High Dynamic Range Images Coding: Embedded and Multiple Description

Alberto Boschetti, Nicola Adami, Pierangelo Migliorati and Riccardo Leonardi
University of Brescia, Brescia, Italy
E-mail: {firstname.lastname}@ing.unibs.it

***Abstract*—The aim of this work is to highlight and discuss a new paradigm for representing high-dynamic range (HDR) images that can be used for both its coding and describing its multimedia content. In particular, the new approach defines a new representation domain that, conversely from the classical compressed one, enables to identify and exploit content metadata. Information related to content are used here to control both the encoding and the decoding process and are directly embedded in the compressed data stream. Firstly, thanks to the proposed solution, the content description can be quickly accessed without the need of fully decoding the compressed stream. This fact ensures a significant improvement in the performance of search and retrieval systems, such as for semantic browsing of image databases. Then, other potential benefits can be envisaged especially in the field of management and distribution of multimedia content, because the direct embedding of content metadata preserves the consistency between content stream and content description without the need of other external frameworks, such as MPEG-21.**

The paradigm proposed here may also be shifted to Multiple description coding, where different representations of the HDR image can be generated accordingly to its content. The advantages provided by the new proposed method are visible at different levels, i.e. when evaluating the redundancy reduction. Moreover, the descriptors extracted from the compressed data-stream could be actively used in complex applications, such as fast retrieval of similar images from huge databases.

I. Introduction

Content analysis and coding are two key issues in the creation of new interactive multimedia applications able to fulfill user requests and expectations (for both professional and non-professional end-users). The possibility to efficiently represent an information source with high quality whilst using a relatively small bandwidth enables, for example, to display video with good quality on mobile devices connected to networks with relatively limited capabilities (e.g. UMTS). Furthermore, the possibility to efficiently describe the multimedia content enables to speed up the searching processes, to increase the quantity and to improve the quality of the information that users may require. In fact, one of the critical issues in current multimedia applications is the actual possibility to efficiently access the desired information. So far, solutions for the problem of data compression and multimedia content indexing have been independently studied. This is probably due to historical reasons related to the two scientific communities and because of the apparently different goals of the two applications.

The state-of-the-art in image and video coding can include references to several standards such as, MPEG-4/AVC and JPEG2K, whereas for the description of multimedia content MPEG-7 is the reference standard. In the classical model adopted by compression systems, the data source is first compressed and then the data-stream is either stored or transmitted. At the decoder side, the end user can access the original source representation or a "lower-quality" version of it, which requires fewer resources in terms of bandwidth or storage space. Anticipating what will be requested in the near future by advanced multimedia applications, R.W. Picard has already introduced a fundamental modification to such model. Her model envisages the possibility to access and edit the multimedia content directly in the compressed domain, without the decoding of the compressed data. This approach is defined as "midstream content access" [15]. According to this, in addition to the three classical parameters that have to be optimized in the coding systems, i.e. reduction of required bandwidth, distortion and computational load, R.W. Picard introduces "the fourth criterion": the minimization of the required work to access and/or modify a particular content included into a compressed data-stream. The target defined by the Piccard's criterion has been indirectly pursued by the proposal of the MPEG-4 standard, which introduces the concept of object coding. Such approach assumes to identify and separate the different components within a scene (such as the background and the foreground objects) and then to independently code each component. Alternatively from the classical coding techniques, the different components (objects) are described by means of mathematical models that evolve over time in order to represent a particular scene. The main aim of the coding systems based on the objects is the reduction of the rate required to represent the content, even if this led to a considerable increase of the encoder complexity (more than exponential). Object coding indirectly satisfies the "fourth criterion" by providing for the first time the possibility to directly access to a specific audio/visual content within the coded representation (and in case, to modify it), in a similar way as in computer graphics, where a scene is generated using models derived from reality. However, even after several years by the definition of the MPEG-4 standard, at the moment the use of coding systems based on objects has not been proposed yet in real applications. A possible reason for the failure of the object coding (as defined in MPEG-4) is the way in which the "objects" are defined and used. If on the one

side models can be effectively used to reproduce a scene in a realistic way, on the other hand it is impractical at the moment to extract accurate models from each real scene, given the existing variety of possible scenes. This consideration suggests that even if the object coding is potentially a very efficient coding approach, both from a source coding point of view and for the content description at high level of abstraction, it is not a feasible approach for the representation of images and/or video of real scenes. For this reason, the only descriptions available for the scene representation are those at a low semantic level (e.g. color, texture, aesthetic elements, etc.), since they can be directly extracted from the image. At the moment, a quick access to such descriptions can be done using the tools provided by the MPEG-7 and MPEG-21 standards. In particular, MPEG-7 offers descriptors and schemes for multimedia content description, whereas MPEG-21 defines particular structures for the interoperable distribution of the contents and their related descriptions.

Usually the content is generated, encoded and stored/distributed; successively, depending on the application, one or more scene descriptors are extracted and linked with the content itself. It is quite evident that this approach has two main disadvantages in terms of efficiency: an additional computational load is required to extract the descriptors, and additional data (additional bandwidth or storage space) are needed to represent the descriptors that partially duplicate the information provided by the compressed data itself. Recently some techniques to jointly perform source coding and content indexing at a low-middle semantic level have been proposed [16]- [1], without the need of complex description models (such as MPEG4 objects).

The main idea of the contribution is to investigate and propose new techniques for a joint source coding and multimedia content description for HDR images. Conversely from the "normal images" that contain only the information required to represent the content on a particular visualization device (monitor, TV, etc.), the high dynamic range images represent the scene captured by an acquisition device. This means that a high dynamic range image contains the information related to the absolute luminance and colors, corresponding to the properties of the psychic phenomena. In this way the dynamic of the captured signal is ideally equal to the real one. Basically, two main motivations could be identified behind the study of such high dynamic signals. The first one is that after the high definition (HD), the high dynamic (already used in particular professional applications) is considered to be the next step to increase the realism and consequently the perceived quality in multimedia applications. The second one, even more important for research purposes, is the possibility to use more detailed information, which is closer to the properties of the psychic phenomena. In particular, the latter aspect should enable a more accurate analysis and the extraction of higher quality descriptors.

To better clarify, next section synthesize the state of the art concerning the three research topics, covered by the here proposed coding paradigm namely: HDR Image coding, Content Description and Joint Image Coding and Content Description. Possible implementations of this new paradigm are then provided in Section III and in Section V.

II. Related works

This contribution deals with three fields: high dynamic range images, content descriptors and scalable joint coding of images and descriptors.

A. High dynamic range images coding

Traditionally, digital images have been represented using 8 bit code for each color component (bpcc). This value has been chosen in order to suit images to the color gamut of the visualization devices (monitors, printers, etc.). For this reason, standard formats, such as TIFF and JPEG, are referred to as "related to the device". However, the light intensity in the real world is characterized by a high dynamic range and the human visual system is able to perceive variations into a relative interval of five orders of magnitude. High dynamic range images aiming at emulating this ability by storing absolute information of luminance and color, and are referred to as "related to the scene". At first, they were used in the computer graphic field and in professional photography and currently also in cinematography. The field that deals with the representation and coding of this kind of images is the so called *High Dynamic Range Imaging*. Nowadays there are few devices able to acquire and display this type of images and the amount of related scientific literature is growing, even if still poor.

Up to now, research efforts have been focused on two main topics: source coding and the representation of high dynamic range signals on traditional devices. When a device able to display a high dynamic range source is not available, it this possible to transform the source itself into a low dynamic range one, using some Tone Mapping Operators (TMOs) which work applying a non-linear local or global processing [17]–[9]. These tools are usually computational expensive and it is consequently better to avoid their use every time an image needs to be displayed.

A solution might be coding the high dynamic range image into a compress stream containing a low dynamic version of the original image, and the signal that has to be added to the low dynamic image in order to regain the high dynamic one [22]. The efficiency of these methods is very low, because of the diversity of the signals involved. To solve these problems, a number of techniques aiming at reducing the prediction error have been developed; these methods basically use polynomial models to expand the values of the low dynamic range image [13].

Moreover, several techniques that code high dynamic range images using traditional methods like JPEG2K have been proposed [11]. In [4] a fast and scalable approach for tone mapping zones of interest in HDR videos is proposed . It combines the benefits of both local and global tone mapping operators, and it is designed for a video-surveillance purpose, in fact it is applied in the context of object detection and

tracking. Due to its adaptivity, it also enhances the visual quality of the image in all light conditions, which facilitated surveillance tasks for both human and automatic operators.

B. Content description

Content description is part of the multimedia analysis field and aims at defining elements able to characterize the content of a media. Using the nomenclature specified in the MPEG7 standard, a Descriptor (D) represents a particular characteristic of a content, such as the average color of an image, while a Descriptor Scheme (SD) concerns the relations between different descriptors and the related multimedia content. There are different types of Descriptors and Descriptor Schemes according to different multimedia content they can describe (images, audio, etc.) and to the semantic level of the information they convey. MPEG7 provides a rich set of D and SD but it does not specify methods to be used in the extraction and the generation of D and SD themselves. Speaking about low-level Descriptors, in the academic literature, there are lots of techniques dealing with their extraction and comparison. Several works have demonstrated the ability of low-level descriptors in defining higher semantic level operations [10] such as clustering [2] and retrieval [6]. Clustering allows the association of images which are similar in meaning, while retrieval is about finding contents into a database according to a query.

These operations associate descriptors to mathematics concepts, like vectors or stochastic processes. Besides the mathematics formats of the input data, it is necessary to decide the approach to the work, in other words if the optimization has to be done for each couple of elements, between each cluster or by using statistical models. Each approach has several advantages and disadvantages, and the choice has to be made according to the operation to do.

Recently, several low-level characteristics linked with high-level concepts have been studied; they would be useful for improving the quality of the available descriptors and for bridging the semantic gap between what the user wants and what he can obtain with a specific request.

In [14] authors identify six fundamental dimensions which allow the aesthetic characterization of an image. This is based on the human visual system, which is composed by "independent modules", each of them focused on a particular task. Starting from such a type of analysis, the following aesthetic primitives have been defined: color (dominant colors, the presence of complementary colors, dynamic), form (clarity of form, silhouette), spatial organization (clarity of organization, golden mean, visual scheme), dynamism, depth (perspective) and identification of human bodies (principal axes). In [5] authors propose a system for the retrieval of similar images, using descriptors related to visual elements and aesthetic criteria. Used characteristics are almost the same as those in [14] but are applied in the wavelet domain.

C. Joint scalable coding and content description

Nowadays several algorithms for image and video coding are available [20]. Each of them exploits redundancies (spatial, temporal and perceptual) to obtain a better compression ratio. In 1994, R.W. Picard proposed a new coding paradigm, which, differently from the classical method, works directly onto the coded bitstream for access and manipulation, without operating the data decompression. Consequently, other than the three classical criterions for quality evaluation of an encoder, which are "bit rate", "distortion" and "computational cost", a new one was added, called *the fourth criterion*. It represents the effort, which must be done to access a given content inside the coded stream. If an encoder supports this new concept, it could be used for applications that in an effective and efficient way could do operations of retrieval, browsing and manipulation of multimedia content. Three advantages are carried by this new modality of work directly onto the coded stream: first is the fast content access, second is the possibility of semantic coding of the multimedia data and third is the opportunity to create a new scalability dimension. Below there are contributions that describe some methods of encoding using descriptors (used for coding, retrieval, etc.). However seems that no methods proposed in literature implement a joint optimization of the four criteria.

[16] proposes a method based on techniques SBIC, CPAM and VQ. The Segmentation-Based Image Coding algorithm splits the image into different homogeneous regions, with no fixed size and each of them is allocated with a number of bits directly dependent from his property. After that, the image channels Y, Cb and Cr are split using the Colored Pattern Appearance Model. In this step, three types of information are extracted for each region: the stimulus strength (SS), the achromatic spatial pattern (ASP) and the chromatic spatial pattern (CSP). The last coding step is the vector quantization, which maps multidimensional vectors into an indexes series of the codebook's word (called codeword). The descriptor of the entire image is build merging the indexes set of SBIC/CPAM/VQ; in fact having these data is very simple to extract the description of each region in terms of shape and color. This technique appears to offer good performance with regard to retrieval of similar content in large databases; but it has not been evaluated in terms of compression ratio achieved. In [18] the image encoder is based on CVPIC technique (Color Visual Pattern Image Coding). The data available at the output of this encoder, for each block in which the image is split, describe the color (in CIEL*a*b* space) and the planarity or the irregularity (presence of edge, corner, gradient). Then, the image descriptor is represented by edge map and color map. Starting from the proposed results, this technique appears to offer good retrieval performance, but it is not possible to evaluate the performance of encoding.

[23] proposes a compression method for videos, where from each shot of the video are extracted the "key-objects". They are particular descriptors that can characterize the entire shot, and they are outlined in terms of color, texture, shape, motion

(from optical flow) and their life cycle. Similarly to the coding standard MPEG4, the metadata extracted for all key-objects are used for building the coded stream. Authors provide no information about real system performance in compression and retrieval work. In [19] a technique for collection of images coding is described. It allows retrieving of content information working directly inside the coded domain. Each image is decomposed in a group of objects associated to semantic indexes (like "tree", "house"). Then, the different areas are split in rectangular blocks and they are coded separately, using a JPEG-like coding method. Hence, the coded stream is made by indexes, by their spatial relations and by the true compressed content. In [1] the authors use Visual CodeBook as a content descriptor for image collection. Through a vector quantization process, applied to a collection of similar images, the VCB is obtained. The bitstream produced for one collection of images contains the descriptor (VCB) and, for each image, the array of indexes for VQ reconstruction and a reminder, used for decoding the image at a certain quality. This method offer good performance both for retrieval/classification of similar images, and for coding efficacy, which is comparable to modern image compression methods, likes JPEG2K.

In [3] authors propose an effective implementation of a scalable encoder for images, with the active use of embedded descriptors during coding. In the proposed example the descriptors are the faces in the picture. According to the proposed method, at first, images areas containing faces are detected and encoded using a scalable method, where the base layer is represented by the corresponding eigenface, and the enhancement layer is formed by the prediction error. The remaining areas are then encoded by using a traditional approach. Simulations show that achievable compression performances are comparable with those provided by conventional, making the proposed approach convenient for source coding and content description.

III. HDRI Multiple Description

In general, a Tone Mapping Operator applied to an HDR image provides a low dynamic range image that can be seen as a particular description of the original content whereas different TMOs provide different descriptions of the same visual content. Considering the problem of reconstructing an HDR image from the knowledge of its different low dynamic range versions can be seen as a Multiple Description issue. Hereafter a scheme able to reconstruct an HDR image starting from three globally tone mapped version is proposed.

A. TM Operator 1

The first tone map operator is very easy to implement, but it makes a fast and reversible tone map of the HDR image. Saying $H(x, y, c)$ the floating point value of the pixel in position (x, y) of the HDR image H in the color component c, the tone mapped value for the same pixel of the same component in the LDR image L is:

$$L(x, y, c) = \frac{H(x, y, c)}{H(x, y, c) + 1} \tag{1}$$

This function, as a tone map operator does, remaps the values from the HDR range $[0, \infty)$ to a limited one $[0, 1)$ by expanding the low–value components and compressing the high–value ones. The expansion/compression of this tone map operator is hence related to the inverse of the HDR value itself.

This operator is perfectly invertible, and the HDR reconstruction $\overline{H}$ is given by:

$$\overline{H(x, y, c)} = \frac{L(x, y, c)}{1 - L(x, y, c)} \tag{2}$$

1) Practical troubles: Practical problems are made by the quantization introduced in the LDR image. In fact, the LDR image is typically stored as a 8–bit image, so its values must be quantized. In theory, according to the Equation 1, $L(x, y, c)$ is never one for a finite HDR input image; however, during the reconstruction quantization effects can shift that value to one.

Even though in such case the reconstructions would not be feasible, we remove a small finite quantity to the $L(x, y, c)$ values when it is one, making always possible the inversion.

B. TM Operator 2

The second operator is explained in [7]. It is a global operator which applies a histogram–like equalization to the HDR image for tone mapping. It is a fast algorithm, which uses a statistical model that approximates the mean square error (MSE) distortion resulting from the combined processes of tone–mapping and compression. Is also provides LDR images with a good visual quality.

As explained in [7], the steps of the tone mapping operation are:

1) HDR image is passed through a logarithm operator.
2) The histogram (base 10) of the luminance is computed, with bin step of $\delta = 0.1$. Here, each bin ($K = 1 \dots N$) has a proper counter (p_K) and a codeword (l_K).
3) From the histogram, a set of remapping slopes (one for each bin) is computed as $s_K = \frac{v_{max} \cdot p_K^{1/3}}{\delta \cdot \sum_{K=1}^{N} p_K^{1/3}}$ This equation is the close form solution of the related optimization problem.
4) The tone map equation, for a generic pixel of the HDR image is:

$$L(x, y, c) = (\log_{10} H(x, y, c) - l_K) \cdot s_K + v_K \tag{3}$$

where l_K, s_K and v_K are respectively the codeword associated to the logarithm of the pixel value, the slopes of that bin, and the mapped LDR value of the bin.

5) The inversion equation is:

$$\overline{\log_{10} H(x, y, c)} = \begin{cases} \frac{L(x,y,c) - v_K}{s_K} + l_K & \text{for} \quad s_K > 0 \\ \sum_l l \cdot p_L(l) & \text{for} \quad s_K > 0 \end{cases} \tag{4}$$

This tone map operator remaps the HDR values to the range $[0, v_{max}]$. The inversion formula is feasible even if quantization of the LDR image is applied. The main idea used in this algorithm is taken from the histogram equalization theory, i.e. the slope of the remapping function is higher where the bins contain more pixel in.

As stated in [7] the reconstruction of the HDR content, starting from the LDR one, is visually better than other global tone mapping operators.

C. TM Operator 3

The third tone map operator is based on the logarithm of the values. It applies a companding operator (like the ones used in the telephone coding, such as $\mu-$law or $A-$law). The key idea for this tone map operator is to enhance the contrast for low values of the HDR image, and decrease it for high values, following a log–scale.

By using this operator, the LDR version of the generic HDR pixel $H(x, y, c)$ is:

$$L(x, y, c) = (\log H(x, y, c) + \Delta) \cdot \alpha \tag{5}$$

where Δ and α are two parameters (offset and scale factor) such that the output of the Equation 5 is never outside the range $[0, 1]$.

The inversion formula is:

$$\overline{H(x, y, c)} = \exp\big(\frac{(L(x, y, c) - \Delta)}{\alpha}\big) \tag{6}$$

1) Practical troubles: When the HDR image has zero values a small offset is added, because it is not possible to compute a logarithm of a negative number. There are no problems with quantization of the LDR image as happens in TM1.

IV. Architecture

The system can be split into two parts: the encoder side and the decoder one. The encoder does the work of multiple description of the original HDR image, producing many LDR images with embedded metadata; the decoder uses one or more LDR image and their metadata to reconstruct a high fidelity HDR image.

A. Encoder

A high–level schema of the encoder is shown in Figure 1.

The HDR image is firstly decoded, then the three tone map operators explained in the previous section are applied. Three LDR images are obtained at this point, and each of them has a proper remapping function embedded in the metadata (e.g. the TMO 2 has the scale vector).

The next step is the joint choice of the best inversion for the HDR reconstruction. Here, a remapping function is created to be applied to the three LDR pictures, in order to build the best approximation of the original HDR image. The L^2 norm has been used, i.e. the value v of the LDR_X is chosen for reconstruct the HDR image if its tone mapped inverse $TMO_X^{-1}(v)$ is the best approximation of the HDR value HDR.

For each LDR image and for each pixel value of them is inserted a measure of distance about its "goodness" in reconstruction, or, alternatively, its "goodness" ranking in reconstruction.

Here is an intuitive example: if the value of a pixel in the HDR image is 2.34 and the reconstructed values starting from the LDR images are respectively 2.35, 2.10 and 2.40, the best approximation in this example is given by the inversion of the tone map of first picture, so it will contain the ranking position 1 for the related LDR pixel value. Then, the second image will contain the ranking position 3, and the third 2.

For the three LDR images, metadata use no more than $2-bit$ for each LDR value, so exactly 64 bytes. Furthermore, it is possible to see that this data can be compressed into intervals (e.x. values from 0 to 10 have ranking 1), so 64 bytes is an upper bound.

B. Decoder

The decoder is shown in Figure 2.

It is possible to discern three types of configuration at the decoder side: only one LDR image is available; all the LDR images are available and only 2 LDR (i.e. not all) images are available.

1) One LDR image available at the decoder: In the first case, the remapping function is applied to the LDR image, in order to retrieve the HDR one. Metadata are used for correctly set the parameters of the reconstruction (for example, is the third tone map is used, offset and amplification are necessary during the reconstruction).

This reconstruction produces the best possible approximation of the HDR image, by using all the information known at the decoder side (that is only one third of the source information).

2) All the LDR images are available at the decoder side: In this case all the information of the encoder is carried to the decoder, and it is possible the production of the best approximation of the original HDR image.

All the tone map operators of the three LDR images are inverted, and, by using the metadata inserted in them, it is possible to build the HDR image using always the first ranking for each pixel, i.e.:

$$\overline{H(x, y, c)} = 1^{st} rank\big(TMO^{-1} LDR_x(x, y, c)\big), x = 1, 2, 3 \tag{7}$$

In case of multiple first ranks, an average between the top tanked values is used.

3) A partial number of LDRs images are available at the decoder side: In this case some information has been lost in the communication channel between encoder and decoder. The reconstruction of the HDR image is then made using the best ranking among the available pixels of the LDR images. The reconstruction is hence:

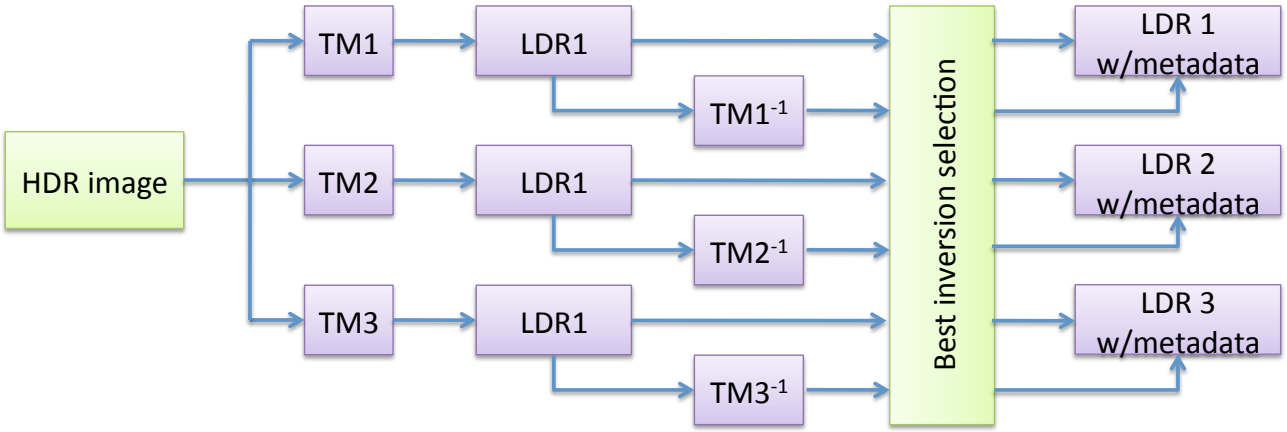

Fig. 1. High–level schema of the encoder for multiple description of HDR images

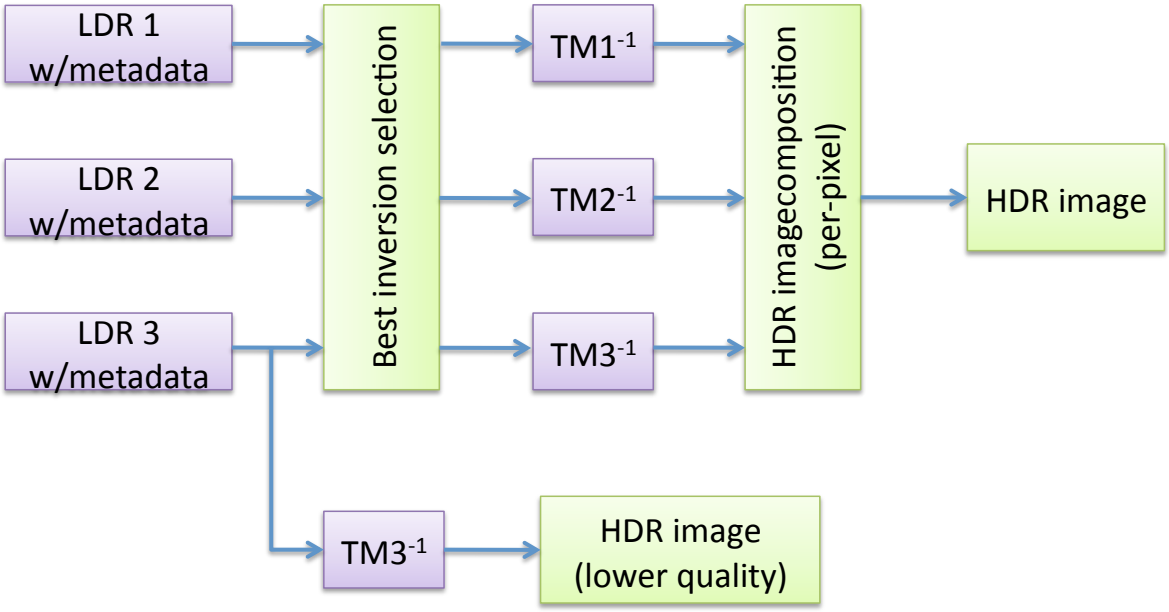

Fig. 2. High–level schema of the decoder

$$\overline{H(x,y,c)} = best\ rank\big(TMO^{-1}LDR_x(x,y,c)\big), x = 1,2,3 \quad (8)$$

In case of equal ranking among 2 or 3 values, their average is used.

C. Example and performances

In Figure 3 a possible result is shown on the *mpi_atrium_3*[1] HDR image. The input HDR image is initially tone mapped with three TMOs, then it is reconstructed using two different approaches (single LDR image inversion and joint inversion with all the candidates).

For measuring the performance of the proposed architecture, the HDR-VDP metric [8] has been used. It compares both visibility and quality in terms of probability, and it is mostly used for testing fidelity between HDR images. High values are associated with images with a bad quality, low values for hi–fidelity images.

As it can be seen in the example, the visual quality increases (in terms of a low percentage of differences between the original and the reconstructed one) if more than one LDR image is used.

In the Figure 4 numeric performance are shown, averaged

[1]Creative Commons 3.0 license, see *http://pfstools.sourceforge.net/hdr_gallery.html*

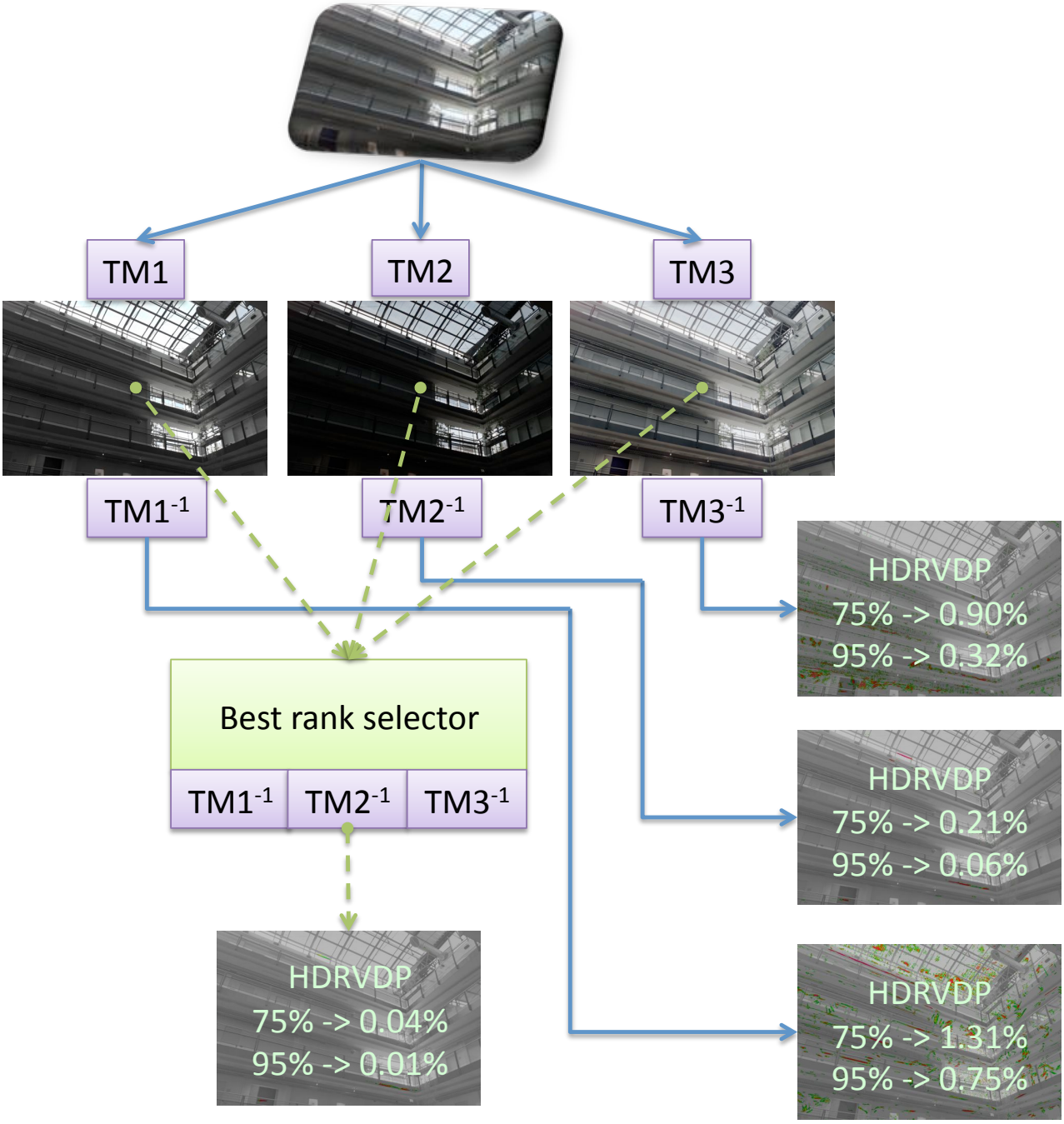

Fig. 3. An example of the proposed architecture, and the performance on the test image.

on a big set of HDR images, when the decoder reconstructs the HDR image with 1, 2 or 3 LDR images, in terms of VDP and MSE (the average is also applied when only 1 or 2 LDR images are used for the reconstruction).

Both HDR VDP and MSE decrease if more than one LDR is taken into consideration at the decoder side, following an exponential–like function.

V. HDR Image Coding With Face Description Embedding

In this section a system for encoding images with embedded face descriptors is proposed. According to this method, regions corresponding to faces in the HDR image are at first roughly encoded by adopting a PCA/eigenfaces based technique, while the residual image (reconstruction error) is compressed with

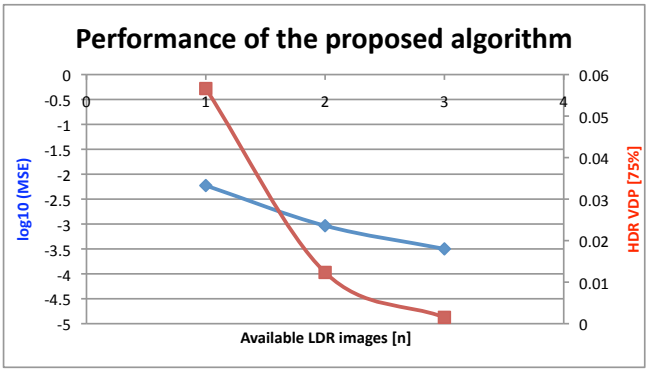

Fig. 4. Performance in terms of VDP and MSE, when the HDR image is reconstructed at the decoder side with 1, 2 and 3 LDR images.

JPEG2K like coding method [12]. By using this architecture, faces (the image descriptors) can be retrieved using only a dot product operation; hence, a fast access to the content description is guaranteed. The overall encoding process is accomplished in two main steps: the first stage concerns all the operations needed to train a learning machine based on eigenfaces (explained in Section V-A), while the real encoding process is realized in a second stage. The produced output is formed by: a series of coefficients and a residual image. The formers are used as descriptors for face recognition and to build a prediction signal of the face; the latter is needed to reconstruct, lossy or lossless, the original input image.

A. Eigenfaces technique

The eigenface technique is nowadays widely used for face recognition purposes. As stated in [21] it is a variation of the PCA (Principal Component Analysis) method, applied to the faces in the images. The goal of this process is to generate a reduced set of eigenvectors which can describe the principal components of the input faces (named eigenfaces).

The result of the Principal Component Analysis technique is given by the most relevant K eigenfaces ($[\varphi_1, \varphi_2, \ldots, \varphi_K]$) and the average face image.

B. Training

The training scheme of the proposed architecture is shown in Figure 5.

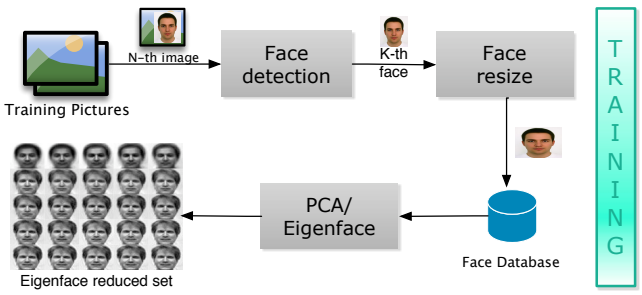

Fig. 5. Training chain

For each image in the training set, the faces contained in it are extracted with a face detection algorithm. The output of this initial step is a list of rectangles: every face in the image is described by its displacement and size within the picture. Successively, the faces are resizes to a standard size ($C \times R$ pixels) and converted to a grayscale image.

When all the faces of the training set have been processed and placed in the database, eigenfaces are generated using the process described in Section V-A. Then, the most relevant K eigenfaces are selected as projection base, together with the averaged face used for input normalization. Each element of this projection base is a vector of $C \times R$ floating point values.

It is very important to train the system by using different faces. It is needed to have at least a few training samples for each sex, race, age, position, illumination and rotation of the possible face. In such way the K eigenvectors of the base contain more information, and they can be used for providing a better reconstruct approximation of different type of faces. It has to be noted that different strategies can be applied to eigenfaces extraction, according to the considered application. In this work we limit to the case where a unique projection base, available both at the encoder and at the decoder, is used for retrieval and coding of all faces detected in images. Another strategy could be, for example, to generate a projection base for any given collection of similar images and include the eigenfaces in the compressed bitstream. In this case the projection base, which is essential to reconstruct the original signal, could also be used to cluster image collection containing similar faces.

C. Encoder

The encoder and decoder scheme are shown in Figure 6[2]. Initially faces are detected in the input image. Assuming that N faces are found; therefore, a list of N rectangles is provided by the face detection system. Each face is then resized (for matching the eigenface size) and converted to its grayscale representation. Successively, every face is projected, after the subtraction of the average–face, on the base of eigenfaces, obtaining a set of K representative coefficients. They are then quantized to the nearest integer and outputted together with the rectangle description. So far, for each face in the input image, a complete description is obtained; in fact, by using only these data, it is possible to create an approximation of the original face (in the exact position within the input image) and also to perform an automatic face recognition. The next block reconstructs all the faces by using the above coefficients and reversing the operations implemented during face projection. Clearly, these reconstructed signals are good approximations of the originals, and so they are used as predictors. Predicted faces are then subtracted from the original ones, generating a residual image. All this operations are performed only on the luminance channel leaving chrominance unaltered.

The last step concerns entropy coding: the 3–channels image is compressed by using JPEG2K while metadata are instead

[2]This anonymized HDR image was taken using the HDR camera Ims chips - HDRC

placed in a XML file, and then compressed using a lossless compression algorithm.

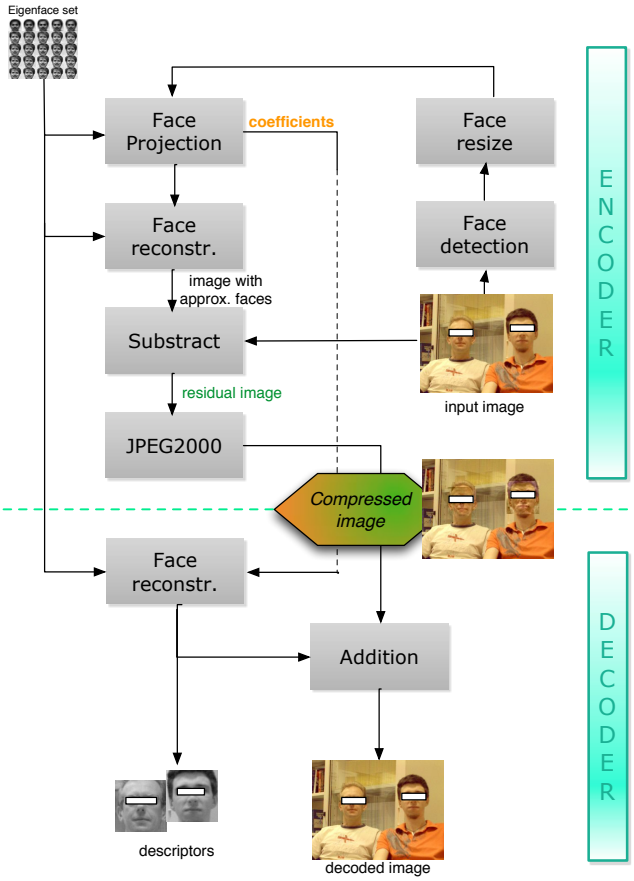

Fig. 6. Scheme of the encoder-decoder

D. Decoder

The decoder side requests less computational power compared to the encoder part. The first step of the decoding chain is composed by the face reconstruction. In order to obtain the face predictors in the image, it is needed to decompress the metadata file, project the faces coefficients on the eigenfaces base and reshape the faces to fit their original size. It has to be noted that face coefficients and locations contained in metadata can also be used for fast content access and retrieval. In order to reconstruct the input image, the residual image is then decompressed. Successively, the predicted faces and residuals are recombined.

VI. Conclusion

In this paper a new paradigm for representing high-dynamic range (HDR) images which can be used for both its coding and describing its multimedia content has been introduced. The proposed approach has tailored in two different contexts: the reconstruction of HDR images forms multiple LDRs and the HDR encoding with the embedding of face descriptors. In both cases the additional information carried by the HDR images are actively used, and many benefits in the encoder/decoder chain are gained.

References

[1] Nicola Adami, Alberto Boschetti, Riccardo Leonardi, and Pierangelo Migliorati. Embedded indexing in scalable video coding. *Multimedia Tools Appl.*, 48:105–121, May 2010.

[2] Sergio Benini, Aldo Bianchetti, Riccardo Leonardi, and Pierangelo Migliorati. Extraction of significant video summaries by dendrogram analysis. In *ICIP*, pages 133–136. IEEE, 2006.

[3] Alberto Boschetti, Nicola Adami, Riccardo Leonardi, and Masahiro Okuda. Image coding with face descriptors embedding. In *2011 IEEE International Conference on Image Processing (IEEE ICIP2011)*, Brussels, Belgium, September 2011.

[4] Alberto Boschetti, Nicola Adami, Riccardo Leonardi, and Masahiro Okuda. An optimal Video-Surveillance approach for HDR videos tone mapping. In *EUSIPCO 2011 (19th European Signal Processing Conference 2011) (EUSIPCO 2011)*, Barcelona, Spain, August 2011.

[5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Studying aesthetics in photographic images using a computational approach. In *In Proc. ECCV*, pages 7–13, 2006.

[6] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40:5:1–5:60, May 2008.

[7] Zicong Mai, Hassan Mansour, Rafal Mantiuk, Panos Nasiopoulos, Rabab Kreidieh Ward, and Wolfgang Heidrich. On-the-fly tone mapping for backward-compatible high dynamic range image/video compression. In *ISCAS*, pages 1831–1834. IEEE, 2010.

[8] Rafał Mantiuk, Scott Daly, Karol Myszkowski, and Hans-Peter Seidel. Predicting visible differences in high dynamic range images - model and its calibration. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Scott J. Daly, editors, *Human Vision and Electronic Imaging X, 17th Annual Symposium on Electronic Imaging (2005)*, volume 5666, pages 204–214, 2005.

[9] Stefano Marsi, Gaetano Impoco, Anna Ukovich, Sergio Carrato, and Giovanni Ramponi. Video enhancement and dynamic range control of hdr sequences for automotive applications. *EURASIP J. Adv. Sig. Proc.*, 2007, 2007.

[10] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:1615–1630, October 2005.

[11] Masahiro Okuda and Nicola Adami. Effective color space representation for wavelet based compression of hdr images. In *Proceedings of the 14th International Conference on Image Analysis and Processing*, ICIAP '07, pages 388–392, Washington, DC, USA, 2007. IEEE Computer Society.

[12] Masahiro Okuda and Nicola Adami. Effective color space representation for wavelet based compression of hdr images. In *ICIAP'07*, pages 388–392, 2007.

[13] Masahiro Okuda and Nicola Adami. Jpeg compatible raw image coding based on polynomial tone mapping model. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, E91-A:2928–2933, October 2008.

[14] Gabriele Peters. Aesthetic primitives of images for visualization. In *Proceedings of the 11th International Conference Information Visualization*, pages 316–325, Washington, DC, USA, 2007. IEEE Computer Society.

[15] R. W. Picard. Content access for image/video coding: "the fourth criterion. 1994.

[16] Guoping Qiu. Embedded colour image coding for content-based retrieval. *Journal of Visual Communication and Image Representation*, 15(4):507 – 521, 2004.

[17] Erik Reinhard and Kate Devlin. Dynamic range reduction inspired by photoreceptor physiology. *IEEE Transactions on Visualization and Computer Graphics*, 11:13–24, January 2005.

[18] Gerald Schaefer and Guoping Qiu. Midstream content access based on color visual pattern coding. In *Storage and Retrieval for Media Databases'00*, pages 284–292, 2000.

[19] M.D. Swanson, S. Hosur, and A.H. Tewfik. Image coding for content-based retrieval. 2727:4–15, March 1996.

[20] Alain Trémeau, Shoji Tominaga, and Konstantinos N. Plataniotis. Color in image and video processing: most recent trends and future research directions. *J. Image Video Process.*, 2008:7:1–7:26, January 2008.

[21] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3:71–86, January 1991.

[22] Greg Ward and Maryann Simmons. Jpeg-hdr: a backwards-compatible, high dynamic range extension to jpeg. In *ACM SIGGRAPH 2006 Courses*, SIGGRAPH '06, New York, NY, USA, 2006. ACM.

[23] HongJiang Zhang, J. Y. A. Wang, and Y. Altunbasak. Content-based video retrieval and compression: a unified solution. In *Proceedings of the 1997 International Conference on Image Processing (ICIP '97) 3-Volume Set-Volume 1 - Volume 1*, ICIP '97, pages 13–, Washington, DC, USA, 1997. IEEE Computer Society.