FACE RECOVERY IN CONFERENCE VIDEO STREAMING USING ROBUST PRINCIPAL COMPONENT ANALYSIS

Wai-tian Tan [o], *Gene Cheung* [#], *Yi Ma* [*]

[o] Hewlett-Packard Laboratories, [#] National Institute of Informatics,
[*] University of Illinois at Urbana-Champaign

ABSTRACT

Irrecoverable data loss is inevitable for low-delay video conferencing over typical loss-prone networks such as the Internet. A semi-super-resolution (SSR) framework has been previously proposed to supply an additional low-resolution (LR) thumbnail to aid error concealment when the high-resolution (HR) image is lost. Super-resolution is an ill-posed problem, however, and previous block-search based SSR methods tend to produce discontinuities in output images, which can be objectionable, especially in human faces where the focus of a viewer usually lies. In this paper, we propose to recover a human face in a lost frame using the same SSR framework, but by operating on the entire face at a time. We leverage a recent work on robust principal component analysis (RPCA), where the "salient" features (human face in our scenario) in a sequence of previous HR frames can be recovered despite the presence of gross but sparse errors. We propose and derive various improved methods to solve the SSR problem using RPCA. Beyond robust recovery of the human face, transformations of the face in previous HR frames are also deduced, so that the recovered face can be appropriately transformed in the lost frame for natural viewing. Experimental results show that our face-based approach gives much improved face recovery compared to previous SSR block searches.

Index Terms— Video streaming, super-resolution, sparse representation

1. INTRODUCTION

Controlling the effects of packet losses in low-delay applications like video conferencing over the Internet is a challenging task. General data recovery methods such as retransmissions and forward error corrections (FEC) are known to be ineffective. In particular, the practical number of retransmissions is limited, especially with large round-trip delay. Possible burst of losses, on the other hand, suggest the need of additional delay for interleaving to efficiently implement FEC.

An alternative resilience approach is to improve loss resiliency of the compressed bitstream itself (at minimal bit overhead), so that usable video information can be salvaged with irrecoverable losses. One such proposal is the *semi-super-resolution* (SSR) framework [1], where a low-resolution (LR) thumbnail of a high-resolution (HR) video image is additionally transmitted, so that in the event that the HR image is lost, the LR thumbnail can be super-resolved for concealment. Super-resolution is an ill-posed problem, however, and previous SSR methods use motion search to find the most similar blocks from previous HR frames to match each up-sampled thumbnail patch. Because block selection decisions are made independently patch-by-patch, the resulting image may suffer from annoying blocking artifacts across block boundaries. This is particular true for human faces in video conference setting, where translational motion assumed in block motion search do not generally hold true, resulting in objectionable visual quality.

In this paper, we present an alternative method to recover a human face in a lost frame in the SSR framework by operating on the entire face. We leverage on a recent work called *robust principal component analysis* (RPCA) [2, 3] where "salient" features (principal components) of a sequence of frames can be recovered, even in the presence of gross but sparse errors. In particular, we formulate and propose several solutions to the SSR problem that exploit RPCA to robustly recover a human face and avoid over-reliance on block search. Beyond robust recovery of the human face, transformations of the face in pervious HR frames are also deduced, so that the recovered face can be appropriately transformed in the lost frame to its proper form for natural viewing. Experimental results show that our face-based approach gives much improved face recovery compared to previous SSR block searches.

The outline of the paper is as follows. We first briefly overview related work in Section 2. We then formulate our problem and outline our solutions in Section 3. Finally, we present experimental results and conclusion in Section 4 and 5, respectively.

2. RELATED WORK

Super-resolution (SR)—recovery of a high-resolution (HR) image from one or more low-resolution (LR) images—has been a much studied problem in signal processing [4] and computer vision communities [5, 6]. Recently, recovery of a HR image from a *mixed-resolution* video, where LR current frame and HR previous frames are available for the construction of a HR current frame, has also been studied [7, 1] in the context of loss-resilient video streaming. In [7, 1], motion compensation at decoder is employed to find the most similar blocks from previous HR frames on a patch-by-patch basis to recover high-frequency content lost during down-sampling of current frame. As discussed earlier, this may result in blocking artifacts that can become objectionable, especially for facial regions, and for lower resolution video.

Our current work is an extension of the RECAP framework [1], but instead of relying on independent block searches, we perform loss recovery in the entire face to minimize blocking artifacts. The added complexity is justified by the relative visual importance of the face, and the selective application only to the face.

Our approach is related to various techniques of using eigenfaces [8], that seek low rank representation of well-aligned faces. To obtain the necessary alignment and to handle partial occlusion (leading to gross noise), we rely instead on robust principal component analysis (RPCA) [2] and its variations [3]. It has been shown [2] that RPCA is a more reliable alternative to PCA, and can be robustly solved to find salient components in large datasets, even in

the face of gross but sparse errors in a fair portion of the data. Fast algorithms [9, 10] that solve the RPCA problem in a computationally efficient manner and scale to large data size (compared to other convex optimization algorithms such as interior point methods [11]) show promise even for real-time applications, such as our conference video streaming scenario. We detail our particular usage of these RPCA tools in the following sections.

3. PROBLEM FORMULATION

Our goal is recovery of human face in a lost high-resolution image $\mathbf{x}_n$ given a surviving LR image (or thumbnail) $\mathbf{y}_n$ and other previous high-resolution frames, as illustrated in Fig. 1 using the last W HR frames. We call this the *semi-super-resolution* problem (SSR). Robust transmission of the thumbnail can be realized using, e.g., source coding methods in [1].

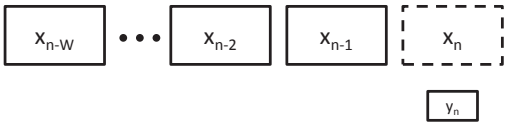


Fig. 1. We seek to combine current LR frame and multiple past HR frames to reconstruct a human face in a lost HR frame.

3.1. Super-Resolution Formulation

We represent the face sub-image and its corresponding thumbnail as column-stacked vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. Without compression, they are related by linear low-pass filter L and down-sampler D by $\mathbf{y} = DL\mathbf{x}$. When $\mathbf{x}$ and $\mathbf{y}$ represent compressed images, a zero mean noise term $\mathbf{n}$ is needed to model quantization noise in lossy compression[1]:

$$\mathbf{y} = DL\mathbf{x} + \mathbf{n} \tag{1}$$

One reasonable choice for reconstructing the lost $\mathbf{x}$ from received thumbnail $\mathbf{y}$ is:

$$\tilde{\mathbf{x}}^{(1)} = \arg\min_{\tilde{\mathbf{x}}} \|DL\tilde{\mathbf{x}} - \mathbf{y}\|_2 \tag{2}$$

Since both D and L are many-to-one functions, the problem is generally ill-posed and $\tilde{\mathbf{x}}^{(1)}$ is not unique. Indeed, as noted by [6] and others, there are in general many $\tilde{\mathbf{x}}$'s that induce a small error term. To reduce the uncertainty and select an $\tilde{\mathbf{x}}$ with other desirable properties, various image priors are incorporated in various ways to regularize the problem.

One example is the addition of a regularization term $J(\tilde{\mathbf{x}})$ to the objective function, where one instead solves:

$$\tilde{\mathbf{x}}^{(2)} = \arg\min_{\tilde{\mathbf{x}}} \left[\|DL\tilde{\mathbf{x}} - \mathbf{y}\|_2 + \lambda J(\tilde{\mathbf{x}})\right] \tag{3}$$

where λ is a tunable parameter to trade off the importance between the HR-LR image error term and the introduced prior. One popular choice for J is *total variations* (TV) [12], where piece-wise smooth signals are preferred and searched. However, fine image details are difficult to recover using TV prior. Another choice for J is a *sparsity* term that counts the number of non-zero coefficients of $\tilde{\mathbf{x}}$ when projected to an over-complete dictionary Ψ [6]. We do not adopt this approach since training a sufficiently descriptive dictionary for our real-time conferencing application is non-trivial. Furthermore, optimizing sparsity (l_0-norm) in general is non-convex and difficult to solve. Approximations have been obtained that admit solution using convex optimization techniques [13]—e.g., iterative re-weighted l_1-norm where each iteration is solvable using standard linear programming algorithms such as interior point method—that are nevertheless computationally expensive.

Instead of adding a term to the cost function, another approach is to explicitly restrict the construction of $\tilde{\mathbf{x}}$, e.g., as a linear combination of atoms in some dictionary Ψ:

$$\tilde{\mathbf{x}}^{(3)} = \arg\min_{\alpha} \|DL\Psi\alpha - \mathbf{y}\|_2 \tag{4}$$

Notice that unlike (3), there is no need to determine the optimal value of the weighting parameter λ, typically found by sweeping it from 0 to ∞ [6]. Thus (4) is potentially simpler to solve computationally than (3). Determining a sufficiently descriptive Ψ is still challenging in general, however. We discuss how we use (4) in our targeted application next.

3.2. Low-Ranked Matrix Prior for SSR

Unlike many super-resolution problems, we have available to the SSR algorithm W previous HR video frames. One approach to construct dictionary Ψ, in order to employ (4), is to seek properly aligned and scaled face sub-images in earlier HR frames. We call this scheme *Linear Aligned Image* (LAI):

$$\Psi^{LAI} = [\mathbf{x}_{n-1}|\ldots|\mathbf{x}_{n-W}].$$

Since L and D in (4) are linear, we have:

$$DL\Psi^{LAI}\alpha = [DL\mathbf{x}_{n-1}|\ldots|DL\mathbf{x}_{n-W}]\alpha,$$

and any optimal α^{LAI} to (4) satisfies:

$$DL\Psi\alpha^{LAI} = Projection(DL\Psi^{LAI}, \mathbf{y}). \tag{5}$$

Since Ψ^{LAI} is of high resolution, it is generally of full rank, and so is $DL\Psi^{LAI}$. Thus, α^{LAI} can be solved readily and the desired reconstruction can be obtained as:

$$\tilde{\mathbf{x}}_n^{LAI} = \Psi\alpha^{LAI} \tag{6}$$

Face images in previous W HR frames are in general not properly aligned or scaled, however. To solve the misalignment problem, given W unaligned image vectors $\mathbf{x}'_i$, we perform *Robust Alignment by Sparse and Low-rank Decomposition* (RASL) [3], an extension of RPCA that incorporates transformations of the observed data into the component analysis, to obtain the desired aligned image vectors $\mathbf{x}_i$. Specifically:

$$\Psi^{LAI} = [\mathbf{x}'_{n-1}|\ldots|\mathbf{x}'_{n-W}] \circ \tau = \mathbf{A} + \mathbf{E} \tag{7}$$

where the transformation τ is determined so that the aligned face sub-images Ψ^{LAI} can be written as a sum of a low-rank matrix $\mathbf{A}$, with sparse error matrix $\mathbf{E}$. Intuitively, $\mathbf{A}$ would contain salient features of the faces, and $\mathbf{E}$ would contain various error, e.g., due to occlusion.

An alternative to LAI is to introduce an additional sparsity prior on α in (4). Specifically, in the *Sparse Linear Image* (SLI) scheme, we use $\mathbf{A}$ from (7) in place of the aligned image data to obtain:

$$\Psi^{SLI} = \mathbf{A}. \tag{8}$$

[1]Though we develop our image model in the SSR context, super-resolution work such as [5, 6] arrive at the same or almost the same image model in their development. Hence we can present here a consistent discussion on formulation for both SSR and super-resolution.

With linearity of L and D, an optimal α^{SLI} to (4) again satisfies:

$$DL\mathbf{A}\alpha^{SLI} = Projection(DL\mathbf{A}, \mathbf{y}). \quad (9)$$

Since $\mathbf{A}$ is low-rank, $DL\mathbf{A}$ is also low-rank, which precludes solving of α^{SLI} uniquely. Instead, we solve:

$$\alpha^{SLI} = (DL\mathbf{A})^{-1} Projection(DL\mathbf{A}, \mathbf{y}),$$

where the -1 in $(DL\mathbf{A})^{-1}$ indicates pseudo-inverse. This choice has the property that $\|\alpha^{SLI}\|_2$ is minimized.

Both LAI and SLI schemes perform projection of thumbnail image $\mathbf{y}$ onto a subspace spanned by either the images themselves (LAI) or their low-rank derivatives (SLI). They are both governed by the linear constraints of (4). One nonlinear approach to significantly reduce $\|DL\tilde{\mathbf{x}} - \mathbf{y}\|_2$ is to replace the "low-pass" portion of an estimate $\tilde{\mathbf{X}}$ by the upsampled thumbnail, as performed in [7] for bi-directionally predicted images. We call this procedure *Low-Pass Replace* (LPR):

$$\tilde{\mathbf{x}}_n^{LPR} = HighPass(\tilde{\mathbf{x}}_n) + Upsample(\mathbf{y}_n) \quad (10)$$

Finally, note that for the sake of our real-time application, our presented methods are composed of simple linear operations, with the exception of RASL, which has already showed promise to be implementable in real-time [10] as discussed in Section 2.

4. EXPERIMENTATION

We use MPEG test sequence *Silent* and *Sean* at CIF resolution and 30 frames per second for comparing various block search approaches to the various methods introduced in Section 3.2. Fig. 2 shows the head portions of frames 43 to 62 of *Silent* sequence, which we assume have been correctly received by a decoder. The pictures show occasional but minor occlusion, and also some mild head movement not uncommon in everyday conversations. Figure 3 shows the concealment using various methods when a later high resolution frame is lost. Our goal is to perform the best possible reconstruction of Fig. 3-(a) that is lost, by using a received LR image in Fig. 3-(b), and the various past images in Fig. 2. The LR is of 1/4 the linear dimension of the HR image for low overheads.

Straight forward methods at concealment include upsampling of received LR image shown in Fig. 3-(b) with bicubic interpolation, which is apparently not visually acceptable. Using a last correct picture provides a visually pleasing, though untrueful reconstruction. One method to exploit the many available past HR frames and the LR frame is to perform block search on blocks in the LR frame and those of the LR frames. Fig. 3-(d) shows the best *independent* block match between upscaled LR picture and the 20 HR pictures using 8×8 blocks, with search range of 12, and quarter-pel resolution. While an effective distortion reduction method, the resulting transformation may results in artifacts such as misalignments and discontinuities. For video pictures at high-definitions, such artifacts tend to be small compared to the facial features and is typically less objectionable. For lower resolution sequence such as CIF *Silent*, the resulting artifacts of Fig. 3-(d) can be annoying. Fig. 3-(e) shows corresponding results for 16×16 blocks. With larger block sizes, the annoying discontinuities are still present, though reduced, and the resulting image bears closer resemblance to Fig. 3-(c) than (a). Fig. 3-(f) shows a hybrid approach where 16×16 blocks are employed as in (e), but adaptive selection is performed on a per block basis to determine if it is preferable to upsample from (b) or perform motion copy. Even with ideal selection using actual MSE difference

Fig. 2. Face portion of frames 43 to 62 (arranged left to right, then top down) of *Silent* sequence showing different head movement and occlusion.

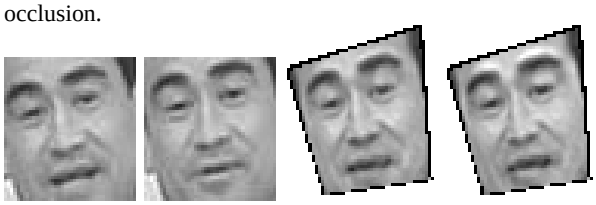
(a) Lost HR (b) received LR (c) last correct (d) 8×8 block

(e) 16×16 block (f) hybrid search (g) LAI

(h) SLI (i) LAI+LPR (j) SLI+LPR

Fig. 3. Various block search methods provides high PSNR but can yield undesirable images. Instead, treating an entire face as a unit tend to produce much more pleasing results.

Fig. 4. Face portion of frames 40 to 55 (arranged left to right, then top down) of *Sean* sequence showing different head movement and occlusion.



(a) Lost HR (b) last correct (c) SLI (d) SLI+LPR

Fig. 5. More faithful reconstruction is achieved by our method, e.g., in the mouth.

between target in Fig. 3-(a) and the candidates, the result in (f) is still not visually pleasant.

Fig. 3-(g) and (h) show, respectively, the results obtained by our LAI and SLI methods in Section 3.2. We notice that the head and eye are in better agreement with our desired target than (c) or any of the earlier frames. The visual quality of LAI and SLI is comparable in terms of having meaningful details, with a slight edge for SLI. We can see a leftover "finger" in the lower right cheek of Fig. 3-(g) that is not present in (h). As we can infer from Fig. 2, this is due to partial occlusion in that position for the first 10 frames by a thumb.

Fig. 3-(i) and (j) show the corresponding results when Low-Pass Replace is applied in addition to LAI and SLI, respectively. We note that further application of LPR further improves images from Fig. 3-(g) and (h) in having a closer pose to the target. This is expected, as the face pose in the thumbnail is unambiguous, and different from the surviving pictures in Fig. 2 even after alignment.

The results for *Sean* is shown in Fig. 5, where a lost HR frame 57 is reconstructed using a thumbnail at 1/4 linear spatial resolution and HR previous frames 40 to 55 shown in Fig. 4. The nearest surviving HR frame is 55 in Fig. 5-(b), which has similar pose to the target except with minor changes in facial expression. Block search results show similar artifacts as before and is omitted for space reasons. The results obtained by our methods SLI and SLI+LPR are shown in Fig. 5-(c) and (d), respectively. We see that our approach has succeeded in producing more faithful reproduction of the target facial expression. Nevertheless, unlike *Silent*, the images produced by SLI and SLI+LPR are similar for *Sean* due to presence of similar facial expression in the surviving images.

For both sequences, there are times when the translation model of block search yields acceptable results. Nevertheless, our more general method is preferred since concealment errors in one frame can propagate into many future frames.

5. CONCLUSION

In this paper, we present and compare a number of semi-super-resolution methods to recover the moving face during a video conference. We develop two methods, LAI, which only requires proper alignment past face sub-image, and a slightly more reliable method, SLI, which exploits earlier RASL [3] work to obtain low rank dictionary. For both methods, we show that further improvement can be obtained by replacing the low-pass component of the reconstruction by the surviving thumbnail.

6. REFERENCES

[1] C. Yeo, W. t. Tan, and D. Mukherjee, "Receiver error concealment using acknowledge preview (RECAP)—an approach to resilient video streaming," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, April 2009.

[2] E. Candés, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," in *Preprint*, December 2009.

[3] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," in *(submitted to) IEEE Transactions on Patern Analysis and Machine Intelligence (PAMI)*, July 2010.

[4] C. A. Segall, R. Molina, and A. K. Katsaggelos, "High-resolution images from low-resolution compressed video," in *IEEE Signal Processing Magazine*, May 2003, pp. 37–48.

[5] J. Yang, J. Wright, Y. Ma, and T. Huang, "Image super-resolution as sparse representation of raw image patches," in *IEEE International Conference on Computer Vision and Pattern Recognition*, Anchorage, AL, June 2008.

[6] W. Dong, G. Shi, L. Zhang, and X. Wu, "Super-resolution with nonlocal regularized sparse representation," in *SPIE Visual Communications and Image Processing*, Huang Shan, China, July 2010.

[7] F. Brandi, R. de Queiros, and D. Mukherjee, "Super-resolution of video using key frames and motion estimation," in *IEEE International Conference on Image Processing*, San Diego, CA, October 2008.

[8] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, 1991.

[9] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," Tech. Rep. UILU-ENG-09-2215, University of Illinois at Urbana-Champaign, October 2009.

[10] C. Qiu and N. Vaswani, "Real-time robust principal components' pursuit," in *Allerton Conference on Communication, Control and Computing*, Monticello, IL, September 2010.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2009.

[12] S. Fariu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super-resolution," in *IEEE Transactions on Image Processing*, January 2006, vol. 15, no.1, pp. 141–159.

[13] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l_1 minimization," in *The Journal of Fourier Analysis and Applications*, December 2008, vol. 14, no.5, pp. 877–905.