generally develops from low, medium, high to ultra-high and tends to be more realistic.

Generally, the video representation is in the trend of being more realistic and facilitating more interactivities. However, the data volume of realistic representation grows explosively, which is thousands or even million times of the conventional 2D video. Thus, video redundancies shall be explored for effective coding.

2.2. Signal and perceptual redundancies in video

Raw captured videos are highly redundant in presenting a realistic 3D world scene. Due to the similarities in an object and high spatial fidelity, spatial neighboring pixels or blocks in an image are highly correlated, which are denoted as spatial correlations, as shown in Fig. 3. Moreover, due to high capturing frame rate, *e.g.*, 60 fps, contents among successive frames are highly correlated, especially for the static regions, which are denoted as temporal correlations. In addition, the 3D world scene is captured simultaneously by a number of cameras with slightly different positions or angles to obtain the 3D depth. The captured images among different views are highly correlated as well, which leads to inter-view correlations. Besides the spatial-temporal-view correlations, there are statistical entropy redundancies based on the probabilities of symbols that would appear. They can be regarded as signal redundancies in videos.

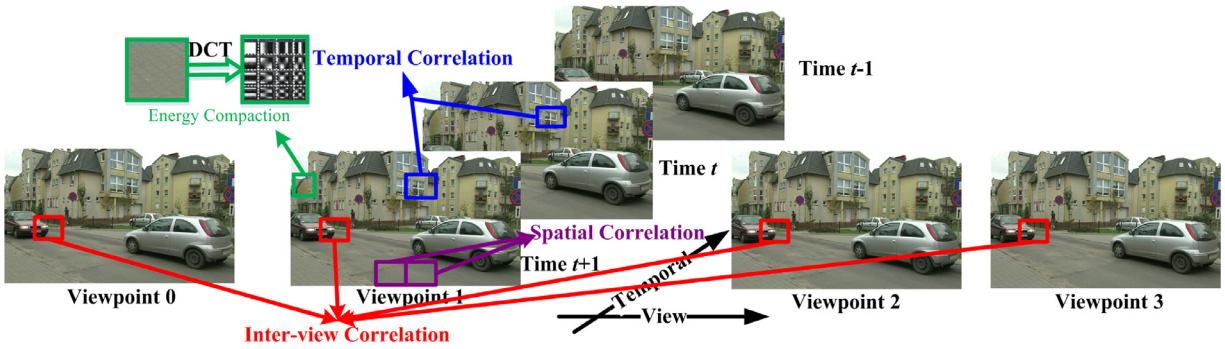


Fig. 3. Correlation and signal redundancies in raw video data.

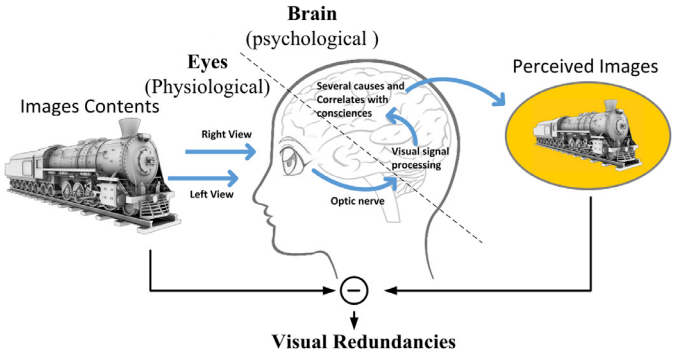


Fig. 4. Physiological and psychological perceptual redundancies.

Since most videos are ultimately perceived by Human Visual System (HVS), not all distortions of videos are noticeable by HVS, which explains the philosophy of perceptual redundancies, as shown in Fig. 4. The HVS comprises two functional parts, the eyes and the brain. Based on the physiological (eyes) and psychological (brain) studies of HVS, many visual properties and redundancies have been revealed and inspired. For example, if several pixel values of an image have a very fine-scale variation, the distortion is usually un-noticeable, leading to the concept of Just Noticeable Difference (JND). These are physiological perceptual redundancies functioned by the eyes. In addition, the perceptual sensitivity varies with the video contents, human conscious and interest, *i.e.* Region-of-Interest (ROI), which is correlated with psychology functioned by the brain. Moreover, new perceptual redundancies are still under further exploration. Video coding aims to exploit and remove the signal and perceptual redundancies as much as possible while maintaining the visual quality. In next section, we review the milestones of video coding standards and their key challenges.

3. Milestones of video coding standards and challenging issues

3.1. Milestones of video coding standards

Worldwide researchers and organizations, such as Motion Picture Expert Group (MPEG) from ISO/IEC and Video Coding Expert Group (VCEG) from ITU-T, make significant contributions to the video coding standardization and advances of coding technologies. Fig. 5 shows the evolution of the coding standards, in which five leading standards (H.261, MPEG-4, H.264/AVC [108], HEVC and VVC in red rectangles) in four generations have been issued in the last three decades. H.264/AVC standard is one of the most successful standards and has been widely used in SD/HD video applications nowadays. Due to the high quality requirements of video applications, the UHD video [31] (such as 4 K, 8 K and beyond) emerges, which increases the traffic load explosively. In 2013, Joint Collaborative Team (JCT) from MPEG and VCEG standardized the third generation of coding standards, called High Efficiency Video Coding (HEVC) [95], targeting at the UHD video applications. Its extensions including 3D Video Coding (3DVC), Scalable Video Coding (SVC) and Screen Content Coding (SCC) were also developed for different scenarios. Beyond HEVC, Joint Video Exploration Team (JVET), consisting of the experts from MPEG and VCEG, was established in 2015 to explore more sophisticated video coding algorithms for the Next Generation Video Coding (NGVC) beyond HEVC, which is Versatile Video Coding (VVC) [36] since July 2018. It targets to bring another 50% bit rate saving for UHD (higher resolution, frame rate, and bit depth) and HDR/WCG videos while maintaining high visual quality to the end user. In addition, Point Cloud Compression (PCC) and 360° panoramic VR video coding standards are also under investigations for the immersive 3D and VR applications. Also, many off-springs or extensions of standards were developed, such as

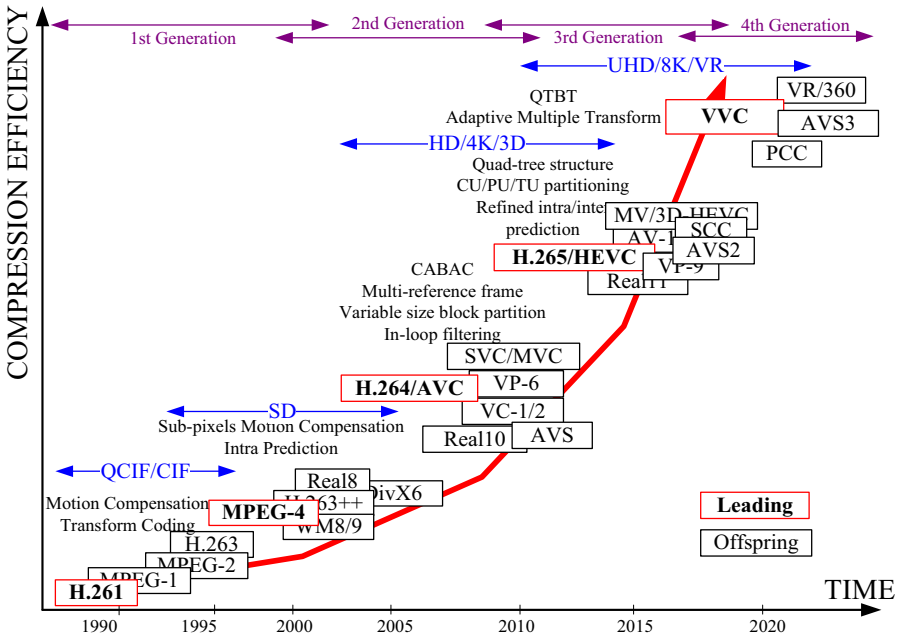


Fig. 5. Evolution of video coding standards.

MV-HEVC, AVS in China, Real11 by RealNetworks, VP8/9 by Google, VC-1/2 by Microsoft, AV-1 by Alliance for Open Media (AOM), *etc.*. There remain a number of common and critical issues that shall be addressed in further upgrading the video coding technologies.

3.2. Key requirements and challenges of video coding optimizations

The main target of video coding is to minimize the bit rate while maintaining the visual quality. There are three key requirements on the video coding [103], including high compression ratio, low complexity, and high visual quality.

3.2.1. Compression ratio

The most important requirement of a coding standard is the compression efficiency. The target of each video coding standard is to double the compression ratio comparing to its predecessor standard. As more and more advanced coding tools have been developed in the new standards, the compression ratio approaches its limitation and is almost saturated. To explore the video redundancies and improve the compression ratio further become vital and extreme challenging.

3.2.2. Coding complexity

The second key requirement is the complexity. The coding complexities of the encoding and decoding algorithms are related to the hardware cost, memory access, computing power and so on, which are in direct proportional to the production, usage and maintenance cost. As more advanced and complicated coding algorithms were adopted in the on-going standards to improve the coding efficiency, it causes dozens or hundreds times of complexity burden as compared to its predecessor. Moreover, as the view-spatial-temporal resolutions, bit depth and the number of views increase, as illustrated in Fig. 2, the computational cost of video codec is multiplicatively raised, probably million times. Thus, to enable the real-time realistic broadcasting, such as UHD, 3D and VR, it is highly desirable to lower the computational cost with optimization techniques.

3.2.3. Visual quality

The third key requirement is the visual quality or QoE since videos shall be compressed as much as possible while maintaining the visual quality. Currently, the image/video quality is mainly measured by Peak-Signal-to-Noise Ratio (PSNR) or Mean Squared Error (MSE) due to its simplicity. However, the PSNR and MSE cannot truly reflect the perceived quality of HVS, which is a complicated non-linear system. Although many important perceptual factors were revealed in psychological and physiological perspectives, the understandings on HVS are still very limited. It is challenging to develop an effective quality metric that can be widely applicable to video coding to explore perceptual redundancies.

Machine learning has the capability of discovering knowledge from unstructured massive data. Many emerging works took advantage of learning algorithms to upgrade the video coding performances. They can generally be divided into three categories according to the three key coding requirements, which are learning based low complexity coding optimization,

Table 1
Block modes in the five leading standards.

	H.261	MPEG-4	H.264/AVC	H.265/HEVC	VVC
Coding block partition	16×16	16×16	16×16 to 4×4 7 types	64×64 to 8×8 Quad-tree structure	256×256 to 8×8 QTBTTT
INTRA prediction	/	1 mode,DC/PCM	4 for 16×16 9 for 4×4	33 directions, DC and Planar, 35 modes	65 directions, DC and Planar, 67 modes
INTER PU	/	/	/	SKIP/Merge, $2N \times 2$ N, $N \times 2$ N, $2N \times N$, nR $\times 2$ N, nL $\times 2$ N, $2N \times$ nU, $2N \times$ nD, $N \times N$	QTBTTT
Reference frame	1	1 for P2 for B	Multi-reference frames, up to 5	Hierarchical, multi-reference frame	Hierarchical, multi-reference frame
TU	8×8 DCT	8×8DCT	4×4 ICT and 2×2 Hardmard Transform	8×8, 16×16 32×32, SQT and NSQT, Integer DCT and DST	4×4, 8×8, 16×16, 32×32 up to 128×128, SQT and NSQT,ICT, IST, AMT

learning based high efficiency coding optimization and high quality coding optimization. They will be discussed in detail in Sections 4–6.

4. Learning based low complexity coding optimization

4.1. Mode decision in video coding

Refined variable block size partitioning is capable of improving the prediction accuracy, which consequently reduces the coding residue and improves the coding efficiency in predictive coding. Table 1 shows the evolution of block modes in standards from MPEG-1, 2 to H.264/AVC, H.265 and the on-going VVC. We can observe that the only one kind of block, *i.e.*, 16×16 denoted as macroblock, is available for the H.261 and MPEG-4. In H.264/AVC, there are seven variable block-size partitioning candidates varying from 16×16 to 4×4, which are 16×16, 8×16, 16×8, 8×8, 4×8, 8×4 and 4×4 [108]. Then, the optimal mode is determined by Rate-Distortion (RD) cost comparison after checking each mode. In H.265/HEVC, the quad-tree structure is adopted for partitioning each Coding Tree Unit (CTU), where the Coding Unit (CU) size varies from 64×64 to 8×8 in quad-tree [95]. For each CU, it could be further partitioned to different Prediction Unit (PU) modes, such as SKIP/Merge, $2N \times 2$ N, $N \times 2$ N, $2N \times N$, nR $\times 2$ N, nL $\times 2$ N, $2N \times$ nU, $2N \times$ nD and $N \times N$ for INTER prediction. The on-going standard VVC of the JVET adopts Quad-Tree plus Binary Tree (QTBT) and Ternary Tree (TT) for block partition [1], which utilizes asymmetric binary and ternary partitioning to split a leaf node into two/three unequal child nodes. Moreover, the CU size ranges from 256×256 to 8×8.

In addition to the variable size block partitioning, the number of prediction modes is also increased. For example, the numbers of INTRA prediction modes are 1, 4/9, 35 and 67 for MPEG-2, H.264/AVC, H.265 and the VVC, respectively. Similarly, more refined modes or parameters are introduced in INTER prediction, reference frame selection, Transform Unit (TU), Motion Estimation (ME) and loop filtering to improve the coding efficiency. More and more block mode candidates are included in the evolution of coding standards from H.264/AVC to VVC, which centuples the coding complexity. Thus, effective mode decision is required.

4.2. Mode decision problem formulation

The mode decision problem in video coding is to select the best one among multiple mode candidates. In addition, multiple decision layers are usually structured in a recursive form. A general mode decision problem is to find the optimal mode set $\{\alpha^*, \beta^*, \gamma^*\}$ by minimizing the RD cost (J) between the candidate ($\mathbf{C}$) and reference ($\mathbf{R}$) blocks, which can be mathematically presented as

$$\{\alpha*, \beta*, \gamma*\} = \arg\min_{\alpha \in \mathrm{A}} \left(\arg\min_{\beta \in \mathrm{B}} \left(\arg\min_{\gamma \in \mathbf{\Gamma}} J(\mathbf{R}, \mathbf{C}(\alpha, \beta, \gamma)) \right) \right), \tag{3}$$

where $\mathbf{C}$ and $\mathbf{R}$ are the candidate and reference blocks, $J(.)$ calculates the RD cost between the two blocks [95], α, β, and γ are mode parameters for candidate block, *e.g.*, CU mode, motion vector and reference frame indices, **A, B** and $\mathbf{\Gamma}$ are sets of mode candidates. Fig. 6 shows the mode decision problems in HEVC for different layers. For example, the CU size decision can be formulated as a recursive binary classification by determining whether the CU is split or not. Some layers, such as PU size decision, INTRA mode and multi-reference frame selection can also be formulated as multi-class classification problems due to multiple candidate modes. Furthermore, multiple decision layers can be formulated as recursive multiclass classifications, which are more complicated. There are five recursive loops in total for INTER coding, which include the CU, PU, reference frame, TU and ME from the top. For HEVC INTRA coding, it has four loops, including CU, PU, angular prediction mode and TU size.

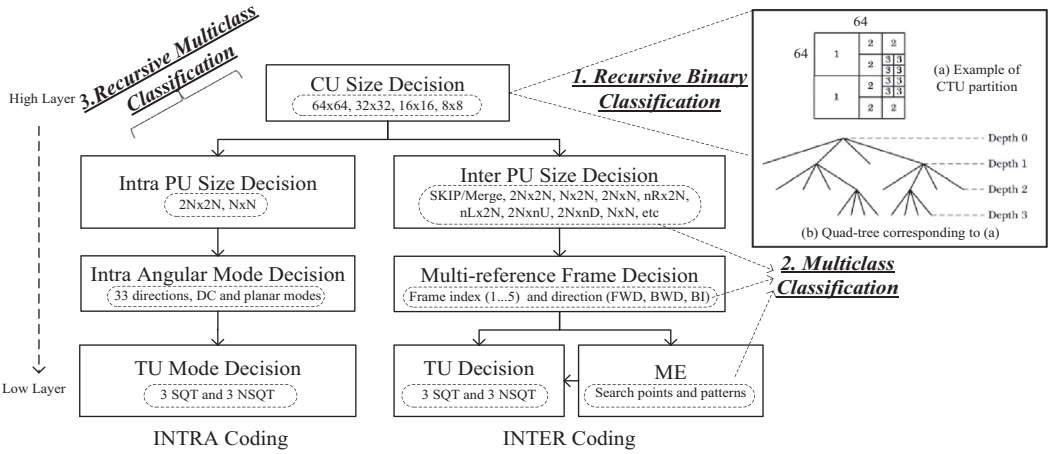


Fig. 6. Mode decision problems in HEVC.

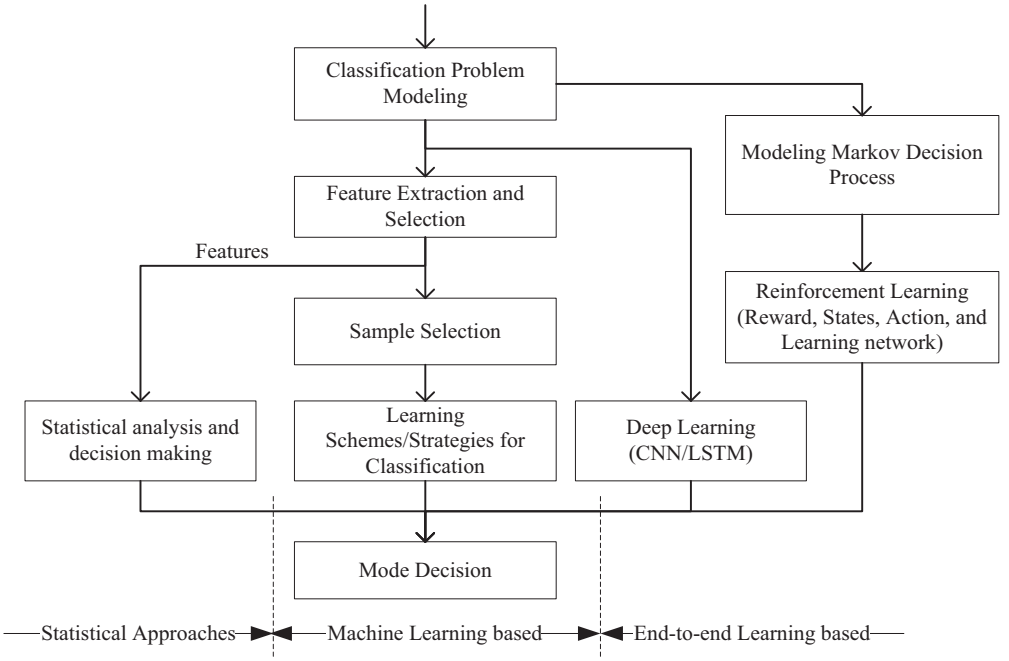


Fig. 7. Workflow of learning based mode decisions.

4.3. Learning based fast mode decision

To adequately address the mode decision problem in video coding, the existing works on fast mode decision can be divided into three categories, including statistical approaches, machine learning based schemes and end-to-end learning based schemes. Fig. 7 shows a general workflow for these three learning based mode decisions.

4.3.1. Statistical approaches

The coding complexity is critical for H.264/AVC and beyond, and a number of statistical approaches have been developed to lower the complexity of video encoder [68,90,129,59,131,91,128]. Liu et al. [68] proposed an INTER mode decision algorithm based on motion homogeneity using Motion Vector (MV) field of 4×4 blocks, where some INTER modes were skipped for motion homogenous regions. Shen et al. [90] proposed an adaptive fast multi-frame selection algorithm that exploited the correlation among the neighboring blocks and information of ME from previous searched reference frames. Zhang et al. [129] developed a statistical early termination model for fast SKIP/DIRECT mode decision in H.264/MVC based on the RD cost distribution. HEVC has more mode candidates and is more complex. To lower the complexity of HEVC mode decision,

Kim et al. [59] adopted the RD cost as the key feature for PU early skip and early termination based on Bayes decision rule. Additionally, the RD cost distribution and spatial correlation [131] were jointly exploited for fast CU mode decision in HEVC INTRA coding. In [91], a fast INTER-mode decision algorithm for HEVC was proposed by jointly exploiting INTER CU correlation of quad-tree structure and the spatio-temporal correlations among neighboring CUs, in which the prediction mode, MV and RD cost were found strongly correlated. In [128], the early termination and early skip models were jointly used for fast mode decision in HEVC INTRA coding, in which RD cost was used as key feature and the coding complexities over different decision layers were jointly minimized subject to acceptable rate-distortion degradation.

Basically, they are statistical approaches for fast mode decision. The general coding framework for these approaches is shown as the left part of Fig. 7. Firstly, it extracts hand-crafted key features, such as the RD cost, MV or sum of weighted CU depth of neighboring CUs, spatial or temporal correlations. Then, empirical statistics-based hard or soft thresholds are then determined for each of these key features subject to high prediction accuracy in the decision making. The advantages of these algorithms are simple, easy for implementation and hardware friendly. Meanwhile, they are usually efficient due to very limited complexity overhead in implementation. However, their drawbacks are 1) only very small number of key features can be exploited in each algorithm, usually 1 to 3 key features, which restricts the discriminability for distinguishing each mode. 2) These features usually work individually and independently. 3) The thresholds of these algorithms are usually determined based on the statistical analyses on a small set, *e.g.*, a number of video frames or sequences, which will reduce the adaptability of the proposed algorithms. Using stricter threshold usually improves the adaptability, but leads to less complexity reduction.

4.3.2. Machine learning based mode decision schemes

Mode decision problem can be casted into a classification problem, and learning algorithms were then explored in classifying modes in video coding. A number of works [23,96,11,8] had explored machine learning based mode decision for H.264/AVC. In these schemes, the seven mode candidates (see Table 1) were divided into several subsets and predicted by trained models. Binary classifiers, including Support Vector Machine (SVM) [23], unsupervised clustering [96], Back Propagation Neural Network (BPNN) [11] and decision tree [8], *etc.*, have been applied to skip some modes or determine the best mode among the seven block candidates in H.264/AVC. Meanwhile, a number of features have been developed, such as RD cost normalized by Sum of Absolute Difference (SAD) [23], spatial and temporal correlation, SKIP mode RD cost [96,8], INTER and INTRA SAD, MV difference and gradient [11].

Compared to H.264/AVC, HEVC has more complex decision problems, which include recursive quad-tree CU mode decision, multi-class PU and TU mode decisions. In addition, HEVC has a much larger number of mode candidates, which makes the classification task more challenging. A number of works have also been proposed for machine learning based HEVC INTER [111,88,130,138,42,72,15,26,110,25,139,101,89,133] and INTRA [22,132,66,123,125] coding optimization in the past few years.

For the HEVC INTER coding, Xiong et al. [111] determined the best CU based on unsupervised K-nearest clustering of Pyramid Motion Divergence (PMD), which was generated from the optical flow of down-sampled frames. The CU size decision in HEVC is a recursive decision process, Shen et al. [88] proposed a CU early termination algorithm for each level of the HEVC quad-tree CU partition by using weighted SVM, where RD cost increments caused by misclassifications were determined as weights in SVM training. Zhang et al. [130] further modelled the quad-tree CU decision as three-level hierarchical binary decision problem, and learned the optimum SVM hyper-planes for early termination and early skip, respectively, at each decision layer. Meanwhile, optimal off-line learning parameters were derived to achieve the trade-off between RD performance and complexity reduction. Zhu et al. [138] enhanced the CU decision based on fuzzy SVM, which considered the sample and feature selections in learning process. Additionally, Bayesian decision rules [42,72], decision tree [15], Neyman-Pearson based rule [26], Markov Random Field (MRF) [110] and Markov Chain Monte Carlo (MCMC) models [25] were also explored to solve the INTER CU size decision problems. Also, boosting and ensemble algorithms of using multiple learning algorithms were also used to improve the prediction accuracy, such as review system using multiple SVMs [138,139] and random forest [101]. Table 2 shows the features, classifier and coding performances of representative leaning based mode decision schemes under different configurations, including All Intra (AI), Low Delay P (LDP), Low Delay B (LDB) and Random Access (RA).

Besides the CU size decision, there are PU, TU size decision and reference frame selection problems in INTER coding. Since the CU size is the outer loop of the mode decision, the PU size is determined conditionally with a given CU. In [139], PU size decision was formulated as a multi-class problem and solved by a multi-class SVM. Since SKIP mode was of high probability to be selected as the optimal mode in each CU, SKIP/Merge mode was predicted by a binary SVM algorithm to early terminate the PU mode decision. In [89], the correlation between variance of residual coefficients and TU size was exploited to reduce the number of TU candidates for a given CU/PU block, then, Bayesian theorem detection was utilized for predicting TU size. The CU, PU, TU and reference frame selection are in different mode decision levels, which can be jointly optimized for the INTER coding. However, the joint classification problem is recursive and hierarchical, which is much more complicated to achieve the optimal.

Features are key to the prediction accuracy of a learning algorithm. Some representative features for the mode decisions are also listed in Table 2, which can be basically categorized into five types: 1) The previous coded information/pre-analysis of the current block, 2) the spatial correlation information, 3) the temporal correlation information, 4) global texture and motion information. Good and representative features will significantly improve the discriminability and prediction accuracy

Table 2
Key features, classifier and performances of leaning based fast mode decision for HEVC and beyond.

Author &Year	Decision Problem	Feature Sets	Feature Number	Classifier	Performance[#]		
					CFG	TS(%)	BDBR (%)
Y. Zhang'15 [130]	INTER CU	(1)CBF. (2)RD cost. (3)Distortion. (4)Coding bits. (5)MV. (6)CU depth. (7)RD costs of neighboring CUs. (8)Skip flag. (9) Quantization Parameter (QP).	9	SVM	LDP	−51.5	1.98
H. S. Kim'16 [42]	CU	Minimum RD cost of (1) INTER and (2) INTRA prediction mode at each CU.	2	Bayesian decision rule	RA LDP LDB AI	−53.6 −48.5 −48.4 −54.2	0.71 0.62 0.63 0.96
J. Moreno' 17 [72]	CU	RD cost	1		RA LD	−38.2 −36.4	0.82 1.69
G. Correa'15 [15]	CU	(1)PU splitting mode. (2)Coding tree depths used in neighboring CTUs already encoded. (3)A division between the RD costs using $2N \times 2N$ and MSM mode, *i.e.* $RD_{2N\times 2N}/RD_{MSM}$. (4)The normalized difference between the RD costs of $2N \times 2N$ and MSM, *i.e.* Norm($RD_{2N\times 2N}$-RD_{MSM}). (5)$RD_{2N\times 2N}$, (6)RD_{MSM}, (7)$RD_{2N\times N}$, (8)$RD_{N\times 2N}$, (9)SkipMergeFlag. (10)MergeFlag.	10	Decision tree	RA	−36.7	0.28
	PU	(1) $RD_{2N\times 2N}/RD_{MSM}$. (2) Norm($RD_{2N\times 2N}$-$RD_{MSM}$). (3) $RD_{2N\times 2N}$ (4) RD_{MSM} (5)RD_{BestPU}. (6)Split flag. (7) RD_{BestPU}/RD_{MSM}. (8) Norm(RD_{BestPU}-RD_{MSM}).	8			−49.6	0.56
	TU	(1)- (3)SADs from luminance Y, chrominance U and residues. (4) RD cost of encoding a TU. (5)-(7) Number of nonzero coefficients for Y, U and residues.	7			−7.2	0.05
Q. Hu'16 [26]	CU PU SKIP	RD cost distribution. Class-conditional probability density function of RD cost.	1 1	Neyman Pearson based rule	RA	−51.4 −45.9	1.20 0.72
K. Goswami '18 [25]	CU	(1)Entropy difference between the current CU and the reference CU. (2)RD cost of skip and non-skip CU. (3)Variance of entropies of child CUs.	3	Bayesian & MCMC	RA LDB	−51.6 −49.5	1.11 1.25
L. Zhu'17 [139]	CU and PU SKIP	(1)SAD between current block and co-located block, (2)CU depth, (3)PU partition, (4)Context skip flag, (5)Merge flag, (6)RD cost, (7)Skip flag, (8)Distortion, (9)CBF, (10)Coding bits, (11)QP.	11	SVM	LDP LDB RA	−68.3 −67.3 −65.6	4.19 3.84 3.66
H. R. Tohidypour '17 [101]	CU in SHVC	(1)Motion information, (2)The lowest RD cost mode of the co-located block in the previous frame of the same EL, (3)The lowest RD cost mode of the parent CU, (4)The lowest RD cost mode of the corresponding block in the BL.	4	Bayesian classifiers	RA	−45.4	1.13
L.Shen'15 [89]	TU	Variance of residual coefficients.	1	Bayesian decision rule	RA LD	−46.0 −46.0	0.5 0.6

(*continued on next page*)

Table 2 (*continued*)

Author &Year	Decision Problem	Feature Sets	Feature Number	Classifier	Performance[#]		
					CFG	TS(%)	BDBR (%)
L. Zhu'16 [133]	CU, Trans-coding	(1) SAD between current and co-located blocks, (2)Coded block pattern, (3) MB partition types, (4)Number of non-zero DCT coefficients, (5) DCT coefficients energy, (6)CU depth, (7)Merge flag, (8)Skip flag from bit-stream, (9) Context skip flag, (10)CBF of current coding, (11)RD cost of current CU. (12)Skip flag of the current CU, (13) Merge flag of the current CU.	13 out of 24	Weighted SVM	LDP RA	−50.2 −49.2	1.98 2.40
F. Duanmu'16 [22]	CU size decision	(1) Sub-CU horizontal and vertical DC difference. (2)CU variance based on pixel level luminance values. (3)CU gradient kurtosis. (4)CU gradient magnitude peak. (5)Zero gradient percentage. (6)CU color number.	6	Decision tree	AI	−52.0	3.65
Y. Zhang'17 [132]	INTRA CU	(1) Variance, (2) variance difference, (3)-(4) RD cost of PLANAR mode divided by distortion or QStep, (5) neighboring RD cost, (6)-(7) neighboring CU or CU plus PU depth, (8)-(9) RD cost and bits of previous coded mode.	3–6 for different layers	SVM	AI	−52.4	1.58
X.Liu'17 [66]	INTRA CU	(1)Neighboring MSE, (2)Direction complexity, (3)Sub-CU's complexity difference, (4)QStep.	4	Dual-SVM	AI	−59.6	1.26
T. Zhang'17 [125]	INTRA CU	(1)Depth difference between current CU and its neighboring CUs, (2)The ratio of HAD costs between the current CU and its neighboring CUs.	2	Linear SVM	AI	−43.1	0.50
	INTRA mode	(1) AGH, (2)AGV	2	Empirical threshold		−15.2	0.18
Z. Liu'16 [70]	CU	(1)Learned Features from 64×64 block of raw pixels* (2)QP	17	CNN	AI	−61.1	2.67
M. Xu'18 [114]	CU	Learned features from 64×64 block of raw pixels*	2688/448	CNN/LSTM	AI LDP	−61.8 −54.2	2.3 1.5
Z. Jin'17 [34]	INTRA QTBT	Learned features from 64×64 block of raw pixels*.	48	CNN	AI	−42.8	0.65
K. Kim'18 [46]	CU	Learned features from 64×64, 32×32, 16×16 block of raw pixels*.	32/32/32	CNN	LDP RA	−60.6 −61.8	3.75 3.91
J. Xu'18 [112]	CU, Trans-coding	(1)MV,(2)MB partition, (3)Bit, (4)Residual	4	LSTM	LDP RA	−59.6 −55.4	1.16 1.53
N.Li'19 [54]	INTER CU	(1)CU Depth. (2)RD cost. (3)Distortion. (4) Skip flag. (5)CU location.	5	RL&NN	LDP	−34.34	0.85

* Features are learned from block of raw pixels, and the number of features are the dimension of input to the first FC layer.

The coding performances are evaluated with Bjontegaard Delta Bit Rate (BDBR) and Time Saving (TS), which are highly co-related to the test video sequences, experimental settings and configurations (CFGs).

of a learning based mode decision, while bad or irrelevant features may have negative impacts on the prediction. In addition, increasing the number of features raises the dimensions of input data, which will increase the complexity overhead of feature extraction and on-line learning complexity. Therefore, feature selection algorithm is required to select more representative features. In [88], a filter based feature selection scheme was presented, where F-score was used as a feature correlation measurement and five top features were selected among ten. In [133], top 13 features were selected from 23 features with the exhaustive search. Although the feature selection overhead is not considered for off-line case, it is time consuming if we have large feature sets. Basically, these features in related works [111,88,130,138,42,72,15,26,110,25,139,101,89,133] are manual handcrafted features and it is very difficult to find reliable and discriminative features for the mode decision. Since there are multiple layers of mode decisions in video coding, including CU size, PU, TU and reference frame selection, *etc.*, different features are usually required for different decision problems and frame types. In addition, due to the diversity of the video contents, such as natural scene, fast motion, surveillance and screen content, features shall be specifically designed to adapt the videos. Therefore, on-line feature selection may be further required in order to have a good trade-off between accuracy and overhead for the mode decision problem.

Sample selection is also an important issue for learning based mode decision to solve the imbalance problem of training data, which had been considered in works [88,130,138,132]. Since the optimal mode can be determined using exhaustive full RDO by the benchmark video encoder, sufficient training data can be labelled. However, similar samples may be generated due to similar contents among frames and videos, which causes sample redundancies. Besides, training samples are imbalancedly distributed among classes and within a class. For example, in the CU size decision the number of large CUs, *e.g.,* the number of 64×64 is much less than those of 8×8 or 16×16 CUs. In addition, the sub-classes of samples are also not evenly distributed within a class, *e.g.* 32×32 and 16×16 are imbalanced in split class. To solve this problem, sub- and up-sampling [132] methods were commonly used to achieve balance data. Also, cost-sensitive learning method has also been used in [88,130,138] by considering the cost types and magnitudes of misclassification, such as RD cost and complexity.

For the INTRA coding, the complexity is lower than the INTER coding; however, the complexity overhead from the feature extraction and learning is more critical. Duanmu et al. [22] adopted the decision tree classifiers to classify block type (*i.e.*, natural image or screen content), CU partitioning and directional blocks. Six dimensional features were exploited. Then, the sequential mode checking process was early terminated when the current mode coding rate was lower than a statistical threshold. Zhang et al. [132] proposed two-stage of SVM based classifications for INTRA CU size decision, where the split, non-split and uncertain were predicted by the first off-line SVM classifier, then, a second stage on-line SVM classifier was used to refine the uncertain prediction from the first one. Different features sets were exploited for each CU decision layer to minimize the complexity overheads. Liu et al. [66] further exploited dual-SVM model for CU depth decision, meanwhile, four dimensional features, including image texture complexity, direction complexity, sub-CUs complexity and Quantization Parameters (QPs), were adopted.

In addition to the CU size decision, there are 35 angular prediction modes in HEVC INTRA coding, which is a multi-class decision problem. In HEVC, instead of checking all 35-mode, Rough Mode Decision (RMD) has been adopted to select a small set (3 to 8 plus MPM modes) for full RDO. H. Zhang et al. [123] proposed a progressive rough mode search (pRMS) to selectively check the potential modes instead of traversing all 35 candidates in RMD. T. Zhang et al. [125] utilized the average gradients in the horizontal (AGH) and vertical direction (AGV) to decide a rough range of block directions. These were statistical approaches. Besides, random forest was used to estimate an INTRA-prediction mode in [83], where only four pixels reflecting a directional property of a block were used as key features to reduce the complexity overhead. It achieved 18.3% and 17.2% complexity reduction for HEVC and VVC, respectively. Few works addressed the INTRA prediction mode using machine learning tools. The main reasons are 1) each INTRA prediction mode is with less complexity, and machine learning complexity overhead will no longer be negligible, and 2) it is more challenging to discriminate the prediction modes since they are eventually highly similar to each other.

The major working flow of these learning based mode decision schemes is shown in Fig. 7. Firstly, model the mode decisions in video coding into classification problems, such as binary, multiple hierarchical binary and multi-class classifications. Secondly, develop a number of handcrafted features and trainable classifiers to solve the classification. The advantage of these works are 1) a number of features are jointly utilized, 2) The prediction accuracy and discriminability can be improved since the classifiers are able to nonlinearly map the input feature vectors into a high dimensional feature space and construct the optimum separating hyper-plane for each class. However, the drawbacks of these algorithms are 1) the features are manually handcrafted. It is difficult to find effective features for each decision problem, which is time-consuming and requires professional domain understandings. Feature selection and extraction are critical issues to the classification performance. 2) There is complexity overhead to extract the features and to learn the optimum hyper-plane with on-line learning mode, which is especially critical when the feature dimensions and the number of training samples are large. 3) Generally, an optimal parameter determination is required to achieve good trade-off between complexity reduction and RD degradation.

4.3.3. End-to-end deep learning based schemes

In recent years, the deep Neural Network (NN) [51] significantly advances prediction performance and has been widely used in visual signal processing and pattern recognition. Due to the difficulties of finding effective features for conventional machine learning based schemes, a number of researchers have devoted their efforts to exploring the end-to-end deep learning based mode decision schemes [69,70,50,55,114,112,34,35]. Liu et al. [69] applied the Convolutional Neural Network (CNN)

to analyze the textures of source picture blocks, and then reduced the maximum number of CU modes in INTRA coding. The QP was introduced in designing the CNN architecture by considering the effect of quantization to the coding costs. Besides, a hardware CNN accelerator was developed for the INTRA CU size decision in [70], where the network configured with only two convolution layers, one pooling hidden layer and two full-connected layers. Instead of training different CNNs, the input blocks 64×64, 32×32, 16×16 and 8×8 of raw pixels were normalized to be 8×8 matrices by using local averaging and sub-sampling, such that they could share one CNN network. Laude et al. [50] explored to replace the RDO process with the CNN classifier for INTRA prediction in each CU, and it was reported only 0.52% BDBR loss. However, complexity analysis was not reported. Xu et al. proposed an early terminated hierarchical CNN to predict INTRA CU partition [55] and a hierarchical Long- and Short-Term Memory (LSTM) network to predict the INTER CU mode [114], where the temporal dependency was further explored. The network shared the convolution layers for feature extraction to lower complexity and Full Connection (FC) layers varied for different CU decisions. The proposed CNN and LSTM have millions of parameters and also require millions of additions and multiplications, which are very complex but not counted in overall complexity [114]. Similarly, in [46], Kim et al. adopted Multilayer Perceptron (MLP) based NN to predict split or non-split for CU depth decision in HEVC INTER- and INTRA-prediction. In [112], Xu et al. proposed a hierarchical LSTM network based CU depth decision to predict the CTU partition for H.264 to HEVC transcoding, where four dimensional handcrafted feature maps including MV, block partition, bit and residue were input to the network. The latest JVET employs QTBT block partitioning structure in the under-going VVC, which improves the coding efficiency at cost of five times or more computational complexity. To lower the INTRA coding complexity of VVC, Jin et al. [34,35] modelled QTBT partition depth range as a 5-class classification problem and applied CNN to predict the CU depth range in QTBT. It reported 43.69% complexity reduction with 0.77% BDBR increase, and meanwhile the complexity overhead of CNN prediction was 4.51% of the encoder. In these works, high dimensional features were learned from block of raw pixels (*e.g.*, 32×32 or 64×64), shown as bottom five rows in Table 2, only the information within a block was used, traditional useful features, such as the spatial-temporal correlations and the RD cost, were not yet considered. Small number of convolutional layers, *e.g.*, 2 to 5, were adopted for CNN to lower the complexity overhead. They were proposed for INTRA coding and works on INTER coding are still rare.

In the end-to-end deep learning based mode decision schemes, the raw pixels/blocks or low level features, such as motion and variance, can be used as input, then middle or high level-features are learned from supervised learning or back-propagation. Finally, more powerful classifiers, such as fully connected MLP, were used to output the results. The advantage of these works [69,70,50,55,114,112,34,35] are 1) thousands of features could be learned from data [114] and there is no need to design professional hand-crafted features. 2) The prediction accuracy can be improved by increasing the number of convolutional layers, *i.e.*, going deeper. Also, it is capable of solving complicated multi-class classification problems. 3) There is plenty of labelled mode data for learning and robust learning models can be built. However, the disadvantages include 1) the end-to-end deep learning is extremely complex and sometimes the complexity is even higher than the encoder for deep networks. Therefore, a good trade-off between the coding gain (complexity reduction or some other performances) and complexity overhead is required. 2) Multiple learned networks are required in compressing different quality videos, *i.e.*, different QPs. 3) Over-fitting problem may exist in the end-to-end learning if the training data is insufficient or improperly selected. 4) The optimal hyper-parameters for the deep NN shall be determined empirically, which is difficult to be explained. 5) Only pixels in the current CU are now used in the learning network regardless the traditional useful features information, such as spatial and temporal correlations. Up to now, external accelerators are required [70] which increases the hardware cost. The coding performance of deep learning based scheme is only comparable to statistical approaches or conventional machine learning based schemes, such that further investigations are essentially required.

Reinforcement Learning (RL) [3] is another hot sub-branch of machine learning, which also gains much attention in both industry and academia. It concerns agents to take actions in an environment by maximizing a cumulative reward, which is different from conventional supervised and unsupervised learning. In [54], Li et al. proposed an end-to-end actor-critic RL based CU early termination scheme for HEVC. Fig. 8 shows the framework of RL based CU decision, where CrtNN and ActNN denote critic NN and actor NN, respectively. It modeled the CU decision process as Markov Decision Process (MDP) and regarded the CU encoding as environment, CU depth predictor as agent, split and non-split as actions and RD loss as reward. The CU decision classifier was learned off-line from the CU decision trajectories based on end-to-end actor-critic RL algorithm. Finally, the learned ActNN was incorporated in the video encoder to predict the CU partition. In fact, the HEVC quad-tree CU structure is difficult to be modelled as the MDP. Also, five handcrafted features and single layer NNs were used to reduce the complexity overhead, which may restrict the coding performance. It is a pioneer work on applying RL to HEVC mode decision. Although the complexity reduction is only comparable to conventional machine learning based scheme which is about 34.34% to 43.33%, RL has powerful potential capability of tackling more complicated decision and control problems. Also, it has advantages of improving the network training with higher performance and flexibility, which deserves further investigations.

4.4. Discussions

The existing learning based fast mode decision methods are mainly on optimizing the CU/PU mode decisions because the CU/PU decisions are the outer-loops and possesses the majority of complexity. In addition, other inner-loop mode decisions rely on the results of the CU/PU decisions. To exploit the advantages of machine learning algorithms and improve the coding effectiveness, it is necessary to bridge the gap between coding algorithms and machine learning algorithms [130]. It is easy

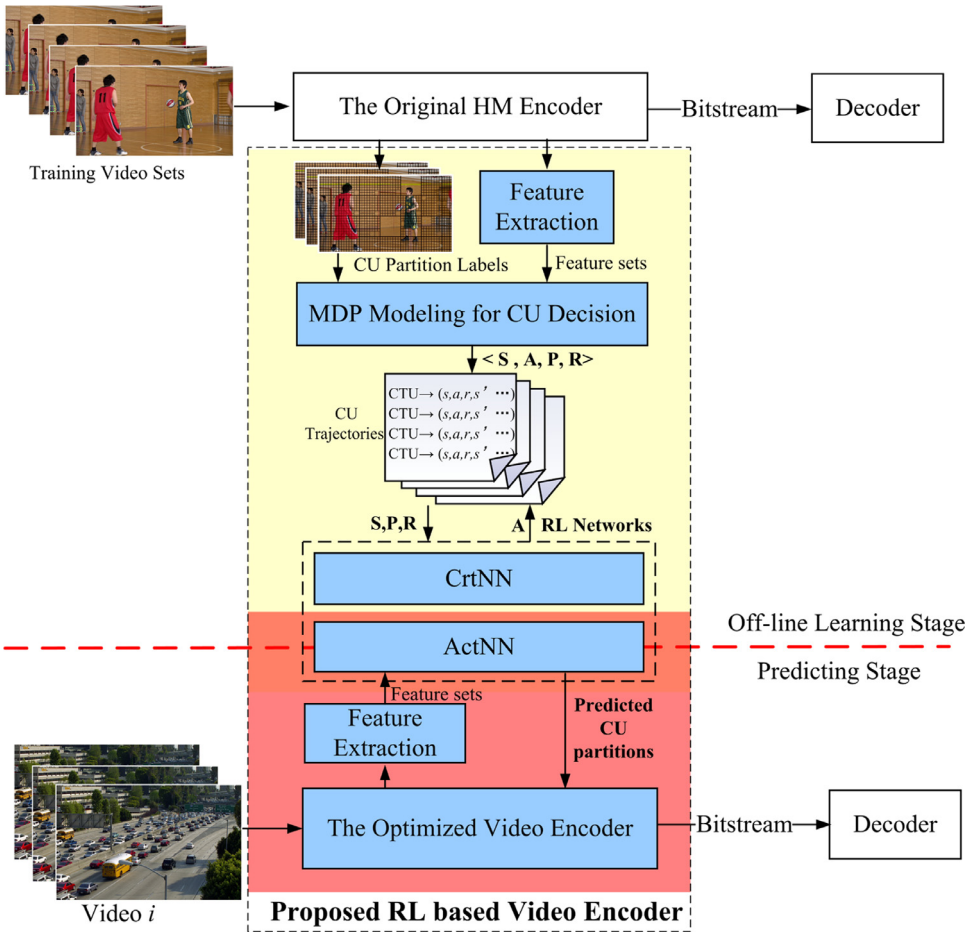


Fig. 8. RL based fast CU mode decision [54].

to formulate CU/PU decision process to fit the binary-class classification that could be well solved by learning algorithms, such as SVM, decision tree and Bayes decision rules. On the contrary, for some modules with much more candidate modes, such as the INTRA angular prediction [123] and ME, it has more difficulties in problem modelling and solving. In addition, their complexities are low for each mode, which reduces the potential complexity reduction in fast mode decision.

With the development of video coding technologies, more refined modes are introduced to achieve higher compression efficiency. For example, the number of INTRA modes increases to 67; more flexible block partitioning structure, *e.g.* QTBTTT, is enabled; available CU size increases to 256×256 in the VVC [95]. In this case, the coding complexity increases exponentially, and low complexity optimization becomes more critical. Meanwhile, each mode decision problem in VVC is more complicated. Therefore, data-driven mode decision and advanced learning tools, such as feed-forward CNN, deep RL [3] and deep NN, are possible good solutions and worthy to be investigated further for these complicated mode decision problems. In addition, the multiple different loops or decision layers can be optimized jointly to minimize the coding complexity.

5. Learning based high efficiency coding optimization

The current video coding standards comply with a block-based hybrid framework, which includes three major components: the predictive coding, transform coding and entropy coding, as shown in Fig. 9. The predictive coding exploits the view-spatial-temporal correlations of video signal, which are the major part of video redundancies. The transform coding adopts the transform to compact the energy in frequency domain, then larger scale is used to quantize the insensitive frequencies of HVS, such as high frequency, in lossy coding. Finally, the entropy coding is used to reduce the statistical entropy redundancies of the signals. In addition, the enhancement algorithms, including pre/post-processing and in-loop filtering, are adopted to improve the quality of the reconstructed video. In this section, we present the learning based high efficiency coding optimization on the above four parts. Table 3 summarizes problem formulation and coding performances for learning based high efficiency coding, where the corresponding optimization modules are shown in Fig. 9.

Table 3
Problem formulation and coding performances for learning based high efficiency coding optimization.

Categories	Problem Formulation	Modules	Learning Tools	BDBR* (%)				Complexity*(times)	
				AI	LDP	LDB	RA	ENC	DEC
Predictive Coding	Predict the spatial and temporal pixel values or patterns, up-sampling; predict MVs, mode index or patterns, *i.e.*, (1) in Fig. 9.	Super-resolution [17]	Sparse coding	/	−4.57			/	/
		Up-sampling [57]	CNN	−5.5	/			7.47	250.21
		Fractional pixel interpolation [65]	GVCNN	/	−2.2	−1.2	−0.9	6.29	1548.6
		INTRA prediction [52]	FC	−3.4	/			91.47	230.11
		INTRA prediction [138]	GAN	−6.75	/			7.0	160
		INTER prediction [137]	CNN	/	/	−1.6	−3.0	1.64	44.7
Transform Coding	Find the optimal transform and quantization kernels to reduce spatial, perceptual redundancies or energy compaction for residual data, *i.e.*, (2).	Transform Index [6]	CNN+FC	−0.2	/			/	
		Transform basis [78]	Annealed learning	−2.10	/	−1.40		/	
		Transform [81] #	SVM	/				/	
		Transform basis# [64]	CNN+FC	/				/	
Entropy Coding	Exploit the probability distribution and statistical redundancies to approaching the entropy, *i.e.*, (3).	Binarization [93]	CNN+FC	−0.33 to −1.13	/			/	
Enhance-ment	Image restoration, filtering, de-noising, residue prediction, *i.e.*, (4) in Fig. 9.	In-loop filter [74]	IFCNN	−4.8	−1.9	/	−2.6	/	/
		In-loop filter [133]	RHCNN	/	−3.38	−4.26	/	<3	/
		In-loop filter [32]	CNN	−4.2	−4.7	−6.0	−5.9	2.14	156.1
		In-loop filter [56]	DenseNet	/	/	/	−6.59	/	
		Post-processing [19]	VRCNN	−4.6	/			/	/
		Reconstruction [122]	MRRNN	−6.7	−7.8	−7.6	/	2.10	240.21

* The BDBR and complexity are average values which are highly correlated with the test sequences, settings and reference models.
# for image coding.

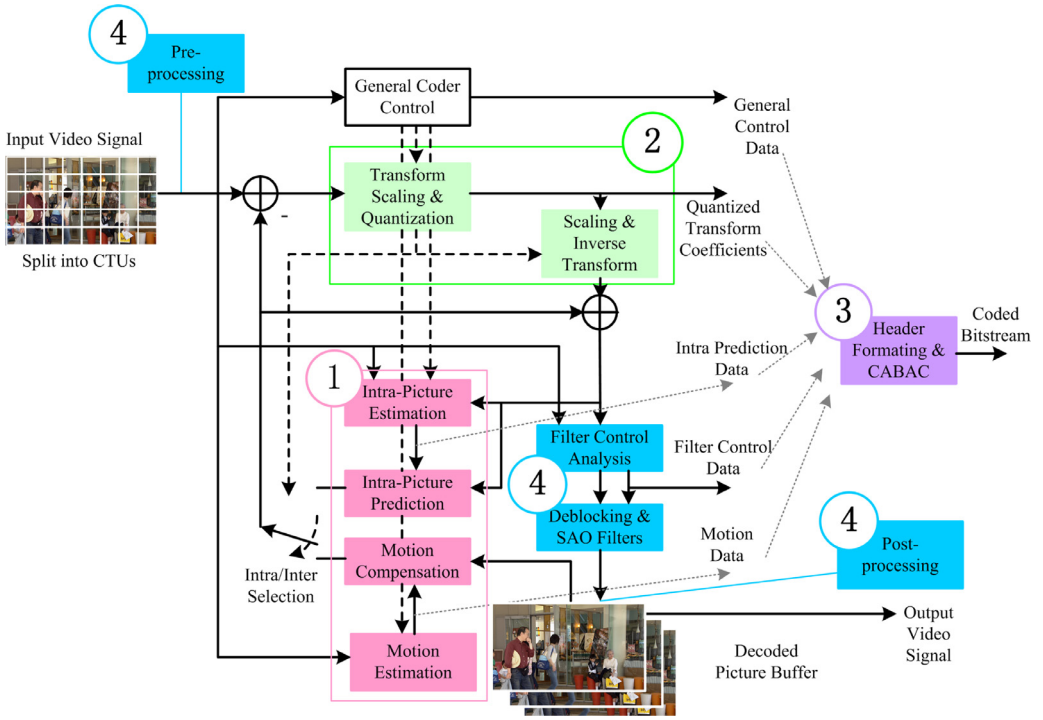


Fig. 9. Learning optimized coding modules for higher compression efficiency [95].

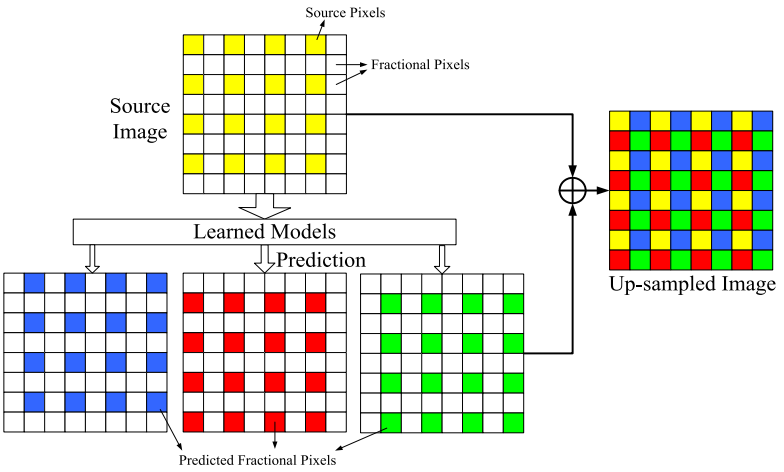
5.1. Learning based predictive coding

Due to high spatial fidelity and capturing frame rate, 2D video contents are strongly correlated in spatial and temporal domains. Predictive coding, *i.e.*, INTRA and INTER predictions, is developed to remove the spatial and temporal redundancies. Only a small number of pattern vectors and the difference between original and predicted data are encoded and transmitted. Therefore, the basic problem of predictive coding can be formulated as finding a mapping function $f()$ to minimize the difference between the original ($\mathbf{X}_O$) and predicted blocks ($\mathbf{X}_P$), which is presented as

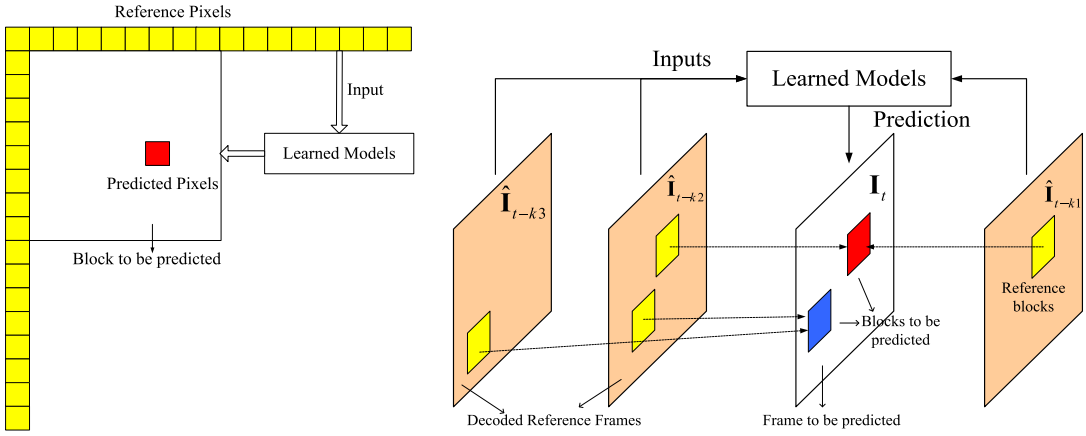
$$\begin{cases} f* = \arg\min_{f} \|\mathbf{X}_O - \mathbf{X}_P\|_{p,q}, s.t.\ r < r_T \\ \mathbf{X}_P = f\left(\hat{\mathbf{X}}_t, \hat{\mathbf{X}}_{t-k}\right) \end{cases}, \tag{4}$$

where $||\cdot||_{p,q}$ is $L_{p,q}$ norm operation. It represents Euclidean norm when q and p are 2 and 1, or Frobenius norm when q and p are 2; r is the coding bit indicating the optimal parameters or side information of mapping function $f()$, r_T is a target bit. $\hat{\mathbf{X}}_t$ and $\hat{\mathbf{X}}_{t-k}$ are reconstructed spatial and temporal blocks with respect to $\mathbf{X}_O$ in time t and t-k. Note that the transform and quantization are not yet considered in Eq. 4. It is to find the optimal mapping function $f^*()$ to predict the block $\mathbf{X_O}$ from the spatial or temporal neighboring blocks by minimizing their difference. Fig. 10 shows typical three key cases in learning based predictive video coding, which include interpolation, INTRA and INTER predictions. Fig. 10(a) is the spatial interpolation and up-sampling prediction case, in which the learned models are used to predict the neighboring or sub-pixels (blue, red and green dots) from the source pixels (yellow). Finally, source and predicted pixels comprise the interpolated image. Fig. 10(b) shows the spatial INTRA prediction, where the learned models are used to predicted the white blocks with the surrounded spatial neighboring pixels (yellow ones). Fig. 10(c) is the INTER prediction from temporal neighboring blocks, which is a temporal interpolation.

Sparse dictionary learning (DL) is able to find a sparse representation of input data in the form of a linear combination of multiple basic elements, which are also denoted as atoms in a dictionary. It has been widely used in image processing for super-resolution/interpolation [100], de-noising and reconstruction [98]. INTER and INTRA predictive coding of the hybrid coding framework can be formulated as an up-sampling problem, in which blocks are down-sampled in predictive coding for low bit rate and then up-sampled to its original resolution at decoder for high visual quality reconstruction. The up-sampling problem can be solved by training over-complete dictionaries to improve the reconstruction quality from the low-quality visual data. In [109], Xiong et al. proposed a sparse spatio-temporal representation for reconstruction of frames. Then, online learning [17] had been utilized to improve the convergence rate of the dictionary learning. Furthermore, multi-scale [99] and progressive dictionary learnings [18] were used to extensively learn spatio-temporal dictionaries, which exploited the inter-layer correlations between base and enhancement layers for quality, spatial and temporal scalable video coding. In



(a) Interpolation/Up-sampling prediction



(b)INTRA prediction (c) INTER prediction

Fig. 10. Three key cases in learning based predictive coding.

addition, the up-sampling problem was also exploited by using the latest CNN [21,44,57,40]. In [21], Dong et al. proposed an image Super-Resolution CNN (SRCNN), which learned an end-to-end mapping between the low and high-resolution images by using a lightweight structured CNN. In [44], Kim et al. increased the network depth to 20 layers inspired by the VGG-net and proposed a very deep convolutional network for image super-resolution, denoted as VDSR. In [57], Li et al. proposed CNN-based block up-sampling scheme for INTRA frame coding, which included block-wise up-sampling in INTRA coding loop and frame-wise up-sampling to refine block boundaries. Based on the down/up-sampling based coding framework, CNN was trained on both the spatial and the temporal dimensions of compressed videos to enhance their spatial resolution [40]. Figs. 11 and 12 show visual quality and PSNR comparisons on the interpolation results from conventional bi-cubic, sparse coding, SRCNN and VDSR. We can observe that the learning based schemes can significantly improve the quality of super-resolved images with different contents, which is about 2.27 dB to 7.06 dB. CNN based schemes [21,44] are able to achieve better results. Since the sub-pixel ME and Motion Compensation (MC) requires sub-pixel interpolation, CNN was adopted to fractional pixel MC [116,117,65] by predicting fractional-pixel from the integer-pixel. In [117], a Fractional-pixel Reference generation CNN (FRCNN) was proposed to generate the reference image for fractional-pixel ME in HEVC, where multiple specified FRCNNs were trained for interpolating the fractional pixels at different locations. It achieved 1.3% to 3.9% BDBR gain on average while the complexities were increased to 6.94 times for encoder and 80.77 times for decoder, respectively. Liu et al. [65] proposed a one-for-all fractional interpolation method based on a deeper and Grouped Variation CNN (GVCNN) to improve the interpolation generality on handling different QPs and sub-pixel locations. It achieved 0.9%, 1.2% and 2.2% average BDBR gain for RA, LDB, and LDP configurations. However, the complexities of the encoder and decoder increased to 6.29 and 1548.58 times of those of the original HEVC, which seems unaffordable.

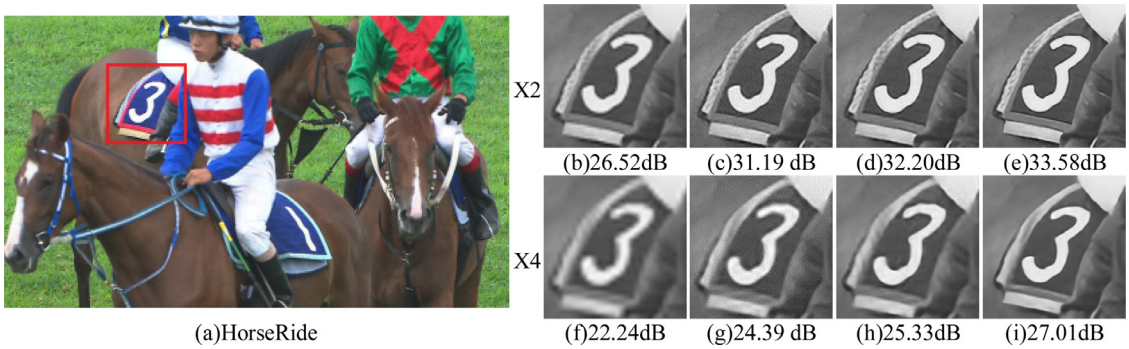


Fig. 11. Visual examples on interpolation with 2 and 4 times scales for HorseRide. (a) Original image, (b)-(i) enlarged region in (a), (b)-(e) are interpolated with scale X2, (f)-(i) are interpolated with scale X4, (b)(f) Conventional, (c)(g) Sparse coding [100], (d)(h)SRCNN [21], (e)(i) VDSR [44].

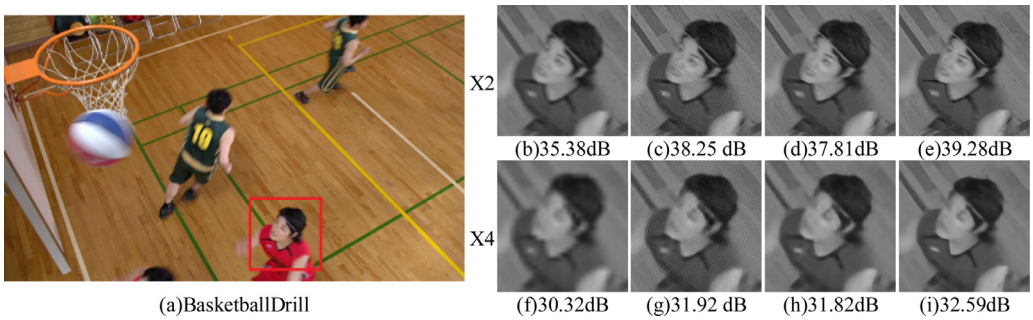


Fig. 12. Visual examples on interpolation with 2 and 4 times scales for BasketballDrill. (a) Original image, (b)-(i) enlarged region in (a), (b)-(e) are interpolated with scale X2, (f)-(i) are interpolated with scale X4, (b)(f) Conventional, (c)(g) Sparse coding [100], (d)(h)SRCNN [21], (e)(i) VDSR [44].

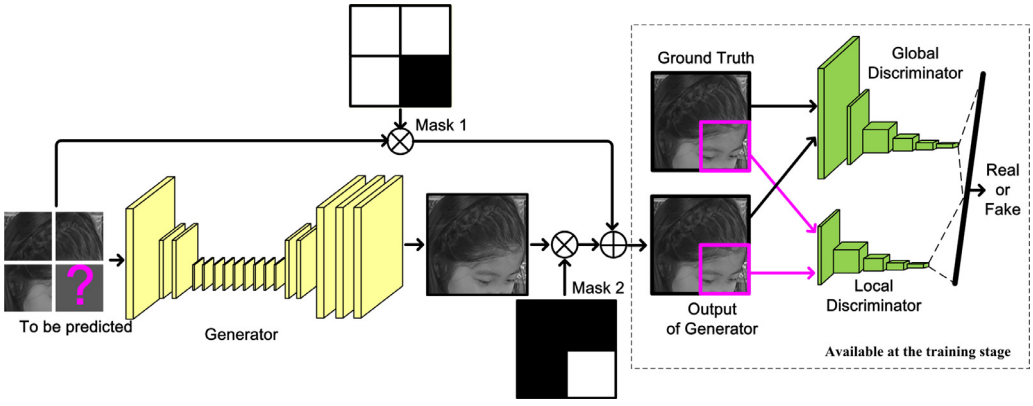


Fig. 13. Architecture of GAN based INTRA prediction [138].

Besides the up-sampling problem, the predictive coding can also be modelled as predicting the spatial and temporal pixel values or patterns, as shown in Fig. 10(b) and (c). Li et al. [52] proposed an efficient multiple lines based prediction method, in which more reference lines were utilized besides the nearest neighboring row and column. A fully connected network was adopted to learn an end-to-end mapping from neighboring reconstructed pixels to the current block [53], where more contextual information of the current block was fed into the fully connected network. One limitation is that the networks shall be trained for each block size. In [138], predicting the pixels from neighboring CUs in the INTRA prediction was modeled as an inpainting problem and accomplished with a Generative Adversarial Network (GAN) model, whose architecture is shown in Fig. 13. The right bottom CTU in the mask was about to be predicted from the left, above and left-above reconstructed CTUs. The global and local discriminators were used to improve learning the INTRA prediction generator. Additional 35 INTRA mode candidates were generated for each CTU and incorporated in encoder and decoder for RD comparison in RDO process. Fig. 14 shows INTRA prediction examples from GAN based and conventional angular prediction in HEVC for 64×64 CU, where SAD is measured. The GAN based scheme has better prediction results in terms of SAD and visual quality. About 6.6% and 6.75% BDBR gains on average were achieved for HEVC and VVC, respectively, which were very

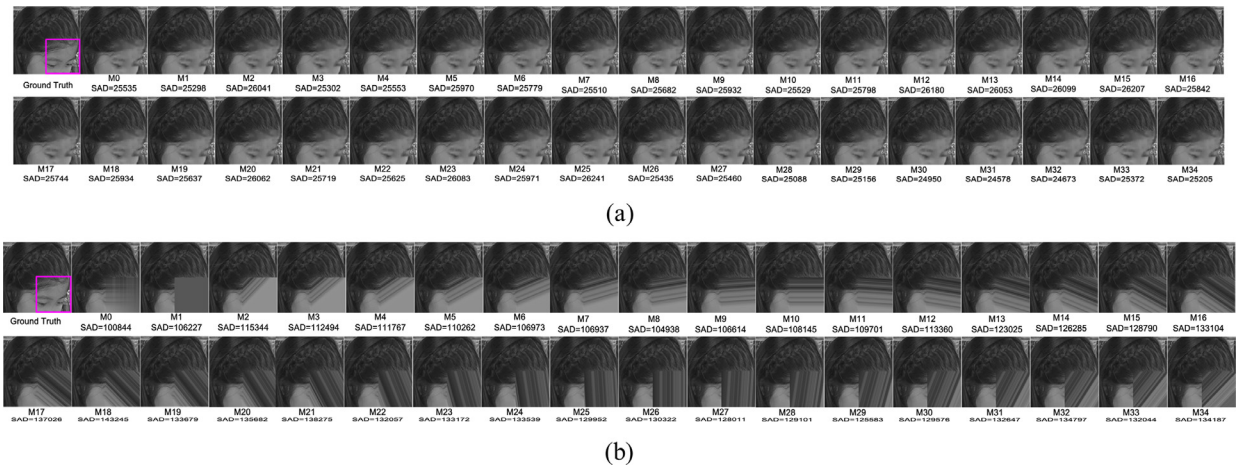


Fig. 14. Examples of Intra prediction results [138]. (a) GAN based INTRA prediction. (b) Angular based INTRA prediction in HEVC.

promising. However, the encoding and decoding complexities increased significantly, which were 7.0 and 160 times on average for HEVC and 2.5 and 257 times on average for VVC, respectively. Chen et al. [9] proposed the concept of VoxelCNN in modelling spatio-temporal coherence for predictive coding, which was then included in ME and hybrid prediction networks. In addition, bidirectional RNN was studied to exploit the temporal information for prediction in video compression [47]. By utilizing more pixel information and advanced prediction, the coding efficiency has been improved for INTRA frames. Zhao et al. [137] employed a CNN to infer the predictive blocks for INTER bi-prediction, while in [136] the reference images were improved to enhance the INTER prediction accuracy. The learning based high efficiency predictive coding for temporal frame is at the initial stage and requires further investigations.

5.2. Learning based transform coding

The transform coding aims to transform an $n \times n$ block residual pixels $\mathbf{X}$ into $n \times n$ coefficients $\mathbf{Y}$ with transform basis $\mathbf{C}$, where $\mathbf{Y}$ is sparse representation of $\mathbf{X}$ by removing the spatial correlations or higher "energy compaction". The transform can be presented as

$$\mathbf{Y} = \mathbf{C}\mathbf{X}\mathbf{C}^{\mathbf{T}}, \tag{5}$$

where $\mathbf{C}^{\mathbf{T}}$ is the a transpose matrix of $\mathbf{C}$. After the transform, the coefficients $\mathbf{Y}$ is quantized for lossy coding with a quantization matrix, which is given by

$$\mathbf{Z} = Q(\mathbf{Y}), \tag{6}$$

where $Q()$ represents the quantization process for a block. $\mathbf{Z}$ is a matrix of quantized coefficients for entropy coding. The reconstructed pixel values $\mathbf{X}'$ after inverse quantization and inverse transform are

$$\mathbf{X}' = \mathbf{D}Q^{-1}(\mathbf{Z})\mathbf{D}^{\mathbf{T}}, \tag{7}$$

where $\mathbf{D}$ is the inverse transform core, which is usually correlated with $\mathbf{C}^{\mathbf{T}}$ for symmetric transform, $Q^{-1}()$ is an inverse quantization**.**

The optimization problem for the transform coding aims to find the optimal transform core $\mathbf{C}$ and quantization $Q()$ to minimize the distortion between $\mathbf{X}$ and $\mathbf{X}'$ subject to limited coding bits of $\mathbf{Z}$, *i.e.,* $r(\mathbf{Z})$, and overhead bits r_c, which can be formulated as

$$\{\mathbf{C}^*, Q^*\} = \underset{\mathbf{C},Q}{argmin}\, ||\mathbf{X} - \mathbf{X}'||_{p,q}, s.t. r(\mathbf{Z}) + r_c \le r_T, \tag{8}$$

where $||\cdot||_{p,q}$ is the $L_{p,q}$ norm. r_c is the number of overhead bits indicating transform basis and quantization, r_T is the target bit. It is find the optimal transform kernel $\mathbf{C}^*$ and quantization Q^* to represent $\mathbf{X}$ in a sparser manner, which suits a dimension reduction problem.

Karhunen-Loève Transform (KLT) is an ideal linear transform that aims to project pixel data onto the eigenvectors. It picks basis vectors one by one by minimizing the distance of the data from the subspace they span. The KLT is data-dependent and source data is needed to estimate the transform core. Then, the core is required to be transmitted to the decoder, which increases the overhead bit r_c. Discrete Cosine Transform (DCT) is a Fourier-related transform using only real numbers, which approaches the KLT's performance for Markov processes. It has been widely used in JPEG and the first generation of coding standards because of the fast decomposition for pipeline implementation and strong energy compaction property. In addition, transform basis is fixed. Based on the DCT, 4×4 Integer Cosine Transform (ICT) was proposed in H.264/AVC by

changing the DCT float-point operations to integer [108]. To improve the energy compaction in transforming larger blocks, Hardmard transform jointly worked with the 4×4 ICT for 16×16 block. In HEVC, different TU sizes (4×4, 8×8, 16×16, and 32×32) and transform cores were proposed based on DCT. The conventional DCT transform assumes the stationary Gaussian distributed signal to obtain the optimal transform, which is not always the optimal for the residue from INTRA prediction. To further improve the performance, content based transform, such as orientation adaptive transform [76] and Mode Dependent Directional Transform (MDDT) [2], were proposed to transform different types of data sources, such as different INTRA prediction modes. The Adaptive Multiple Transform (AMT) [6] was introduced in VVC by testing a set of transform cores in the coding loop and the optimal basis was selected based on the RD cost comparison. Then, an index of the transform basis was explicitly transmitted to the decoder, which could also be improved with learning based prediction [77].

In [85], the MDDT algorithm was modified by introducing l_0-norm regularized optimization in order to obtain robust learning algorithm and enforce the sparsity-constraint in the optimization process. Similarly, in [78], the residual blocks were classified into a number of classes and transformed individually. Then, an annealing based learning technique was adopted to improve the performance. These algorithms are similar to the AMT, and the major difference is using the learning algorithm to predict the best transform basis $\mathbf{C}$ and reduce the overhead index bit r_c. A number of alternative attempts [37,38,81,64] explored to transform the residue in a sparser or more effective manner. In [37], a novel dictionary learning based transform was proposed, in which residues were transformed to the number of transformed coefficients for dictionary construction. In [38], a cascaded sparse/DCT two-layer representation was proposed for coding prediction residues in HEVC, in which a dictionary was trained to present the patterns of structured residual signals for low bit rates and DCT representation was cascaded to reduce overhead bits at higher bit rates. In [81], the SVM was used to approximate the DCT coefficients, in which the data was modelled within the given level of accuracy based on the property that SVM selected a limited number of samples in training, noted as supported vectors. It was applied for image coding and outperformed the JPEG. In [64], CNN was used to simulate DCT-like transform, in which CNN based transform was used to non-linearly map the block pixels into sparse coefficients and a CNN based inverse transform was used to map the coefficients to block pixels. Three convolutional layers and one FC layer were used for CNN training. Distortion and bit rate were considered in the loss function for network training. Considerable improvements over JPEG were reported at low bit rate. However, the existing CNN networks were processed with raw block pixels, fixed block size and image coding, further integrations and investigations will be interesting.

Kuo et al. proposed a feed-forward data-driven subspace approximation with augmented kernels transform, shortened as Saak transform [48], in which the kernels were derived from the second-order statistics of inputs. Besides, an adjusted bias was added to annihilate activation's nonlinearity, which is Saab transform [49]. The Saak and Saab transforms are variants of Principal Component Analysis (PCA) for dimension reduction. Data labels and back-propagation are not required. The Saak and Saab transforms have been successfully tested to image classification and forgeries and achieved promising results. Their applications to video coding will be interesting and worthy investigations.

The transform coding is a type of dimension reduction problem following predictive coding. Data dependent transform is a promising trend. Up to now, it has been proved that the learning based transform coding outperforms the JPEG and INTRA coding. However, more investigations on their effectiveness to INTER coding are still required. As such, how to improve the effectiveness and integrate the learning based transform coding with quantization and predictive coding will be open problems in the near future. Meanwhile, the genericity versus specificity of the learning based transform is required to be considered.

5.3. Learning based enhancement algorithms

A distorted image or block ($\mathbf{y}$) can be presented as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \varepsilon, \tag{9}$$

where $\mathbf{x}$ is the original image/block, $\mathbf{H}$ is the degradation matrix, $\boldsymbol{\varepsilon}$ is an additive noise. The distortion in (9) may have different forms as the matrix $\mathbf{H}$ changes. To recover the original image $\mathbf{x}$ from different types of distortions, such as blocking, ringing and blur artifacts, the general enhancement problem is to recover the optimum $\mathbf{x}^*$ by

$$\mathbf{x}^* = \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_{q,p} + \lambda\Omega(\mathbf{x}), \tag{10}$$

where $\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_{q,p}$ is the fidelity term of $L_{q,p}$ norm, $\Omega(\mathbf{x})$ is the constraint term to limit the optimum solutions range, and λ is a parameter for trade-off. If $\mathbf{H}$ is an identity matrix, $\mathbf{y}$ only contains the additive noise and (10) becomes an image de-noising problem; if $\mathbf{H}$ is the blurring operator, (10) denotes a de-blurring problem; If $\mathbf{H}$ is the down-sampling operator or a composite operator combining blurring and down-sampling, (10) becomes an image super resolution problem; If $\mathbf{H}$ is a sampling matrix or a mask, (10) represents the image inpainting problem. The better the recovered $\mathbf{x}$, the better coding performance may achieve when the enhancement is applied to video coding. Based on the types of image optimization problem and the modules applied in video coding, the enhancement algorithms can be divided into two categories, i.e. in-loop filtering and pre-/post-processing.

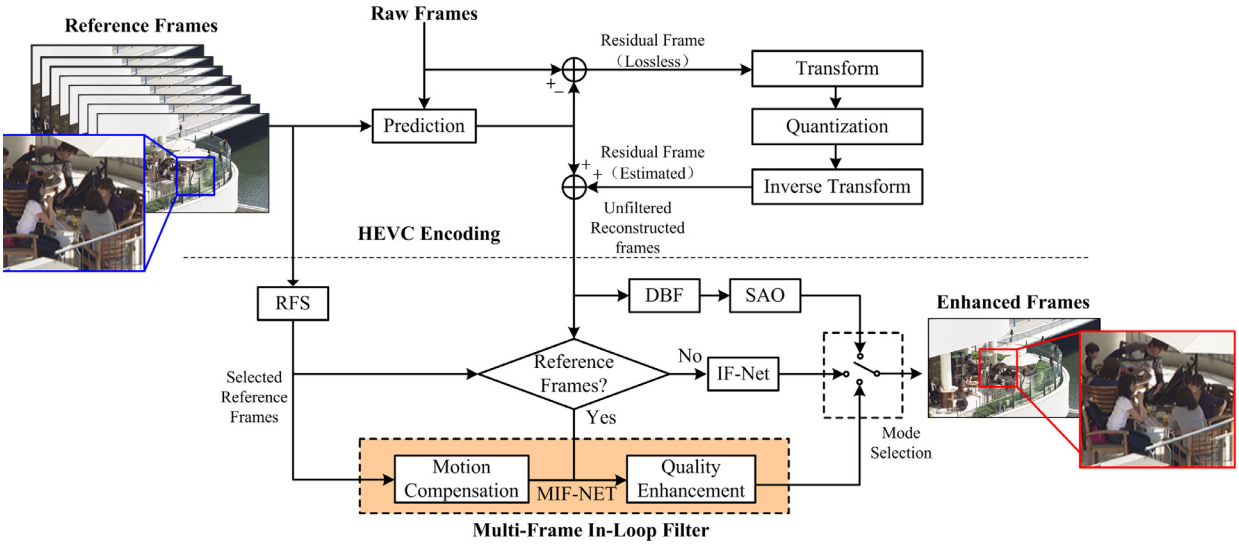


Fig. 15. Framework of MIF algorithm [56].

The first category focuses on the in-loop filtering module of the codec, where the filtered blocks will be used as reference for subsequent blocks in the encoding or decoding processes. In [12], a self-learning based image de-blocking framework was proposed, where the de-blocking was formulated as a Morphological Component Analysis (MCA) based image decomposition task. In this scheme, the low and high-frequency parts were decomposed by using Block Matching and 3D (BM3D) filtering algorithm. Park et al. [74] presented an In-loop Filtering using CNN (IFCNN) to improve the quality of reconstructed image, in which the signaling bits were not required because using the same trained weights in both encoders and decoder. To further improve the coding efficiency, a Residual Highway CNN (RHCNN) was proposed for in-loop filtering [133], where the RHCNN consists of several residual highway units and convolutional layers. The RHCNN was employed as a high-dimensional filter to enhance the quality of reconstructed frames besides the de-blocking and Sample Adaptive Offset (SAO) filters in HEVC, which reduces bit rate 3.38% on average at cost of about 3 times encoding complexity. Jia et al. [32] designed a multi-dimensional CNN model for loop filtering and content-aware multi-model CNN was performed for different regions. It achieves 4.2% to 6.0% BDBR gains for INTRA and INTER configurations while the encoding and decoding complexities were increased to 2.14 times and 156.10 times. In addition, Li et al. [56] proposed a Multi-frame In-loop Filter (MIF) scheme for HEVC based on DenseNet, in which multiple successive temporal frames were jointly used to enhance the filtering performance. In addition, there is a mode selection between conventional SAO, in-loop filter and the MIF, as shown in Fig. 15. These schemes enhanced the quality of the reconstructed blocks in the loop of coding process, which not only improved the quality of reconstructed images, but also improved the coding efficiency when the enhanced image/block was used as reference for subsequent INTRA or INTER prediction. These in-loop enhancements may have sequential data dependency among blocks, which has negative impacts on the parallelism. In addition, the in-loop enhancement will be recalled for many times in RDO, which increases the coding and decoding complexity dramatically.

The other category of works is pre-/post-processing algorithm for image enhancement. In [19], a CNN-based post-processing algorithm for HEVC was presented, in which a Variable filter size Residue-learning CNN (VRCNN) was adopted to improve the coding performance and accelerate network training. Different from the conventional schemes on optimizing the encoder, Wang et al. [105] focused on improving the video quality at the decoder side, where a very deep CNN was applied to eliminate the distortions and enhanced the quality of HEVC-compressed videos. At the receiver side, the low quality face regions were recovered with the face patches in the database. A deep CNN was used to mimic the reverse function of video coding in [61], which established an end-to-end mapping from the decoded frame to an enhanced one. Zhang et al. [134] proposed an Adaptive Residual Network (ARN) for high-quality image restoration in video coding, where short-cuts were used to reduce the model complexity. In [140], the CNN models were used to improve the quality of synthesized images, which was applied to both in-loop View Synthesis Optimization (VSO) in 3D-HEVC encoder and the out-loop reconstruction at the decoder. In [122], a novel quality enhancement method was proposed by using a Multi-reconstruction Recurrent Residual Network (MRRN), where a recursive residual structure was designed to capture the multi-scale similarity of compression artifacts. Compared with the in-loop filter, pre-/post-processing module will be conducted in the reconstruction phase out of the coding loop. They can be used for both encoder and decoder, or only at the decoder side. In addition, the data dependency is weaker since it is out of the coding loop, which facilitates the block-wise parallelism in video coding. However, the coding gains are lower than those of in-loop schemes. In addition, due to various distortion levels caused by quantization, multiple learning models shall be trained for different QPs [133,56,19,122], which could be improved.

5.4. Discussions

The key to the high efficiency coding is predicting the sample values, patterns or coefficients for de-correlation. Dictionary learning and deep learning are two powerful tools that can be exploited to improve the prediction and the coding efficiency. However, deep learning based optimizations are in fact difficult to be theoretically explained and reproduced, even with the same training data and parameter settings. In addition, the related learning based works mainly focused on the optimizations for prediction, enhancement and transform coding. Entropy coding is rarely addressed. Only one recent work focused on entropy coding [93], where the NN was used to predict the probability distribution of the syntax INTRA modes for multi-level arithmetic engine. It saved about 0.33% and 1.13% bit rate under different settings. The major reason is that it is more difficult to model the entropy coding problem and fully exploit the advantages of learning algorithms.

The learning based schemes, especially those using the deep learning, were able to achieve promising coding gains but caused much more complexity overhead and hardware cost, which increased multiple times for the encoder and hundreds or thousands times for the decoder although the GPU acceleration was enabled [65,138]. Low complexity or low cost adaptation for these learning based schemes is highly desired for standardizations and practical usage. Since videos are not only used for viewing, but may also be used for recognition tasks, such as face recognition, robotic vision or retrieval, in addition to maintaining high visual quality, key features correlated with recognition tasks shall be also preserved [20]. In this case, coding optimization by considering multi-objectives and features deserves future investigations.

A number of researchers are also seeking the possibility of proposing new learning based coding framework that differs from the hybrid block based coding and adapts to different environments, such as distributed coding for mobile or cloud computing environments [94]. In addition to applying the CNN to existing coding modules, new end-to-end compression framework [33] was also exploited. The coding efficiency of these coding algorithms was reported better than INTRA coding and there is a large potential to achieve higher gain.

6. Learning based visual quality assessment (VQA)

The objective of video coding is to minimize the distortion (D) or maximize the quality (Q) subject to bit rate (R) constraints, which can be presented as

$$\min D,\ s.t. R \leq R_T, \tag{11}$$

where R_T is a target bit rate. Nowadays, the distortion D is still measured with MSE while the quality Q is measured with PSNR, which are based on the pixel-by-pixel difference between the original and reconstructed images. PSNR and MSE are simple but can hardly reflect the real perceived quality of HVS. To fully exploit the perceptual redundancies in videos, perceptual video encoder is a possible tentative solution, to which developing a perceptual quality metric Q that is consistent with HVS becomes the key. Up to now, many Visual Quality Assessment (VQA) metrics have been developed, such as SSIM [106], FSIM [124], Multi-Scale SSIM (MS-SSIM) [107], MOVIE [84] and so on. However, there is still no such perceptual quality metric that is universally accepted as compared with PSNR. HVS is a complicated non-linear system. Although many important perceptual factors are revealed in psychological and physiological perspectives, the understandings on HVS and human brain are still very limited and under explorations. It is challenging to develop an effective visual quality metric consistent with human perception. Machine learning provides new opportunities by mining visual factors from data and achieving data-driven solutions. In this section, VQAs are categorized into two major classes, subjective and objective VQAs, and analyzed in detail.

6.1. Subjective VQA and labelled datasets

In subjective VQA, a group of subjects are invited to give scores on the quality of a series of distorted images or videos under given procedures and testing environments. Then, the processed Mean Opinion Score (MOS) or Differential MOS (DMOS) from the subjects is regarded as the ground truth quality, which reflects the visual responses of HVS. To make the subjective VQA more rigid and rational, Video Quality Experts Group (VQEG) established in 1997 by the ITU-T and ITU-R focus on the subjective and objective visual quality studies. A series of recommendations were released to identify subjective testing methods, procedures and environments, such as ITU-R BT.500 [29] and BT.710 [30] for TV and HDTV images, BT.1788 [79] for video quality, BT.1438 [28] for stereoscopic images, and BT.2021 [80] for 3D video systems.

There are two major contributions of subjective quality assessment experiments. Firstly, they facilitate better understanding of the HVS mechanism and reveal the new visual properties. For example, Contrast Sensitive Function (CSF), visual attention and ROI, visual sensitivity, JND, depth perception, visual masking effects, binocular fusion and rivalry, *etc.*, had been investigated in last few decades. They serve as the groundwork for the feature and model design towards the objective quality metrics. The other contribution is labelling the quality scores of distorted image/videos, which will be the data sources and ground truth labels for designing and verifying the objective quality metrics.

Generally, one dataset consists of a number of source images/videos with diverse contents and spatial-temporal characteristics, denoted as N. Then, they are distorted with a mount of different types of distortions (such as compression, blur, or white noise) and degrees, denoted by P and Q, respectively. The total number of distorted image/videos is $N \times P \times Q$. Each

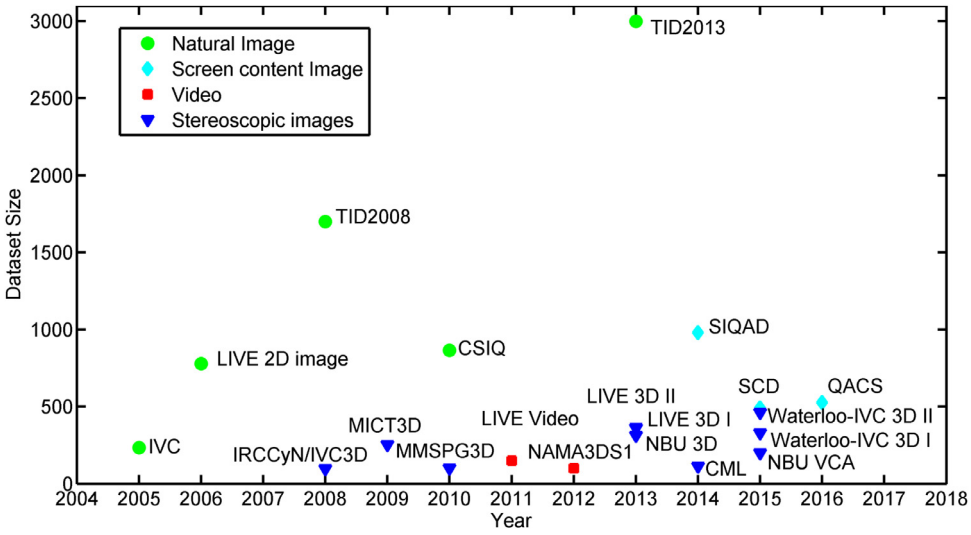


Fig. 16. Scales of distorted images/videos in typical VQA datasets.

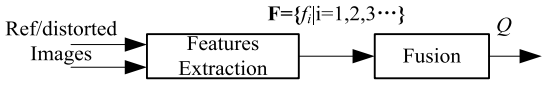
of them will be viewed and scored by a number of human subjects, *e.g.*, 35, and the statistical mean quality after excluding outliers will be the quality label of one distorted image/video. Fig. 16 shows the number of distorted image/videos in typical VQA datasets since 2000, which includes datasets for natural image, Screen Content Image (SCI), stereoscopic image and videos. We can observe that the largest TID2013 dataset has around 3000 labelled images, and most image datasets are less than 1000. In addition, the video datasets are even less and their labelled data is only 100 to 150, which is a very small number. Although more and more datasets are labelled and released in recent years, they can hardly be used jointly due to different properties (SD/HD, 2D/3D, image/video, and SCI/natural contents), distortions, type of scales (quality/depth, MOS/DMOS, normalization) and settings (displays, testing environments, viewing conditions, procedures). In addition, the image or video MOSs are usually obtained from mean values of discrete five-grade scale quality 5 to 1, corresponding to "Excellent", "Good", "Fair", "Poor" and "Bad", which may not be accurate enough in measuring the coding distortion levels, *e.g.*, 51 for HEVC and 100 for JPEG in compression. Fine-grained subjective quality assessment [126] and datasets [104,62] are an interesting and rising research topic, however, it is much more laborious than the coarse-grained.

In fact, it is laborious, expensive and time-consuming to perform the subjective VQA labeling large number of images or videos [92]. Nevertheless, the subjective quality cannot be applied to other unknown videos and in-loop visual signal processing. Therefore, objective quality metrics are desired. As for designing an objective quality metric, the understandings of visual perception mechanism are key issues to develop effective features. Also, the dataset will be important data source for training, validating and testing the learning based VQA. Creating large and fine-grained datasets [126,92] is challenging but vital necessary in developing reliable and capable VQA models.

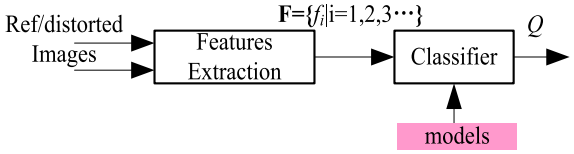
6.2. Machine learning based VQA

The other category is the objective quality assessment, which estimates the quality scores of distorted videos. According to the availability of the reference, VQA metrics can be categorized as Full Reference (FR), Reduced Reference (RR), and No Reference (NR). FR requires a 'perfect' quality image/video, *i.e.,* the reference is fully available, RR needs partial side information and NR doesn't have any reference information. MSE and PSNR are FR VQA metrics, which have been widely used in video coding due to their simplicity. They are straightforward but not accuracy enough to present the perceived quality in HVS. With more understandings on the HVS, many featured VQAs have been developed to represent distortion in images and videos.

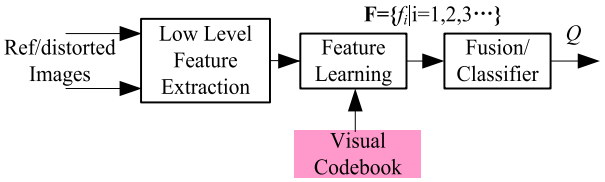
According to the feature and fusion algorithms, VQA metrics can be classified into four categories: handcrafted feature based, handcrafted feature plus learning based, feature learning based and end-to-end learning based approaches, as shown in Fig. 17. Table 4 summarizes the features and learning algorithms for typical VQA metrics in the four types. In the first category, the PSNR and MSE adopted the squared pixel difference between distorted and reference images as the only feature. SSIM [106] introduced luminance, contrast, and structure comparison and combined the three indices to assess image quality. In addition, Visual Information Fidelity (VIF) [87] and gradient similarity [63] were further considered. These are image quality metrics. A Video Quality Metric (VQM) [75] was proposed by adopting seven key features, including spatial information loss, shift of edges from/to diagonal orientation, spread of chroma, spatial gain from enhancement, temporal impairments, and localized color impairments. They were then combined with weighted summation and weights were empirically determined. Seshadrinathan et al. proposed MOtion-based Video Integrity Evaluation index (MOVIE) [84] for video quality assessment, in which spatial quality was estimated using the Gabor coefficients and then tuned with motion infor-



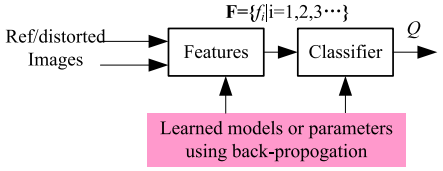
(a) Handcraft feature based approaches.



(b) Handcraft feature plus learning based approaches.



(c) Feature Learning based approaches.



(d) End-to-end deep learning based approaches.

Fig. 17. The pipeline of learning based VQA prediction.

Table 4
Features and learning methods of some typical VQA metrics.

Types	Metrics	Key Features	Learning
1st	PSNR	Squared pixel difference	\
	SSIM [106]/MSSIM [107]	Luminance, contrast, structure	\
	VIF [87]	Visual information fidelity	\
	VQM [75]	Spatial loss, shift of edges from/to diagonal orientation, chroma spread, spatial gain, temporal impairments, localized color impairments	\
2nd	Lin [60]	Gabor	SVR
	Xue [115]	GM, LoG	SVR
	DIIVINE [71]	Wavelet coefficients, GGD	SVM,SVR
	Vega [102]	Bitrate, multiple frames, video motion, noise ratio, scene complexity, blur mean, blockiness, motion intensity	RBM
	Yang [119]	HoG of wavelet coefficients, magnitude, variance and entropy, depth perception map	DBN
	VMAF [58]	AN-SNR, DLM, VIF, MCPD	SVM
3rd	Shao [86]	Dictionary learning	\
	Zhang [135]	Dictionary learning	\
	CORNIA [120]	Visual codebook	SVM
4th	[39,7,24,127,4,118,41,43,45]	CNN/RNN	

mation. Perceptually weighted MSE [27] was used in pooling by considering different importance of image/video regions. The working pipeline of this category VQA metrics is extracting handcrafted features and empirically fused for quality prediction, as shown in Fig. 17(a). They are usually simple and straightforward, but are hard to adapt to various image/video contents. As more brain-inspired features are disclosed, incorporating more features can improve the prediction accuracy and adaptation. However, it becomes more and more challenging to fuse high dimensional features.

In the second category, the machine learning algorithm has been introduced to address the problem of fusing multiple features. Lin et al. [60] developed a FR quality metric for stereoscopic images. It extracted Gabor features for left-view and right-view, which were combined together via binocular rivalry model. Finally, it learned a mapping function between the features and quality score by using Support Vector Regression (SVR). Xue et al. [115] exploited Gradient Magnitude (GM) and Laplacian of Gaussian (LoG), and then learned a regression model using SVR too. Moorthy et al. [71] developed a two-step framework for NR image quality assessment named DIIVINE. It firstly classified image distortion types by using SVM based on wavelet coefficients and Generalized Gaussian Distribution (GGD), and then predicted image quality score using SVR. Literature [102] proposed a learning based RR video quality assessment for live streaming, in which unsupervised Restricted Boltzmann Machines (RBMs) were learned at the server side and NR measurements were performed at the client side by using transmitted RBM models. Yang et al. [119] applied Deep Belief Network (DBN) for blind metric evaluating stereo images by fusing 2D handcrafted features and 3D depth perception map. Li et al. [58] proposed Video Multi-Method Assessment Fusion (VMAF) algorithm that predicted the quality score by fusing four methods, including Anti-noise SNR (AN-SNR), Detail Loss Measure (DLM), VIF and Mean Co-Located Pixel Difference (MCPD), with a SVM classifier. It claimed better reflection on human perception of video quality than other widely used objective metrics and has been used in industry, such as video streaming in Netflix. The work pipeline of these works is shown in Fig. 17(b). Firstly, the handcrafted features are extracted and trainable learning algorithms are used to map features to a final quality score. The performance of these metrics highly relies on the effectiveness of the extracted handcrafted features. Due to limited labeled data for model training, over-fitting may easily occur, and the effectiveness on cross-database validation needs to be further improved. The available video datasets are even less and few literatures are done on learning based VQA for videos.

The third category is feature learning based approaches, in which the features are learned from data. They attempt to solve the difficulties on extracting effective features, as shown in Fig. 17(c). Shao et al. [86] learned dictionaries to represent the latent structure of images by using a set of basis vectors, *i.e.,* learning the representative features. Then, quality indices generated from using sparse coefficient vectors were fused. Ye et al. [120] introduced an unsupervised learning framework, called CORNIA, for NR image quality assessment, in which a visual codebook was learned from unlabeled image patches and then the quality was predicted using supervised linear SVM. Zhang et al. [135] restructured video with plurality of temporal layers and learned temporal dictionaries to represent the flickering artifacts in synthesized 3D video. They are capable of learning more discriminative features beyond the handcrafted features. Meanwhile, features are learned in representing images/videos and the quality labels are not required. However, higher level features, such as sparse coefficient vectors [86], phase and amplitude similarities [135], shall also be manually designed. This process can be regarded as a feature transformation.

With the breakthrough of deep learning in image recognition, researchers have tried to apply deep learning to VQA recently [39,7,24,127,4,118,41,43,45], *i.e.,* the fourth category as shown in Fig. 17(d). Kang et al. [39] employed CNN in general-purpose NR image quality assessment, which combined feature learning and quality regression as a holistic and end-to-end way. Bosse et al. [7] adopted the Siamese network in image quality assessment, which was trained for FR but could be used for NR by extracting part of the network. Fan et al. [24] proposed a general-purpose NR image quality assessment based on multi-expert CNN. It classified distortion types via CNN and trained specific expert-CNN for each distortion type. Then, the outputs of each expert-CNN were aggregated to be the final image quality score. Zhang et al. [127] proposed an FR video quality assessment by employing transfer learning with CNN. They pre-trained the network on distorted images and then transferred it to videos due to small-scale video quality database. Bampis et al. [4] formulated the continuous-time video quality prediction as a time-series prediction problem and predicted the quality score by using RNNs. Yan et al. [118] proposed a two-stream CNN based NR image quality predictor where one stream focused on the image intensity and the other learned structure features from gradient. Kim et al. [41] proposed a CNN based NR quality assessment for omnidirectional image. It first predicted quality score for each patch and aggregated the scores together as score of the omnidirectional image. Then a discriminator was learned using adversarial learning to assess the predicted score with the help of human perception guide. These end-to-end learning based VQA metrics can learn features and mapping function from raw visual data automatically and simultaneously. Also, these end-to-end schemes can significantly improve the prediction performance by fitting the MOS data well. However, the HVS mechanisms behind are difficult to be explained and the learned knowledge is not transferable or applicable to other tasks if without re-training.

In addition, one of the most challenging issues is the deep learning based approaches require very large amount of labelled data for training [127,43]. The learning model may be difficult to handle various contents and distortions if the training dataset is not sufficient or fails to adequately represent real world videos [92]. Besides, Shortage of data may probably cause serious over-fitting problem. In fact, only very limited labelled images are available in quality assessment and it is even more critical for video datasets as mentioned in Section 6.1. One tentative solution is data augmentation by labeling the quality of each patch [24], instead of an entire image. This augmentation is usually under the assumption that all the patches in an image or video are with the same quality, which may not hold true. Another attempt is using an existing metric to generate sufficient quality scores for training [45]. It may also be bias since the deep NN is essentially simulating

an existing metric instead of HVS perception. Moreover, in the current stage, models are learned from a large data, but validated and tested on very small number of data. How to solve the data availability problem or how to learn from small data deserves more works. Transfer learning [127,97,73] that transfers learned knowledge from a domain with plenty of data is a possible solution and worthy further investigations.

6.3. Further discussions

When applying the VQA algorithm in the video coding modules as the quality objective, the adaptation of the VQA algorithm from image/video based to block based is required [113]. Then, the adaptation on rate-distortion theory shall be also re-considered since it was initially designed based on the MSE [82,67]. One shortcut way is to build a mathematical relation between VQA and MSE before applying to video coding. However, this approximation decreases the VQA accuracy. In addition, the current VQAs are specific for different applications, such as image [24,118], video [58,127,4], stereo/3D [119,120] or omnidirectional virtual reality [41]. Thus, the perceptual video encoding algorithms will be specifically designed by using these metrics. A general purpose VQA that is applicable to different applications is preferred but hard to achieve. Another challenging issue is the computational complexity. Since more advanced feature extraction tools and learned classifiers are adopted in the quality prediction, the computational complexity increases significantly. High frequent calling the learning based VQA algorithms, especially the deep learning based schemes, in the RDO will make the coding algorithm extremely complex. How to integrate the VQA, especially the learning based VQA with better performance, into the video coding with acceptable complexity is worthy to be studied.

7. Conclusions and future works

7.1. Conclusion

In this article, we present a systemic survey on the recent advances and challenges associated with machine learning video coding optimization, which aims to provide researchers with a strong foundation and open the horizon for data-driven video signal processing. This survey is mainly presented from three key aspects, including learning based low complexity optimization, learning based high efficiency coding optimization and learning based visual quality assessment. In each part, the problem formulation, workflows, key technical advances, advantages and challenge issues are presented. These learning based video coding optimizations can encode videos in a smarter manner and significantly improve the coding performances for some coding modules. Problem formulation is the key to bridge the gap between coding algorithms and machine learning algorithms, which determines the capabilities of exploiting the advantages of learning algorithms and maximizing the coding effectiveness. For better adaptation and effectiveness, feature extraction and selection, instance/sample selection, and cost function shall be properly designed while using the learning tools. End-to-end deep learning is a new emerging tool that can significantly improve the prediction and classification accuracy in video coding. More investigations on problem formulation and adaptation are still required. Meanwhile, this complexity overhead is an important issue that shall be well addressed for practical applications. In summary, learning based coding optimizations do have many advantages and potentials, which will be a promising direction for academic and industrial communities.

7.2. Future works

Based on the review of the related works, there rise a number of promising research directions for future work:

1. Intelligent/smart video coding: not only encode the video in a smarter manner to enhance the coding efficiency, but also enable the encoded videos with more advanced cognitive information, such as face/human recognition, object/even detection, captioning and commenting for intelligent video applications. Merging the analytics and recognition tasks with the encoding tasks not only can re-use the video information more efficiently, but also can reduce the complexity from decoding for recognition, analytics and retrieval. In this case, the video coding shall consider preserving the key features on analytics or recognition.
2. Learning algorithm is one of the key factors on improving the coding performance. It is worth applying advanced learning algorithms, such as active learning, ensemble learning, reinforcement learning, transfer learning and deep learning, to video coding to tackle more complicated decision making problems in the new generation video coding standards, *i.e.,* VVC and beyond.
3. Perception based video coding that explores the perceptual redundancies is a promising research area worthy of future consideration. One of the most essential issues is to develop a well-recognized objective visual perceptual model in term of accuracy, complexity and adaptability, where more understandings on HVS and large scale subjective visual data are required for model training and reliable validation. Meanwhile, rational integration of learning based perception models as cost criterion to each video coding module will be challenging in terms of coding complexity and adaptability. Effective implementations and perception based RDO theory need further investigations.
4. Deep learning based video coding optimization is expected to continuously attract significant research interest, including the prediction, filtering, enhancement, transform, control as well as new coding framework. Higher coding

gains can be expected from the potential technical improvements. However, the computational complexity cost is a critical issue that shall be well considered in the optimization.

5. Low hardware and low cost implementation for machine learning based video coding, especially the deep learning based schemes, will be an important issue for practical usage.
6. The machine learning models, especially for deep learning, highly rely on the training data, *i.e.*, data dependency. For the off-line learning models, they are fixed after training, which may degrade the coding performance once they are applied to a new type of video contents or distributions. It is important to introduce the on-line or transfer learning to update the learned models. Also, it is necessary to consider the trade-off between genericity and specificity of learning models.

In the future, we believe learning based coding optimization is a promising research direction for video coding. More learning algorithms and optimization techniques will be continuously investigated and introduced in video coding to achieve lower complexity, higher coding efficiency, higher visual quality as well as more intelligent functionalities.

Declaration of Competing Interest

I declared that I have no conflict of interest with this submission.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61672443, 61772344 and 61871372, in part by Guangdong Natural Science Foundation for Distinguished Young Scholar under Grant 2016A030306022, in part by the Key Project for Guangdong Provincial Science and Technology Development under Grant 2017B010110014, in part by RGC General Research Fund (GRF) 9042322, 9042489 (CityU 11200116,11206317), Shenzhen International Collaborative ResearchProject under Grant GJHZ20170314155404913, in part by the Shenzhen Science and Technology Program under Grant No. JCYJ20170811160212033 and JCYJ20180507183823045, in part by Guangdong International Science and TechnologyCooperative Research Project under Grant 2018A050506063, in part by Membership of Youth Innovation Promotion Association, Chinese Academy of Sciences under Grant 2018392.

References

[1] J. An, H. Huang, K. Zhang, Y.-.W. Huang, S. Lei, Quad-tree plus binary tree structure integration with JEM tools, *JVET of VQEG and MPEG*, doc. JVET-B0023 (2016).
[2] A. Arrufat, P. Philippe, O. Déforges, Mode-dependent transform competition for HEVC, in: Proc. IEEE ICIP, Quebec City, QC, 2015, pp. 1598–1602.
[3] K. Arulkumaran, M.P. Deisenroth, M. Brundage, A.A. Bharath, Deep reinforcement learning: a brief survey, IEEE Signal Process. Mag. 34 (6) (2017) 26–38.
[4] C.G. Bampis, Z. Li, I. Katsavounidis, A.C. Bovik, Recurrent and dynamic models for predicting streaming video quality of experience, IEEE Trans. Image Process. 27 (7) (2018) 3316–3331.
[5] J. Berent, P.L. Dragotti, Plenoptic manifolds, IEEE Signal Process. Mag. 24 (6) (2007) 34–44.
[6] T. Biatek, V. Lorcy, P. Philippe, Adaptive transforms for inter-predicted residuals in post-HEVC video coding, in: Proc. DCC, Snowbird, UT, 2017 433-433.
[7] S. Bosse, D. Maniry, K.R. Müller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Trans. Image Process. 27 (1) (2018) 206–219.
[8] P. Carrillo, T. Pin, H. Kalva, Low complexity H.264 video encoder design using machine learning techniques, Proc. ICCE (2010) 461–462.
[9] Z. Chen, T. He, X. Jin, F. Wu, Learning for video compression, arXiv:1804.09869v1 [cs.MM], (2018).
[10] Z. Chen, Y. Li, Y. Zha, Recent advances in omnidirectional video coding for virtual reality: projection and evaluation, *Signal Process.* 146 (2018) 66–78.
[11] J.C. Chiang, W.C. Chen, L.M. Liu, K.F. Hsu, W.N. Lie, A fast H.264/AVC-based stereo video encoding algorithm based on hierarchical two stage neural classification, *IEEE J. Select. Topics Signal Process.* 5 (2) (2011) 309–320.
[12] Y.W. Chiou, C.H. Yeh, L.W. Kang, C.W. Lin, S.J.F. Jiang, Efficient image/video deblocking via sparse representation, in: *Proc*. IEEE VCIP, San Diego, CA, 2012, pp. 1–6.
[13] Cisco Visual Networking Index: forecast and methodology 2016-2021 (2017) [online] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html.
[14] Cisco Visual Networking Index: global mobile data traffic forecast update, 2016-2021 White Paper, (2017) [online] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.
[15] G. Correa, P.A. Assuncao, L.V. Agostini, L.A.S. Cruz, Fast HEVC encoding decisions using data mining, IEEE Trans. Circuits Syst. Video Technol. 25 (4) (2015) 660–673.
[16] Q. Dai, J. Wu, J. Fan, F. Xu, X. Cao, Recent advances in computational photography, Chin. J. Electron. 28 (1) (2019) 1–5.
[17] W. Dai, Y. Shen, X. Tang, J. Zou, H. Xiong, C.W. Chen, Sparse representation with spatio-temporal online dictionary learning for promising video coding, IEEE Trans. Image Process. 25 (10) (2016) 4580–4595.
[18] W. Dai, Y. Shen, H. Xiong, X. Jiang, J. Zou, D. Taubman, Progressive dictionary learning with hierarchical predictive structure for low bit-rate scalable video coding, IEEE Trans. Image Process. 26 (6) (2017) 2972–2987.
[19] Y. Dai, D. Liu, F. Wu, A convolutional neural network approach for post-processing in HEVC intra coding, in: Proc. MMM, 2017, pp. 28–39.
[20] L. Ding, Y. Tian, H. Fan, Y. Wang, T. Huang, Rate-performance-loss optimization for inter-frame deep feature coding from video, IEEE Trans. Image Process. 26 (12) (2017) 5743–5757.
[21] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, *IEEE Trans. Patt. Anal. Mach. Intell.* 38 (2) (2016) 295–307.
[22] F. Duanmu, Z. Ma, Y. Wang, Fast mode and partition decision using machine learning for intra frame coding in HEVC screen content coding extension, IEEE J. Emerg. Sel. Topics Circuits Syst. 6 (4) (2016) 517–531.
[23] E.M. Enriquez, A.J. Moreno, M.A. Pellon, F.D. Maria, A two level classification based approach to inter mode decision in H.264/AVC, IEEE Trans. Circuits Syst. Video Technol. 21 (11) (2011) 1719–1732.
[24] C. Fan, Y. Zhang, L. Feng, Q. Jiang, No reference image quality assessment based on multi-expert convolutional neural networks, IEEE Access 6 (2018) 8934–8943.
[25] K. Goswami, B.G. Kim, A design of fast high efficiency video coding (HEVC) scheme based on Markov chain Monte Carlo model and Bayesian classifier, IEEE Trans. Indust. Electron. 65 (11) (2018) 8861–8871.

[26] Q. Hu, X. Zhang, Z. Shi, Z. Gao, Neyman-Pearson-based early mode decision for HEVC encoding, IEEE Trans. Multimedia 18 (3) (2016) 379–391.
[27] S. Hu, L. Jin, H. Wang, Y. Zhang, S. Kwong, C.-C.J. Kuo, Objective video quality assessment based on perceptually weighted mean squared error, IEEE Trans. Circuits Syst. Video Technol. 27 (9) (2017) 1844–1855.
[28] ITU-R BT.1438-0. Subjective assessment of stereoscopic television pictures (2001).
[29] ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television pictures (2002).
[30] ITU-R BT.710-4. Subjective assessment methods for image quality in high-definition television (2001).
[31] ITU-R Rec. BT.2020. Parameter values for ultra-high definition television systems for production and international programme exchange (2012).
[32] C. Jia, S. Wang, X. Zhang, S. Wang, J. Liu, S. Pu, S. Ma, Content-aware convolutional neural network for in-loop filtering in high efficiency video coding, IEEE Trans. Image Process. (2019) in press, doi:10.1109/TIP.2019.2896489.
[33] F. Jiang, W. Tao, S. Liu, J. Ren, X. Guo, D. Zhao, An end-to-end compression framework based on convolutional neural networks, IEEE Trans. Circuits Syst. Video Technol. (2017) in press, doi:10.1109/TCSVT.2017.2734838.
[34] Z. Jin, P. An, L. Shen, C. Yang, CNN oriented fast QTBT partition algorithm for JVET intra coding, in: Proc. IEEE VCIP, St. Petersburg, FL, 2017, pp. 1–4.
[35] Z. Jin, P. An, L. Shen, Fast QTBT partition algorithm for JVET intra coding based on CNN, in: Proc. PCM, 2017, pp. 59–69.
[36] , Joint call for proposals on video compression with capability beyond HEVC, *JVET of VQEG and MPEG*, Doc. JVET-H1002(v6), 8th Meeting, 2017.
[37] J.W. Kang, Structured sparse representation of residue in screen content video coding, Electron. Lett. 51 (23) (2015) 1871–1873.
[38] J.W. Kang, M. Gabbouj, C.-C.J. Kuo, Sparse/DCT (S/DCT) two layered representation of prediction residuals for video coding, IEEE Trans. Image Process. 22 (7) (2013) 2711–2722.
[39] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, in: Proc. CVPR, 2014, pp. 1733–1740.
[40] A. Kapperler, S. Yoo, Q. Dai, A.K. Katsaggelos, Super-resolution of compressed videos using convolutional neural networks, in: Proc. IEEE ICIP, 2016, pp. 1150–1154.
[41] H. Kim, H. Lim, Y. Ro, Deep virtual reality image quality assessment with human perception guider for omnidirectional image, IEEE Trans. Circuits Syst. Video Technol. (2019) 1-1. (In press).
[42] H.-.S. Kim, R.-.H. Park, Fast CU partitioning algorithm for HEVC using an online-learning based Bayesian decision rule, IEEE Trans. Circuits Syst. Video Technol. 26 (1) (2016) 130–138.
[43] J. Kim, A. Nguyen, S. Lee, Deep CNN-based blind image quality predictor, *IEEE Trans. Neur. Netw.* 30 (1) (2019) 11–24.
[44] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: Proc. IEEE CVPR, Las Vegas, Nevada, USA, 2016, pp. 1646–1654.
[45] J. Kim, S. Lee, Fully deep blind image quality predictor, IEEE J. Sel. Top. Sign. Proces. 11 (1) (2017) 206–220.
[46] K. Kim, W. Ro, Fast CU depth decision for HEVC using neural networks, IEEE Trans. Circuits Syst. Video Technol. (2018) in press, doi:10.1109/TCSVT.2018.2839113.
[47] C.Y.S. Kin, and B. Coker, Video compression using recurrent convolutional neural networks, Stanford University, (2017) [online] http://cs231n.stanford.edu/reports/2017/pdfs/423.pdf.
[48] C.C.J. Kuo, Y. Chen, On data-driven Saak transform, J. Vis. Commun. Image R. 50 (2018) 237–246.
[49] C.C.J. Kuo, M. Zhang, S. Li, J. Duan, Y. Chen, Interpretable convolutional neural networks via feedforward design, J. Vis. Commun. Image R. 60 (2019) 346–359.
[50] T. Laude, J. Ostermann, Deep learning-based intra prediction mode decision for HEVC, in: Proc. PCS, Nuremberg, 2016, pp. 1–5.
[51] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (May 2015) 436–444.
[52] J. Li, B. Li, J. Xu, R. Xiong, W. Gao, Fully connected network based intra prediction for image coding, IEEE Trans. Image Process. 27 (7) (2018) 3236–3247.
[53] J. Li, B. Li, J. Xu, R. Xiong, Efficient multiple line based intra prediction for HEVC, IEEE Trans. Circuits Syst. Video Technol. 28 (4) (2018) 947–957.
[54] N. Li, Y. Zhang, L. Zhu, W. Luo, S. Kwong, Reinforcement learning based coding unit early termination algorithm for high efficiency video coding, J. Visual Commun. Image R. 60 (2019) 276–286.
[55] T. Li, M. Xu, X. Deng, A deep convolutional neural network approach for complexity reduction on intra-mode HEVC, in: Proc. IEEE ICME, 2017, pp. 1255–1260.
[56] T. Li, M. Xu, R. Yang, X. Tao, A DenseNet based approach for multi-frame in-loop filter in HEVC, 2019 Data Compression Conference (DCC),, 2019 Mar. 26-29.
[57] Y. Li, D. Liu, H. Li, L. Li, F. Wu, H. Zhang, H. Yang, Convolutional neural network-based block up-sampling for intra frame coding, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2018) 2316–2330.
[58] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, M. Manohara, Toward a Practical Perceptual Video Quality Metric, Netflix TechBlog, 2016.
[59] K. Lim, J. Lee, S. Kim, S. Lee, Fast PU skip and split termination algorithm for HEVC intra prediction, IEEE Trans. Circuits Syst. Video Technol. 25 (8) (2015) 1335–1346.
[60] C. Lin, Z. Chen, N. Liao, Full-reference quality assessment for stereoscopic images based on binocular vision model, in: Proc. IEEE VCIP, Chengdu, China, 2016.
[61] R. Lin, Y. Zhang, H. Wang, X. Wang Q. Dai, Deep convolutional neural network for decompressed video enhancement, in: Proc. DCC, Snowbird, UT, 2016 617-617.
[62] J.Y. Lin, R. Song, C.-.H. Wu, T. Liu, H. Wang, C.-.C. Jay Kuo, MCL-V: a streaming video quality assessment database, J. Visual Commun. Image R. 30 (2015) 1–9.
[63] A. Liu, W. Lin, M. Narwaria, Image quality assessment based on gradient similarity, IEEE Trans. Image Process. 21 (4) (2012) 1500–1512.
[64] D. Liu, H. Ma, Z. Xiong, F. Wu, CNN-based DCT-like transform for image compression, in: Proc. MMM, 2018, pp. 61–72.
[65] J. Liu, S. Xia, W. Yang, M. Li, D. Liu, One-for-All: grouped variation network-based fractional interpolation in video coding, IEEE Trans. Image Process. 28 (5) (2019) 2140–2151.
[66] X. Liu, Y. Li, D. Liu, P. Wang, L.T. Yang, An adaptive CU size decision algorithm for HEVC intra prediction based on complexity classification using machine learning, IEEE Trans. Circuits Syst. Video Technol. 29 (1) (2019) 144–155.
[67] Y. Liu, J. Liu, A. Argyriou, S. Ci, Binocular-combination-oriented perceptual rate-distortion optimization for stereoscopic video coding, IEEE Trans. Circuits Syst. Video Technol. 28 (8) (2018) 1949–1959.
[68] Z. Liu, L. Shen, Z. Zhang, An efficient intermode decision algorithm based on motion homogeneity for H.264/AVC, IEEE Trans. Circuits Syst. Video Technol. 19 (1) (2009) 128–132.
[69] Z. Liu, X. Yu, S. Chen, D. Wang, CNN oriented fast HEVC intra CU mode decision, in: IEEE ISCAS, Montreal, QC, 2016, pp. 2270–2273.
[70] Z. Liu, X. Yu, Y. Gao, S. Chen, X. Ji, D. Wang, CU partition mode decision for HEVC hardwired intra encoder using convolution neural network, IEEE Trans. Image Process. 25 (11) (2016) 5088–5103.
[71] A.K. Moorthy, A.C. Bovik, Blind image quality assessment: from natural scene statistics to perceptual quality, IEEE Trans. Image Process. 20 (12) (2011) 3350–3364.
[72] A.J. Moreno, E.M. Enriquez, F.D. de Maria, Bayesian adaptive algorithm for fast coding unit decision in the high efficiency video coding (HEVC) standard, Signal Process. Image Commun. 56 (2017) 1–11.
[73] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowledge Data Eng. 22 (10) (2010) 1345–1359.
[74] W.-S. Park, M. Kim, CNN-based in-loop filtering for coding efficiency improvement, in: Proc. IEEE IVMSP, 2016, pp. 1–5.
[75] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Trans. Broadcast. 50 (3) (2004) 312–322.
[76] S. Puri, S. Lasserre, P. Le Callet, Improved coefficient coding for adaptive transforms in HEVC, in: *Proc.* PCS, Nuremberg, 2016, pp. 1–5.

[77] S. Puri, S. Lasserre, P. Le Callet, CNN-based transform index prediction in multiple transforms framework to assist entropy coding, in: Proc. EUSIPCO, Kos, Greece, 2017.
[78] S. Puri, S. Lasserre, P. Le Callet, Annealed learning based block transforms for HEVC video coding, in: Proc. IEEE ICASSP, Shanghai, 2016, pp. 1135–1139.
[79] R-REC-BT.1788. Methodology for the subjective assessment of video quality in multimedia applications (2007).
[80] R-REC-BT.2021-1.subjective methods for the assessment of stereoscopic 3DTV systems (2015).
[81] J. Robinson, V. Kecman, Combining support vector machine learning with the discrete cosine transform in image compression, IEEE Trans. Neural Net. 14 (4) (2003) 950–958.
[82] K. Rouis, M.C. Larabi, J.B. Tahar, Perceptually adaptive lagrangian multiplier for HEVC guided rate-distortion optimization, IEEE Access 6 (2018) 33589–33603.
[83] S. Ryua, J.W. Kang, Machine learning-based fast angular prediction mode decision technique in video coding, IEEE Trans. Image Process. (2018) in press, doi:10.1109/TIP.2018.2857404.
[84] K. Seshadrinathan, A.C. Bovik, Motion tuned spatio-temporal quality assessment of natural videos, IEEE Trans. Image Process. 19 (2) (2010) 335–350.
[85] O.G. Sezer, R. Cohen, A. Vetro, Robust learning of 2-D separable transforms for next-generation video coding, in: Proc. DCC, Snowbird, UT, 2011, pp. 63–72.
[86] F. Shao, K. Li, W. Lin, G. Jiang, M. Yu, Q. Dai, Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties, IEEE Trans. Image Process. 24 (10) (2015) 2971–2983.
[87] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.
[88] X. Shen, L. Yu, CU splitting early termination based on weighted SVM, EURASIP J. Image Video Process. 2013 (4) (2013) 1–11.
[89] L. Shen, Z. Zhang, X. Zhang, P. An, Z. Liu, Fast TU size decision algorithm for HEVC encoders using Bayesian theorem detection, Signal Process. Image Commun. 32 (2015) 121–128.
[90] L. Shen, Z. Liu, Z. Zhang, G. Wang, An adaptive and fast multi-frame selection algorithm for h.264 video coding, IEEE Signal Process. Lett. 14 (11) (2007) 836–839.
[91] L. Shen, Z. Zhang, Z. Liu, Adaptive inter-mode decision for HEVC jointly utilizing inter-level and spatio-temporal correlations, IEEE Trans. Circuits Syst. Video Technol. 24 (10) (2014) 1709–1722.
[92] Z. Sinno, A.C. Bovik, Large-scale study of perceptual video quality, IEEE Trans. Image Process. 28 (2) (2019) 612–627.
[93] R. Song, D. Liu, H. Li, F. Wu, Neural network-based arithmetic coding of intra prediction modes in HEVC, in: Proc. IEEE VCIP, St. Petersburg, FL, USA, 2017, pp. 1–4.
[94] X. Song, X. Peng, J. Xu, G. Shi, F. Wu, Cloud-based distributed image coding, IEEE Trans. Circuits Syst. Video Technol. 25 (12) (2015) 1926–1940.
[95] G. Sullivan, J. Ohm, W. Han, T. Wiegand, Overview of the high efficiency video coding (HEVC) standard, IEEE Trans. Circuits Syst. Video Technol. 22 (12) (2012) 1649–1668.
[96] Y.H. Sung, J.C. Wang, Fast mode decision for H.264/AVC based on rate distortion clustering, IEEE Trans. Multimedia 14 (3) (2012) 693–702.
[97] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, in: *Proc. ICANN*, LNCS 11141, 2018, pp. 270–279.
[98] Y. Tang, W. Gong, Q. Yi, W. Li, Combining sparse coding with structured output regression machine for single image super-resolution, *Inform. Sci.* 430-431 (2018) 577–598.
[99] X. Tang, H. Xiong, X. Jiang, Multiscale online dictionary learning for quality scalable video coding, in: Proc. DCC, Snowbird, UT, 2014 428-428.
[100] R. Timofte, V. De Smet, L. Van Gool, A+: adjusted anchored neighborhood regression for fast super-resolution, in: Proc. ACCV, 9006, Singapore, SG, 2014, pp. 111–126.
[101] H.R. Tohidypour, H. Bashashati, M.T. Pourazad, P. Nasiopoulos, Online-learning based mode prediction method for quality scalable extension of the high efficiency video coding (HEVC) standard, IEEE Trans. Circuits Syst. Video Technol. 27 (10) (2017) 2204–2215.
[102] M.T. Vega, D.C. Mocanu, J. Famaey, S. Stavrou, A. Liotta, Deep learning for quality assessment in live video streaming, IEEE Signal Process. Lett. 24 (6) (2017) 736–740.
[103] – Vision, Applications and requirements for high efficiency video coding (HEVC), Doc. N11872, *MPEG*, 2011.
[104] H. Wang, I. Katsavounidis, J. Zhou, J. Park, S. Lei, X. Zhou, M.O. Pun, X. Jin, R. Wang, X. Wang, Y. Zhang, J. Huang, S. Kwong, C.-C.J. Kuo, VideoSet: a large-scale compressed video quality dataset based on JND measurement, J. Visual Commun. Image R. 46 (2017) 292–302.
[105] T. Wang, M. Chen, H. Chao, A novel deep learning-based method of improving coding efficiency from the decoder-end for HEVC, in: Proc. DCC, Snowbird, UT, 2017, pp. 410–419.
[106] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[107] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multi-scale structural similarity for image quality assessment, 37th IEEE Asilomar Conf. Signals, Syst. Computers, Nov. 2003.
[108] T. Wiegand, G. Sullivan, G. Bjøntegaard, A. Luthra, Overview of the H.264/AVC video coding standard, IEEE Trans. Circuits Syst. Video Technol. 13 (7) (2003) 560–576.
[109] H. Xiong, Z. Pan, X. Ye, C.W. Chen, Sparse spatio-temporal representation with adaptive regularized dictionary learning for low bit rate video coding, IEEE Trans. Circuits Syst. Video Technol. 23 (4) (2013) 710–728.
[110] J. Xiong, H. Li, F. Meng, S. Zhu, Q. Wu, B. Zeng, MRF-based fast HEVC inter CU decision with the variance of absolute differences, IEEE Trans. Multimedia 16 (8) (2014) 2141–2153.
[111] J. Xiong, H. Li, Q. Wu, F. Meng, A fast HEVC inter CU selection method based on pyramid motion divergence, IEEE Trans. Multimedia 16 (2) (2014) 559–564.
[112] J. Xu, M. Xu, Y. Wei, Z. Wang, Z. Guan, H. Fast, 264 to HEVC Transcoding: a deep learning method, IEEE Trans. Multimedia (2018) in press, doi:10.1109/TMM.2018.2885921.
[113] L. Xu, W. Lin, L. Ma, Y. Zhang, Y. Fang, K.N. Ngan, S. Li, Y. Yan, Free-energy principle inspired video quality metric and its use in video coding, IEEE Trans. Multimedia 18 (4) (2016) 590–602.
[114] M. Xu, Y. Liu, H. Hu, F. He, Reducing complexity of HEVC: a deep learning approach, IEEE Trans. Image Process. 27 (10) (2018) 5044–5059.
[115] W. Xue, X. Mou, L. Zhang, A. Bovik, X. Feng, Blind image quality assessment using joint statistics of gradient magnitude and laplacian features, IEEE Trans. Image Process. 23 (11) (2014) 4850–4862.
[116] N. Yan, D. Liu, H. Li, B. Li, L. Li, F. Wu, Convolutional neural network based fractional pixel motion compensation, IEEE Trans. Circuits Syst. Video Technol. 29 (3) (2019) 840–853.
[117] N. Yan, D. Liu, H. Li, F. Wu, A convolutional neural network approach for half-pel interpolation in video coding, in: Proc. IEEE ISCAS, 2017, pp. 1–4.
[118] Q. Yan, D. Gong, Y. Zhang, Two-Stream convolutional networks for blind image quality assessment, IEEE Trans. Image Process. 28 (5) (2019) 2200–2211.
[119] J. Yang, Y. Zhao, Y. Zhu, H. Xu, W. Lu, Q. Meng, Blind assessment for stereo images considering binocular characteristics and deep perception map based on deep belief network, Inform. Sci. 474 (2019) 1–17.
[120] P. Ye, J. Kumar, L. Kang, D. Doermann, Unsupervised feature learning framework for no-reference image quality assessment, in: Proc. CVPR, 2014, pp. 1098–1105.
[121] youtube.com Traffic Statistics, Alexa Internet, Amazon.com (2018).
[122] L. Yu, L. Shen, H. Yang, L. Wang, P. An, Quality enhancement network via multi-reconstruction recursive residual learning for video coding", IEEE Signal Process. Lett. 26 (4) (2019) 557–561.
[123] H. Zhang, Z. Ma, Fast intra mode decision for high efficiency video coding (HEVC), IEEE Trans. Circuits Syst. Video Technol. 24 (4) (2014) 660–668.

[124] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: a feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378–2386.
[125] T. Zhang, M.T. Sun, D. Zhao, W. Gao, Fast intra mode and CU size decision for HEVC, IEEE Trans. Circuits Syst. Video Technol. 27 (8) (2017) 1714–1726.
[126] X. Zhang, W. Lin, S. Wang, J. Liu, S. Ma, W. Gao, Fine-grained quality assessment for compressed images, IEEE Trans. Image Process. 28 (3) (2019) 1163–1175.
[127] Y. Zhang, X. Gao, L. He, W. Lu, R. He, Objective video quality assessment combining transfer learning with CNN, *IEEE Trans. Neural Netw. Learn. Syst.* (2019) 1–15.
[128] Y. Zhang, N. Li, S. Kwong, G. Jiang, H. Zeng, Statistical early termination and early skip models for fast mode decision in HEVC INTRA coding, *ACM Trans. Multimedia Comput. Commun., Appl.* 15 (3) (2019) 23 Article 70pages.
[129] Y. Zhang, S. Kwong, G. Jiang, X. Wang, M. Yu, Statistical early termination model for fast mode decision and reference frame selection in multiview video coding, IEEE Trans. Broadcast. 58 (1) (2012) 10–23.
[130] Y. Zhang, S. Kwong, X. Wang, H. Yuan, Z. Pan, L. Xu, Machine learning based coding unit depth decisions for flexible complexity allocation in high efficiency video coding, IEEE Trans. Image Process. 24 (7) (2015) 2225–2238.
[131] Y. Zhang, S. Kwong, G. Zhang, Z. Pan, Y. Hui, G. Jiang, Low complexity HEVC intra coding for high quality mobile video communication, IEEE Trans. Indust. Inform. 11 (6) (2015) 1492–1504.
[132] Y. Zhang, Z. Pan, N. Li, X. Wang, G. Jiang, S. Kwong, Effective data driven coding unit size decision approaches for HEVC intra coding, IEEE Trans. Circuits Syst. Video Technol. 28 (11) (2018) 3208–3222.
[133] Y. Zhang, T. Shen, X. Ji, Y. Zhang, R. Xiong, Q. Dai, Residual highway convolutional neural networks for in-loop filtering in HEVC, IEEE Trans. Image Process. 27 (8) (2018) 3827–3841.
[134] Y. Zhang, L. Sun, C. Yan, X. Ji, Q. Dai, Adaptive residual networks for high quality image restoration, IEEE Trans. Image Process. 27 (7) (2018) 3150–3163.
[135] Y. Zhang, H. Zhang, M. Yu, S. Kwong, Y.S. Ho, Sparse representation based video quality assessment for synthesized 3D videos, IEEE Trans. Image Process. (2019) (in press), doi:10.1109/TIP.2019.2929433.
[136] L. Zhao, S. Wang, X. Zhang, S. Wang, S. Ma, W. Gao, Enhanced CTU-level inter prediction with deep frame rate up-conversion for high efficiency video coding, in: Proc. IEEE ICIP, 2018, pp. 206–210.
[137] Z. Zhao, S. Wang, S. Wang, X. Zhang, S. Ma, J. Yang, Enhanced bi-prediction with convolutional neural network for high efficiency video coding, IEEE Trans. Circuits Syst. Video Technol. (2018) in press, doi:10.1109/TCSVT.2018.2876399.
[138] L. Zhu, S. Kwong, Y. Zhang, S. Wang, X. Wang, Generative adversarial network based intra prediction for video coding, IEEE Trans. Multimedia (2019) to be published, doi:10.1109/TMM.2019.2924591.
[139] L. Zhu, Y. Zhang, Z. Pan, R. Wang, S. Kwong, Z. Peng, Binary and multi-class learning based low complexity optimization for HEVC encoding, IEEE Trans. Broadcast. 63 (3) (2017) 547–561.
[140] L. Zhu, Y. Zhang, S. Wang, H. Yuan, S. Kwong, H.H.S. Ip, Convolutional neural network based synthesized view quality enhancement for 3D video coding, IEEE Trans. Image Process. 27 (11) (2018) 5365–5377.