A ROBUST MOTION ESTIMATION METHOD USING WARPING FOR VIDEO FRAME ENLARGEMENT

Ying Chen, Mark J. T. Smith

Purdue University
West Lafayette, IN 47907
cheny@purdue.edu

ABSTRACT

This paper addresses the challenge of high quality enlargement of coded video frames. The proposed method uses a combination of hierarchical processing, forward warping and backward warping to adaptively determine the high definition output pixels. Experimental results show that the proposed method can achieve an average of 2-4dB PSNR improvement over conventional frame interpolation methods while preserving robust performance in the face of complex motion.

Index Terms— Interpolation, Warping, Robustness, Motion Estimation, Uncovered Background

1. INTRODUCTION

Motion compensated predictive (MCP) video coding algorithms, like H.264, have been well engineered over many years by the international community and can achieve a very high level of performance [1]. These algorithms are therefore very attractive for use in many applications. One such application in particular is compression and storage of lecture videos. With the rapidly growing interest in digitally recorded course work and distance learning and the interest in being able to display lecture video at a variety of spatial resolutions, both high quality and efficient spatial scalability become important. Algorithms like H.264 are ideal from a compression perspective but are not engineered for spatial enlargement. Frame enlargements are typically handled by interpolation methods [2].

Recently, an H.264-based framework was considered for lecture video where a high spatial resolution reference frame of video was retained at the receiver and H.264 was used to code the sequence at low spatial resolution [3, 4]. Periodically, a new reference frame is sent to accommodate scene changes or significant drift in image content. The authors introduced a method for displaying the low resolution coded video at high spatial resolution by warping [5] the reference frame into the spatially interpolated video. This approach effectively allowed high definition spatial information to be retained in the decoding/scaling process while at the same time avoiding having to make any changes to the H.264 algorithm. The authors showed that the video could be enlarged with high spatial quality and that the performance was significantly better than conventional interpolation methods typically employed to enlarge frames.

While this approach has a number of attractive features, the authors observed difficulty when moderate to high motion was present. Errors in the (reconstruction side) motion estimation resulted in unnatural distortions. This problem was mitigated to some extent by the lecture video scenario, which inherently has lower motion than many sequences, and also by the relatively neutral background that often accompanies lecture videos. The inherent issue is that the warping method in [3] is not able to represent uncovered background effectively. When the amount of uncovered background is very small, distortions often go unnoticed. But for high motion and for general sequences with spatially rich backgrounds, the resulting spatial distortions (which occur in high definition) are highly visible.

In this paper, we attempt to address this robustness issue by improving the warping strategy. The new algorithm maintains the property of high spatial definition of the enlarged video frames but is shown to avoid the kind of spatial distortions that surfaced previously under high and complex motion conditions. Moreover, the new algorithm is not restricted to lecture video scenarios, although lecture video is still our motivating application.

This paper is organized as follows. In Section 2, a brief overview of Adaptive Control Grid Interpolation is presented, which serves as background for the subsequent discussions in Section 3, where the new enlargement algorithm is proposed. Experimental results are presented in Section 4, followed by a summary and conclusions in Section 5.

2. ADAPTIVE CONTROL GRID INTERPOLATION

Adaptive Control Grid Interpolation (ACGI) was explored and developed by Monaco et al. [6] for image morphing. The method that was proposed and futher developed by Frakes et al.[5] is a block-based model that uses an optical flow equation within the blocks. The motion representation has

sufficient degrees of freedom to handle complex motion and provides a reasonably compact motion vector representation.

The general method operates on frame pairs (e.g. X_1 and X_2), like convention MCP methods. If X_1 and X_2 are the source frame and target frame respectively, the goal is to warp X_1 into X_2.

Let $X_1[i,j]$ denote the intensity of the pixels in frame X_1 at position $[i,j]$. We assume in this model that pixel intensities remain fixed but move from frame to frame, where the movement can be represented by displacement vectors. So at each position,

$$X_2[i,j] = X_1[i + d_1[i,j], j + d_2[i,j]] \tag{1}$$

where $d_1[i,j]$ is the horizontal motion displacement component at position $[i,j]$ and $d_2[i,j]$ is the vertical motion displacement component. Those two components can be characterized as a linear combination of 4 components, which are independent basis functions in $\theta[i,j]$

$$\begin{aligned} d_1[i,j] &= \alpha_1\theta_1[i,j] + \alpha_2\theta_2[i,j] + \alpha_3\theta_3[i,j] + \alpha_4\theta_4[i,j] \\ &= \alpha^T\theta[i,j] \\ d_2[i,j] &= \beta_1\theta_1[i,j] + \beta_2\theta_2[i,j] + \beta_3\theta_3[i,j] + \beta_4\theta_4[i,j] \\ &= \beta^T\theta[i,j]. \end{aligned} \tag{2}$$

The displacement vectors define the particular motion model, which leads to a number of models that one can consider, such as affine, bilinear, perspective, and so on. In our model the basis functions are associated with a block region R defined by its four corners. Each block then has associated with it a total of 8 parameters to estimate. Frame X_2 is also partitioned into blocks, denoted R. Within each block the mean square error equation is used

$$\sum_{i,j \in R} (X_2[i,j] - X_1[i + \alpha^T\theta[i,j], j + \beta^T\theta[i,j]])^2 \tag{3}$$

to estimate the parameters α and β so that the 8 parameters minimize the squared error associated with the block. To simplify equation (3), we can approximate it using a Taylor series expansion

$$\begin{aligned} \sum_{i,j \in R} (X_2[i,j] - X_1[i,j] - \frac{\partial X_1[i,j]}{\partial i}\alpha^T\theta[i,j] \\ - \frac{\partial X_1[i,j]}{\partial j}\beta^T\theta[i,j])^2. \end{aligned} \tag{4}$$

The optimization process uses an iterative gradient-based method in a quad tree framework. If the error in a block is above a predetermined threshold, we split the block into 4 smaller blocks. This quad tree splitting is repeated until the mean square error threshold criterion is satisfied.

3. ROBUST MOTION VECTOR DETERMINATION

In the warping enlargement method, when the motion in the sequence is relatively low, the resultant quality of the enlarged video appears very sharp and is close to the quality of the original video sequences[3]. As we progress in the sequence and subsequent frames deviate more and more from the first reference frame, the likelihood increases of incurring distortion. This distortion manifests itself as unnatural warping of contours and edge boundaries, and is illustrated in the example shown in Figure 1(a). The figure is a warped frame taken from the *Stefan* sequence. The distortion around the left arm is quite visible. As the deviations between the current and reference frame become larger, the algorithm is challenged to find accurate displaced pixels and thus the motion vectors can be unreliable until the next reference frame is encountered.

To address the robustness of the motion vectors, we exploit the presence of the updated reference frames. More specifically, forward warping is applied between the first reference frame and the current frame. Then backward warping is applied between the current frame and the second reference frame. This provides two high definition warped versions of the current frame. Because uncovered background in the forward motion estimation process is typically available in the backward motion estimation process and vice versa, using both high definition images together can dramatically reduce warping errors. In some cases, both forward and backward warped images may contain pixels in the same region that are corrupted. Thus, we consider a third choice, that of the interpolated low resolution image.

Now with three enlarged images, all of the same frame, the task is to choose on a pixel-by-pixel basis the best from the three enlarged candidate frames. How to do this is an issue to address, considering that there is an inherent ambiguity associated with using a mean square error(MSE) approach. To elaborate on this point, consider that we compute the squared error between the warped high resolution frame and the interpolated frame over each block. One might expect that if the mean square error associated with that region were high, then warping distortion had occurred. However, improvement in sharpness derived from the warping (when done correctly) will show up as error energy in the MSE. Consequently it may often be difficult to tell from the MSE if the source of the error was from warping distortion or from low-frequency-high-frequency spatial differences.

To resolve this ambiguity, we perform the MSE calculations in the downsampled domain. In this domain, we don't have differences in sharpness contributing to the MSE. Thus, the proposed method is first to downsample the warped frame Z which is $2M \times 2N$, so that it has the same resolution as the low resolution frames ($M \times N$).

In the low resolution domain, we compare the pixel values from the downsampled warped frame Z_d and the one from the low resolution video frame X. If the pixel value difference

between Z_d and X is large, we assume that the corresponding 4 pixels in the original warped frame are not correct.

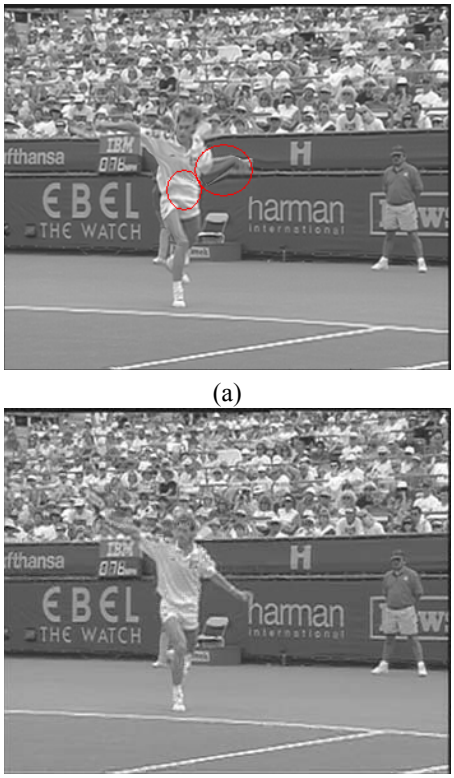
This idea is applied to both forward- and backward-warped images, given that every Nth frame we will have a high definition reference frame. For each frame between any two reference frames, we first bilinearly interpolate the low resolution frame to high resolution, the result of which we denote X_I. Then we use the preceding reference frame and warp it into X_I (forward warping), which we denote Z_F. Similarly, we use the succeeding reference frame and warp it to X_I (backward warping), resulting in Z_B. Generally speaking, when the current frame is temporally close to the preceding reference frame, the forward warped frame is most accurate. Likewise, when the current frame is close to the succeeding reference frame, the backward warped frame is better. After performing the block MSE calculations, we select Z_B or Z_F depending on which has the lower MSE. In the event that the block MSEs for both Z_B and Z_F are above threshold, we use the block derived from bilinearly interpolating X. While it is true that this block will lack the desired high frequency detail, the infrequency of our choosing the interpolated case combined with the blocks being relatively small leads to an overall enlarged image that appears sharp and void of distortion.

4. EXPERIMENTAL RESULTS

We tested three types of video sequence. One set represents the motion expected in the "talking head" lecture video case. The sequences we used to represent this case are Akiyo, Salesman, Foreman, Carphone, and News. Each has low to moderate motion. Another class of sequences we tested consists of video with high spatial detail, such as the Flower, Tempete, and Bus sequences. The last sequences we considered are those with high motion. Here we included the sequences Stefan and Table.

For each test sequence, every Nth frame was designated as a high definition reference frame. For the frames in between we first low pass filter and downsample them to get the low resolution frame. The downsample factor in this case was 2. The low resolution frames and reference frames are used as the input. The warping and pixel selection process describe in the preceding section was then applied. The resulting enlarged output sequence was then compared to the original frames which serve as ground truth and from which we compute the PSNR. For each test sequence, 50 frames were considered in the comparison.

Some caution should be exercised in accepting the PSNR blindly as a measure of quality. The nature of the approach we've taken considers high definition warping features to be perfectly acceptable as long as the warping is not visually objectionable. Geometric features in the warped frames could be displaced by a pixel or two from what appears in the original. This, because it is not perceptible, is acceptable even though the PSNR will be reduced. As a result, we examine both subjective and PSNRs in our comparisons.



(a)

(b)

Fig. 1. Subjective comparison of the 28th frame of the Stefan sequence. (a) Depicts the output of the warping method presented in [3]. (b) Shows the output of the method presented here.

As we can see from Figure $1(a)$, the warped frame overall looks very sharp but in some local areas, like around the left arm, there are geometric distortions. While the image and the distorted regions are all very sharp (compared to an interpolated frame), the nature of the distortion is still objectionable. In contrast, the output of the proposed algorithm (Figure 1(b)) preserves the object geometry and provides a sharp enlargement. To illustrate the visual quality improvement, in Figure 2 we show the original frame, the bilinear interpolated frame and our result for the 28th frame of the *Stefan* sequence. Figure 2(d) illustrates the choice of the motion vectors. The black pixels mean the pixel values were chosen from forward warping, white means they were chosen from the backward warping and gray means they were chosen from the bilinearly interpolated frame.

Choice of the threshold can have an impact on the performance and can allow for a tradeoff between geometric distortion and spatial dispersion (i.e. interpolation blur). The lower the threshold the more biased that algorithm is toward selecting blocks derived from the interpolated frame. And the higher the threshold, the more the bias is directed toward

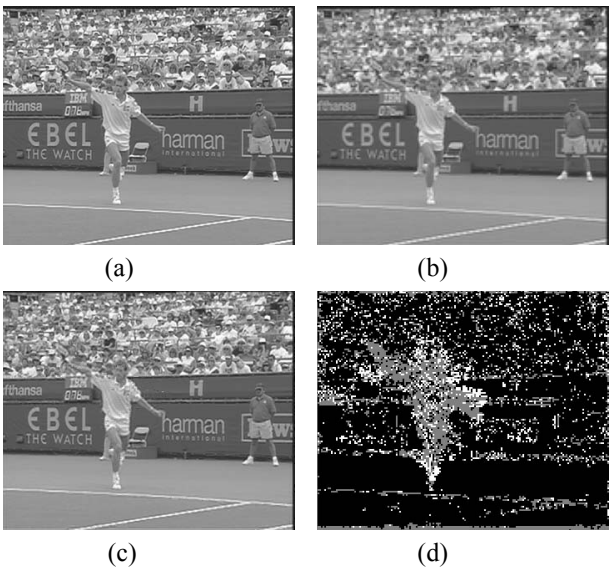


Fig. 2. Frame 28 from the stefan sequence (a) original 28th frame, (b) bilinear interpolated, (c) warped frame with post-processing, (d) indicator of choice of motion vectors.

Table 1. PSNR comparison for talking head video sequences using different methods

	Akiyo	Sales	Foreman	Carphone	News
Bilinear	33.33	29.59	29.64	30.33	28.57
Bicubic	34.07	30.12	30.11	30.76	29.50
Xin	33.51	29.23	28.60	29.87	28.26
New	34.24	33.27	31.11	32.88	33.63

the warping. In cases of a high threshold one can sometimes observe geometric distortion as well as improved PSNR. For the proposed algorithm we have generally observed that the PSNR agrees with our subjective assessments when comparative tests are performed.

For comparative purposes, we present in the tables numerical PSNR assessments for our algorithm along with several competing methods. Shown in the table are the PSNR results for bilinear interpolation, bicubic interpolation, the edge-directional interpolation method of Xin [7], and the algorithm proposed in this paper. Tables 1 and 2 list the resulting PSNR for three sets of video sequences.

As we can see, for all three sets of video sequences, the proposed method outperforms bilinear interpolation, bicubic interpolation and Xin's interpolation method. For the talking head video sequences, the proposed method outperforms the others by an average of 2 dB. For sequences with high details such as Tempete and Flower, the proposed method achieves a $3-4$ dB improvement on average. The reason that the proposed method performs particularly well for video sequences with high detail is that it inherently retains the high frequency information while such information is lost in methods that interpolate from a decimated frame.

Table 2. PSNR comparison for video sequence with high details and with moderate to high motion.

	Sales	Temp	Flower	Bus	Ste	Table
Bilinear	29.59	26.10	22.03	25.01	25.57	29.06
Bicubic	30.12	26.64	22.39	25.64	26.33	29.76
Xin	29.23	25.57	21.41	24.48	24.50	28.76
New	33.27	29.40	26.77	26.04	28.08	30.23

5. SUMMARY AND CONCLUSIONS

Our proposed method using a post processing approach to reduce the effects of motion vectors whose accuracy is unreliable. The method employs the use of a combination of forward warping, backward warping, and a hierarchical motion structure that adaptively selects pixels from the forward warped frames, backward warped frames and the bilinearly interpolated frames. Significant improvement in robust performance is observed compared to the warping approach presented in [3]. The new method achieves robust enlargements even though direct warping breaks down and outperforms the competing methods both subjectively and objectively.

An issue that has yet to be considered is efficient implementation and the application of this approach to low bit rate spatially scalable video coding. These are issues for future research.

6. REFERENCES

[1] Thomas Wiegand et al, "Overview of the h.264/avc video coding standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, July 2003.

[2] Anil K. Jain, "Fundamentals of digital image processing," *New Jersey: Prentice-Hall*, pp. 253–255, 1989.

[3] Y. Chen and M. J. Smith, "High quality spatial interpolation of video frames using an adaptive warping method," *IEEE Digital Signal Processing Workshop*, Sept 2006.

[4] Y. Chen et al., "A low bit-rate video coding approach using modified adaptive warping and long-term spatial memory," *VCIP*, Jan 2007.

[5] D. Frakes et al., "Application of an adaptive control grid interpolation technique to morphological vascular restruction," *IEEE Trans. on B.E.*, vol. 50, Feb 2003.

[6] Joseph Monaco, *Generalized Motion Models for Video Applications*, Ph.D. thesis, Georgia Institute of Technology, 1997.

[7] Xin Li et al., "New edge-directed interpolation," *IEEE Trans. on Image Processing*, vol. 10, Oct 2001.