DICTIONARY-BASED MULTIPLE FRAME VIDEO SUPER-RESOLUTION

Qiqin Dai, Seunghwan Yoo, Armin Kappeler, and Aggelos K. Katsaggelos

Dept. of EECS, Northwestern University, Evanston, IL, USA

ABSTRACT

In this paper, we propose a multiple-frame super-resolution (SR) algorithm based on dictionary learning and motion estimation. We adopt the use of multiple bilevel dictionaries which have also been used for single-frame SR. Multiple frames compensated through sub-pixel motion are considered. By simultaneously solving for a batch of patches from multiple frames, the proposed multiple-frame SR algorithm improves over single frame SR. We also propose a novel dictionary learning algorithm based on which dictionaries are trained from consecutive video frames, rather than still images or individual video frames, which further improves the performance of the developed video SR algorithm. Extensive experimental comparisons with state-of-the-art SR algorithms verifies the effectiveness of our proposed multiple-frame SR approach.

Index Terms— Video super-resolution, dictionary learning, sparse coding, optical flow estimation.

1. INTRODUCTION

Video SR, namely estimating high-resolution (HR) frames from low-resolution (LR) input sequences, has become one of the fundamental problems in image processing and has been extensively studied since the original work by Tsai and Huang [1].

There are mainly two categories of SR methods. In the first category, LR frames are modeled as down-sampled and degraded by blur and noise versions of the HR frames. With model-based SR methods, the original HR frames, blur kernel, noise level and motion field are estimated, either simultaneously [2–4], or separately [5]. In the second category, the LR frames are modeled as directly down-sampled version of the HR frames. Example based or learning based methods [6–9] are then proposed to estimate the HR frames from the LR inputs, without explicit estimation of the blur kernel or noise level. In this paper, we focus on the second category of methods.

Some important results have been reported applying Dictionary Learning (DL) techniques to the SR of images [6, 7, 10], deblurring [11] and to the denoising of images and image sequences [12,13]. However, according to our knowledge, the only work reported in the literature on the application of DL to video SR is the work in [8]. According to it, block-based motion estimation is performed among input and LR key-frames and DL is only applied for single frame SR when the motion compensation error is larger than the threshold. The approach reported in [8] however has neither sub-pixel precision in motion estimation nor an advanced DL SR technique.

In this work we borrow two ideas from single frame SR, namely, the bi-level coupled dictionary [6, 7, 14, 15] and the multiple-dictionary [9] ideas. Based on these we propose a multiple-dictionary multiple frame video SR algorithm utilizing sup-pixel accurate motion estimation. With the proposed SR approach the estimated optical flow is utilized to obtain multiple-frame high accuracy registration and so that an HR frame is reconstructed from multiple LR frames.

The multiple-frame SR performance is further improved by training dictionaries from consecutive video frames. Most of the DL techniques use still images or individual video frames to train the dictionaries [6–13, 16] . However, this causes some inconsistency in multiple-frame SR since we are super-resolving videos while the dictionaries are trained from still images. The proposed training from video algorithm incorporates temporal information of frame patches into the dictionaries, and makes the training and testing phases consistent. In-depth and comprehensive experiments prove that our proposed SR framework outperforms state-of-the-art SR frameworks, such as Bilevel [14, 15], Enhancer [17] and Bayesian [2] on 4K (2160×3840) sequences.

2. DICTIONARY BASED MULTIPLE-FRAME SUPER-RESOLUTION APPROACH

Given a LR sequence $\{I_1^l, \ldots, I_k^l, \ldots\}$, the goal of SR is to estimate the corresponding HR sequence $\{I_1^h, \ldots, I_k^h, \ldots\}$. Since each frame is primarily correlated with its neighbors and to also reduce computation, when we are super-resolving I_k^h, only the $(M + N)$ adjacent frames $\{I_{k-M}^l, \ldots, I_{k+N}^l\}$ are used. Clearly when $N = 0$, causal processing is performed. In this section, we introduce an approach to find the sparse representation of a sample LR patch y_k by incorporating the motion information from its neighboring frames. The core idea of this approach originates from the fact that image registration through motion compensation provides multiple

This work was supported in part by a grant from Samsung DMC R&D Center.

observations of the same scene, enabling the SR algorithm to take advantage of the details lost in the k^{th} frame but present in past or future frames.

2.1. Multiple Bi-level Dictionary Learning

The first task needed to be addressed is the coupled learning of matching HR and LR dictionaries over a large database of training HR images. Each HR image I_j^h in the training database is degraded by blur and noise and down-sampled resulting in the corresponding LR image $\widetilde{I}_j^l$. Each LR image $\widetilde{I}_j^l$ is up-sampled using bicubic interpolation so that I_j^h and I_j^l have the same size. In the rest part of this paper, we regard LR frame I_j^l as bicubic interpolation upscaled original LR frame $\widetilde{I}_j^l$ for convenience. I_j^h and I_j^l are then divided into patches of size $W \times W$. All these patches are lexicographically ordered to form the columns of matrices Y^{tr} and X^{tr}, respectively. In the dictionary learning phase, we aim at finding HR and LR dictionaries D^h and D^l such that the sparse representation of any HR patch over D^h is identical to that of the corresponding LR patch over D^l. In order to do so, Yang *et al.* [14, 15] formulated the following bi-level optimization problem

$$\begin{aligned} \min_{D^h, D^l} \quad & \left\| X^{tr} - D^h Z \right\|_F^2 \\ \text{s.t.} \quad & Z = \arg\min_{A^{tr}} \left\| FY^{tr} - FD^l A^{tr} \right\|_F^2 + \lambda \sum\nolimits_j \left\| \alpha_j^{tr} \right\|_1 \\ & \left\| D^h(:,k) \right\|_2 \leq 1, \left\| D^l(:,k) \right\|_2 \leq 1. \end{aligned} \tag{1}$$

where a_j^{tr} expresses the sparse representation of the j^th HR/LR patch pair via D^h and D^l, respectively, A^{tr} is the matrix containing all the ordered α_j^{tr} as its columns, $\|.\|_F$ and $\|.\|_1$ represent the Frobenius and the l_1 norm, respectively, λ is the regularization parameter which controls the sparsity of the sparse coefficient a_j^{tr}, and F is a linear operator which extracts features from the LR patches. Finally, the notation $D^h(:,k)$ and $D^l(:,k)$ denotes the k^{th} columns of matrices D^h and D^l, respectively.

In the testing phase, given an observed LR patch y, we first solve the following LASSO problem

$$z = \arg\min_{\alpha} \left\| Fy - FD^l \alpha \right\|_F^2 + \lambda \left\| \alpha \right\|_1 . \tag{2}$$

Then the sparse coefficient vector z is applied to the HR dictionary D^h to obtain the HR patch x corresponding to y, that is,

$$x = D^h z. \tag{3}$$

With the bi-level dictionary learning technique, in the training phase, when updating the sparse coefficient matrix A^{tr} in the lower level, the optimization is consistent with the optimization in the testing phase (Equation 2), thus guaranteeing the reconstruction accuracy. Improved SR results have been obtained with this bilevel formulation compared to the previous formulation in [6, 7]

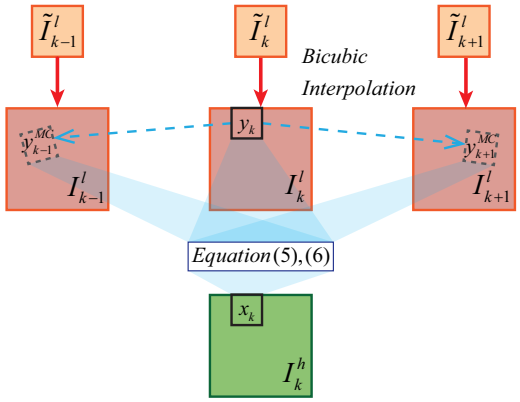

Fig. 1. The multiple-frame SR algorithm for $M = N = 1$.

Because of the divergent structures and textures in images of different styles, using a general coupled dictionary is often not good enough to super-resolve all variations of image patches. Considering the fact that image patches, according to their appearance, can be classified into different categories (such as textures, flat regions, edges, etc.), we train a coupled dictionary for each such category. The heuristic clustering strategy in [9] is integrated in our framework. More specifically, K-Means clustering is performed on sampled LR training patches Y^{tr} after applying the feature filter F. Let Y_c^{tr} be the clustered LR training patches belonging to cluster c, and X_c^{tr} its corresponding HR training patches. The coupled dictionary $(D_c^l\ D_c^h)$ is then trained on X_c^{tr} and Y_c^{tr} based on Equation (1).

After learning the C coupled dictionaries $(D_1^l\ D_1^h), \ldots, (D_C^l\ D_C^h)$, during the testing phase, for a sample LR patch y, the most appropriate dictionary c^* is determined via

$$c^* = \arg\min_{c=1\ldots C} \left\| O_c - Fy \right\|_2^2 , \tag{4}$$

where O_c is the centroid of the columns of the c^{th} LR dictionary. Here, the Euclidean distance between the centroid and the LR patch is used as the similarity measure. The best dictionary pair $(D_{c^*}^l\ D_{c^*}^h)$ is then used to find the HR version of y (denoted by x) by solving Equation (2).

2.2. A Multiple-Frame Super-Resolution Algorithm

A dictionary based multiple-frame super resolution algorithm (with $M = N = 1$) is shown in Fig. 1. The three consecutive LR frames are shown in pink while the HR frame corresponding to the middle LR frame is depicted in green. We want to estimate the patch x_k which is the HR version of the patch y_k, by combining information from patch y_k, the motion compensated patches $y_{k-M}^{MC}, \ldots, y_{k-1}^{MC}, y_{k+1}^{MC}, \ldots, y_{k+N}^{MC}$ and the pre-trained multiple coupled dictionaries $(D_c^l\ D_c^h)$.

In this approach, the optical flow estimation is performed to find the motion compensated versions of y_k in the past and

future frames, denoted by $\left\{y_{k+i}^{MC}\right\}_{i=-M}^{i=N,i\neq 0}$. To super-resolve y_k in the k^{th} frame, the most appropriate LR dictionary indexed by c^*, out of the C possible choices, is found via Equation (4). Then the best dictionary pair $(D_{c^*}^l\ D_{c^*}^h)$ is picked to find the HR version of y_k according to

$$\min_{\substack{\alpha_k,\alpha_{k+i}^{MC}\\ i=-M,\ldots,N,i\neq 0}} \left\|Fy_k - FD_{c^*}^l\alpha_k\right\|_2^2 + \sum_{i=-M,i\neq 0}^{N}\left\|Fy_{k+i}^{MC} - FD_{c^*}^l\alpha_{k+i}^{MC}\right\|_2^2 + \lambda(\|\alpha_k\|_1 + \sum_{i=-M,i\neq 0}^{N}\left\|\alpha_{k+i}^{MC}\right\|_1) + \sum_{i=-M,i\neq 0}^{N}\gamma_i\left\|D_{c^*}^h\alpha_k - D_{c^*}^h\alpha_{k+i}^{MC}\right\|_2^2 \tag{5}$$

$$x_k = D_{c^*}^h\alpha_k, \tag{6}$$

where α_{k+i}^{MC} is the sparse representation of y_{k+i}^{MC}. The first two terms in Equation (5) ensure the fidelity to the LR observations (similarly to Equation (2)). The middle two terms are l_1 regularizers promoting the sparse representation of the LR patches by the LR dictionaries. The last term enforces the similarity of the reconstructed HR patches from past and future frames to the current frame, controlled by the $(M+N)$ parameters γ_i. Only α_k is used in Equation (6) to reconstruct the HR patch in the current frame.

3. TRAINING DICTIONARIES FROM VIDEOS

Typically, for the dictionary learning process, all training patches are sampled from still images or individual video frames. This causes some inconsistency in training and testing, since we are trying to super-resolve videos while the dictionaries are trained from still images. Also the optimizations for training (Equation (1)) and testing (Equation (5) are not consistent.

We therefore propose a new dictionary training algorithms based on consecutive video frames and motion estimation. During the training phase, a number of consecutive video frames from the training videos are used. The original HR frames $\left\{I_{k+i}^h\right\}_{i=-M}^{i=N}$, are degraded to generate the LR frames $\left\{I_{k+i}^l\right\}_{i=-M}^{i=N}$. Motion estimation is then performed to find the corresponding patches $\left\{y_{k+i}^{MC}\right\}_{i=-M}^{i=N,i\neq 0}$ $\left(\left\{x_{k+i}^{MC}\right\}_{i=-M}^{i=N,i\neq 0}\right)$ of y_k (x_k) in the backward and forward frames.

The patch clustering strategy from Section 2.1 is applied to cluster each training patch set $\left(y_k, \left\{y_{k+i}^{MC}\right\}_{i=-M}^{i=N,i\neq 0}, x_k\right)$ based on the LR patches y_k after applying the feature filter F. Let $\left(Y_{k\ c}^{tr}, \left\{Y_{k+i\ c}^{MC\ tr}\right\}_{i=-M}^{i=N,i\neq 0}\right)$ be the clustered consecutive LR training patches belonging to cluster c and $X_{k\ c}^{tr}$ the corresponding HR training patches to $Y_{k\ c}^{tr}$. The coupled dictionary $(D_c^l\ D_c^h)$ for the multiple-frame testing is then trained on $X_{k\ c}^{tr}$ and $\left(Y_{k\ c}^{tr}, \left\{Y_{k+i\ c}^{MC\ tr}\right\}_{i=-M}^{i=N,i\neq 0}\right)$ according to the bilevel dictionary learning in Equation (7). Although the objective function in Equation (7) is highly nonlinear and nonconvex, it can be optimized by optimizing alternatively over $\left(A_k^{tr}\ \left\{A_{k+i}^{MC\ tr}\right\}_{i=-M}^{i=N,i\neq 0}\right)$, D_c^h and D_c^l while keeping other terms fixed.

Notice that the lower level optimization of Equation (7) in the training phase is now consistent with the optimization in the testing phase (Equation (5)).

4. EXPERIMENTAL RESULTS

Our proposed algorithm extends the bilevel dictionary learning [14, 15] in two aspects: from single dictionary to multiple dictionaries and from single frame to multiple frames. We first show that each extension is beneficial by comparing the SR performances of single dictionary singe frame SR (Bilevel), multiple dictionaries single frame SR (MDSF), single dictionary multiple frames SR (SDMF), multiple dictionary multiple frames (MDMF) and MDMF with video training (MDMFVT), i.e., the proposed approach. We also compare the performance of the proposed algorithm with state-of-the-art video SR algorithms, such as Enhancer [17] and Bayesian [2].

We performed an extensive set of experiments utilizing frames of a 4K video database [18]. There is high demand of upscaling videos of lower resolutions to 4K resolution, due to the proliferation of 4K monitors. Upscaling of 1080P (1080×1920) resolutoin to 4K videos is a representative ex-

$$\begin{aligned}
&\min_{D_c^h,D_c^l} \quad \left\|X_{k\ c}^{tr} - D_c^h Z_k\right\|_F^2 \\
&\text{s.t.} \quad Z_k, Z_{k+i}^{MC}{}_{i=-M,\ldots,N,i\neq 0} = \arg\min_{\substack{A_k^{tr},A_{k+i}^{MC\ tr},\\ i=-M,\ldots,N,i\neq 0}} \left\|FY_{k\ c}^{tr} - FD_c^l A_k^{tr}\right\|_F^2 + \sum_{i=-M,i\neq 0}^{N}\left\|FY_{k+i\ c}^{MC\ tr} - FD_c^l A_{k+i}^{MC\ tr}\right\|_F^2 \\
&\qquad + \sum\nolimits_j \lambda(\left\|\alpha_{k\ j}^{tr}\right\|_1 + \sum_{i=-M,i\neq 0}^{N}\left\|\alpha_{k+i\ j}^{MC\ tr}\right\|_1) + \sum_{i=-M,i\neq 0}^{N}\gamma_i\left\|D_c^h A_k^{tr} - D_c^h A_{k+i}^{MC\ tr}\right\|_F^2 \\
&\left\|D_c^h(:,j)\right\|_2 \leq 1, \left\|D_c^l(:,j)\right\|_2 \leq 1,
\end{aligned} \tag{7}$$

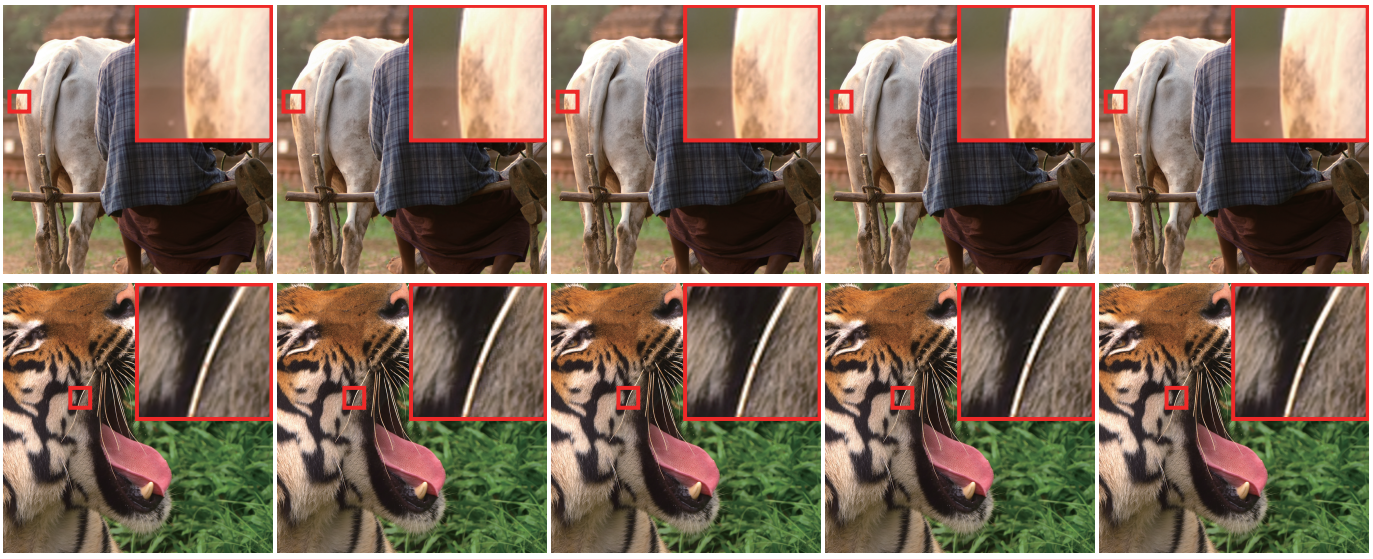

Fig. 2. Visual Comparison of SR results on scene 25 and scene 33. Our proposed algorithm can generate natural-looking frames without noticeable visual artifacts. Because the testing frames have high resolution, results are better viewed in zoomed PDF.

ample used in this paper, resulting in an upscale factor of 2. 50 scenes are used for training and 6 for testing. In the training phase of these experiments, 800,000 patches sets are sampled from the center frame and the motion compensated neighboring frames for training the dictionary from videos, while the same 800,000 patches in center frames are used for training the dictionary from images. The patch size is 5×5 and no feature filter is applied to the LR patches. The regularization were chosen experimentally to be $\lambda = 2$, $\gamma_i = 0.003$. Every dictionary for the SDSF and SDMF methods has 512 atoms and the dictionary for the MDSF and MDMF has 8 sub-dictionaries with 512 atoms each. For the multiple-dictionary in the testing phase, we solve the LASSO problem with only one sub-dictionary, therefore the comparison is fair.

In the testing phase, 6 consecutive video frames are super-resolved by each method. For those multiple-frame SR methods, the current LR frame, together with one LR backward and one LR forward frames (i.e., $M = N = 1$), are utilized to estimate the current HR frame. We tested a number of optical flow estimation algorithms [19]. Based on their comparison we are using the method in [20] in all reported experiments. Results comparing these methods are shown in Table 1.

Table 1 shows that DL based multiple-frame SR methods (SDMF, MDMF) outperform single frame SR (Bilevel, MDSF). We can also see that multiple-dictionary SR methods improve over single-dictionary SR methods. Finally, the proposed approach (MDMFVT) provides the best SR performance compared to all other methods, including the state-of-the-art Bayesian [2] and Enhancer [17], proving the effectiveness of the proposed video training algorithm.

Scene	#2	#8	#18	#25	#33	#37
Bicubic	45.32	38.18	41.43	44.40	40.22	48.39
Bilevel [14, 15]	46.17	39.94	43.04	46.69	42.95	49.31
SDMF	46.87	40.08	43.41	47.52	43.08	50.91
MDSF	46.84	40.34	43.37	47.37	43.27	50.05
MDMF	47.72	40.59	43.82	48.45	43.68	52.54
Enhancer [17]	46.01	40.26	43.66	46.21	43.00	50.21
Bayesian [2]	46.49	39.94	43.19	46.08	42.51	49.71
Proposed	**48.14**	**40.98**	**44.32**	**49.19**	**44.49**	**52.84**

Table 1. Experimental result comparing different method in PSNR.

From the visual comparison results in Figure 2, we notice that all these SR methods produce sharper images than bicubic interpolation, however artifacts are introduced. Next we notice that our proposed method has fewer artifacts and shaper edges compared to all other methods.

5. CONCLUSIONS

In this paper we presented a novel video SR framework based on dictionary learning and motion estimation by optical flow. According to it, the HR patches are estimated from multiple corresponding LR patches in multiple frames, making the dictionary-based reconstruction algorithm more accurate. The dictionary training algorithm that utilizes multiple frames of the training videos further improved the SR performance by making the training and testing phases consistent. We performed experiments with 4K videos and showed that our method outperforms the state-of-the-art algorithms.

6. REFERENCES

[1] RY Tsai and Thomas S Huang, "Multiframe image restoration and registration," *Advances in computer vision and Image Processing*, vol. 1, no. 2, pp. 317–339, 1984.

[2] Ce Liu and Deqing Sun, "A bayesian approach to adaptive video super resolution," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 209–216.

[3] C Andrew Segall, Rafael Molina, and Aggelos K Katsaggelos, "High-resolution images from low-resolution compressed video," *Signal Processing Magazine, IEEE*, vol. 20, no. 3, pp. 37–48, 2003.

[4] Stefanos P Belekos, Nikolas P Galatsanos, and Aggelos K Katsaggelos, "Maximum a posteriori video super-resolution using a new multichannel image prior," *Image Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1451–1464, 2010.

[5] Qi Shan, Zhaorong Li, Jiaya Jia, and Chi-Keung Tang, "Fast image/video upsampling," in *ACM Transactions on Graphics (TOG)*. ACM, 2008, vol. 27, p. 153.

[6] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma, "Image super-resolution via sparse representation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861–2873, 2010.

[7] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma, "Image super-resolution as sparse representation of raw image patches," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[8] Byung Cheol Song, Shin-Cheol Jeong, and Yanglim Choi, "Video super-resolution algorithm using bi-directional overlapped block motion compensation and on-the-fly dictionary training," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 3, pp. 274–285, 2011.

[9] Shenlong Wang, D Zhang, Yan Liang, and Quan Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2216–2223.

[10] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*, pp. 711–730. Springer, 2012.

[11] Ryo Nakagaki and Aggelos K Katsaggelos, "A vq-based blind image restoration algorithm," *Image Processing, IEEE Transactions on*, vol. 12, no. 9, pp. 1044–1053, 2003.

[12] Matan Protter and Michael Elad, "Image sequence denoising via sparse and redundant representations," *Image Processing, IEEE Transactions on*, vol. 18, no. 1, pp. 27–35, 2009.

[13] Michael Elad and Michal Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *Image Processing, IEEE Transactions on*, vol. 15, no. 12, pp. 3736–3745, 2006.

[14] Jianchao Yang, Zhaowen Wang, Zhe Lin, Xianbiao Shu, and Thomas Huang, "Bilevel sparse coding for coupled feature spaces," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2360–2367.

[15] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang, "Coupled dictionary training for image super-resolution," *Image Processing, IEEE Transactions on*, vol. 21, no. 8, pp. 3467–3478, 2012.

[16] Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3501–3508.

[17] Infognition, "Video enhancer," 2010, http://www.infognition.com/videoenhancer/.

[18] Harmonic Inc, "Harmonic 4k footage," 2014, http://www.harmonicinc.com/resources/videos/4k-video-clip-center.

[19] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011.

[20] Marius Drulea and Sergiu Nedevschi, "Total variation regularization of local-global optical flow," in *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*. IEEE, 2011, pp. 318–323.