Saliency guided Wavelet compression for low-bitrate Image and Video coding

Souptik Barua[1], Kaushik Mitra[2] and Ashok Veeraraghavan[1]
[1]Rice University, USA, [2]Indian Institute of Technology Madras, India

***Abstract*—We propose an improved saliency guided wavelet compression scheme for low-bitrate image/video coding applications. Important regions (faces in security camera feeds, vehicles in traffic surveillance) get degraded significantly at low bitrates by existing compression standards, such as JPEG/JPEG-2000/MPEG-4, since these do not explicitly utilize any knowledge of which regions are salient. We design a compression algorithm which, given an image/video and a saliency value for each pixel, computes a corresponding saliency value in the wavelet transform domain. Our algorithm ensures wavelet coefficients representing salient regions have a high saliency value. The coefficients are transmitted in decreasing order of their saliency. This allows important regions in the image/video to have high fidelity even at very low bitrates. Further, our compression scheme can handle several salient regions with different relative importance. We compare the performance of our method with the JPEG/JPEG-2000 image standards and the MPEG-4 video standard through two experiments: face detection and vehicle tracking. We show improved detection rates and quality of reconstructed images/videos using our Saliency Based Compression (SBC) algorithm.**

Saliency, Wavelet transform, Image coding, Video coding

I. Introduction

High quality images and videos need to be significantly compressed before being transmitted over a low bitrate channel. For example, in aerial surveillance, the captured high resolution images/videos are compressed by a large factor before being transmitted over a wireless network to a base station several miles away. Important, or salient regions are often small and hence severely degraded at low bitrates, since compression standards such as JPEG/JPEG-2000 for images, and MPEG-4 for videos, do not explicitly handle salient regions. Existing saliency guided compression schemes ([1], [2], [3], [4]) suffer from one or more of the following problems: they modify the original values of transmitted coefficients to incorporate saliency; cannot handle multiple salient regions with different relative importances; or are computationally expensive.

We propose a saliency guided compression scheme which preserves the quality of salient regions even at low bitrates. The proposed SBC scheme utilizes the wavelet transform for compression, similar to the JPEG-2000 encoder [5]. The motivation for this is that salient regions are typically localized in space (and time), and can be compactly represented by wavelets. In our scheme, a detector first identifies salient regions. Then, our algorithm computes a saliency value for each wavelet coefficient. This value, which we will call wavelet saliency, prioritizes the transmission of wavelet coefficients that represent salient regions, thus preserving their quality even at high compression rates. Another major advantage of our method is that we approximate salient regions using rectangles, and thus incur only a nominal overhead in transmitting the saliency map.

The contributions of this paper are as follows:

- Develop an algorithm which, given an image/video and its corresponding saliency map, computes saliency values of its wavelet transform coefficients. Coefficients are transmitted according to saliency, thus preserving the quality of important regions.
- Show improved reconstruction and detection performance of the proposed SBC scheme at low bitrates ($<$ 0.4bits per pixel) on two classes of salient objects: a) faces and b) moving vehicles, against JPEG/JPEG-2000 and MPEG-4 AVC encoders respectively.

II. Related work

The JPEG [6] and MPEG-4 [7] compression standards use DCT to encode images and videos respectively, with the latter also using motion estimation to remove temporal redundancy. In this paper, by MPEG-4, we imply the H.264 or MPEG-4 AVC coder. The wavelet transform based JPEG-2000 standard introduced two modes of Region-of-Interest (ROI) encoding: a) general scaling based method (GSBM) and b) maximum shift or MAXSHIFT method [1]. GSBM needs to transmit the ROI's shape, while MAXSHIFT allows for only 2 saliency levels: ROI and background. Bitplane-by-Bitplane shift (BbBShift) [8] and Partial Significant Bitplane shift (PSBShift) [9] allow for multiple salient regions with different saliency values, but are not compatible with JPEG-2000. Our algorithm on the other hand can be seamlessly integrated with JPEG-2000.

Sanchez et al. [2] use a foveation approach to prioritize coefficients belonging to ROIs. A Gaussian priority distribution is calculated for each sub-band as a measure of wavelet saliency. Our method needs to compute the wavelet saliency for only a few sub-bands, and uses simple addition operations. A number of visual attention guided compression algorithms have been proposed: Harding and Robertson [10] use it to perform intelligent DCT encoding; Guo and Zhang [3] use a multiresolution spatiotemporal saliency detection model; Hadizadeh and Bajic [4] add a saliency distortion metric in the MPEG-4 codec; while Shen et al [11] use local motion and edge information. These human visual system based schemes may underperform when multiple salient objects are

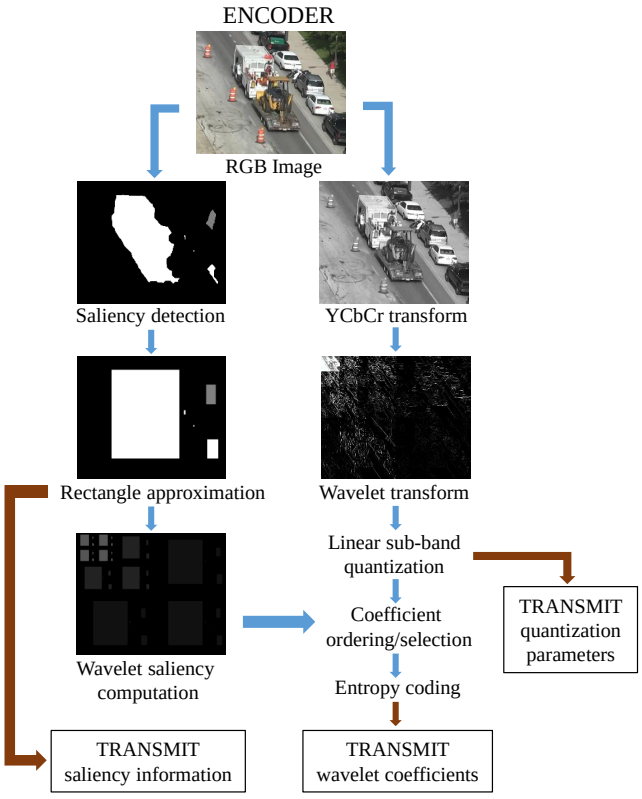

Fig. 1. The proposed SBC encoder progresses along two paths: First path generates a wavelet saliency map from a spatial saliency map. Second path generates quantized wavelet transform coefficients. The coefficients are ordered by saliency, entropy coded and transmitted as per bitrate constraints. The saliency value, and vertices of different salient rectangles are also transmitted. The decoder performs the inverse steps to reconstruct the image.

scattered throughout the frame, or when the decompressed images are viewed not by a person but by a machine for tasks such as detection or recognition. Our algorithm can flexibly use any saliency detector depending on the end-application, and incorporate multiple salient objects with different relative importances during compression.

III. Saliency Guided Wavelet Compression

We present our complete compression framework (Figure 1) in this section for images, but this readily extends to videos. The different steps are as follows:

A. Raw saliency map generation

The first step is to generate a spatial saliency map. We use a fast state of the art saliency detector: the detector can be a generic object detector (BING [12]), or it can be a context-aware object detector (Viola-Jones detector for faces [13], GMM tracking [14] for objects in motion), which detects regions with different levels of importance. Every pixel is assigned a positive integer to indicate its relative importance in the image, which we will refer to as the pixel's saliency value. The regions which the detector identifies as highly important are assigned a higher positive integer compared to the non-salient regions, such as the background.

B. Saliency map overhead reduction

Transmitting the raw saliency map with arbitrarily shaped ROIs, even in compressed form, adds a significant overhead. For example, lossless PNG compressed saliency maps of faces from the UMD Faces dataset [15] still incurs an average overhead of 0.01 bits per pixel (bpp). We overcome this problem by approximating each salient region with a rectangular bounding box. Every pixel inside the bounding box has the same saliency value. We only need to transmit the opposing vertices of the rectangle and its associated saliency value, incurring a nominal overhead. In the example given above, our bounding box approximation adds an overhead of only 2.4×10^{-5}bpp.

C. Wavelet saliency computation

The next step is to translate the rectangular spatial saliency map to a wavelet domain saliency map so as to decide which coefficients will be transmitted first. We define the wavelet saliency for a wavelet coefficient as the sum of the image pixel values at all locations where the corresponding wavelet basis function is non-zero.

Given an image I and its spatial saliency map s_I, we first resize the original image dimensions to the nearest values M and N such that both $M, N \equiv 0(\mod 2^K)$, where K is the number of levels of wavelet decomposition desired. This is done to avoid odd-dimensional sub-bands, which will require additional zero padding. The wavelet saliency s_w^k is recursively computed for LL bands at each level of decomposition k ($k \in \{0, 1, 2, \cdots, K-1\}$) as follows:

$$s_w^{k+1}(i,j) = \sum_{i'=2i-1}^{2i} \sum_{j'=2j-1}^{2j} s_w^k(i',j') \qquad (1)$$

where $i = 1, 2, \cdots, M/2^k$ and $j = 1, 2, \cdots, N/2^k$. The base case of the recursion is $s_w^0(i,j) = s_I(i,j)$, the spatial saliency value at (i,j). This is effectively a Haar wavelet transform of the LL band. We assign identical wavelet saliency values to the LH, HL and HH bands as illustrated in Figure 2. This fact, coupled with the recursive nature of (1), makes the computation fast. For example, suppose the background is assigned a spatial saliency value of 1. The wavelet saliency values, except at the salient/non-salient boundaries, will be $4, 16, \cdots, 4^K$ at successive levels of decomposition, since the Haar wavelet's support is 4 pixels. The choice of value for the salient region is made depending on how much importance we wish to give it. We typically use a spatial saliency value of $4^k + 1, k \in \mathbb{N}$ for the salient region. This value implies that k levels of wavelet coefficients $(K, K-1, \cdots, K-k+1)$ from the salient region will be transmitted ahead of the K^{th} level of background coefficients. Hence we say that the salient region is k levels more salient than the background.

D. Wavelet transform

The original uncompressed RGB image is first converted to YCbCr color space in the same way as the JPEG-2000 standard [5]. We then compute the Haar wavelet transform of the YCbCr image. Again like JPEG/JPEG-2000, the chrominance channels are subsampled by a factor of 2, both horizontally and

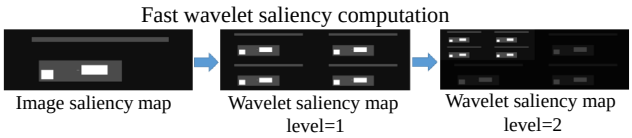

Fig. 2. The wavelet saliency computation procedure. We recursively compute the Haar wavelet transform of the LL band at each decomposition level, and copy the saliency values to the corresponding locations in the LH, HL and HH bands.

vertically, by setting the LH, HL and HH wavelet coefficients at the finest scale to zero.

E. Quantization, ordering and entropy coding

The wavelet transform coefficients are then quantized on a sub-band basis, along the lines of JPEG-2000. However, we got the best results by simple linear quantization. Each sub-band is mean subtracted and then scaled to an 8-bit integer. A fraction of the coefficients is then selected for transmission, according to bitrate constraints. The chosen coefficients are sequentially written onto a binary file, which is entropy coded using LZ77 algorithm, followed by Markov chain based range entropy coding [16] and transmitted. The means and scaling factors of the chosen coefficients are also transmitted, with a nominal overhead.

F. Decoder

The bounding box information received is used to recreate the spatial saliency map. An identical algorithm is used to compute the wavelet saliency map. The received wavelet transform coefficients are decoded, then placed in their correct location in the wavelet decomposition structure using the wavelet saliency map. For every sub-band, we then rescale each coefficient using the corresponding scale coefficient and add the corresponding mean value. The inverse wavelet transform is performed, followed by a YCbCr to RGB transform to reconstruct the image at the decoder.

IV. Experiments and Results

We evaluate our saliency guided compression scheme SBC on two datasets: a) the UMD remote faces image dataset [15]; and b) the VIRAT video dataset [17]. We use the Haar wavelet basis for all wavelet transform computations. We compute 6 levels of wavelet decomposition for the image compression experiment, and 4 levels of wavelet decomposition for successive batches of 16 frames for the video compression experiment. We used the JPEG and JPEG-2000 codecs available in MATLAB. For MPEG-4 AVC we used the popular FFMPEG codec [18]. All subsequent figures are best viewed in color.

A. Image compression on UMD faces dataset

The salient region in this dataset is faces. We detect faces using the Viola-Jones detector [13].

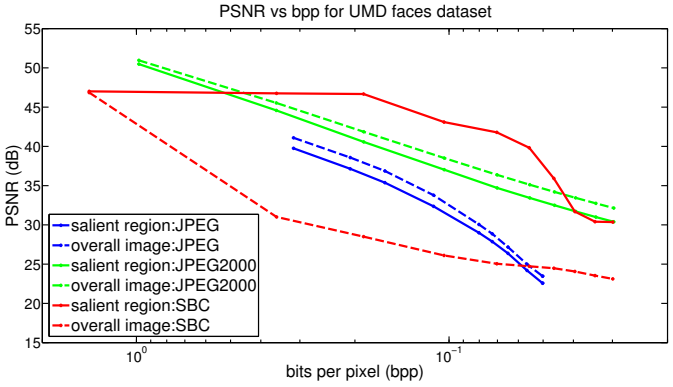

Fig. 3. Average PSNR(dB) of 48 reconstructed UMD face dataset images compressed using JPEG, JPEG-2000 and SBC at different bpp. SBC outperforms JPEG-2000 in the bpp range 0.04 to 0.4bpp

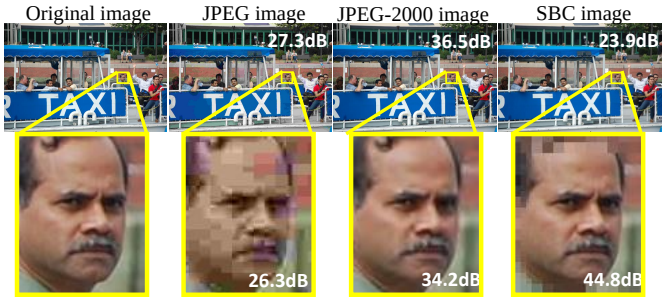

Fig. 4. Reconstructed image from the UMD faces dataset for JPEG, JPEG-2000 and SBC at 0.07bpp. Our reconstruction has lesser artifacts in the face region.

Reconstruction performance comparison: Figure 3 plots the average PSNR (dB) of 48 6-megapixel images compressed using JPEG, JPEG-2000 and SBC, as a function of bits per pixel (bpp). In the low bitrate regime of 0.04-0.4bpp, we are significantly better than JPEG, whereas we outperform JPEG-2000 by 4.5dB on average. Figure 4 shows the improved reconstructed result on an image chosen from the dataset. JPEG and JPEG-2000 both display visible artifacts in the face region at the selected bitrate of 0.07bpp.

Detection performance comparison: We next run a face detection experiment on the reconstructed images. As reported in Table I, SBC has a better true face detection rate (TDR) than both JPEG (17% higher) and JPEG-2000 (2.5% higher)

	JPEG		*JPEG − 2000*		*SBC*	
	TDR	*FPR*	*TDR*	*FPR*	*TDR*	*FPR*
bpp	(%)	($\times 10^{-4}$%)	(%)	($\times 10^{-4}$%)	(%)	($\times 10^{-4}$%)
0.02	-	-	84.8	2.5	**88.7**	1.5
0.04	37.2	0.2	86.0	2.5	**88.7**	1.6
0.06	59.5	0.5	88.0	2.8	**89.5**	2.1
0.08	82.9	0.9	88.0	2.8	**89.5**	2.3
0.10	**89.5**	1.6	89.5	2.4	89.5	2.3

TABLE I
True Detection rate (%) and False positive rate (%) measured at different bpp for UMD faces dataset.

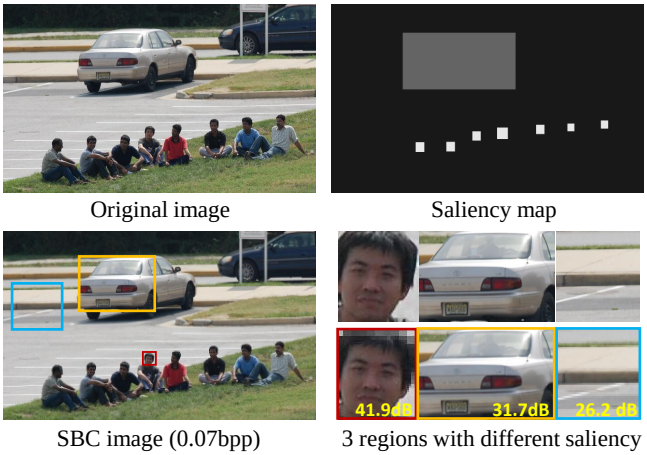

Fig. 5. This figure demonstrates SBC's ability to handle images with multiple levels of saliency. The top row shows the original image and its corresponding 3-level saliency map. The bottom row shows the decompressed SBC image on the left. We pick out 3 regions of different saliency values from the original image and the SBC compressed image as shown on the bottom right. The PSNR values (face: 41.9dB, car: 31.7dB and pavement: 26.2dB) show that reconstruction quality of a region is proportional to its saliency.

below 0.1bpp. Though the false positive rate (FPR) is higher than JPEG, it is significantly lesser than JPEG-2000.

Multiple levels of saliency: SBC can handle many salient regions with different saliency values. We demonstrate this in Figure 5 on an image chosen from the UMD faces dataset. We observe from the corresponding saliency map that the chosen image has 3 levels of saliency: the faces have the highest saliency value (17), the parked car has medium saliency (5), while the rest of the image is non-salient (1). The intuition behind the choice of saliency values has been explained in III C. In this case, it means that the faces are effectively two wavelet levels more salient ($17 > 4^2$), whereas the car is one wavelet level more salient ($5 > 4^1$) than the rest of the image. The image compressed using SBC is shown in the bottom row. We also calculate the PSNR in 3 regions having different saliency values: a face, rear of the parked car and a portion of the pavement. The PSNR values computed for each region indicate that the fidelity of a region in the reconstructed image is directly related to its saliency value.

B. Video compression on VIRAT dataset

We use the VIRAT dataset, which is a collection of surveillance videos, for evaluating video compression performance. We define salient regions in these videos as objects in motion. We used an adaptive mixture model based saliency detector [14] to track moving objects. The salient objects are assigned saliency values as powers of 8 plus one (for e.g, $9, 65, \cdots$), since the Haar wavelet support spans 8 pixels for videos as opposed to 4 pixels for images.

Reconstruction performance comparison: We plot the PSNR against bpp for SBC and MPEG-4 compressed videos in Figure 6. SBC recovers salient regions on average 3.2dB better than MPEG-4 for bitrates above 0.15bpp. Below this bitrate, the computationally expensive motion compensation based MPEG-4 algorithm is more efficient.

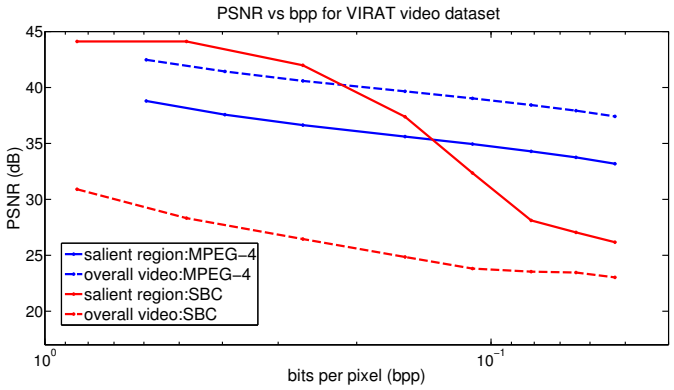

Fig. 6. PSNR (dB) plotted as a function of video bitrate measured in bits per pixel (bpp) for a VIRAT video. SBC performs on average 3.2dB better than MPEG-4 above a bitrate of 0.15bpp. Below this bitrate, MPEG-4's computationally expensive motion compensation encoding performs better than our 3D-wavelet approach.

V. Conclusion and Discussion

We proposed an improved saliency guided compression method based on the space-time localization property of the wavelet transform. Given the image/video saliency regions, we design an algorithm to compute the wavelet saliency map. The algorithm associates higher values for wavelet coefficients that represent the salient regions and transmits these ahead of other wavelet coefficients. We performed reconstruction and detection experiments to show the efficacy of the proposed algorithm. In the face image dataset, salient regions are reconstructed with quality on average 5dB more than the best compression standard for same bitrate. We also get higher rate of correct face detections on the reconstructed images than other algorithms. In the video dataset, the salient regions are reconstructed with better quality than MPEG-4 above 0.15bpp.

In the future we intend to replace Haar wavelets with CDF biorthogonal wavelets [5], to more efficiently compress data. We further intend to quantify the improved runtime performance, which arises due to using a fast saliency detector such as BING [12] (which detects salient regions at 300fps) and wavelet transforms.

References

[1] Charilaos Christopoulos, Joel Askelof, and Mathias Larsson, "Efficient methods for encoding regions of interest in the upcoming JPEG2000 still image coding standard," *Signal Processing Letters, IEEE*, vol. 7, no. 9, pp. 247–249, 2000.

[2] Victor Sanchez, Anup Basu, and Mrinal K Mandal, "Prioritized region of interest coding in JPEG2000," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, no. 9, pp. 1149–1155, 2004.

[3] Chenlei Guo and Liming Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *Image Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 185–198, 2010.

[4] Hadi Hadizadeh and I Bajic, "Saliency-aware video compression," *Image Processing, IEEE Transactions on*, vol. 23, no. 1, pp. 19–33, 2014.

[5] Athanassios Skodras, Charilaos Christopoulos, and Touradj Ebrahimi, "The JPEG 2000 still image compression standard," *Signal Processing Magazine, IEEE*, vol. 18, no. 5, pp. 36–58, 2001.

[6] Nasir Ahmed, T Natarajan, and Kamisetty R Rao, "Discrete cosine transform," *Computers, IEEE Transactions on*, vol. 100, no. 1, pp. 90–93, 1974.

[7] Atul Puri, Xuemin Chen, and Ajay Luthra, "Video coding using the h. 264/mpeg-4 avc compression standard," *Signal processing: Image communication*, vol. 19, no. 9, pp. 793–849, 2004.

[8] Zhou Wang and Alan C Bovik, "Bitplane-by-bitplane shift (BbBShift)-a suggestion for JPEG2000 region of interest image coding," *Signal Processing Letters, IEEE*, vol. 9, no. 5, pp. 160–162, 2002.

[9] Lijie Liu and Guoliang Fan, "A new JPEG2000 region-of-interest image coding method: partial significant bitplanes shift," *Signal Processing Letters, IEEE*, vol. 10, no. 2, pp. 35–38, 2003.

[10] Patrick Harding and Neil Roberston, "Task-based visual saliency for intelligent compression," in *Signal and Image Processing Applications (ICSIPA), IEEE International Conference on*. IEEE, 2009, pp. 480–485.

[11] Liquan Shen, Zhi Liu, and Zhaoyang Zhang, "A novel h. 264 rate control algorithm with consideration of visual attention," *Multimedia tools and applications*, vol. 63, no. 3, pp. 709–727, 2013.

[12] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *IEEE CVPR*, 2014.

[13] Paul Viola and Michael J Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[14] Chris Stauffer and W Eric L Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on.* IEEE, 1999, vol. 2.

[15] Rama Chellappa, Jie Ni, and Vishal M Patel, "Remote identification of faces: Problems, prospects, and progress," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1849–1859, 2012.

[16] Jacob Ziv and Abraham Lempel, "A universal algorithm for sequential data compression," *IEEE Transactions on information theory*, vol. 23, no. 3, pp. 337–343, 1977.

[17] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al., "A large-scale benchmark dataset for event recognition in surveillance video," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*. IEEE, 2011, pp. 3153–3160.

[18] Fabrice Bellard, M Niedermayer, et al., "Ffmpeg," *Available from: http://ffmpeg. org*, 2012.