INTER-FRAME PREDICTION USING MOTION HINTS

Aous Thabit Naman, Rui Xu, and David Taubman

School of Electrical Engineering and Telecommunications,
The University of New South Wales, Australia.

ABSTRACT

We recently proposed a novel approach that employs motion hints for inter-frame prediction. Motion hints are a loose and global description of motion communicated as metadata; they specify motion but they leave it to the client/decoder to find the exact locations where motion is applicable. This work proposes a multi-scale approach for identifying these exact locations, which are then used with the available reference frames to generate an inter-frame prediction. The proposed approach is localized and robust to noise and illumination changes. The scheme of this work is applicable to close-loop prediction, but it is more useful in open-loop prediction scenarios, such as using prediction in conjunction with remote browsing of surveillance footage, communicated by a JPIP server. We show that, with reasonably accurate motion, it is possible to produce good inter-frame predictions visually and in terms of PSNR.

***Index Terms*—** Teleconferencing, Video Surveillance, Video Signal Processing, Image Communication, Motion Compensation

1. INTRODUCTION

In conventional video coding schemes, such as H.261 to H.265 and MPEG1 to MPEG4, the encoder specifies precisely how to construct a predictor and which reference frames to use. This is sub-optimal for interactive applications, because, in such applications, it is useful for the client to use other frames as references; these other frames are obtained from browsing forward, backward, zooming in, etc., and can potentially be of higher quality.

To address these shortcomings, we propose the use of motion hints; a motion hint is a reasonably-accurate description of motion but with a loose description of where it is applicable. Motion hints are communicated as metadata associated with a video sequence; for example, tracking metadata used in surveillance footage implicitly communicates motion hints. A motion hint, in this case, is some geometrical shape (we use a quadrilateral in this work) that encloses an object and tracks its movement across many frames of the video sequence.

Motion hints provide a global description of motion over specific domains; for example, the domain for a tracking motion hint is any region that includes the object being tracked. The motion hint is *global* to the extent that the domain is included in many (perhaps all) frames, enabling the client to predict a given object from any frame in its cache that contains the object. We first proposed the use of metadata to communicate motion information in [1], where it is employed in the context of JSIV [2].

The challenge in the proposed approach is to find the subset of the domain, whose motion is described by a motion hint; we refer to this as the valid region for the motion hint. In our earlier work [1], we used pixel matching in the image domain to find the valid region. Such an approach is not ideal because it is sensitive to noise and changes in illumination. In this work, we employ a multi-scale approach that uses the Laplacian pyramid. In addition to increased robustness to noise and illumination changes, the method proposed here can deduce the valid region of a motion hint within one frame, using as little as one other reference frame, although additional reference frames help. By contrast, the approach in [1] uses exactly 2 reference frames. The proposed algorithm is also localized, needing only local data to make its decisions, which is useful for parallel implementation.

There are many applications for the proposed approach. In one example, considered in this work, a JPIP server [3, 4] might choose either to send every frame from a surveillance footage or to send every other frame, along with motion hints that a client can use to predict the missing frames. Another application is frame rate upsampling (FRUS) [5, 6]. Our work stands out from conventional FRUS techniques in that, in our work, the client has the motion vectors for the frame being predicted but does not have the exact region where this motion is applicable, while in conventional FRUS, the client/decoder has to estimate these motion vectors.

Fundamentally, the problem addressed by this paper can be understood as that of separating foreground from background regions, where the foreground and background motions themselves are given – these are the motion hints. In the context of closed-loop video coding, Orchard [7] and Kim *et al.* [8] also proposed methods that can be used to partially segment frames based on multiple candidate motions. A key distinction in our case is that the valid regions are evaluated at the reference frames, rather than the predicted frame itself; this allows us to assemble evidence from multiple reference frames. Foreground object detection in video has also received considerable attention from researchers – see [9,10] for recent examples. In our problem, however, the foreground/background separation process must be carried out within a decoder, using only the reference frames, which are themselves corrupted by compression artifacts.

2. METADATA, MOTION HINTS, AND PREDICTION

In this work, each frame has associated metadata. To keep things simple, we choose a video sequence which consists of a single tracked object; the metadata associated with each frame communicates a motion hint for the foreground object and another for the background region. Each motion hint is represented by a quadrilateral that explicitly identifies the domain of the hint within each frame[1]. This can be extended to any number of potentially overlapping domains, quadrilateral or otherwise.

We write $\mathcal{D}_k^{\mathcal{F}}$ and $\mathcal{D}_k^{\mathcal{B}}$ for the foreground and background domains within frame f_k. We also write $v_k^{\mathcal{F}}[\mathbf{n}]$ for the likelihood that point $\mathbf{n}$ belongs to the foreground motion's valid region (i.e., the object being tracked). Ideally, $v_k^{\mathcal{F}}[\mathbf{n}]$ equals 1 when point $\mathbf{n}$ belongs to the foreground and 0 otherwise, but we allow $v_k^{\mathcal{F}}[\mathbf{n}]$ to take other values in the range $[0, 1]$ to reflect the degree of confidence associated with our inferences. Obviously, $v_k^{\mathcal{F}}[\mathbf{n}] = 0$ for all $\mathbf{n} \notin \mathcal{D}_k^{\mathcal{F}}$. We can define a similar quantity, $v_k^{\mathcal{B}}[\mathbf{n}]$, to represent the likelihood that point $\mathbf{n}$ belongs to the background's valid region. In the simple setup of this work, where there is only one foreground object, the

[1] The background quadrilateral covers the entire frame.

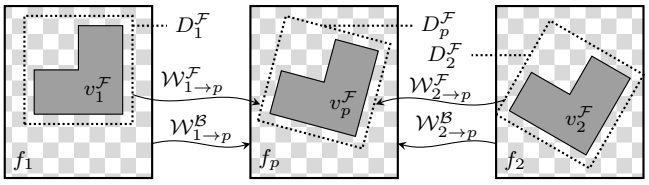

Fig. 1: Mapping foreground and background domains.

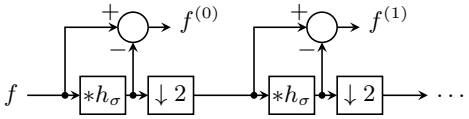

Fig. 2: Laplacian pyramid used in the multiscale approach. $\downarrow 2$ represent down-sampling by 2 in each direction.

background likelihood is derived using $v_k^{\mathcal{B}}[\mathbf{n}] = 1 - v_k^{\mathcal{F}}[\mathbf{n}]$. Section 3 elaborates on the determination of $v_k^{\mathcal{F}}[\mathbf{n}]$.

In this work, each quadrilateral domain is split into two triangles and the associated motion information is derived from the affine transformation between corresponding triangles in successive frames. Richer models could be explicitly communicated, so as to describe complex motion over the foreground and/or background domains, but here we restrict our attention to the affine flows implied by the quadrilateral vertices. In any event, we write $\mathcal{W}_{k\to l}^{\mathcal{F}} = \mathcal{W}(\mathcal{D}_k^{\mathcal{F}}, \mathcal{D}_l^{\mathcal{F}})$ for the overall motion compensation operator that maps locations within the domain $\mathcal{D}_k^{\mathcal{F}}$ in frame f_k to locations within the corresponding domain $\mathcal{D}_l^{\mathcal{F}}$ in frame f_l.

To generate a prediction $f_{\to p}$ for frame f_p, we map two reference frames f_r, where $r \in \{1, 2\}$, and their foreground likelihoods $v_r^{\mathcal{F}}$ to the coordinates of frame f_p using their respective foreground mapping operators $\mathcal{W}_{r\to p}^{\mathcal{F}}$ to obtain $f_{r\to p}^{\mathcal{F}}$ and $v_{r\to p}^{\mathcal{F}}$. We also map the reference frames f_r and their background likelihoods $v_r^{\mathcal{B}}$ using their respective background motion operators $\mathcal{W}_{r\to p}^{\mathcal{B}}$ to obtain $f_{r\to p}^{\mathcal{B}}$ and $v_{r\to p}^{\mathcal{B}}$. Figure 1 shows some of the mappings used here.

A foreground predictor $f_{\to p}^{\mathcal{F}}$ for frame f_p is obtained from a weighted average of the foreground predictions, as follows

$$f_{\to p}^{\mathcal{F}}[\mathbf{n}] = \frac{\sum_{r\in\{1,2\}} \left(\delta + v_{r\to p}^{\mathcal{F}}[\mathbf{n}] \cdot f_{r\to p}^{\mathcal{F}}[\mathbf{n}]\right)}{\sum_{r\in\{1,2\}} \left(\delta + v_{r\to p}^{\mathcal{F}}[\mathbf{n}]\right)} \tag{1}$$

where δ is a small positive value whose purpose is only to ensure that the (1) is well-defined everywhere, and all computations are performed point-wise at each location $\mathbf{n}$. A background predictor $f_{\to p}^{\mathcal{B}}$ for frame f_p is similarly obtained.

The idea in forming an overall prediction for f_p, denoted by $f_{\to p}$, is to prefer the foreground as long as it has high likelihood. Accordingly, we define $v_{\to p}^{\mathcal{F}}[\mathbf{n}] = \max_{r\in\{1,2\}}\{v_{r\to p}^{\mathcal{F}}[\mathbf{n}]\}$, and form $f_{\to p}$ as follows:

$$f_{\to p}[\mathbf{k}] = f_{\to p}^{\mathcal{F}}[\mathbf{k}] \cdot v_{\to p}^{\mathcal{F}}[\mathbf{k}] + f_{\to p}^{\mathcal{B}}[\mathbf{k}] \cdot \left(1 - v_{\to p}^{\mathcal{F}}[\mathbf{k}]\right) \tag{2}$$

3. FOREGROUND LIKELIHOOD ESTIMATION

3.1. Image Domain Foreground Likelihood Estimation

Here, we review the image domain method from [1], before discussing the proposed multi-scale approach. The foreground likelihood is estimated at each reference frame. Here, we consider estimating $v_1^{\mathcal{F}}$ for reference frame f_1 using two other reference frames, f_0 and f_2. The basic idea here is that location $\mathbf{n}$ belongs to the foreground if the predictor obtained using the foreground motion model produces a better estimate for that location than the predictor obtained using the background model. Specifically, we define two errors:

$$\begin{aligned} D_{\mathcal{B}}[\mathbf{n}] &= \min\left\{(f_1[\mathbf{n}] - f_{0\to 1}^{\mathcal{B}}[\mathbf{n}])^2, (f_1[\mathbf{n}] - f_{2\to 1}^{\mathcal{B}}[\mathbf{n}])^2\right\} \\ D_{\mathcal{F}}[\mathbf{n}] &= (f_1[\mathbf{n}] - \tfrac{1}{2} f_{0\to 1}^{\mathcal{F}}[\mathbf{n}] - \tfrac{1}{2} f_{2\to 1}^{\mathcal{F}}[\mathbf{n}])^2 \end{aligned} \tag{3}$$

Then, we set $v_1'[\mathbf{n}]$ to 1 wherever $D_{\mathcal{F}}[\mathbf{n}] \le D_{\mathcal{B}}[\mathbf{n}]$, and 0 otherwise. Finally, $v_1^{\mathcal{F}}$ is obtained by subjecting v_1' to a 5×5 uniform low-pass filter (moving average) to reduce sensitivity to noise. Of course, this estimate is only evaluated inside $\mathcal{D}_1^{\mathcal{F}}$, since $v_1^{\mathcal{F}}$ is zero elsewhere.

Equation (3) can be easily extended to exploit color information by replacing each scalar $f[\mathbf{n}]$ with a color vector $\mathbf{f}[\mathbf{n}]$; the square operation in (3) becomes the square of the Euclidean length.

3.2. Multiscale Foreground Likelihood Estimation

To estimate the foreground likelihood $v_1^{\mathcal{F}}$, the multiscale approach needs only one other reference frame, say f_2, although Section 3.2.3 shows how the evidence from additional reference frames can be included into the estimated likelihood. The multiscale approach can also estimate $v_1^{\mathcal{F}}$ using the foreground motion information alone (i.e., without reference to the background motion), although Section 3.2.4 shows how the evidence from multiple motion models can be combined to improve $v_1^{\mathcal{F}}$.

The multi-scale approach is based on a Laplacian pyramid representation of f_1 and $f_{2\to 1}^{\mathcal{F}}$, as shown in Figure 2. Unlike the pyramid proposed by [11], the one used here cannot readily be used to reconstruct f_1 and $f_{2\to 1}^{\mathcal{F}}$ from their detail images $f_1^{(d)}$ and $f_{2\to 1}^{\mathcal{F},(d)}$, but requires fewer filtering operations. In this work, h_σ is a 7×7 Gaussian low-pass filter with $\sigma = 1.5$, which admits a very small amount of aliasing in exchange for a small region of support.

We estimate the foreground likelihood progressively, starting from a coarse scale and moving to the finest scale of the pyramid. At a coarse scale, small mismatches between $f_1^{(d)}$ and $f_{2\to 1}^{\mathcal{F},(d)}$ have little effect; therefore, it is quite possible that the whole domain is marked as foreground. As we progress from one scale to the next, finer details are discovered and the likelihood estimates are combined, using the method described in Section 3.2.2. To begin, though, we describe the generation of foreground likelihood estimates for one scale in isolation.

3.2.1. Intra-Scale Foreground Likelihood Estimation

We follow a probabilistic approach to foreground likelihood estimation. Let $\mathcal{F}$ denote the true foreground region. For any given scale d, we estimate a log-likelihood ratio $l^{(d)}[\mathbf{n}]$, representing the ratio between the probability that $\mathbf{n}$ belongs to $\mathcal{F}$ and the probability that $\mathbf{n}$ belongs to its complement $\overline{\mathcal{F}}$, conditioned on a set of observations $\Theta^{(d)}$. That is,

$$l^{(d)} = \log \frac{P\{\mathcal{F}|\Theta^{(d)}\}}{P\{\overline{\mathcal{F}}|\Theta^{(d)}\}} = \log \frac{P\{\Theta^{(d)}|\mathcal{F}\}}{P\{\Theta^{(d)}|\overline{\mathcal{F}}\}} \tag{4}$$

The second equality is based on the reasonable and widely used assumption that $P\{\mathcal{F}\} = P\{\overline{\mathcal{F}}\}$. The set of observations $\Theta^{(d)}$ are:

Ternary Feature Maps: A ternary feature map $T^{(d)}[\mathbf{n}]$ is derived from two quantities, $f_+^{(d)}[\mathbf{n}] = \max\{f^{(d)}[\mathbf{n}], 0\}$ and $f_-^{(d)}[\mathbf{n}] = \min\{f^{(d)}[\mathbf{n}], 0\}$, using

$$T^{(d)}[\mathbf{n}] = \begin{cases} 1, & f^{(d)}[\mathbf{n}] > (f_+^{(d)} * h_{\mathrm{T}})[\mathbf{n}] \\ -1, & f^{(d)}[\mathbf{n}] < (f_-^{(d)} * h_{\mathrm{T}})[\mathbf{n}] \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $h_{\mathrm{T}}[\mathbf{n}]$ is a symmetric low-pass FIR filter; here, we set $h_{\mathrm{T}} = 1.5h_\sigma$, $\sigma = 2$. The ternary feature maps are discussed in more detail in [12]. This step generates $T_1^{(d)}$ and $T_{2\to 1}^{\mathcal{F},(d)}$ from $f_1^{(d)}$ and $f_{2\to 1}^{\mathcal{F},(d)}$.

Structural Measures: We employ a novel measure $A^{(d)}[\mathbf{n}]$ of the local structure at each location $\mathbf{n}$ in a specific detail image $f^{(d)}[\mathbf{n}]$, whose values range from 0 (unstructured "noise") and 1 (highly structured image features). $A^{(d)}[\mathbf{n}]$ is derived from the degree of non-uniformity amongst the magnitudes of the DFT coefficients obtained over a small window centered about $\mathbf{n}$, within the ternary feature map. A much more comprehensive discussion of this structure measure and its properties is the subject of [12].

Quantization, typical in lossy compression, produces some artifacts which can have their own structural characteristics. We make the structural measure more robust to these artifact by burying them in noise that has a magnitude comparable to the expected amount of quantization. Specifically, for a given detail image $f^{(d)}$ that suffers from quantization noise with variance σ_q^2, we calculate a noisy detail image $f_n^{(d)}$ using

$$f_n^{(d)}[\mathbf{n}] = \begin{cases} f^{(d)}[\mathbf{n}], & |f^{(d)}[\mathbf{n}]| \geq \Delta_n \\ U(0,\sigma_n^2)[\mathbf{n}], & \Delta_n \geq |f^{(d)}[\mathbf{n}]| > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Here, $U(0,\sigma_n^2)$ is a uniformly-distributed pseudo-random number, over the interval $[-\Delta_n, \Delta_n]$, where $\Delta_n = \sqrt{3\sigma_n^2}$, and the variance σ_n^2 is set to $\alpha_n \sigma_q^2$. We find that $\alpha_n = 3$ gives a good compromise between removing artifacts and keeping relevant features, although values of α_n as small as 1 can also give acceptable results. We are interested in frames that have been compressed using JPEG2000 (e.g., for JPIP streaming applications); in this setting, σ_q^2 can be derived from the quantization parameters and number of discarded bit-planes for wavelet subbands whose passbands overlap that of the Laplacian detail band in question. For the results presented here, σ_q^2 is derived experimentally.

The noisy detail images $f_{n,1}^{(d)}$ and $f_{n,2\to1}^{\mathcal{F},(d)}$ are used to generate noisy ternary feature maps, which are then used to generate structural measures $A_1^{(d)}$ and $A_{2\to1}^{\mathcal{F},(d)}$. From these, we obtain the derived quantities A_μ, which identifies locations that are highly structured in both frames, and A_Δ, which identifies the structural dissimilarity between the frames; specifically,

$$A_\mu[\mathbf{n}] = A_1^{(d)}[\mathbf{n}] \cdot A_{2\to1}^{\mathcal{F},(d)}[\mathbf{n}], \quad A_\Delta[\mathbf{n}] = \frac{A_{2\to1}^{\mathcal{F},(d)}[\mathbf{n}]}{A_1^{(d)}[\mathbf{n}] + A_{2\to1}^{\mathcal{F},(d)}[\mathbf{n}]} \tag{7}$$

Shape Factor: We calculate a cross-correlation $\rho^{(d)}[\mathbf{n}]$ between corresponding regions in the ternary feature maps, using

$$\rho^{(d)}[\mathbf{n}] = \frac{\sum_{\mathbf{j}\in\mathcal{R}_\mathbf{n}} (T_1^{(d)}[\mathbf{j}] \cdot T_{2\to1}^{\mathcal{F},(d)}[\mathbf{j}])}{\sqrt{\sum_{\mathbf{j}\in\mathcal{R}_\mathbf{n}} (T_1^{(d)}[\mathbf{j}])^2 \cdot \sum_{\mathbf{j}\in\mathcal{R}_\mathbf{n}} (T_{2\to1}^{\mathcal{F},(d)}[\mathbf{j}])^2}} \tag{8}$$

where $\mathcal{R}_\mathbf{n}$ is a small region around location $\mathbf{n}$; we use a disc of radius $r = 3$. Since intensity information has already been removed from the ternary feature maps, we think of $\rho^{(d)}[\mathbf{n}]$ as an indication of *shape* similarity between the two frames in the neighbourhood of $\mathbf{n}$.

Power Ratio: A power ratio $M_\Delta^{(d)}[\mathbf{n}]$ is calculated using

$$M_\Delta^{(d)}[\mathbf{n}] = \frac{\sum_{\mathbf{j}\in\mathcal{R}_\mathbf{n}} (f_{2\to1}^{\mathcal{F},(d)}[\mathbf{j}])^2}{\sum_{\mathbf{j}\in\mathcal{R}_\mathbf{n}} (f_1^{(d)}[\mathbf{j}])^2 + \sum_{\mathbf{j}\in\mathcal{R}_\mathbf{n}} (f_{2\to1}^{\mathcal{F},(d)}[\mathbf{j}])^2} \tag{9}$$

This ratio gives an indication of how similar the power are in the two detail images.

Now, we are in a position to estimate the log-likelihood ratio, which is given by

$$l_1 = \log \frac{P\{\Theta|\mathcal{F}\}}{P\{\Theta|\overline{\mathcal{F}}\}} = \log \frac{P\{\rho, M_\Delta, A_\Delta, A_\mu|\mathcal{F}\}}{P\{\rho, M_\Delta, A_\Delta, A_\mu|\overline{\mathcal{F}}\}} \tag{10}$$

where the superscript (d) has been removed for brevity. To reduce complexity, we decompose the joint probability function in the numerator $P\{\rho, M_\Delta, A_\Delta, A_\mu|\mathcal{F}\}$ into the multiplication $P\{\rho|M_\Delta, A_\Delta, A_\mu, \mathcal{F}\}$, $P\{M_\Delta|A_\Delta, A_\mu, \mathcal{F}\}$, $P\{A_\Delta|A_\mu, \mathcal{F}\}$, and $P\{A_\mu|\mathcal{F}\}$. We similarly decompose the denominator. Then, we remove the conditionally independent terms, and assume that $P\{A_\mu|\mathcal{F}\} = P\{A_\mu|\overline{\mathcal{F}}\}$, to get

$$l_1 \approx \underbrace{\log \frac{P\{\rho|A_\mu, \mathcal{F}\}}{P\{\rho|\overline{\mathcal{F}}\}}}_{\text{shape}} + \underbrace{\log \frac{P\{M_\Delta|A_\mu, \mathcal{F}\}}{P\{M_\Delta|A_\mu, \overline{\mathcal{F}}\}}}_{\text{power}} + \underbrace{\log \frac{P\{A_\Delta|\mathcal{F}\}}{P\{A_\Delta|\overline{\mathcal{F}}\}}}_{\text{dissimilarity}} \tag{11}$$

The log-likelihood ratio becomes a combination of three terms. The shape term is sensitive to geometric features such as edge orientation, being based on the ternary features, and hence independent of the level of image contrast. The probability density functions (PDFs) $P\{\rho|A_\mu, \mathcal{F}\}$ and $P\{\rho|\overline{\mathcal{F}}\}$ are obtained experimentally using the Kodak test set. To collect statistics for $\mathcal{F}$, we match each image with noisy and slightly shifted copies of itself, while for $\overline{\mathcal{F}}$ we match each image with all other images in the set.

The power term is sensitive to differences in the amount of local image contrast, between the detail images, but relatively insensitive to shape. The dissimilarity term is sensitive to differences in the degree of apparent structure between the detail images. The PDFs for each of these terms are also derived experimentally.

3.2.2. Inter-Scale Foreground Likelihood Propagation

We combine the intra-scale log-likelihoods $l_1^{(d)}$ progressively, starting from the coarsest resolution and working towards the finest, forming combined log-likelihood ratios $s_1^{(d)}$, according to

$$s_1^{(d)} = l_1^{(d)} + \max\{1 - 1.5 \cdot A_\mu^{(d)}, 0\} \cdot (s_1^{(d+1)})_{\uparrow 2} \tag{12}$$

where $(\cdot)_{\uparrow 2}$ denotes upsampling by 2 and interpolation using a 7×7 cubic spline interpolator. The idea here is that if the scale under consideration has enough evidence (high A_μ value), then we can ignore the coarser resolution information; otherwise, we accumulate the information (multiply the probabilities) from the lower resolution.

The overall foreground likelihood for frame f_1 is obtained from $s_1^{(0)}$ using a simple transducer function, given by

$$v_1^{\mathcal{F}}[\mathbf{n}] = \begin{cases} 0, & s_1^{(0)}[\mathbf{n}] \leq 0 \\ s_1^{(0)}[\mathbf{n}]/5, & 5 \geq s_1^{(0)}[\mathbf{n}] > 0 \\ 1, & s_1^{(0)}[\mathbf{n}] > 5 \end{cases} \tag{13}$$

3.2.3. Extension to Multiple Components and Multiple Frames

Suppose the foreground likelihood for f_1 can be estimated using two other references frames, say f_0 and f_2. In this case, we have two sets of observations $\Theta_0^{(d)}$ between $f_{0\to1}^{\mathcal{F},(d)}$ and $f_1^{(d)}$ and $\Theta_2^{(d)}$ between $f_{2\to1}^{\mathcal{F},(d)}$ and $f_1^{(d)}$; therefore, the log-likelihood ratio is

$$l_{\Theta_0,\Theta_2}^{(d)} = \log \frac{P\{\mathcal{F}|\Theta_0^{(d)}, \Theta_2^{(d)}\}}{P\{\overline{\mathcal{F}}|\Theta_0^{(d)}, \Theta_2^{(d)}\}} = \log \frac{P\{\Theta_0^{(d)}, \Theta_2^{(d)}|\mathcal{F}\}}{P\{\Theta_0^{(d)}, \Theta_2^{(d)}|\overline{\mathcal{F}}\}} \tag{14}$$

Treating $\Theta_0^{(d)}$ and $\Theta_2^{(d)}$ as independent observations,

$$l_{\Theta_0,\Theta_2}^{(d)} = \log \frac{P\{\Theta_0^{(d)}|\mathcal{F}\} \cdot P\{\Theta_2^{(d)}|\mathcal{F}\}}{P\{\Theta_0^{(d)}|\overline{\mathcal{F}}\} \cdot P\{\Theta_2^{(d)}|\overline{\mathcal{F}}\}} = l_{\Theta_0}^{(d)} + l_{\Theta_2}^{(d)} \tag{15}$$

Evidently, this approach can be extended to any number of reference frames and/or color components. In practice, we sum the intra-scale log-likelihoods before combining them.

3.2.4. Including Background Information

For the case of a single foreground object considered in this work, we have $\overline{\mathcal{F}} = \mathcal{B}$. For each color component of each frame, the

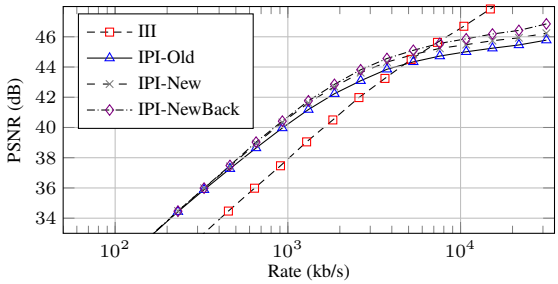

Fig. 3: A comparison of the performance of various schemes for the "Book" sequence. Note the logarithmic horizontal axis.

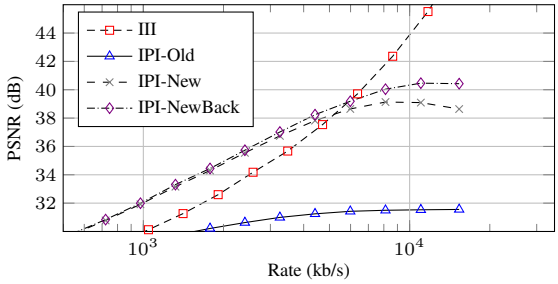

Fig. 4: A comparison of the performance of various schemes for the "Synthetic" sequence. Note the logarithmic horizontal axis.

background motion hint can provide a new set of observation. Consider the background observations, $\Theta_{\mathcal{B}}^{(d)}$, obtained between a pair of reference frames for which the foreground observations are $\Theta_{\mathcal{F}}^{(d)}$. Following the same line of argument as before, we have

$$l_{\Theta_{\mathcal{F}},\Theta_{\mathcal{B}}}^{(d)} = \log \frac{P\{\Theta_{\mathcal{F}}^{(d)}|\mathcal{F}\} \cdot P\{\Theta_{\mathcal{B}}^{(d)}|\mathcal{F}\}}{P\{\Theta_{\mathcal{F}}^{(d)}|\mathcal{B}\} \cdot P\{\Theta_{\mathcal{B}}^{(d)}|\mathcal{B}\}} = l_{\Theta_{\mathcal{F}}}^{(d)} - l_{\Theta_{\mathcal{B}}}^{(d)} \quad (16)$$

where $l_{\Theta_{\mathcal{B}}}^{(d)} = P\{\Theta_{\mathcal{B}}^{(d)}|\mathcal{B}\}/P\{\Theta_{\mathcal{B}}^{(d)}|\mathcal{F}\} = P\{\Theta_{\mathcal{B}}^{(d)}|\mathcal{B}\}/P\{\Theta_{\mathcal{B}}^{(d)}|\overline{\mathcal{B}}\}$.

Due to occlusion, special care should be exercised when there are two or more observations for the background. Occluded regions in the background have a negative log-likelihood, which can be misleading. With this in mind, if any background log-likelihood value is negative, we replace it with the smaller of 0 and the largest background log-likelihood value.

$$l_{\Theta_{\mathcal{B}}}^{(d)}[\mathbf{n}] = \begin{cases} l_{\mathcal{B}_1}^{(d)}[\mathbf{n}] + l_{\mathcal{B}_2}^{(d)}[\mathbf{n}], & l_{\mathcal{B}_1}^{(d)}[\mathbf{n}],\, l_{\mathcal{B}_2}^{(d)}[\mathbf{n}] \geq 0 \\ \max\{l_{\mathcal{B}_1}^{(d)}[\mathbf{n}], l_{\mathcal{B}_2}^{(d)}[\mathbf{n}]\}, & l_{\mathcal{B}_1}^{(d)}[\mathbf{n}] \cdot l_{\mathcal{B}_2}^{(d)}[\mathbf{n}] < 0 \\ 2 \cdot \max\{l_{\mathcal{B}_1}^{(d)}[\mathbf{n}], l_{\mathcal{B}_2}^{(d)}[\mathbf{n}]\}, & l_{\mathcal{B}_1}^{(d)}[\mathbf{n}],\, l_{\mathcal{B}_2}^{(d)}[\mathbf{n}] < 0 \end{cases} \quad (17)$$

In practice, we subtract the background log-likelihoods from the intra-scale log-likelihoods before combining them.

4. RESULTS

Here, we consider a JPIP server, serving a video sequence. For a given data rate, the server can send every frame to a client; we identify this option as "III." Alternatively, the server can send every other frame at a better quality, letting the client predict the missing frames. The client can use the image domain method of Section 3.1, identified as "IPI-Old." Alternatively, the client can use the multiscale method of Section 3.2, either with or without the background motion model, identified as "IPI-NewBack" and "IPI-New." All color components are used in all the approaches. For all prediction scenarios, frames f_{2i}, are delivered while frames f_{2i+1} are predicted. Foreground likelihoods are generated for each reference frame f_{2i}, using the surrounding reference frames f_{2i-2} and f_{2i+2}.

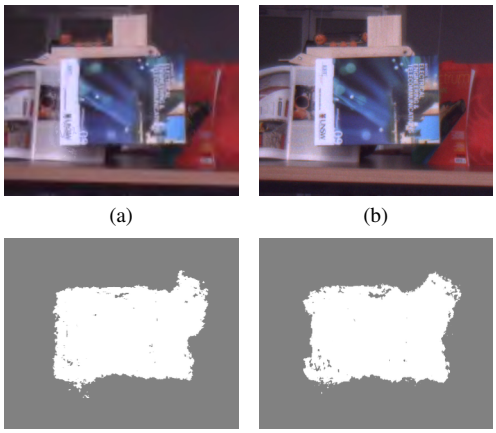

Fig. 5: Prediction using the IPI-NewBack method. (a) Predicted foreground object and its surrounding region at frame PSNR of around 39dB from the "book" sequence. (b) Original region at full quality. (c) One of the foreground likelihoods used in prediction. (c) The other likelihood.

We consider two video sequences, "book" and "synthetic". Both videos are 1024×768 compressed into JPEG2000 using Kakadu[2] to have 5 resolutions, 20 quality layers, and 32×32 codeblocks. The "book" is a real color sequence in YUV420 format shot at 25 frames/s; we are using 51 frames only. "synthetic" is a synthetic grey-scale sequence; we are using 45 frames and assuming 25 frames/s. In both sequences, the foreground object undergoes an affine motion. All results are reported in terms of luminance PSNR. The stated data rates include all encoded data but do not include JPEG2000 headers; moreover, the small cost of the quadrilaterals used to communicate the tracked foreground object is ignored – this is interpreted as tracking metadata that would be communicated for other reasons in a surveillance application.

Figures 3 and 4 show the performance of the various schemes for the "book" and "synthetic" test sequences. It can be seen that using prediction produces the best results up to around 45dB for the "book" and 39dB for "synthetic", beyond which it is better to send independently coded frames. It can also be seen that "IPI-NewBack" performs better than the other methods; especially, against "IPI-Old," which was proposed in our earlier work [1]. "IPI-New" also outperforms "IPI-Old," even though it does not use the background motion hint to determine the foreground likelihood.

Figure 5 shows predicted regions using "IPI-NewBack" from the "book" sequence when the quality is around 39dB. It can be seen that for the "book" sequence the quality is good.

5. CONCLUSIONS

In this work, we have demonstrated that using motion hints can produce good predictions if the motion can be described by these hints. These motion hints can be communicated using metadata. The proposed multi-scale approach works better than an earlier image domain approach; it is also more flexible in that it can utilize observations from other frames, other color components, and backgrounds. We find that the foreground likelihood estimates improve with the availability of more observations. In the context of browsing surveillance footage using a JPIP server, the proposed method can produce a better browsing experience without needing extra bandwidth.

[2]Kakadu Software ver. 7.1, `http://www.kakadusoftware.com/`

6. REFERENCES

[1] A.T. Naman, D. Edwards, and D. Taubman, "Efficient communication of video using metadata," *18th Proc. IEEE Int. Conf. Image Proc. 2011*, pp. 589–592, September 2011.

[2] A.T. Naman and D. Taubman, "JPEG2000-based scalable interactive video (JSIV) with motion compensation," *Image Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2650 –2663, September 2011.

[3] ISO/IEC 15444-9, "Information technology – JPEG 2000 image coding system – Part 9: Interactivity tools, APIs and protocols," 2004.

[4] D. Taubman and R. Prandolini, "Architecture, philosophy and performance of jpip: internet protocol standard for JPEG 2000," *Int. Symp. Visual Comm. and Image Proc.*, vol. 5150, pp. 649–663, July 2003.

[5] Demin Wang, Liang Zhang, and A. Vincent, "Motion-compensated frame rate up-conversion – part I: Fast multi-frame motion estimation," *Broadcasting, IEEE Transactions on*, vol. 56, no. 2, pp. 133 –141, June 2010.

[6] Demin Wang, A. Vincent, P. Blanchfield, and R. Klepko, "Motion-compensated frame rate up-conversion – part II: New algorithms for frame interpolation," *Broadcasting, IEEE Transactions on*, vol. 56, no. 2, pp. 142 –149, June 2010.

[7] M.T. Orchard, "Predictive motion-field segmentation for image sequence coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 3, no. 1, pp. 54 –70, February 1993.

[8] Jae Hoon Kim, A. Ortega, Peng Yin, P. Pandit, and C. Gomila, "Motion compensation based on implicit block segmentation," *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pp. 2452 –2455, October 2008.

[9] P. Smith, T. Drummond, and R. Cipolla, "Layered motion segmentation and depth ordering by tracking edges," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 4, pp. 479 –494, April 2004.

[10] K.A. Patwardhan, G. Sapiro, and V. Morellas, "Robust foreground detection in video using pixel layers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 4, pp. 746 –751, april 2008.

[11] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *Communications, IEEE Transactions on*, vol. 31, no. 4, pp. 532 – 540, apr 1983.

[12] A.T. Naman and D. Taubman, "A soft measure for identifying structure from randomness in images," *20th Proc. IEEE Int. Conf. Image Proc. 2013*, September 2013, Accepted for publication.