COMPRESSION USING SELF-SIMILARITY-BASED TEMPORAL SUPER-RESOLUTION FOR FULL-EXPOSURE-TIME VIDEO

Mihoko Shimano [o], Gene Cheung [#], Imari Sato [#]

[o] University of Tokyo, PRESTO, JST, [#] National Institute of Informatics

ABSTRACT

In order to allow sufficient amount of light into the image sensor, videos captured in poor lighting conditions typically have low frame rate and frame exposure time equals to inter-frame period—commonly called full exposure time (FET). FET low-frame-rate videos are common in situations where lighting cannot be improved a priori due to practical (e.g., large physical distance between camera and captured objects) or economical (e.g., long duration of night-time surveillance) reasons. Previous work in computer vision has shown that content at a desired higher frame rate can be recovered (to some extent) from the captured FET video using self-similarity-based temporal super-resolution. From an end-to-end communication standpoint, however, the following practical question remains: what is the most compact representation of the captured FET video at encoder, given that a higher frame rate reconstruction is desired at the decoder? In this paper, we present a compression strategy, where, for a given targeted rate-distortion (RD) tradeoff, FET video frames at appropriate temporal resolutions are selected for encoding using standard H.264 tools at encoder. At the decoder, temporal super-resolution is performed on the decoded frames to synthesize the desired high frame rate video. We formulate the selection of individual FET frames at different temporal resolutions as a shortest path problem to minimize Lagrangian cost of the encoded sequence. Then, we propose a computation-efficient algorithm based on monotonicity in predictor's temporal resolution to find the shortest path. Experiments show that our strategy outperforms an alternative naïve approach of encoding all FET frames as is and performing temporal super-resolution at decoder by up to 1.1dB at the same bitrate.

Index Terms— Video compression, super-resolution, self similarity

1. INTRODUCTION

Appropriate exposure time (and subsequently frame rate) for a captured video is influenced by the lighting condition of the scene of interest. If the lighting condition is poor, then exposure time must be long to permit sufficient amount of light into the photographic film or image sensor, avoiding undesirable underexposure effects. Clearly, exposure time cannot be longer than inter-frame period in a video, hence a longer required exposure time can lead to lower video frame rate, which is usually not desirable.

Quite often, lighting condition cannot be improved a priori due to practical or economical constraints. For example, the scene of interest can be too far from the camera to physically insert light sources before capture. Another example is a night-time surveillance / observation situation, where prolonged illumination in a large area will lead to unacceptably high cost, or disturbance to the captured animals in their nocturnal habitat. Thus, it is often unavoidable to handle captured videos with full exposure time (FET) and lower frame rate than desirable, where the exposure time of each frame equals the inter-frame period of the captured video. See Fig. 1 for an illustration.

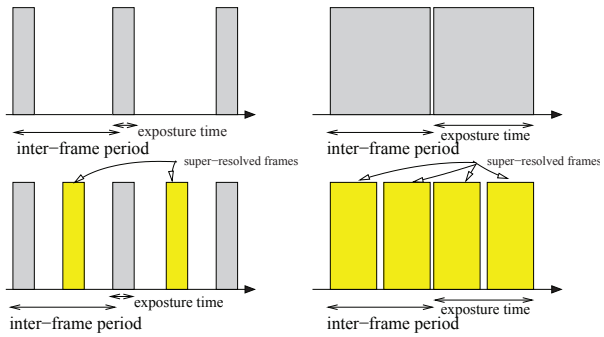

Fig. 1. Illustration of temporal super-resolution for: a) sub-exposure-time video, and b) full-exposure-time (FET) video. Super-resolved frames are colored yellow.

While FET is in general not a desirable property, to synthesize video at a required higher frame rate using FET video as input, previous work in computer vision [1] has proposed *temporal super-resolution* using *self-similarity* (TSR-SS) to exploit the FET property. The key observation is that the same motion blur patterns due to long exposure time often reappear at multiple timescales. By identifying motion blur patches in one FET video frame (using Maximum A Posteriori (MAP) estimate [1]), two corresponding frames at twice the frame rate can be constructed by halving the blurs at appropriate spatial locations. Specifically for FET video, [1] has shown that TSR-SS out-performs previous temporal super-resolution methods such as optical flow [2] by up to 6dB in PSNR of the reconstructed high-frame-rate video.

For a typical end-to-end communication scenario, where encoding is performed at video capture location and decoding is performed at viewer's display, however, the following practical question remains: what is the most compact representation of the captured FET video at encoder, given that a higher frame rate reconstruction is required at decoder? One obvious solution is to perform temporal super-resolution at encoder to up-sample captured FET video to high-frame-rate content prior to compression. However, this means compression must be performed in higher-than-capture frame rate, expending large number of coding bits. An alternative is to encode captured FET video as is at encoder, relying on decoder to perform temporal super-resolution on compressed frames to recover required high-frame-rate content. However, quantization noise due to compression can severely affect super-resolution, harming constructed high-frame-rate video quality.

In this paper, we present instead an adaptive compression strategy for FET video, where FET video frames are selectively encoded *at appropriate temporal resolution* to optimize rate-distortion (RD)

tradeoff. The key idea is to encode at low temporal resolution frames that can be easily synthesized via super-resolution at decoder (to save bits), and to encode at high temporal resolution frames that are difficult to synthesize after quantization (to preserve quality). We formulate the selection of individual FET frames at different temporal resolutions as a shortest path problem in a directed acyclic graph (DAG) to minimize Lagrangian cost of the encoded sequence. We then develop a fast shortest-path search algorithm based on assumption on monotonicity in predictor's temporal resolution. Experiments show that our adaptive strategy outperforms the alternative approach of encoding all FET frames at encoder and performing temporal super-resolution at decoder by up to 1.1dB at the same bitrate.

The outline of the paper is as follows. After a brief discussion on related work in Section 2, we overview TSR-SS in Section 3. We formulate our RD optimization problem in Section 4, and detail our shortest-path algorithm in Section 5. We present results and conclusion in Section 6 and 7, respectively.

2. RELATED WORK

Frame interpolation in time was studied in the context of variable frame rate for low-bitrate video coding [3, 4]. While similar in motivation to our temporal super-resolution problem, such methods typically rely on motion compensation or optical flow analysis [2], which do not perform well for FET video with motion blurs at low frame rates. By comparison, our previously proposed TSR-SS [1] has shown noticeably superior performance for FET video.

Our problem of selecting FET video frames at the "best" temporal resolutions for video encoding is a new entry into a family of RD-optimizing dependent resource allocation problems. It was first studied in the seminal work [5] on bit allocation for dependent frames in motion-compensated video coding. [6] later studied the problem of selecting the appropriate quantization parameters (QP) and skipped frames[1] (to save bits) in video to minimize resulting distortion subject to a rate constraint. Recently, [7] studied the problem of selecting subsets of texture and depth maps of multiview images for differential encoding at appropriate QPs to minimize synthesized view distortions at decoder subject to a total rate constraint. Our current work differs from these previous work in two respects. First, out unique problem setting on FET video coding demands a selection of frames at different temporal resolutions to optimize RD performance, which has not been previously studied. Second, our fast solution search is developed based on monotonicity in predictor's temporal resolution, which is also new.

3. SELF SIMILARITY FOR TEMPORAL SUPER-RESOLUTION

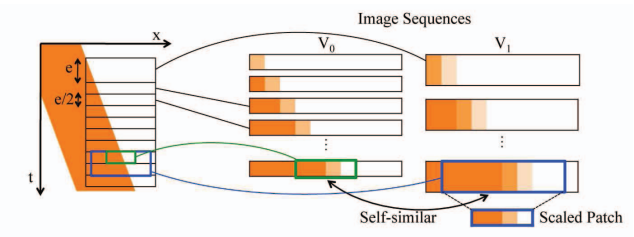

Fig. 2. Self-similarity. A x-t plane for single scan line, and 1-D image sequences with different temporal resolutions.

[1]Skipped frames can be construed as one form of temporal resolution. We focus here instead on optimally selecting temporal resolutions of FET video frames, where, unlike a skipped frame, a lower resolution FET frame is a pixel-by-pixel average of two higher resolution FET frames.

In natural images, self-similar texture patches tend to redundantly recur many times inside the image, both within the same scale as well as across different scales. In the same way, self-similarity exists in the spatio-temporal domain of videos, for example, if objects in a scene follow similar trajectories with constant but different velocities. On the basis of this observation, previously we proposed a method to temporally super-resolved video from a single FET video by exploiting self-similarity, *i.e.*, a self-similar appearance that represents integrated motion of objects during each exposure time of videos with different temporal resolutions [1].

For simplicity, let us consider self-similarity in the case of 1D image sequences of a uniformly moving object. Fig. 2 illustrates two FET image sequences: V_0 has exposure time (also inter-frame period) $e/2$, half of that of V_1, e. Consider, for example, a small 1D image patch of V_0 with exposure time $e/2$. In a patch of V_1 captured with exposure time e at the same position in x-t plane, the same object moves twice the distance. If the spatial size of the patch of V_1 is twice that of the patch of V_0, the patch of V_0 becomes similar to the corresponding patch of V_1 because the object moves twice the distance during the exposure time of V_1. This self-similar relationship can be extended to a 2D-image patch. Utilizing this self-similar relationship between such different temporal resolution frames which are created from the original captured frames, TSR-SS can create a high-frame-rate video from the scaled self-similar patches.

4. PROBLEM FORMULATION

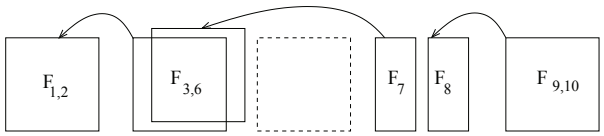

Fig. 3. Example of FET video encoding. MR (captured) frames $F_{(1,2)}$ and $F_{(9,10)}$ are coded as is. LR frame $F_{(3,6)}$ is average of captured frames $F_{(3,4)}$ and $F_{(5,6)}$. SR F_7 and F_8 are synthesized from MR frame $F_{(7,8)}$.

We formulate our problem of selecting FET frames at appropriate temporal resolutions as a formal combinatorial optimization in this section. We first overview the degrees of freedom in the optimization unique for our problem setting. Each captured FET (medium temporal resolution or MR) frame $F_{(i-1,i)}$ can be super-resolved via TSR-SS at encoder into *super-resolved (SR) frames* F_{i-1} and F_i. It can also be combined with captured frame $F_{(i-3,i-2)}$ into an *low resolution (LR) frame* $F_{(i-3,i)}$. (In turn, LR frame $F_{(i-3,i)}$ can be super-resolved into MR frames $F_{(i-3,i-2)}$ and $F_{(i-1,i)}$.) In general, encoding SR frames gives high quality but requires many coding bits, while encoding LR frames gives poor synthesized frame quality at decoder but requires few bits. The problem is to find a combination of SR frames, MR frames and LR frames at encoder that gives the best RD tradeoff. Fig. 3 shows one example of such frame combination.

4.1. DAG Representation of Frame Selection

Denote by $\mathcal{F} = \{F_{(1,2)}, F_{(3,4)}, \ldots, F_{(2N-1,2N)}\}$ the N captured MR frames. Denote by $\mathcal{F}^s = \{F_1, F_2, \ldots, F_{2N}\}$ the $2N$ corresponding SR frames synthesized at encoder prior to compression. In addition, let averages of neighboring captured frames (LR frames) be $\mathcal{F}^a = \{F_{(1,4)}, F_{(3,6)}, \ldots, F_{(2N-3,2N)}\}$, where $F_{(i-3,i)}$ is the pixel-by-pixel average of MR frames $F_{(i-3,i-2)}$ and $F_{(i-1,i)}$. The goal is to identify which subset of frames in $\mathcal{F}$, $\mathcal{F}^s$ and $\mathcal{F}^a$ should be coded for a targeted RD tradeoff.

We construct a directed acyclic graph (DAG) to represent selections of frames for coding as follows. For the three sets of SR frames $\mathcal{F}^s$, MR frames $\mathcal{F}$ and LR frames $\mathcal{F}^a$, we create three rows of nodes from top to bottom, one node x for each frame F_x, as shown in Fig. 4. We line up nodes i, $(i-1,i)$ and $(i-3,i)$ representing F_i, $F_{(i-1,i)}$ and $F_{(i-3,i)}$, where i is an even index, in one column; these frames correspond to content of F_i at different temporal resolutions. In addition, we create start and end nodes s and t at the left and right end of the DAG.

We draw an edge $e_{x\to y}$ from node x to y to represent the case when frame F_y is differentially coded using F_x as predictor. Thus, all the potential predictor-predictee relationships between two frames are shown as edges. Essentially, nodes i and $(j,i), \forall j$, can be predictors for nodes $i+1$ and $(i+1,k), \forall k$. In addition, start node s has edges to nodes of candidate first frames of the coded sequence (I-frame)—frame F_1, $F_{(1,2)}$ and $F_{(1,4)}$. End node t has edges stemming from nodes of candidate last frames of the sequence, frame F_{2N}, $F_{(2N-1,2N)}$ and $F_{(2N-3,2N)}$.

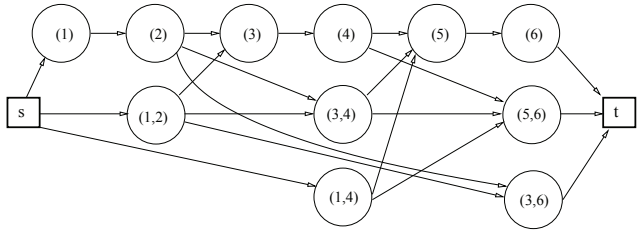

Fig. 4. DAG example for six SR frames, and start and end node s, t.

Because a motion-compensated frame has one predictor frame, which in turn has one predictor frame, all the way back to the first frame that is encoded as an I-frame, a *path* $\mathbf{p}$ from s to t through the constructed DAG represents a selection of frames for encoding of the sequence. If we can then assign cost to the edges representing rate and distortion costs, then we can formulate the frame selection problem as as minimum cost path problem. Towards that goal, we first discuss rate and distortion model.

4.2. Rate Model

For a given path $\mathbf{p} = \{s, p_1, p_2, \ldots, p_{L_p}, t\}$, where $L_\mathbf{p}$ is the length of path $\mathbf{p}$ minus s and t, we can write the encoding rate $R(\mathbf{p})$ of the coded sequence described by path $\mathbf{p}$ as as sum of individual selected frames:

$$R(\mathbf{p}) = \sum_{l=1}^{L_\mathbf{p}} r^c(p_l, p_{l-1}) \tag{1}$$

where $p_0 = s$, and the coded rate $r^c(p_1, p_0)$ of the first frame depends only on the selection of the first frame F_{p_1}. The coded rate $r^c(p_l, p_{l-1})$ of the l-th frame depends only on the selection of l-th frame F_{p_l} and its predictor $F_{p_{l-1}}$. This Markovian rate dependency model has been shown to be useful in [7].

4.3. Distortion Model

We can similarly write the resulting distortion $D(\mathbf{p})$ of the super-resolved sequence at decoder, given encoded frame selection path $\mathbf{p}$, as a simple sum:

$$D(\mathbf{p}) = \sum_{l=1}^{L_\mathbf{p}} d(p_l, \mathbf{p}) \tag{2}$$

where $d(p_l, \mathbf{p})$ is the distortion corresponding to coded frame F_{p_l}. If coded frame F_{p_l} is a SR frame, then distortion $d(p_l, \mathbf{p})$ is simply the coded distortion $d^c(p_l, p_{l-1})$ that depends on the selection of its predictor $F_{p_{l-1}}$ only [7]. If F_{p_l} is a MR frame or a LR frame, then we must account for synthesized distortions of all SR frames to be generated at decoder of which F_{p_l} contains content. Let $(a(p_l), b(p_l))$ be the range of SR frames of which frame F_{p_l} averaged over, given F_{p_l} is a MR frame $F_{(i-1,i)}$ or an LR frame $F_{(i-3,i)}$. We can write distortion $d(p_l, \mathbf{p})$ of coded frame F_{p_l} as:

$$d(p_l, \mathbf{p}) = \begin{cases} d^c(p_l, p_{l-1}) & \text{if } F_{p_l} \in \mathcal{F}^s \\ \sum_{k=a(p_l)}^{b(p_l)} d^s(k, \mathbf{p}) & \text{o.w.} \end{cases} \tag{3}$$

where $d^s(k, \mathbf{p})$ is the synthesized distortion of super-resolved SR frame F_k at decoder given selected frame path $\mathbf{p}$. Because TSR-SS uses all coded frames in the sequence to detect self-similarity, $d^s(k, \mathbf{p})$ cannot be further simplified to depend only on local p_l and p_{l-1}, as done for distortion $d^c(p_l, p_{l-1})$ of coded SR frame F_{p_l}.

4.4. Shortest Path in DAG

Having defined rate cost $R(\mathbf{p})$ and distortion cost $D(\mathbf{p})$ for a given frame selection path $\mathbf{p}$, we can formalize our path selection problem as a Lagrangian optimization problem:

$$\min_{\mathbf{p}\in\mathcal{P}} \Theta(\mathbf{p}) = D(\mathbf{p}) + \lambda R(\mathbf{p}) \tag{4}$$

where $\mathcal{P}$ is the set of feasible paths from s to t, and λ is the Lagrangian multiplier. We can alternatively write (4) as a sum of individual Lagrangian costs $\theta(p_l, \mathbf{p})$ of coded frame F_{p_l}:

$$\begin{aligned} \Theta(\mathbf{p}) &= \sum_{l=1}^{L_\mathbf{p}} \theta(p_l, \mathbf{p}) \\ \theta(p_l, \mathbf{p}) &= d(p_l, \mathbf{p}) + \lambda r^c(p_l, p_{l-1}) \end{aligned} \tag{5}$$

By assigning each individual Lagrangian cost $\theta(p_l, \mathbf{p})$ to edge e_{p_{l-1}, p_l} of path $\mathbf{p}$, we see now that a shortest cost path in the DAG, where the cost of the path $\mathbf{p}$ is the sum of individual edge costs $c(e_{p_{l-1}\to p_l})$'s, corresponds to the optimal frame selection in (5). We discuss how we find this shortest path in the next section.

5. FAST SHORTEST PATH ALGORITHM

There are two complications when trying to solve the shortest path problem in (5). The first is that Lagrangian cost $\theta(p_l, \mathbf{p})$ of edge $e_{p_{l-1}\to p_l}$ depends on the *entire path* $\mathbf{p}$ rather than just the two end nodes of the edge, p_{l-1} and p_l. The second is that even if edge cost $c(e_{p_{l-1}\to p_l})$ can be determined solely as function of p_{l-1} and p_l, finding the shortest path for a large number of nodes and edges can be computationally expensive. We address the two issues in order.

5.1. Iterative Shortest Path Procedure

To address the first concern, we propose an iterative procedure where, in each iteration, a shortest path (SP) in the DAG is found given *fixed* edge costs; i.e, each cost $c(e_{x\to y})$ of edge $e_{x\to y}$ is fixed given nodes x and y, independent of other nodes. Edge costs are adjusted after each iteration. The procedure terminates when two consecutive iterations return the same SP. Details of the procedure is shown in `Iterative SP Procedure`.

We explain the rationale behind the procedure as follows. Step 2 initialize each edge $e_{x\to y}$ to be the smallest $\theta(y, \mathbf{p}^i)$ possible, since all other nodes besides x and y are SR nodes, resulting in the smallest synthesized distortions. This provides the procedure an opportunity to find a path away from initial path of all SR nodes. Subsequently, for each discovered SP $\mathbf{p}$, each edge cost of SP will only increase according to selected nodes in $\mathbf{p}$. This means the procedure is guaranteed to converge. Moreover, the calculated cost of the converged

`Iterative SP Procedure`

1: Initialize previous path $\mathbf{p}'$ to be path with all SR nodes.
2: Initialize $c(e_{x\to y})$ to be $\theta(y, \mathbf{p}^i)$, where path $\mathbf{p}^i$ contains x and y, and all other nodes are SR nodes.
3: Find SP $\mathbf{p}$ given fixed edge costs.
4: **if** $\mathbf{p}' \neq \mathbf{p}$ **then**
5: $\quad c(e_{p_{l-1}\to p_l}) \leftarrow \max\{\theta(p_l, \mathbf{p}), c(e_{p_{l-1}\to p_l})\}$.
6: $\quad \mathbf{p}' \leftarrow \mathbf{p}$. Goto Step 3.
7: **end if**

SP is an upper-bound of the Lagrangian cost of the true SP due to the max operation in Step 5.

5.2. Monotonicity in Predictor's Temporal Resolution

For given fixed edge costs, we can speed up the SP search with the assumption of *monotonicity in predictor's temporal resolution*. The observation is that finer temporal resolution of a frame containing content of SR frame F_i, say $F_{(i-1,i)}$, is in general a better predictor for future frames $F_{(i+1,k)}$'s, $\forall k$, than coarser temporal resolution of the same frame, say $F_{(i-3,i)}$. So if the local cost of choosing the finer temporal resolution of the frame is already less than the coarser resolution, then the coarser resolution is globally sub-optimal.

In more rigorous details, let $\psi(x)$ be the shortest sub-path from s to node x. We state the above observation formally as follows:

Lemma 1 *If $\psi(x) < \psi(y)$, where F_x and F_y are fine and coarse temporal resolution of the same frame, then F_y is globally sub-optimal.*

Proof of Lemma 1 *We prove by contradiction. Suppose there exists SP $\mathbf{p}^o$ that includes node $p_l^o = y$. We construct a new path $\mathbf{p}^* = \{\psi(x), p_{l+1}^o, \ldots, p_{L_{\mathbf{p}^o}}^o, t\}$. By assumption, $\psi(x) < \psi(y)$. Further, given F_x and F_y are fine and coarse temporal resolution of the same frame, $c(e_{x\to p_{l+1}^o}) \leq c(e_{y\to p_{l+1}^o})$ by monotonicity of predictor's temporal resolution. Hence the cost of path $\mathbf{p}^*$ is strictly less than cost of SP $\mathbf{p}^o$, a contradiction.*

The lemma can be used in the following way to speed up the SP search. Shortest sub-paths to nodes in the DAG are calculated column-by-column from left to right. At each column of nodes, if $\psi(x) < \psi(y)$ where F_x is a finer temporal resolution of F_y, then node y is sub-optimal and can be pruned from DAG right away.

6. EXPERIMENTATION

Fig. 5. SR frame in `intrsct3` at QP=35.

To test the performance of our proposed compression scheme (`adapt`), we first shot a 300-frame 384×216 FET video sequence `intrsct3` as data for MR frame set $\mathcal{F}$. We then created SR frame set $\mathcal{F}^s$ via TSR-SS, and $\mathcal{F}^a$ by averaging neighboring MR frames. Assuming an Group of Pictures (GOP) size of 30 captured frames, we then encoded combinations of $\mathcal{F}$, $\mathcal{F}^s$ and $\mathcal{F}^a$ at different QPs using H.264 to calculate encoding rates r^c's and distortions d^c's. For a given QP, we varied multiplier value λ when finding the best frame combinations for encoding via SP to obtain an RD curve. We repeated the experiment for different QPs; `adapt` is the convex hull of all adaptive curves. See Fig. 5 for an example of a TSR-SS synthesized SR frame.

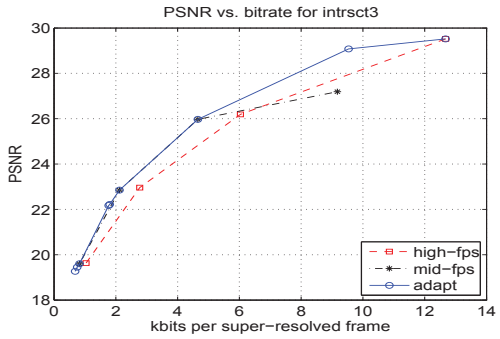

Fig. 6. RD performance of various compression schemes. x-axis is bits per super-resolved frame, and y-axis is video quality in PSNR.

In Fig. 6, we compare our scheme to simple schemes when the entire FET video is first up-sampled via TSR-SS before encoding (`high-fps`), and when the FET video is encoded as is, then is super-resolved temporally via TSR-SS at decoder (`mid-fps`). We first observe that `mid-fps` performed better than `high-fps` at lower bitrate, while `high-fps` performed better at higher bitrate. This is intuitive, since we know `high-fps` preserves TSR-SS quality but consumes more bits. We next observe that `adapt` outperformed both `high-fps` and `mid-fps` at higher bitrate: by up to 1.1dB and 1.8dB, respectively. This shows that adaptive selection of FET frames at different temporal resolution is important.

7. CONCLUSION

In this paper, we propose an adaptive encoding strategy to select full-exposure-time (FET) frames at different temporal resolutions to optimize RD performance. Results show our adaptive scheme outperformed naïve schemes by up to 1.1dB in PSNR. For future work, we are considering adaptive selection of QPs at a frame level together with temporal resolutions.

8. REFERENCES

[1] M. Shimano, T. Okabe, I. Sato, and Y. Sato, "Video temporal super-resolution based on self-similarity," in *The Tenth Asian Conference on Computer Vision*, Queenstown, New Zealand, November 2010.

[2] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. ECCV 2004 (LNCS3024)*, 2004, pp. 25–36.

[3] T. Chen, "Adaptive temporal interpolation using bidirectional motion estimation and compensation," in *IEEE International Conference on Image Processing*, Rochester, NY, September 2002.

[4] J. Zhai, K. Yu, J. Li, and S. Li, "A low complexity motion compensated frame interpolation method," in *IEEE International Symposium on Circuits and Ssytems*, Kobe, Japan, May 2005.

[5] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," in *IEEE Transactions on Image Processing*, September 1994, vol. 3, no.5.

[6] S. Liu and C.C. J. Kuo, "Joint temporal-spatial bit allocation for video coding with dependency," in *IEEE Transactions on Circuits and Systems for Video Technology*, January 2005, vol. 15, no.1, pp. 15–26.

[7] G. Cheung and V. Velisavljević, "Efficient bit allocation for multiview image coding & view synthesis," in *IEEE International Conference on Image Processing*, Hong Kong, September 2010.