OPTIMIZED FRAME STRUCTURE FOR INTERACTIVE LIGHT FIELD STREAMING WITH COOPERATIVE CACHING

Wei Cai[o], Gene Cheung[#], Taekyoung Kwon[o], Sung-Ju Lee[*]

[o] Seoul National University, [#] National Institute of Informatics, [*] Hewlett-Packard Laboratories

ABSTRACT

Light field is a large set of spatially correlated images of the same static scene captured using a 2D array of closely spaced cameras. Interactive light field streaming is the application where a client continuously requests successive light field images along a view trajectory of her choosing, and in response the server transmits appropriate data for the client to correctly reconstruct desired images. The technical challenge is how to encode captured light field images into a reasonably sized frame structure a priori (without knowing eventual clients' view trajectories), so that during streaming session, expected server transmission rate can be minimized, while satisfying client's view requests. In this paper, we design efficient frame structures, using I-frames, redundant P-frames and distributed source coding (DSC) frames as building blocks, to optimally trade off storage size of the frame structure with expected server transmission rate. The key novelty is to optimize structures in such a way that decoded images in caches of neighboring cooperative peers, connected together via a secondary network such as ad hoc WLAN for content sharing, can be reused to further decrease the server-to-client transmission rate. We formulate the structure design problem as a Lagrangian minimization, and propose fast heuristics to find near-optimal solutions. Experimental results show that the expected server streaming rate can be reduced by up to 83% compared to an I-frame-only structure, at less than twice the storage required.

Index Terms— light field, interactive streaming, cooperative caching

1. INTRODUCTION

Light field [1] is a large set of spatially correlated images of the same static scene taken from a 2D array of closely spaced cameras. Because conventional display terminals show only one image at a time, typically a client browses the light field data by observing single images in succession across time [2]. *Interactive light field streaming* (ILFS) [3] captures this media interaction between streaming server and client: a client continuously requests successive light field images along a view trajectory of her choosing, and in response the server transmits appropriate data for the client to correctly reconstruct desired images for display.

The technical challenge for ILFS is to encode captured light field images into a reasonably sized frame structure a priori, so that during actual streaming session, the expected server transmission rate to the client interactively selecting views is minimized. This is important if the server-client connection is over an expensive and/or bandwidth-limited link such as Wireless Wide Area Network (WWAN). The problem is challenging because at encoding time, the exact view trajectory that a client will take at stream time is unknown, making it difficult to employ *differential coding* to reduce the transmission rate. Differential coding, typical in coding of single-view video (with temporal dimension), assumes a previous frame F_{i-1} of time instant $i-1$ is available at decoder for prediction of target image F_i of instant i, so that only (quantized) differential $F_i - F_{i-1}$ needs to be encoded. If view trajectory in ILFS (with spatial dimension and no temporal dimension) is not known at encoding time, then no frame can be assumed to be available at decoder with certainty for prediction of the target image, and traditional differential coding cannot be applied as is. A simple alternative strategy is to forego differential coding and encode every light field image as an independently coded I-frame. However, this results in a large server transmission rate because no inter-frame correlation is exploited for coding gain.

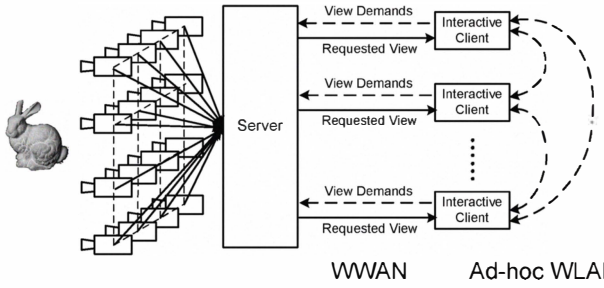


Fig. 1. System overview. A 2D array of closely spaced cameras capture spatially correlated images. Server encodes images into a frame structure. A client interactively requests images along her chosen trajectory from the server, while sharing displayed images with neighboring peers locally.

In this paper, we derive new frame structures, using I-frames, redundant P-frames [4] and distributed source coding (DSC) frames [5] as building blocks, to optimally trade off storage size of the structure with expected server transmission rate. The key novelty over previous ILFS work is in optimizing structures in such a way that decoded images in caches of neighboring cooperative peers (*cooperative cache*), connected locally via a secondary network such as ad hoc WLAN, can be reused to decrease server transmission rate. Scenarios where the clients are locally connected together while engaging in ILFS with the server include 3D visualization in art museums or cultural heritage sites [6], where light field images of valued objects like statues or temples were captured and prepared a priori. When guests visit these sites, they can enrich their visual experience with alternative views of the same objects on their handheld devices from different viewing angles and under different lighting conditions. See Fig. 1 for an illustration.

To impart intuition for the structure design problem, consider first the case where the server can perform encoding in real-time during a streaming session, and tranmission in the peer-to-peer (P2P) secondary network has negligible cost and delay compared to the

978-1-61284-350-6/11/$26.00 ©2011 IEEE

primary network. The minimum server transmission rate in this case is to transmit at "the rate of innovation"; i.e., only uncorrelated information that is not already contained in peers' cache needs to be transmitted. For example, if peer X requests image $C_{i,j}$ from the server, the most "similar" image $C_{x,y}$ in all peers' cache is first forwarded to X, and the server sends only differential $C_{i,j} - C_{x,y}$.

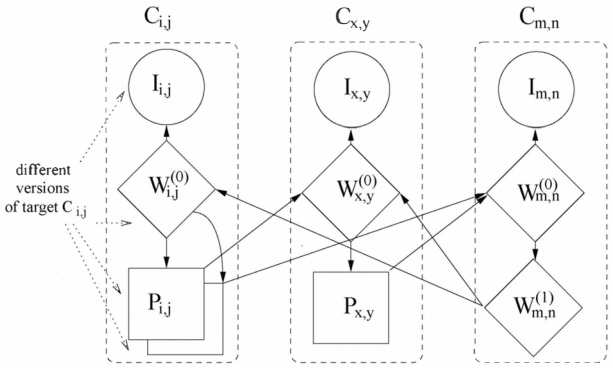


Fig. 2. Example of ILFS frame structure. I-, DSC and P-frames are denoted by circles, diamonds and squares, respectively. In this example, image $C_{i,j}$ has four coded versions: I-frame $I_{i,j}$, DSC frame $W_{i,j}^{(0)}$, and two P-frames $P_{i,j}(x,y)$ and $P_{i,j}(m,n)$.

Of course, our problem setting requires encoding of light field images prior to stream time. To exploit correlation between similar image $C_{x,y}$ in cooperative cache and requested image $C_{i,j}$, we construct a *redundant* structure at server—redundant in that a light field image can be represented by more than one coded version—as follows. An independently coded I-frame version $I_{i,j}$ of target $C_{i,j}$ is first encoded. A differentially coded P-frame version $P_{i,j}$ is then encoded using a "merged" version $W_{x,y}^{(0)}$ of $C_{x,y}$ as predictor. See Fig. 2 for an illustration. If $W_{x,y}^{(0)}$ is not available at peers' cache, $I_{i,j}$ is transmitted from server. If $W_{x,y}^{(0)}$ is available at peers' cache, then $W_{x,y}^{(0)}$ is forwarded to peer X via P2P network, and $P_{i,j}$ is transmitted from server, where $|P_{i,j}| < |I_{i,j}|$. This results in two decoded versions of $C_{i,j}$, $I_{i,j}$ and $P_{i,j}$, depending on the availability of $W_{x,y}^{(0)}$ in peers' cache. To reconciliate these two versions into one $W_{i,j}^{(0)}$, so that merged $W_{i,j}^{(0)}$ can be used as unique predictor for other images (as done for $C_{x,y}$), DSC is deployed. Essentially, transform coefficient bit-planes of decoded versions ($I_{i,j}$ and $P_{i,j}$ in the example) will be the same except for a few least significant bits (LSB), and DSC encodes enough LSB bit-planes so that the same target can be decoded no matter which decoded version is used as predictor [5].

Clearly, the above construction creates a redundant P-frame $P_{i,j}$ for each similar image $C_{x,y}$ to $C_{i,j}$, increasing storage but potentially decreasing server transmission rate. The crux is to design structures that select the right amount of redundancy to optimally trade off storage size with server transmission rate. We show that by optimizing this tradeoff, we reduce the expected server transmission rate by up to 83% compared to an I-frame-only structure, at less than twice the storage required.

The outline of the paper is as follows. We first review related work in Section 2. We then overview our ILFS system and assumptions in Section 3. We formulate our structure design problem as a Lagrangian minimization in Section 4, and present a fast heuristic algorithm as a solution in Section 5. Results and concluding remarks are presented in Section 6 and 7, respectively.

2. RELATED WORK

We discuss previous coding schemes for ILFS and discuss related work on cooperative networks that exploit peers' cooperation for system-wide performance gain in different application scenarios.

2.1. Coding Structures for ILFS

The uncertainty of which predictor frame is available for differential coding of a target image during encoding time is a major source of difficulty for ILFS, and novel coding structures have been proposed to address this [7, 8]. [7] assumed a user only switches to a adjacent view during an ILFS session, and hence one out of a small subset of adjacent frames must be available at decoder for prediction of the target image during a view-switch. [7] then proposed to differentially encode one SP-frame for each predictor frame, so that the server can transmit an SP-frame corresponding to the predictor frame residing in the decoder during stream time. The identical construction property of SP-frames ensures the same reconstruction of the target image no matter which SP-frame (corresponding to the predictor frame in the decoder cache) was actually transmitted. For the same assumption of adjacent view switches, [8] proposed to use DSC instead, where the number of LSB bit-planes that need to be transmitted depends on the quality of the *side information*, i.e., the largest difference between the predictor frame at decoder and the target image. The key difference between [7, 8] and our work is that we assume *random access* is also possible in ILFS, where a non-adjacent image can be selected by a client (see example user interface in [2] where random access images can be selected naturally). For these random access images, we optimize structures to exploit content in cooperative cache to reduce server transmission rate.

[4, 5] have studied redundant frame structures for interactive multiview video streaming (IMVS), where a user can periodically select one out of many views available at server as the streaming video is played back in time. Though the notion of frame redundancy is similar, we focus here instead on exploitation of content in cooperative caches to reduce server transmission rate for ILFS.

2.2. Cooperative Multi-homed Networks

We stress that our assumption of devices connected to multiple networks simultaneously, such as WWAN to server and ad hoc WLAN to neighboring peers, is a common one in the literature [9, 10, 11] and in practice (e.g., smart phone), where different optimizations are performed exploiting the multi-homing property. [9] shows that aggregation of an ad hoc group's WWAN bandwidths can speed up individual peer's infrequent but bursty content download like web access. [10] proposes an integrated cellular and ad hoc multicast architecture, where the cellular base station delivered packets to proxy devices with good channel conditions, and then proxy devices utilize local ad hoc WLAN to relay packets to other devices. Recently, [11] utilizes a secondary ad hoc WLAN network for local recovery of WWAN broadcast / multicast packets lost during WWAN transmission, exploiting peers' cooperation. Our proposal extends this body of work on cooperative multi-homed networks to ILFS, by exploiting correlation between requested images and content residing in peers' caches to lower server transmission rate.

We note that our proposed redundant frame structures for ILFS is applicable to multi-homed wireless networks motivated in this paper, as well as heterogeneous wired networks. For example, a set of clients connected together via a campus LAN want to access a light field dataset located in a faraway network location.

3. SYSTEM OVERVIEW

3.1. System Overview

The system model we consider for ILFS is shown in Fig. 1. Cameras in a $M \times M$ 2D array capture images from a scene of interest and send these uncompressed pictures to a media server. The server encodes these captured images offline into an optimized redundant frame structure $\mathcal{S}$ of I-, P- and DSC frames for storage.

A client interested in ILFS is connected to the server via WWAN (Wireless Wide Area Network). In addition, clients are also connected to their one-hop neighbors via ad hoc WLAN (Wireless Local Area Network). For each client's view request, the server can send the required data directly via WWAN (*direct* mode). It can instruct the client to retrieve data from a neighboring peer (*indirect* mode). It can also instruct the client to first retrieve a reconstructed image from a neighboring peer, then send pre-encoded differential(s) between the requested image and the neighbor's reconstructed image (*mixed* mode). Hence the secondary network provides image sharing to alleviate heavy server-client transmission in indirect and mixed modes.

3.2. View Interaction Model

An ILFS client remains in a streaming session for a random number of view switches L before departing. As often done in lifetime modeling, we will assume random variable L follows a Poisson distribution:

$$f(L) = \frac{\mu^L e^{-\mu}}{L!} \tag{1}$$

where the mean lifetime $E[L]$ is μ.

Captured images are arranged into a 2-D grid. Let $C_{i,j}$ be the image captured by camera on row i and column j. We assume all clients start an ILFS session at view C_{x^I,y^I}. There are two kinds of movement for each client: *walk* and *jump*. Walk movement means the client selects adjacent views to the current view, resulting in a contiguous view trajectory over time. In other words, having observed image $C_{i,j}$, the client requests one of its adjacent views, $C_{i\pm1,j\pm1}$. The probability for a client to select the walk movement is denoted by p_w. We assume that the probabilities of switching to adjacent views are the same; thus, given the number of adjacent views is N_{adj}, the probability to each adjacent view is $\frac{p_w}{N_{adj}}$.

Jump movement means a client switches to further-away views than adjacent views in the light field. Let the probability of a client selecting jump movement be $p_J = 1 - p_w$. We assume all non-adjacent images have the same access probability $\frac{p_J}{M \times M - N_{adj} - 1}$.

3.3. Cooperative Peer Model

We assume that the average number of one-hop neighboring ILFS clients U participating in cooperative caching at any given time is known. We assume also that the cache size of each client is sufficiently large, so that until the client departs from her ILFS session, every displayed image is cached. The lifetime of each neighbor in the system is also modeled by random variable L. We assume bandwidth for the ad hoc WLAN is sufficiently large for all U peers in the immediate neighborhood to share their images when needed. Thus, bandwidth constraint in the ad hoc WLAN is not explicitly modeled.

4. PROBLEM FORMULATION

We formulate our frame structure design problem as a Lagrangian minimization in this section. We first discuss how frames in a given structure are used during an ILFS session in Section 4.1. Using the interaction and cooperative peer models discussed in previous section, we then derive image display and caching probabilities (the likelihood that an image $C_{i,j}$ is requested and a coded version is cached at neighboring peers) in Section 4.2 and 4.3, respectively. Using the derived probabilities, we define storage and server transmission costs for a given structure $\mathcal{S}$ in Section 4.4 and 4.5, respectively. Finally, we define our Lagrangian minimization in Section 4.6.

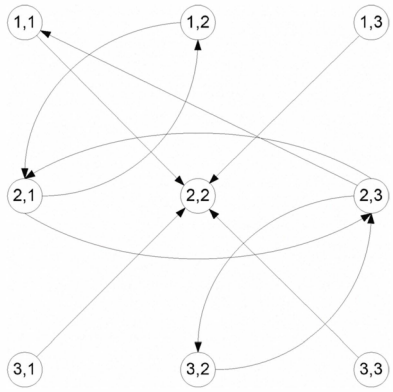


Fig. 3. An Example of Frame Structure

4.1. Coded Frames in Frame Structure

Each light field image $C_{i,j}$ can be redundantly encoded into structure $\mathcal{S}$ as I-frame, DSC frame and multiple P-frames, denoted by $I_{i,j}$, $W_{i,j}$, and $P_{i,j}(x,y)$, respectively. For a given $\mathcal{S}$, there are $K_{i,j}$ P-frames $P_{i,j}(x,y)$'s, each encoded using coded version $W^{(0)}_{x,y}$ of image $C_{x,y}$ as predictor, and the $K_{i,j}$ P-frames are ordered in increasing sizes: $|P_{i,j}(x^1_{i,j}, y^1_{i,j})| \leq \ldots \leq |P_{i,j}(x^{K_{i,j}}_{i,j}, y^{K_{i,j}}_{i,j})|$. As an example, Fig. 3 shows a frame structure for a 3×3 light field, where a single node (i,j) denotes *all* coded frames for image $C_{i,j}$. Each edge from (i,j) to (x,y) indicates a P-frame $P_{i,j}(x,y)$ has been constructed for image $C_{i,j}$ using coded version of non-adjacent $C_{x,y}$ as predictor. (P-frames using adjacent images as predictors are not shown.) For example, image $C_{3,1}$ has a P-frame $P_{3,1}$ using coded version of $C_{2,2}$ as a predictor.

When a client requests image $C_{i,j}$ from server, server first checks if a neighboring peer has a merged version $W^{(0)}_{i,j}$ in cache. If so, server instructs client to retrieve $W^{(0)}_{i,j}$ from neighboring peer in indirect mode. If $W^{(0)}_{i,j}$ is not in cooperative cache, then server checks, in the order of P-frame sizes, if a merged version $W^{(0)}_{x^k_{i,j}, y^k_{i,j}}$ of a predictor $C_{x^k_{i,j}, y^k_{i,j}}$ of P-frame $P_{i,j}(x^k_{i,j}, y^k_{i,j})$ is available. If so, server instructs client to retrieve $W^{(0)}_{x^k_{i,j}, y^k_{i,j}}$ from neighboring peer and transmits differentially coded P-frame $P_{i,j}(x^k_{i,j}, y^k_{i,j})$ in mixed mode. As an example, in Fig. 3, if client requests $C_{3,1}$ and $W^{(0)}_{2,2}$ is available in cooperative cache, then $W^{(0)}_{2,2}$ is shared, and server sends differentially coded $P_{3,1}(2,2)$. In this mode, server sends in addition DSC frame $W^{(0)}_{i,j}$, so that all reconstructed versions of $C_{i,j}$ can merge into one unique decoded version. Because all predictors for $W^{(0)}_{i,j}$ are slightly different coded versions of $C_{i,j}$, $W^{(0)}_{i,j}$ contains no motion information and encodes only a few LSB bit-planes, resulting in very small frame size $|W^{(0)}_{i,j}|$. If none of the predictors of P-frames $P_{i,j}$'s are in cooperative cache, then server transmits

I-frame $I_{i,j}$ and DSC-frame $W_{i,j}^{(0)}$ in direct mode.

Instead of encoding multiple P-frames $P_{i,j}$'s, each using a different predictor $W_{x,y}^{(0)}$, an alternative is to encode a *single* DSC frame $W_{i,j}^{(1)}$ with multiple predictors. In Fig. 2, DSC frame $W_{m,n}^{(1)}$ is encoded using $W_{i,j}^{(0)}$ and $W_{x,y}^{(0)}$ as predictors. $W_{i,j}^{(1)}$ essentially encodes a set of motion information for each predictor, then encodes enough LSB bit-planes of the transform coefficients of the motion residuals, so that unique decoding is guaranteed no matter which predictor is available at decoder. Because $W_{i,j}^{(1)}$ encodes motion information, it is much larger in size than $W_{i,j}^{(0)}$; we call DSC frames $W_{i,j}^{(0)}$ and $W_{i,j}^{(1)}$ *type* 0 *DSC* and *type* 1 *DSC*, respectively [5]. The advantage of using type 1 DSC $W_{i,j}^{(1)}$ over multiple P-frames $P_{i,j}$'s is storage saving, since only one frame is encoded. The disadvantage is transmission rate, since a fairly large type 1 DSC $W_{i,j}^{(1)}$ needs to be transmitted no matter which predictor is available at the decoder.

4.2. Image Display Probabilities

We model transition from images to images in ILFS using a discrete-time Markov chain. Specifically, we construct a $N \times N$ transition matrix $\mathbf{A}$, where $a_{i*M+j,x*M+y}$ is the view transition probability of a client selecting image $C_{x,y}$ after viewing $C_{i,j}$. From earlier discussion on view interaction model, each entry in $\mathbf{A}$ can be written as:

$$a_{i*M+j,x*M+y} = \begin{cases} 0 & \text{if } x = i, y = j \\ \frac{p_w}{N_{adj}} & \text{if } |x-i| \leq 1, |y-j| \leq 1 \\ \frac{p_J}{M \times M - N_{adj} - 1} & \text{o.w.} \end{cases} \tag{2}$$

Let $1 \times N$ *initial probability vector* be $\mathbf{g}$, where g_{i*M+j} is the probability that client selects image $C_{i,j}$ as starting view. $\mathbf{g}$ has only one non-zero entry: $g_{x^I*M+y^I}$ corresponding to initial starting view C_{x^I,y^I} has value 1. We can hence calculate the image display probability $\mathbf{p}(l)$ after l view transitions by computing $\mathbf{g}\mathbf{A}^l$:

$$\mathbf{p}(l) = \mathbf{g}\mathbf{A}^l \tag{3}$$

where $p_{i,j}(l) = p_{i*M+j}(l)$ is the probability of image $C_{i,j}$ being displayed after exactly l transitions.

4.3. Image Caching Probabilities

Once an image is decoded and displayed at a peer, it is stored in the peer's cache, which can then be shared by neighboring peers via ad-hoc WLAN. The probability for a coded version of $C_{i,j}$ to be cached by a neighbor, $q_{i,j}$, is subject to two factors: the image display probability $p_{i,j}(l)$ after l view transitions, and the number of neighboring peers U. Given each one of U neighbors has a random lifetime L, the expected current "age" l (number of completed view transitions) of a live neighbor when an ILFS client selects an image is:

$$E[l] = E[E[l|L]] = E[L/2] = \mu/2 \tag{4}$$

A live neighbor of age $\mu/2$ would have cached $C_{i,j}$ if $C_{i,j}$ was viewed within $\mu/2$ view transitions. We can now write $q_{i,j}$ as 1 minus the probability that none of the U neighbors have switched to image $C_{i,j}$ in $\mu/2$ view switches:

$$q_{i,j} = 1 - \left(\prod_{l=0}^{\mu/2} 1 - p_{i,j}(l) \right)^U \tag{5}$$

4.4. Storage Cost

The storage cost of structure $\mathcal{S}$ in the server can be calculated by a sum of all frames in the structure $\mathcal{S}$ as following:

$$B(\mathcal{S}) = \sum_{F_{i,j} \in \mathcal{S}} |F_{i,j}| \tag{6}$$

For given image $C_{i,j}$, size of an I-frame, DSC frame and P-frame are $|I_{i,j}|$, $|W_{i,j}|$ and $|P_{i,j}^{x,y}|$, respectively. Size of an P-frame $|P_{i,j}^{x,y}|$ depends in general on the correlation between the target image $C_{i,j}$ and the predictor image $C_{x,y}$, which in turn depends on the Euclidean distance between (i,j) and (x,y). $|I_{i,j}|$, $|W_{i,j}|$ and $|P_{i,j}^{x,y}|$ can be obtained empirically using codecs such as H.263 [12] for I- and P-frames and [5] for DSC frames.

To summarize, we can write $|F_{i,j}|$ simply as follows:

$$|F_{i,j}| = \begin{cases} |I_{i,j}| & \text{if } F_{i,j} \text{ is a I-frame} \\ |W_{i,j}^{(0)}| & \text{if } F_{i,j} \text{ is type 0 DSC frame} \\ |W_{i,j}^{(1)}| & \text{if } F_{i,j} \text{ is type 1 DSC frame} \\ |P_{i,j}(x,y)| & \text{if } F_{i,j} \text{ is a P-frame} \end{cases} \tag{7}$$

4.5. Server Transmission Cost

We can now derive the server transmission cost $C(\mathcal{S})$ from the server to a client over a ILFS session as follows. A ILFS client can have a lifetime of L view transitions with probability $f(L)$, and for each transition l of L total transitions, it can be either in walk or jump movement, resulting in transition cost $tr_w(l)$ and $tr_J(l)$, respectively:

$$C(\mathcal{S}) = \sum_L f(L) \left[\sum_{l=0}^{L} p_w tr_w(l) + (1-p_w) tr_J(l) \right] \tag{8}$$

For walk transition cost $tr_w(l)$, for each possible chosen image $C_{i,j}$ with probability $p_{i,j}(l)$, it incurs a server transmission cost if the image does not already reside in cooperative cache with probability $1 - q_{i,j}$. Given that walk movement implies that the presently observed frame is an adjacent view of the requested view, we approximate the transmission cost to be the average size $|P_{i,j}^{adj}|$ of $P_{i,j}$'s using adjacent frames as predictors, plus DSC frame $W_{i,j}^{(0)}$ of $C_{i,j}$:

$$tr_w(l) \approx \sum_{i,j} p_{i,j}(l)(1-q_{i,j}) \left(|P_{i,j}^{adj}| + |W_{i,j}| \right) \tag{9}$$

For jump transition cost $tr_J(l)$, we assume the client has not previously viewed the requested image (and hence does not reside in her own cache). If image $C_{i,j}$ does not reside in cooperative cache either, then server checks, in increasing order of size of P-frames $P_{i,j}$'s (if multiple P-frames are used instead of type 1 DSC frame), if any one of predictors $W_{x_{i,j}^k, x_{i,j}^k}^{(0)}$'s is in cooperative cache. If so, it incurs cost $pr_{i,j}(k)$ if the k-th predictor is the first predictor found. If not, it incurs cost $np_{i,j}$.

$$tr_J(l) \approx \sum_{i,j} p_{i,j}(l)(1-q_{i,j}) \left[\sum_{k=1}^{K_{i,j}} pr_{i,j}(k) + np_{i,j} \right] \tag{10}$$

If k-th predictor is found in cooperative cache, then corresponding P-frame $P_{i,j}(x_{i,j}^k, y_{i,j}^k)$ (if multiple P-frames are used) or type

1 DSC frame $W_{i,j}^{(1)}$ (if type 1 DSC frame is used), and type 0 DSC frame $W_{i,j}^{(0)}$, are transmitted from server in mixed mode:

$$\begin{aligned} pr_{i,j}(k) &= \left[\prod_{h=1}^{k-1}(1-q_{x_{i,j}^h,y_{i,j}^h})\right] q_{x_{i,j}^k,y_{i,j}^k}\left[sp_{i,j}(k)+|W_{i,j}^{(0)}|\right] \\ sp_{i,j}(k) &= \begin{cases} |P_{i,j}(x_{i,j}^k,y_{i,j}^k)| & \text{if multiple } P_{i,j}\text{'s for } C_{i,j} \\ |W_{i,j}^{(1)}| & \text{o.w.} \end{cases} \end{aligned} \quad (11)$$

If none of the $K_{i,j}$ predictors of P-frames $P_{i,j}$'s are in cooperative cache, then I-frame $I_{i,j}$ and type 0 DSC frame $W_{i,j}^{(0)}$ must be transmitted from server in direct mode:

$$np_{i,j} = \left[\prod_{h=1}^{K_{i,j}}(1-q_{x_{i,j}^k,y_{i,j}^k})\right]\left(|I_{i,j}|+|W_{i,j}^{(0)}|\right) \quad (12)$$

4.6. Optimization Problem Definition

We can now formally define the search for the optimal redundant frame structure for ILFS as a combinatorial optimization problem: find structure $\mathcal{S}$, using I-, P- and DSC frames as building blocks, in feasible space[1] Φ that possesses the smallest possible expected transmission cost $C(\mathcal{S})$ while a storage constraint $B(\mathcal{S})$ is observed. We denote this optimization problem as:

$$\min_{\mathcal{S}\in\Phi} C(\mathcal{S}) \quad \text{s.t. } B(\mathcal{S}) \le \overline{B} \quad (13)$$

Constrained optimizations such as (13) are usually difficult, and so we focus next on solving the corresponding unconstrained Lagrangian optimization for given Lagrange multiplier λ instead:

$$\min_{\mathcal{S}\in\Phi} J(\mathcal{S}) = C(\mathcal{S}) + \lambda B(\mathcal{S}) \quad (14)$$

5. REDUNDANT FRAME STRUCTURE DESIGN

To find a structure $\mathcal{S}$ that minimizes Lagrangian cost (14) for given λ, we present a greedy algorithm in this section where in each iteration step, the Lagrangian cost is locally minimized.

5.1. Algorithm Overview

We first overview the algorithm. We first initialize a structure $\mathcal{S}$ with an I-frame $I_{i,j}$ and a type 0 DSC frame $W_{i,j}^{(0)}$ for every image $C_{i,j}$ in the light field. This guarantees $\mathcal{S}$ is feasible. Then, for each iteration, for each image $C_{i,j}$ we *nominate* a candidate P-frame $P_{i,j}(x,y)$ with predictor $W_{x,y}^{(0)}$—one that reduces $\mathcal{S}$'s Lagrangian cost the most. Among candidates of all images $C_{i,j}$'s, we select the best candidate $P_{i,j}(x,y)$ as the one that can most reduce the structure's Lagrangian cost. We implement the best candidate $P_{i,j}(x,y)$ either as a new P-frame representation of image $C_{i,j}$, or as a type 1 DSC-frame by *merging* all the existing P-frames of image $C_{i,j}$ (if any) plus $P_{i,j}(x,y)$ to a DSC frame $W_{i,j}^{(1)}$. The procedure of nominating, selecting and implementing candidate P-frames continues until no more new P-frames can be found that can further reduce Lagrangian cost.

[1] A feasible structure is one where any possible request by client for image $C_{i,j}$ can be fulfilled, even if the image is not available in cooperative cache.

5.2. Algorithm Complexity Reduction

To speed up the proposed algorithm, we discuss two simplifications to reduce computation complexity. We observe that solving (8) requires two nested loops of large number of iterations (for all L's, $L \in \mathcal{Z}^+$, such that $f(L) > 0$). To reduce its complexity, we solve instead the following *quantized* version, where $f(L)$ is divided into Φ equal-size probability ranges, and within each range θ we compute the expected lifetime l_θ as representative of that range. We can now write $C(\mathcal{S})$ as:

$$C(\mathcal{S}) \approx \sum_{\phi=1}^{\Phi}\frac{1}{\Phi}\left[\sum_{\theta=1}^{\phi} p_w tr_w(l_\theta) + (1-p_w)tr_J(l_\theta)\right] \quad (15)$$

(15) amounts to quantization of $f(L)$ into Φ discrete points of equal probability, and $C(\mathcal{S})$ is evaluated only at those Φ points. Complexity of (15) is now only $O(\Phi^2)$, where we choose Φ to be a small integer.

The second observation is that Lagrangian objective $J(\mathcal{S})$ in (14) is a sum of local Lagrangian terms for individual light field images $C_{i,j}$'s. To see that, we first note that the storage cost term $B(\mathcal{S})$ in (6) is a sum of frame representations of individual images $C_{i,j}$'s. For transmission cost $C(\mathcal{S})$, we can rewrite (8), (9) and (10) by rearranging the order of summations, so that transmission cost is also a sum of individual contributions from different images $C_{i,j}$'s:

$$\begin{aligned} C(\mathcal{S}) &= \sum_{i,j}\sum_{L} f(L)\left[\sum_{l=0}^{L} p_w tr_{w,i,j}(l) + (1-p_w)tr_{J,i,j}(l)\right] \\ tr_{w,i,j}(l) &= p_{i,j}(l)(1-q_{i,j})\left(|P_{i,j}^{adj}| + |W_{i,j}|\right) \\ tr_{J,i,j}(l) &= p_{i,j}(l)(1-q_{i,j})\left[\sum_{k=1}^{K_{i,j}} pr_{i,j}(k) + np_{i,j}\right] \end{aligned} \quad (16)$$

The corollary of the second observation is that when searching for a P-frame candidate $P_{i,j}(x,y)$ for image $C_{i,j}$, we only need to compare the change in Lagrangian cost *for this image* $C_{i,j}$ only, instead of the entire structure $\mathcal{S}$. The amount of computation required is hence drastically reduced.

6. EXPERIMENTATION

6.1. Experimental Setup

To validate the performance of our discovered frame structures, we set up the following experiments. For light field data, we downloaded a 9×9 light field image sequence `bunny` from [2], each image of size 1024×1024. To encode I- and P-frames, we used an open source H.263 encoder [12], and for type 0 and 1 DSC frames, we used the same codec in [5]. Quantization parameters were set so that the Peak Signal-to-Noise (PSNR) of the encoded frames was around 32dB. Default values for parameters of the ILFS setting were set as follows: walk movement probability was $p_w = 0.65$, average lifetime of a ILFS peer was $\mu = 40$ switches (about half the light field images), average number of one-hop neighboring peers was $U = 4$, Lagrange multiplier in (14) was $\lambda = 0.02$. Depending on the particular experiment performed, one parameter was varied to observe its effect on performance.

We compare performance of our generated structures (`opt`) outputted from our optimization to three fixed frame structures. `I-only` encodes only one I-frame $I_{i,j}$ for each light field image $C_{i,j}$ and performs no cooperative caching. `P-adj` encodes in addition four P-frames $P_{i,j}$'s for image $C_{i,j}$, one for each adjacent

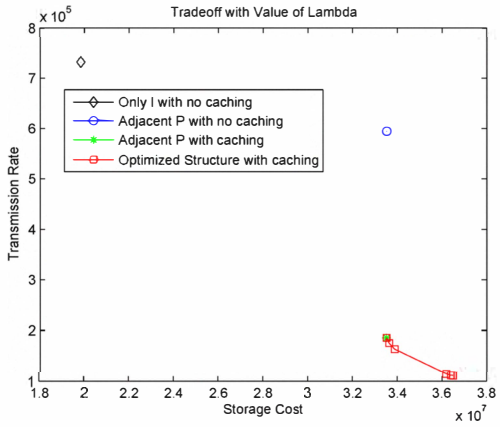
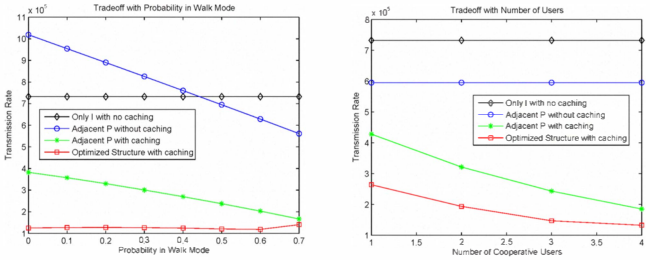


Fig. 4. Expected transmission as function of storage cost for different frame structures.



(a) transmission rate vs. p_w (b) transmission rate vs. U

Fig. 5. Expected transmission rate of different frame structures for different walk movement probability p_w, and different number of cooperative neighboring peers U, respectively.

image (horizontal or vertical) of $C_{i,j}$, and a type 0 DSC frame $W_{i,j}^{(0)}$ for merging. `P-adj-nc` performs no cooperative caching, while `P-adj-c` performs cooperative caching.

6.2. Experimental Results

In Fig. 4, we see the tradeoff between expected transmission rate (expected number of bits transmission per ILFS session) and storage (number of bits) for different frame structures. For `opt`, we varied λ from 0.002 to 0.064 to induce different tradeoffs. We first see that `P-adj-nc`, similar to structures proposed in [7, 8], reduced transmission rate by 17% compared to `I-only`. With cooperative caching, however, `P-adj-c` further reduced transmission rate by 66% compared to `P-adj-nc`. The overhead for `P-adj-nc` and `P-adj-c` is an increase in storage by 70% over `I-only`. As λ decreased, we see that `opt` can reduce transmission rate by 40% compared to `P-adj-c`. Notice that even at the right-most point of `opt`, the storage requirement is less than twice the size of `I-only`, which is quite reasonable in practice.

In Fig. 5(a), we see the performance of different frame structures in expected server transmission rate as function of walk movement probability p_w. As expected, as p_w increased, the likelihood of an adjacent image being selected increased, and transmission rate of `P-adj-c` and `P-adj-nc` decreased. In contrast, the value of non-adjacent P-frames (and type 1 DSC frames) decreased as p_w increased, and the performance of `opt` worsened slightly. Note also that for small p_w, `P-adj-nc` actually performed worse than `I-only`, due to the overhead in type 0 DSC frames $W_{i,j}^{(0)}$'s, a point that was overlooked in previous work [7, 8].

In Fig. 5(b), we see the transmission rate of different frame structures as function of number of one-hop neighboring peers U. As expected, more peers meant better performance for `P-adj-c` and `opt` that exploited cooperative cache. The important observation here is that even if there is only one cooperative peer, the improvement of `opt` over other structures is significant.

7. CONCLUSION

In this paper, we discuss the frame structure design problem for interactive light field streaming (ILFS). Unlike previous work on ILFS, we design structure so that decoded images residing in neighboring peers' cache can be shared, either for display directly or as predictors to the desired images, so that the server transmission rate can be further reduced. Using I-frames, redundant P-frames and two versions of distributed source coding (DSC) frames, we formulated the structure design problem as a Lagrangian minimization problem. We presented a greedy strategy to grow a structure so that Lagrangian cost is locally minimized at every iteration. Experimental results show that our generated structure can reduce server transmission rate by up to 83% compared to the I-frame-only structure, at less than twice the storage required.

8. REFERENCES

[1] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. SIGGRAPH'96*, August 1996, pp. 31–42.

[2] "Stanford Light Field Archive," http://lightfield.stanford.edu/lfs.html.

[3] P. Ramanathan, M. Kalman, and B. Girod, "Rate-distortion optimized interactive light field streaming," in *IEEE Transactions on Multimedia*, June 2007, vol. 9, no.4, pp. 813–825.

[4] G. Cheung, A. Ortega, and N.-M. Cheung, "Generation of redundant coding structure for interactive multiview streaming," in *Seventeenth International Packet Video Workshop*, Seattle, WA, May 2009.

[5] N.-M. Cheung, A. Ortega, and G. Cheung, "Distributed source coding techniques for interactive multiview video streaming," in *27th Picture Coding Symposium*, Chicago, IL, May 2009.

[6] K. Ikeuchi, M. Sakauchi, H. Kawasaki, and I. Sato, "Constructing virtual cities by using panoramic images," in *International Journal of Computer Vision*, July-August 2004, vol. 58, no.3.

[7] Prashant Ramanathan and Bernd Girod, "Random access for compressed light fields using multiple representations," in *IEEE International Workshop on Multimedia Signal Processing*, Siena, Italy, September 2004.

[8] A. Aaron, P. Ramanathan, and Bernd Girod, "Wyner-Ziv coding of light fields for random access," in *IEEE International Workshop on Multimedia Signal Processing*, Siena, Italy, September 2004.

[9] P. Sharma, S.-J. Lee, J. Brassil, and K. Shin, "Aggregating bandwidth for multihomed mobile collaborative communities," in *IEEE Transactions on Mobile Computing*, March 2007, vol. 6, no.3, pp. 280–296.

[10] Randeep Bhatia, Li (Erran) Li, Haiyun Luo, and Ram Ramjee, "ICAM: Integrated cellular and ad hoc multicast," in *IEEE Transactions on Mobile Computing*, August 2006, vol. 5, no. 8, pp. 1004–1015.

[11] X. Liu, G. Cheung, and C.-N. Chuah, "Structured network coding and cooperative wireless ad-hoc peer-to-peer repair for WWAN video broadcast," in *IEEE Transactions on Multimedia*, 2009, vol. 11, no.4.

[12] ITU-T Recommendation H.263, *Video Coding for Low Bitrate Communication*, February 1998.