Machine Learning Approach to Blocking Effect Reduction in Low Bitrate Video

Ana Stojkovikj[1], Dejan Gjorgjevikj[2], Zoran Ivanovski[1]

[1] FEEIT Skopje Macedonia
anazstojkovic@gmail.com, zoran.ivanovski@feit.ukim.edu.mk
[2] FINKI Skopje Macedonia
dejan.gjorgjevikj@finki.ukim.mk

Abstract. This work presents an approach for blocking artifacts removal in highly compressed video sequences using an algorithm based on dictionary learning methods. In this approach only the information from the frame content is used, without any additional information from the coded bit-stream. The proposed algorithm adapts the dictionary to the spatial activity in the image, by that avoiding unnecessary blurring of regions of the image containing high spatial frequencies. The algorithms effectiveness is demonstrated using compressed video with fixed block size of 8x8 pixels.

Keywords: Image compression, Video compression, Coding schemes, Blocking artifacts, Super-resolution, Dictionary learning methods, Machine learning methods.

1 Introduction

Digital video is essential part of human interaction today. Its widespread was made possible by introduction of fast and efficient standards for video compression. The most popular and most widely used today is H.264/MPEG-4 AVC, while the new HEVC standard is still in the phase of slow acceptance by the industry due to its complexity. The compression algorithms used in most standards are prone to introduction of artifacts in the final compressed video sequence that can be especially noticeable at low bitrates. The nature of different types of artifacts, as well as the reasons for their introduction, is described in details in [1,2]. Among the different types of artifacts probably the most perceptually annoying are the blocking artifacts. To cope with this problem, compression standards for digital video of the H.264 series have embedded deblocking filter. Another widely used approach is post-processing, performed on the decompressed video sequence. In that direction many algorithms for reduction of blocking artifacts were proposed [3,4,5,6,7]. They use spatial filtering techniques [4] in the area where blocking effect appears or techniques in which the discontinuity in the luminance level is modeled with 2D linear function [5,6]. In [4], three filtering modes depending on the spatial activity and the characteristics of the human visual system (HVS model) are proposed. The algorithm depends on the coding information extracted from the bit-stream. These algorithms treat only the fixed blocking effect

S. Loshkovska and S. Koceski (eds.), *ICT Innovations 2015*,
Advances in Intelligent Systems and Computing 399,
DOI: 10.1007/978-3-319-25733-4_18

introduced at the boundaries of the block, and not the blocking effect inside the block. In [7,8], fixed blocking artifacts, as well as displaced ones, that are result from motion compensation between frames, are effectively treated and reduced. In [7] a technique that utilizes 1D spatial filtering is proposed. It is implemented in two phases, detection of presence of the blocking artifacts and adaptive directional filtering. In [8], a fast algorithm for detection and reduction of displaced and fixed blocking-artifacts that considers only the luminance samples of the frame was proposed. Compared to [8], [7] is more computationally expensive, due to the fact that spatial filtering is applied on all 64 pixels in the block. Although many algorithms for adaptive filtering were proposed, still one of the major problems in these algorithms is introduction of blurring in the areas with high spatial activity.

Another very pronounced artifact of video compression is blurring due to the high frequencies suppression in the quantization phase of the compression algorithm. This artifact is usually coped with using image restoration and super-resolution techniques. Many algorithms for single image super-resolution are based on the concept of joint dictionary learning and sparse representation [9,10,11,12]. These techniques are effective in boosting of high frequencies and, thus, sharpening the image. However, when applied to images containing blocking artifacts they often increase the visibility of the artifacts.

In our approach an algorithm similar to those utilized for super-resolution is used. The algorithm aims to restore the compressed frame, with an intention of reducing the blocking artifacts and increasing the high-frequency content at the same time. Its novelty is in combining the adaptive filtering approach [8] and the dictionary learning methods via sparse representation of an image patch [9,10,11].

In the Section 2 of the paper, a short overview of the nature of different compression artifacts is presented, after which the proposed algorithm is described. Experimental results are presented in Section 3 and Section 4 contains conclusions and directions for future research.

2 Proposed Algorithm

In order to better present the proposed algorithm, a short description of the blocking artifacts nature is presented in the following text.

The utilization of blocks, as base units in processes of transformation, quantization and motion estimation generates unreal discontinuities in the block boundaries in the reproduced frame of the video sequence. These discontinuities can be classified into three sub-categories, usually designated as mosaic effect, staircase effect and false edge [1]. Mosaic effect appears in regions with low spatial activity, i.e. smooth regions. On a block level, in the process of quantization, very often almost all alternate components (AC) from the DCT transform are quantized to zero, therefore, in the reconstruction stage blocks are reconstructed from the DC components. The fusion of these reconstructed blocks produces mosaic effect, and it is characterized with abrupt changes of the luminance level at the block edges. Staircase effect appears along a diagonal line or curve, in the form of fake vertical and horizontal edges at the block

boundaries. False edge appears in the vicinity of real edge, and it is due to the motion estimation and compensation between frames in the video.

Restoration of compressed images is a real challenge due to the existence of compression artifacts. Since there is no available information about the uncompressed image, there is a need of a priori knowledge that can be obtained using machine learning approaches. Most intuitive approach in knowledge based image restoration is the dictionary learning approach that is widely used in single-frame SR approaches.

The proposed algorithm was implemented to work with image blocks of size 8x8 pixels; nevertheless, the same approach is applicable for different block sizes. In this paper only grayscale frames (Y component) are considered. The algorithm can be easily extended to consider color frames.

The approach consists of three steps, shown in Fig.1 a). In the first step, image patch of size 8x8 pixels is extracted from the area around each pixel of the frame from the compressed video. For each extracted patch the procedure in the second step, shown in Fig.1 b), is applied separately for horizontal and vertical direction.

In order to make a better distinction between the different types of compressed image patches, we trained three separate dictionaries depending on the spatial activity in the region around the pixel of interest. In the first step, the spatial activity is calculated and then depending on the activity one of the following cases applies. In case when spatial activity is very high, the extracted image patch remains unchanged and there is no need for reconstruction. If the activity is not very high, recovery patch is estimated using one out of three dictionaries, depending on the level of measured activity as described in subsection 2.1. After selecting one of the three dictionaries, a sparse representation of the recovery patch is estimated, as a linear combination of the available dictionary pairs. Iterative estimation of the sparse representation is performed by minimizing the error between the extracted compressed image patch and the estimate of the patch. As a minimizing function, L_2 norm with regularization term is used. In the third step back projection is performed by averaging the luminance of the overlapping areas of neighboring pixels. At the end the frames restored carrying out the procedure in horizontal and in vertical direction are averaged.

2.1 Measuring the Local Spatial Activity

The proposed algorithm uses three types of dictionaries. The selection of the dictionary to be used is determined by the values of the parameters calculated from the luminance values of the neighboring pixels, following the approach of the filtering algorithm described in [8]. For the vertical direction these parameters ($L_{i,j}$, $R_{i,j}$ and $D_{i,j}$) are calculated as shown by the equations (1), (2) and (3). Similar equations are used for the horizontal direction.

$$D_{i,j} = f_{i,j} - f_{i,j+1} \tag{1}$$

$$L_{i,j} = \sum_{m=1}^{3} \left| f_{i,j-m} - f_{i,j-m+1} \right| \tag{2}$$

$$R_{i,j} = \sum_{m=1}^{3} \left| f_{i,j+m} - f_{i,j+m+1} \right| \tag{3}$$

Here $f_{i,j}$ is the luminance value of the pixel at the coordinates i and j from the compressed image. The value of $D_{i,j}$ reflects luminance difference at the border between columns j and $j+1$, and the values of $L_{i,j}$ and $R_{i,j}$ reflect the activity in the region of size 3 pixels left and right of the border, respectively.

The same thresholds as in [8] were used, in order to distinguish which dictionary to use. As shown in Fig.1 b), the first dictionary is used in image regions with low spatial activity, where blocking artifacts are most noticeable. The second dictionary is used in regions with medium spatial activity, weak edges and textures. The third dictionary is used for regions with high spatial activity, sharp edges and clear textures.

Very high values of these measurements imply occurrence of natural edge, in which case the image pattern should be left unchanged.

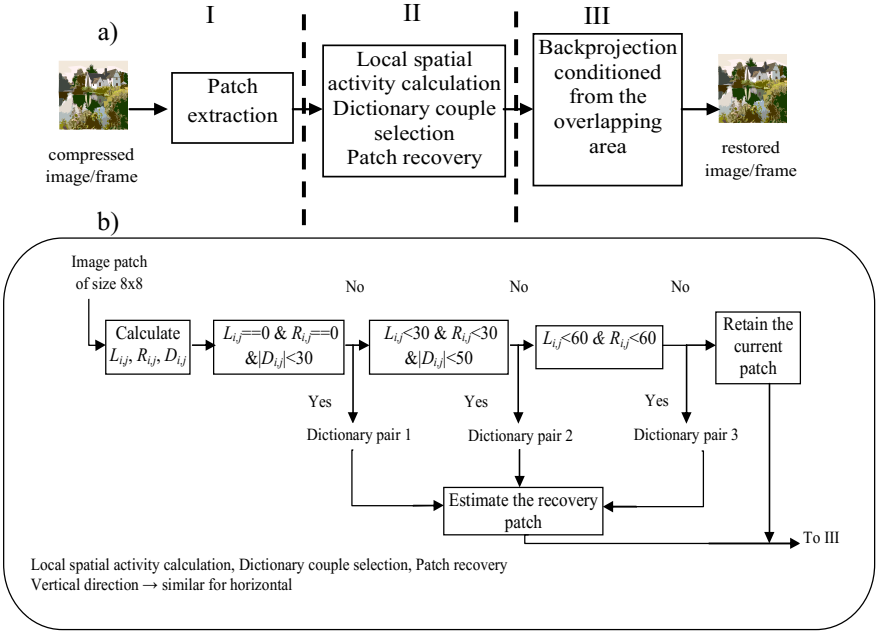

Fig. 1. Block diagram of the proposed algorithm

2.2 Training Process

The aforementioned dictionaries are constructed during the training process. 51 cropped images of size 256x256, taken from frames of 10 different low bitrate videos were used for training. Different types of dynamic and static scenes, with big content variety, were considered.

In order to employ the idea for joint dictionary learning that is usually applied in single-image super-resolution, we used the same concept as in [9] and [10]. Every dictionary is a set of pairs of patches - dictionary pairs. Each pair consists of a patch extracted from the uncompressed image and a corresponding patch from the compressed image. All patches in a dictionary extracted from uncompressed frames are

forming a subset denoted as $\mathbf{D}_{\mathrm{u}}$, and the corresponding parts of the dictionary pairs, extracted from the compressed frames are forming a subset denoted as $\mathbf{D}_{\mathrm{c}}$. Training set of dictionary pairs will be denoted with $\mathbf{P} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{\mathrm{W}}$, where $\mathbf{X}^u = \{\mathbf{x}_i\}_{i=1}^{\mathrm{W}}$ represents the subset of uncompressed image patches, and $\mathbf{Y}^c = \{\mathbf{y}_i\}_{i=1}^{\mathrm{W}}$, is the subset consisted of compressed image patches. W is the number of patterns in the set. The sparse representation is denoted with $\mathbf{Z}$.

Joint Dictionary Learning. Joint dictionary learning in the training stage is usually performed with utilization of (a) K-SVD algorithm, or (b) k-means algorithm, or simply by (c) alternate minimization of particular cost function of two variables, the estimated set $\{\mathbf{D}_{\mathrm{u}}, \mathbf{D}_{\mathrm{c}}\}$ and estimated sparse representation $\mathbf{Z}$.

Joint Dictionary Learning Using L2 norm Minimization. The estimation of the dictionary is achieved by minimizing the cost functions of the form:

$$\mathbf{D}_{\mathrm{u}} = \arg\min_{\{\mathbf{D}_{\mathrm{u}}, \mathbf{Z}\}} \|\mathbf{X}^u - \mathbf{D}_{\mathrm{u}}\mathbf{Z}\|_2^2 + \lambda\|\mathbf{Z}\|_1 \tag{4}$$

$$\mathbf{D}_{\mathbf{c}} = \arg\min_{\{\mathbf{D}_{\mathrm{c}}, \mathbf{Z}\}} \|\mathbf{Y}^c - \mathbf{D}_{\mathrm{c}}\mathbf{Z}\|_2^2 + \lambda\|\mathbf{Z}\|_1 \tag{5}$$

by forcing the uncompressed and compressed representations to share same minimization code, as proposed in [9]. Minimization procedure, for both dictionary pair, and sparse representation is performed iteratively with appropriate alternation of the variables (method (c) from above). This type of dictionary learning was performed with the Matlab package developed in [13] that utilizes Quadratically Constrained Quadratic Programming Package.

Joint Dictionary Learning Using Joint k-means Algorithm. Another approach for dictionary learning known as joint k-means clustering (JKC) is presented in [10]. In this approach the main idea is to jointly cluster both types of image patches, i.e. image patches from the compressed frame and the appropriate image patches from the uncompressed frame.

The procedure is similar to the classical k-means clustering. For k clusters, we can define a set of cluster centers $\{\mathbf{c}_j\}_{j=1}^{k}$, where each center $\mathbf{c}_j$ consists of uncompressed and compressed parts, $\mathbf{c}_j^x$ and $\mathbf{c}_j^y$, respectively. According to the algorithm joint patch vector $\mathbf{xy}_i = (\mathbf{x}_i, \mathbf{y}_i)$ belongs to certain cluster if both $\mathbf{x}_i$ and $\mathbf{y}_i$ share the same center. The algorithm is consisted of four steps with the two alternating steps (cluster assignment and cluster re-centering), as follows:

1. Arbitrarily initialize the k centers.
2. (Cluster assignment) For each $i \in \{1, \ldots N\}$, $L(i) = j'$ if both $\mathbf{c}_{j'}^x$ and $\mathbf{c}_{j'}^y$, are the closest centers to $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively; otherwise $L(i) = 0$.
3. (Cluster re-centering) For each $j \in \{1, \ldots k\}$, a related cluster is defined as $\mathcal{C}_{\mathrm{j}} = \{\mathbf{z}_{\mathrm{i}} \text{ s.t. } \mathbf{L}(i) = j\}$ and the joint center $(\mathbf{c}_j^x, \mathbf{c}_j^y)$ is recomputed.
4. Repeat steps 2 and 3 until $\boldsymbol{L}$ no longer changes.

In this procedure, $\boldsymbol{L}$ is a vector of labels that contains, element by element, the index of the assigned cluster. We set $\boldsymbol{L}$=0 for those vectors that do not find any placement, i.e. do not belong to the same neighborhood (cluster) of compressed and uncompressed patches.

Additionally, for each obtained dictionary pair, in order to counter-balance the negative effect of the pruning, simple geometrical transformations of the patches should be considered. These are: rotation of 90°, 180° and 270°, horizontal and vertical reflection, as well as the two types of diagonal reflection.

2.3 Patch Recovery and Image Restoration Process

In this step of the proposed approach, the aim is to estimate the recovery patch by using the sparse representation as a linear combination from the patches in the $\mathbf{D}_u$ subset of the dictionary. The coefficients of the sparse representation $\boldsymbol{\alpha}$ are estimated by solving the optimization problem, as shown below. After that, estimation of the recovery patch $\boldsymbol{x}$ is performed using estimated coefficients. The procedure is as follows:

Input: The appropriate trained dictionary consisted of $\mathbf{D}_u$ and $\mathbf{D}_c$ and the extracted patch $\boldsymbol{y}$ for each pixel of the compressed frame.

1. Subtract the DC component from the particular image patch.
2. Solve the optimization problem defined with: $\min_{\boldsymbol{\alpha}} \|\mathbf{D}_c\boldsymbol{\alpha} - \hat{\boldsymbol{y}}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$.
5. Estimate the restoration patch $\boldsymbol{x}=\mathbf{D}_u \cdot \boldsymbol{\alpha}$.
6. Backprojection: put the estimated patch back into the restored image $\widehat{\mathbf{X}}$ by averaging all estimated values for each pixel. Multiple values are estimated for each pixel due to overlapping blocks.

Output: Restored image $\widehat{\mathbf{X}}$.

3 Results

For the performance testing of the proposed approach, nine different video sequences were taken from the Consumer Video Library database site [14]. They were compressed to constant bitrates in the range of 512 to 1200 kbps, and from each sequence one frame was extracted and converted to grayscale. The original uncompressed sequences labeled with 3, 4, 7 and 9 are VGA sequences (640x480p), and the sequences labeled with 1, 2, 5, 6 and 8 are HD videos (1920x1080p). Most of the testing frames were taken from parts of the videos (sequences labeled with 1, 2, 4, 6 and 8) where the scene was static and the camera wasn't moving. In sequence 3, the scene is static and there is a considerable zooming present, while in the sequences 7 and 9, the scenes are very dynamic and the camera is not moving. The sequence labeled with 5 has a very dynamic scene and moving camera (football terrain). Content from natural scene is considered in sequences 1, 2, 3, 7 and 8, and the sequences labeled with 1, 4,

5, 7 and 8 are abundant with details. Faces, as most searched content in an image, are considered in sequences labeled with 4, 6 and 9.

As a measure of quality we have used Peak Signal to Noise Ratio (PSNR) and Mean structural similarity index (MSSIM). These measures are frequently used when the objective and subjective quality are discussed, despite the fact that they do not correspond to the amount of blockiness in a particular image. In order to measure the amount of blockiness in the restored frame we have used Blockiness Measure (BM), as proposed in [15].

In all tests the regularization factor λ was estimated using extensive search in the range [0, 1]. Visually most pleasing results were achieved using λ=0.1.

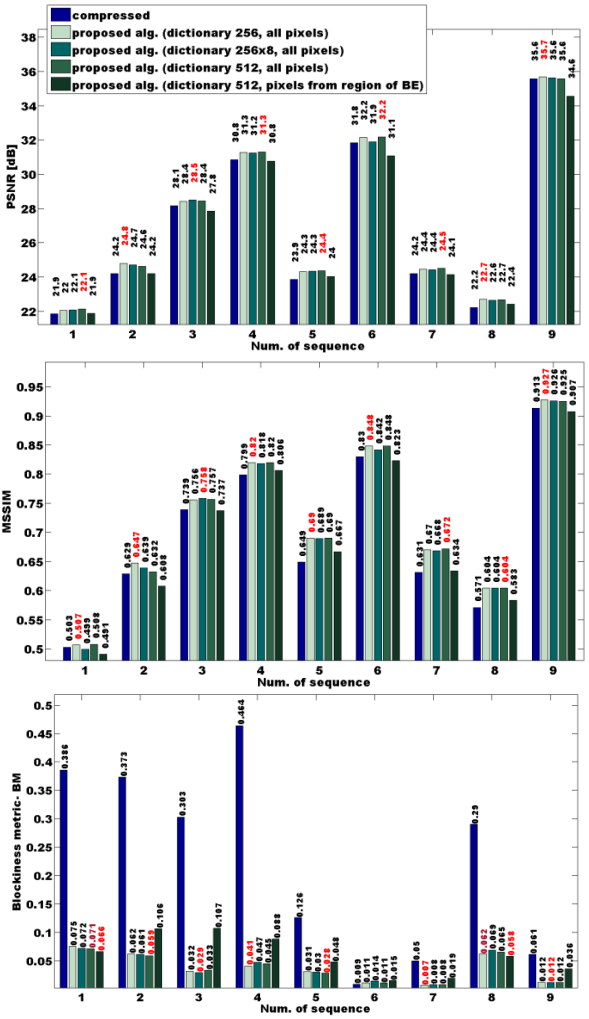

Fig. 2. PSNR, MSSIM and BM values for the compressed frames, and the restored ones with four versions of the proposed algorithm

Two variants of the proposed approach were considered in the performance testing. In the first variant the restoration is performed for each pixel in the frame, and in the second the restoration is performed only for the pixels where the blocking effect (region of BE) was detected, with the detection procedure described in [8]. Results from this comparison are shown in Fig.2. It can be noticed that when algorithm is applied to each pixel, the performance is better in terms of measured quality as well as visual quality.

Two different algorithms for dictionary learning were considered. The results of using dictionaries constructed by algorithm labeled with (c), and the algorithm labeled with (b), (both described in Section 2) were compared.

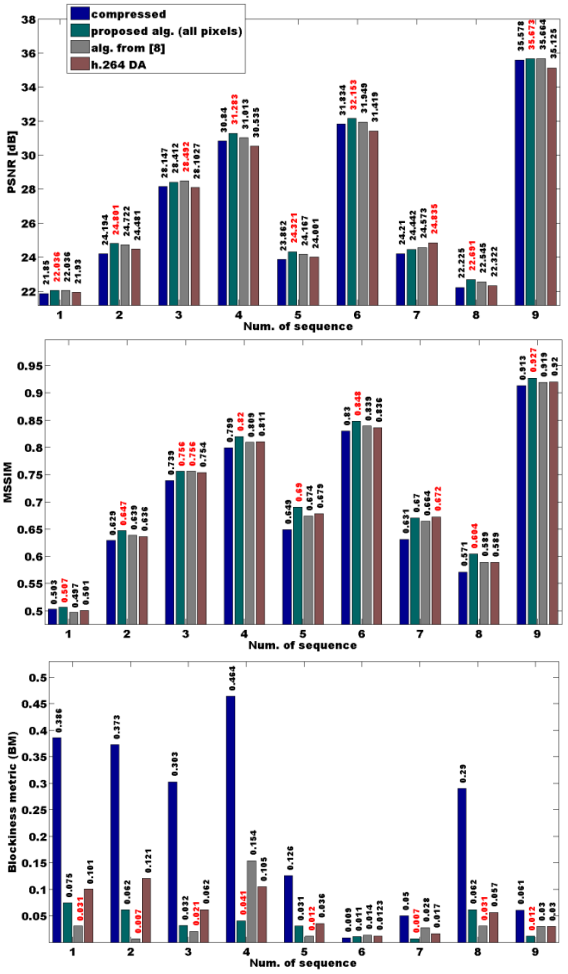

Fig. 3. PSNR, MSSIM and BM values for the compressed frames, the restored frames with the proposed algorithm, the restored frames with the algorithm from [8] and the restored frames with h.264 DA

Because the differences in the obtained results are almost unnoticeable, only numerical values of the quality improvement for the algorithm (b) are presented here.

The size of the dictionary and the variability of image patterns may affect the quality of the restored image. For that purpose dictionaries of size 256, 256x8 (the acquired 256 patterns were geometrically transformed in 8 versions) and 512 were trained. The results are presented in Fig.2. It can be noticed that the improvement in quality compared to compressed frames is achieved in all cases, except for sequence 6 when the quality is measured using BM. This is due to the fact that compressed sequence 6 does not contain significant amount of blocking artifacts. The restoration procedure for this sequence introduced smoothing of some textured regions (ex. grass, leaves etc.), however, the level of smoothing is very low and the restored frame is as pleasant as the compressed one. For all other sequences the blockiness is reduced, but it is not completely eliminated, as can be seen from the values of the BM presented in the graphics from Fig.2 and Fig.3, as well from the results presented in Fig.4.

Fig. 4. Frames from sequence 1, 4, 6, 7, 8, 9 (column-wise); Frames: compressed, restored with dictionary of size 256, restored with dictionary 256x8; restored using [8]; restored using the h.264 deblocking alg. (row-wise)

If we compare the numerical results of different size dictionaries of different sizes, presented in graphics on Fig.2, we can notice that in most cases they have higher values when the images are restored using dictionaries of size 512. It can also be noticed that results achieved with dictionaries of size 256, don't differ too much from those achieved with dictionaries of size 512. This fact brings us to a conclusion that in cases where a particular dictionary is descriptive enough, increasing its size does not affect significantly the video quality. On the other hand, the usage of smaller dictionaries is more efficient in terms of computations and time consumption. Also, from Fig.2, it can be noticed that in the cases when all versions of geometrical appearance of the patch prototypes are considered in the dictionary, PSNR and MSSIM have smaller values compared to those when using dictionaries of sizes 256 and 512, suggesting that dictionaries without geometrical variations are more effective. At the same time, adding the geometrical variations to the dictionaries increases the variability of patterns that are used in the restoration procedure, and due to this fact, the final estimate of the image has more details and distinguishable edges, thus better visual quality. In what follows only the results obtained using the dictionary 256 will be presented.

Comparison results of different algorithms can be seen in Fig. 4. The fourth and the fifth row of Fig.4 show results obtained using the algorithm proposed in [8] and the in-loop adaptive deblocking algorithm implemented in h.264 (h.264 DA), [16], applied as a post-processing algorithm, respectively.

The numerical results for the approach proposed in [8], and h.264 DA (mode 4 - strongest filtering), in comparison with the proposed algorithm are presented in Fig.3. As can be seen in Fig.3, our approach has achieved better results than [8] in 7 out of 9 sequences in PSNR terms and in all 9 sequences in MSSIM terms. In terms of BM the proposed algorithm outperforms [8] only in four cases (sequences labeled with 4, 6, 7 and 9). For the rest 5 sequences, the reduction of the blocking effect is obvious, but the numerical values show that the performance of the proposed and the algorithm from [8] are comparable. Considering h.264 DA, the proposed algorithm shows better performance in terms of PSNR and MSSIM for all sequences except for sequence 7. In this sequence, considering that the camera is moving and also the movement of the bees is rapid, applying stronger filtering with h.264 DA produced smoother outcome in which the blocking artifacts were reduced, while some details were lost. This caused higher PSNR and MSSIM values, compared to the results achieved with the proposed algorithm.

4 Conclusion and Future Work

In this paper an algorithm for adaptive restoration using dictionary learning methods, targeting blockiness reduction in highly compressed videos, was presented. From the presented results it can be concluded that higher values of PSNR and MSSIM for the proposed algorithm are result of the performed restoration, which cannot be obtained using only adaptive low-pass filtering. The presented results also demonstrate significant blocking-effect reduction. The overall performance of the proposed algorithm is comparable and, in some cases, superior to the algorithm proposed in [8] and h.264 DA. Considering the computational cost of the algorithms, the proposed algorithm is

computationally more expensive compared to other two algorithms. However, it is a choice of tradeoff between the achieved higher quality and performance speed.

The future research will focus on expansion of the algorithm to work with color videos and different sizes of compression blocks. The research will also address the problem of computational complexity through optimization of the descriptive power of the dictionary.

References

1. Zeng, K., Zhao, T., Rehman, A., Wang, Z.: Characterizing Perceptual Artifacts in Compressed Video Streams. In: Proc. of SPIE, Human Vision and Electronic Imaging XIX, vol.9014 (2014)
2. Randhawal, K.S., Kumar, P.: A Novel Approach for Blocking Artifacts in Compressed Video Streams. In: International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, vol.2. (2012)
3. Kong, H.S., Vetro, A., Sun, H.: Edge map guided adaptive post-filter for blocking and ringing artifacts removal. In: Proc. of International Symposium on Circuits and Systems (ISCAS), vol.3, pp. III-929-932. (2004)
4. Tai, S.C., Chen, Y.Y., Sheu, S.F.: Deblocking Filter for Low Bit Rate MPEG-4 Video. In: IEEE Trans. Circuits Syst. Video Technol., vol.15, no.6, pp.733-741. (2005)
5. Liu, S., Bovik, A.C.: Efficient DCT-Domain Blind Measurement and reduction of Blocking Artifacts. In: IEEE Trans. Circuits Syst. Video Technol., vol.12, no.12, pp.1139-1149. (2002)
6. Petrovski, A., Kartalov, T., Ivanovski, Z., Panovski, Lj.: Blind Measurement and Reduction of Blocking Artifacts. In: 48th International Symposium ELMAR on Multimedia Signal Processing and Communications, pp.73-76. (2006)
7. Kochovski, B., Kartalov, T.,Ivanovski, Z., Panovski, Lj.: An Adaptive Deblocking Algorithm for Low Bitrate Video. In: Proc. of IEEE 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP), pp.888-893. (2008)
8. Petrov, A., Kartalov, T., Ivanovski, Z.: Blocking Effect Reduction in Low Bitrate Video on a Mobile Platform. In: Proc. of IEEE 16th International Conf. on Image Processing (ICIP), pp.3937-3940. (2009)
9. Yang, J., Wright, J., Huang, T., Ma, Y.: Image Super-resolution via Sparse Representation. In: IEEE Trans. on Image Processing, vol.19, no.11, pp.2861-2873. (2010)
10. Bevilacqua, M.: Algorithms for Super-resolution of Images and Videos Based on Learning Methods. In: Image Processing, University of Rennes 1. (2014)
11. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi Morel, M.-L.: Compact and Coherent Dictionary Construction for Example-based Super-resolution. In: Proc. of IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp.2222-2226. (2013)
12. Thiagrajan, J., Ramamurty, K., Spanias, A.: Multilevel Dictionary Learning for Sparse Representation of Images. In: Proc. of IEEE DSP/SPE Workshop, pp.271-276. (2011)
13. Lee, H., Battle, A., Raina, R.., Ng, A.: Efficient Sparse Coding Algorithms. In: Advances in Neural Information Processing Systems, pp.801-808. (2007)
14. Consumer Digital Video Library, http://www.cdvl.org/
15. Wang, Z., Bovik, A.C., Evans, B.L.: Blind Measurement of Blocking Artifacts in Images. In: Proc. of International Conf. on Image Processing, vol.3, pp.981-984. (2000)
16. List, P., Joch, A., Lainema, J., Bjontegaard, G., Karczewicz, M.: Adaptive Deblocking Filter. IEEE Trans. Circuits Syst. Video Technol, vol.13, no.7, pp.614-619. (2003)