A semiautomatic saliency model and its application to video compression

Vitaliy Lyudvichenko, Mikhail Erofeev, Yury Gitman, Dmitriy Vatolin

***Abstract*—This work aims to apply visual-attention modeling to attention-based video compression. During our comparison we found that eye-tracking data collected even from a single observer outperforms existing automatic models by a significant margin. Therefore, we offer a semiautomatic approach: using computer-vision algorithms and good initial estimation of eye-tracking data from just one observer to produce high-quality saliency maps that are similar to multi-observer eye tracking and that are appropriate for practical applications. We propose a simple algorithm that is based on temporal coherence of the visual-attention distribution and requires eye tracking of just one observer. The results are as good as an average gaze map for two observers.**

While preparing the saliency-model comparison, we paid special attention to the quality-measurement procedure. We observe that many modern visual-attention models can be improved by applying simple transforms such as brightness adjustment and blending with the center-prior model. The novel quality-evaluation procedure that we propose is invariant to such transforms.

To show the practical use of our semiautomatic approach, we developed a saliency-aware modification of the x264 video encoder and performed subjective and objective evaluations. The modified encoder can serve with any attention model and is publicly available.

***Index Terms*—Eye-Tracking, Saliency, Video Compression, Visual Attention, x264.**

I. Introduction

Visual saliency modeling is a promising approach to improving the quality of many existing image- and video-processing applications, such as description [1], quality measurement [2], retargeting [3] and compression [4]. Among them, saliency-aware video compression probably delivers the most practical value.

According to Cisco's forecast [5], the amount of video in worldwide mobile traffic will increase 11-fold between 2015 and 2020 (from 55% to 75% of all traffic), whereas connection speeds are expected to grow just 3-fold by 2020. Globally, the spread of IPTV, especially video on demand, and new video formats (UHD and 360 video) will cause IP video traffic to consume up to 80% of all global IP traffic by 2019 [6]. The saliency-aware approach can compress videos efficiently by exploiting features of the human visual system without switching to a new video-encoding standard. Modern video codecs still don't take attention non-uniformity into account, so psycho-visual video optimization could be the main vector of further video-encoding research.

Major advances in saliency prediction resulted from image-saliency models. Nevertheless, in the case of video, what determines the regions of interest are the features of object movement rather than visual features of the objects themselves. Unfortunately, motion descriptors are an area of computer vision that is less studied than image descriptors, since they represent more-complex phenomena. Therefore, the quality of existing automatic approaches is worse than even single-observer eye tracking [7]. Acceptable saliency-map quality for practical video-processing applications is possible through explicit eye tracking of multiple observers and averaging of the collected results. Doing so, however, can be very expensive and time consuming.

On the other hand, constant frame changes prevent an observer from surveying all of the video content (unlike in the image case). Thus, forcing observers to focus on a single object leads to temporal coherence in the visual-attention distribution. In summary for the case of video, making a good initial saliency prediction is difficult (owing to the complex motion structure), but making further predictions or improving initial ones is simple when employing well-predictable inert temporal saliency structure.

Therefore, we propose a semiautomatic approach that trades off between two previous approaches. Our model is initialized using eye-tracking data from just one observer; this data is sufficiently accurate but temporally incoherent. A temporal-propagation algorithm enforces the temporal coherence of the model, whose quality thus becomes similar to eye-tracking data from two observers.

As in [8], we discovered that simple manipulations of saliency maps, such as brightness correction and addition of a center-prior image, can improve the quality of all tested saliency models. Therefore, we propose a method for finding the transformations that maximize the quality of a given saliency model. We employed this method in our comparison of video saliency maps predicted by 15 saliency models; it greatly improved the performance of all tested models. The comparison revealed that automatic models are only slightly better than the simplest center-prior model, making them impractical for saliency-aware video compression. We therefore believe our semiautomatic model is an optimal choice for such applications in terms of the benefit-to-expense ratio.

Previously, no state-of-the-art attention-aware video encoder was publicly available, even though such an encoder would be highly practical and a relatively straightforward implementation. This absence forces researchers to cobble together pipelines [4], [7], [9], [10] and use suboptimal implementations of the reference standard [4], [9]; in addition, it limits the fairness of the video-compression comparisons among different saliency models. Therefore, we propose a saliency-aware modification

of the x264 [11] encoder that enables anyone to effectively embed any visual-attention model into the compression pipeline. The encoder is publicly available from our project page at http://compression.ru/video/savam/.

We evaluated the proposed encoder and saliency model using a subjective experiment in which we obtained a 23% bit-rate savings compared with regular x264. Also, we present an objective evaluation of the encoder for other models.

The remainder of this discussion is organized as follows: In Section II we provide an overview of existing approaches to visual-attention modeling and saliency-aware compression. Section III introduces our semiautomatic visual-attention model. In Section IV we describe a method of model-transformation fitting and demonstrate its benefits by comparing 15 saliency models, and in Section V we present an attention-aware modification of the x264 encoder and evaluate it both subjectively and objectively.

II. Related work

A. Visual-attention modeling

To the best of our knowledge, no other research has attempted to construct saliency maps semiautomatically. Therefore, the most related efforts involve entirely automatic models of visual attention. All existing visual-attention models can be classified under two main approaches: bottom up and top down [12].

The bottom-up approach assumes the image properties drive attention. In [3], the saliency of a point is considered to be the uniqueness of a small surrounding area. The authors of [13] use the same definition of saliency, but they also perform postprocessing on the basis of pixel reciprocity and association of pixels into objects. In [14], *saliency* refers to the uniqueness of certain image frequencies and is extracted in the Fourier domain. This idea expands to the case of video in [4] through the use of a multiscale pyramid of quaternion Fourier transforms for the initial image and motion-strength maps. The authors of [15] propose a general algorithm to extract saliency from local image features. They transform the feature map into a Markov chain, marking the edges using a normalized measure of distinctiveness as well as the spatial distance between nodes. The saliency map is the equilibrium distribution obtained from the random-walk algorithm.

The top-down approach assumes the viewer's goals and experience are the main drivers of attention; thus, it requires recognition of objects familiar to human experience along with an understanding of their relationships. In our estimation, the most remarkable model of top-down attention appears in [16]. Here, the authors use face, person and car detection together with multiple bottom-up features to train a per-pixel SVM classifier. They then consider the distance to the SVM hyperplane to be the saliency value. Although their proposed approach obviously cannot consider complex spatial relationships, it nevertheless demonstrates high scores in different comparisons [8], [13]. Recent deep-learning advances increase the accuracy of object recognition and classification. In [17], the authors apply convolutional neural networks to detect salient objects, taking into account both the local and global features of images.

Although Yarbus in [18] describes the important role of top-down mechanisms in determining eye movements, these mechanisms remain poorly studied; at this point, corresponding models can only produce comparable results relative to bottom-up ones.

B. Saliency-based compression

The main idea of saliency-based compression is bit allocation in favor of salient regions. There are several implementations of this idea. We propose classifying them according to the following criteria:

- Model of visual attention underlying the method
- Reference encoder: MPEG-1 [9]; MPEG-4 [4], [9]; or H.264 [4], [10], [19]–[22]
- Method of bit-allocation control: implicit [4], [9], [10] (video preprocessing before encoding; e.g., non-uniform blur) or explicit (modifying internal encoder data; e.g., setting saliency-specific quantization-parameter (QP) values for macroblocks) [19]–[22]
- Evaluation methodology: either researchers can claim that videos encoded using their methods have lower bit rates than the reference video at the same visual quality [4], [9], [10], [19], [20] or they can conclude that their proposed encoders provide better visual quality than a reference at the same bit rate [21], [22]. We believe the second strategy is slightly more reliable because checking bit rates is easy, but confirming that two different videos have same visual quality is difficult.
- Method of visual-quality measurement: objective [20] or subjective [22]

In [9], the researchers propose a saliency-based video-compression framework based on the Itti-Koch-Niebur (IKN) saliency model. They followed a constant-quality, variable-bit-rate strategy and showed that compared with MPEG-1 and MPEG-4 encoders, application of nonuniform blur (guided by the saliency map) to the entire video before compression using the reference encoder significantly reduces the output bit rate. Improving on this work, [19] replaces nonuniform blur with explicit individual selection of the QP values for the macroblocks.

To achieve high time efficiency in the saliency-based video-compression framework that [20] proposes, the authors perform temporal propagation of the saliency map computed from a single frame and proceeding to successive frames. (In Section III we show how a similar propagation approach can improve the quality of the gaze map from a single observer.) The authors explicitly control bit allocation by setting individual QP values.

The method of bit allocation proposed in [21] attempts to avoid compression artifacts in non-salient regions that could grab viewer attention and thus change the initial saliency map.

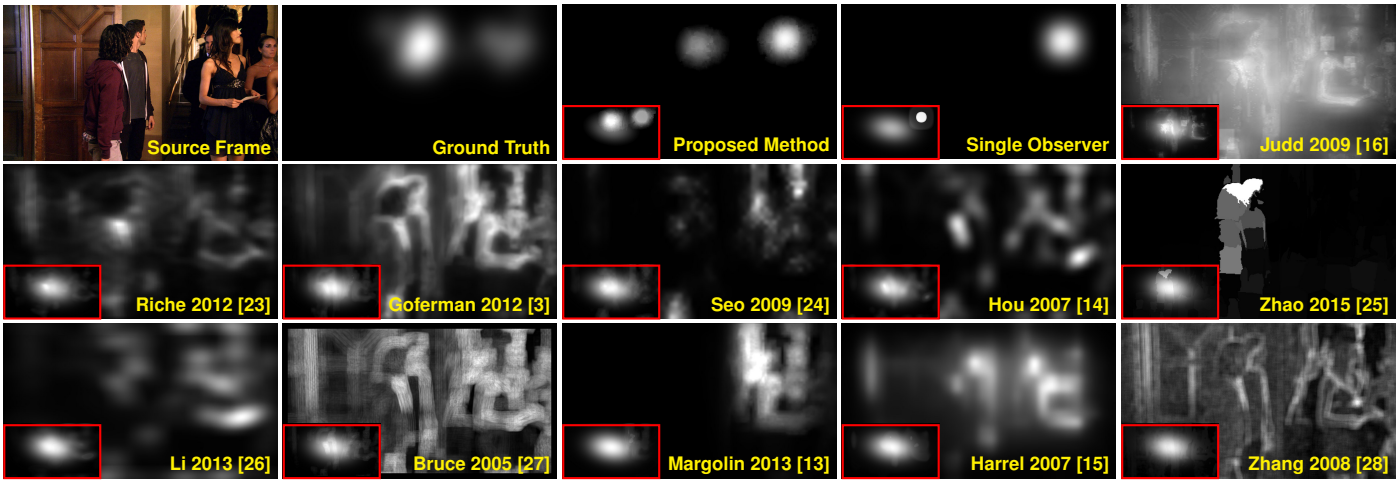

Fig. 1. Saliency maps predicted by different methods. The icons in the lower-left corners are the same images prepared for comparison (see Section IV). Histograms for all images are normalized for the sake of visibility. The weakness of automatic methods is clearly visible relative to eye tracking for even a single observer.

III. Proposed Visual-Attention Model

According to [8], existing visual-attention models offer little improvement over the center-prior model. Moreover, results presented in [7], along with our results (Figure 3), indicate that none of these models can compete with eye tracking, even for a single observer.

Because we intend to apply our saliency framework to video compression, we require high-quality saliency maps. The exact dependence of similarity to ground truth on the number of observers is obtained in [8]. In accordance with these results, the only way to achieve such high quality is to use the eye-tracking procedure for many observers—a very labor-intensive and unreasonable prospect.

One solution is a semiautomatic approach that uses fixation points from just one observer together with some postprocessing that employs spatiotemporal features of attention.

Most of the time the spatial distribution of visual attention is strongly nonuniform. At least the artistic content has several foci of attention, but everywhere else is significantly less salient. Confirming this observation is the fact that gazes from just a few observers make good saliency predictions [7], [8]. Despite the spatial nonuniformity, saliency maps have a high degree of temporal uniformity (e.g., the same object has similar saliency in adjacent frames). This phenomenon can be explained in the context of physiology. Thus, an observer's next eye movement can be determined by short-term memory of the scene, because human short-term memory retains a representation of the environment for some time [29].

On the basis of these observations, we employ the temporal uniformity of saliency to restore the nonuniform spatial structure using a temporal saliency-propagation algorithm:

$$\mathbf{R}_t = \beta \mathbf{P}_t^+ + (1-\beta)\mathbf{P}_t^-, \tag{1}$$

where $\mathbf{R}_t$ is t-th frame of the the propagation result and $\mathbf{P}_t^+$ and $\mathbf{P}_t^-$ are forward and backward terms, respectively, defined as follows:

$$\mathbf{P}_t^{\pm}(p) = \alpha \mathbf{P}_{t\mp1}^{\pm}(p + v_t^{\pm}(p)) + (1-\alpha)\mathbf{S}_t(p). \tag{2}$$

Here, $p \in \mathbb{R}^2$, $\mathbf{S}_t$ is a source sequence of saliency maps (acquired from single-observer eye tracking), $\vec{v}_t^{\pm}(p)$ is a motion vector field from $\mathbf{S}_{t\mp1}$ to $\mathbf{S}_t$, and α and β are algorithm parameters.

Computation of the vectors $\vec{v}^{\pm}(p)$ uses the motion-estimation algorithm described in [30]. This algorithm is faster than common approaches to computing optical flow because of its block structure, so the temporal propagation can be used for real-time encoding applications (the term $\mathbf{P}^-$ should be excluded in that case). A dense optical flow is unnecessary, because the generated saliency maps are intended for video compression at the block level.

This propagation technique is especially helpful for scenes with multiple saliency foci. Figure 1 shows an example. The technique also helps fill frames for which no fixation data has been collected because of blinking or saccades.

We used the training sequence [7] to estimate the dependence of saliency-map quality on the parameters α and β. The results, illustrated in Figure 2, show the method achieves the best quality with $\alpha = 0.8$ and $\beta = 0.45$. Such a high α value indicates that an output saliency map strongly depends on its adjacent frames. This observation is coherent with the temporal uniformity of visual attention. We used these optimal parameter values in all our experiments. It is worth noting that the estimated function in Figure 2 has a sole distinctive local maximum. In other words, the optimal parameter values for different observers are approximately the same; adjusting the parameters individually is unnecessary.

IV. Evaluation Methodology

As we mentioned, the simple center-prior model—which doesn't consider any image context—is only up to 10% worse than state-of-the-art saliency models (see the results on Figure 3). This small difference is because the quality of

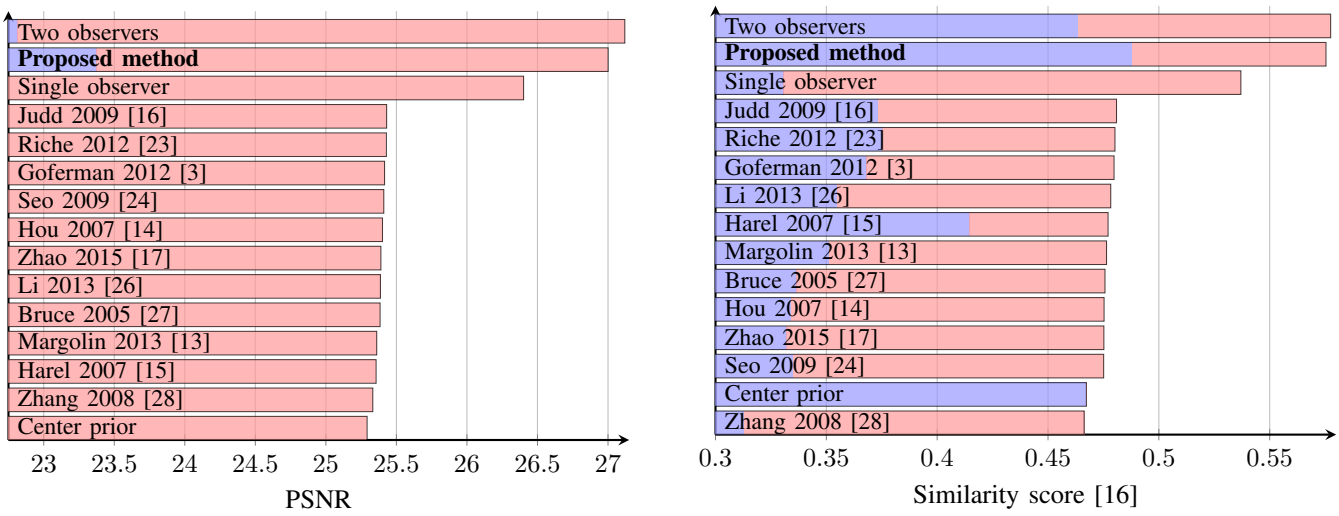

Fig. 3. Objective-evaluation results of visual-attention models using different measures. The blue parts of the bars show the initial model quality; the red parts show the cumulative increase after applying the simple transformations (see Section IV). These transformations yield a significant gain for automatic models as well as for models that use eye-tracking data. In particular, some automatic models can outperform others only after the fitting procedure. All automatic models, however, are still unable to compete with eye tracking, even for a single observer.

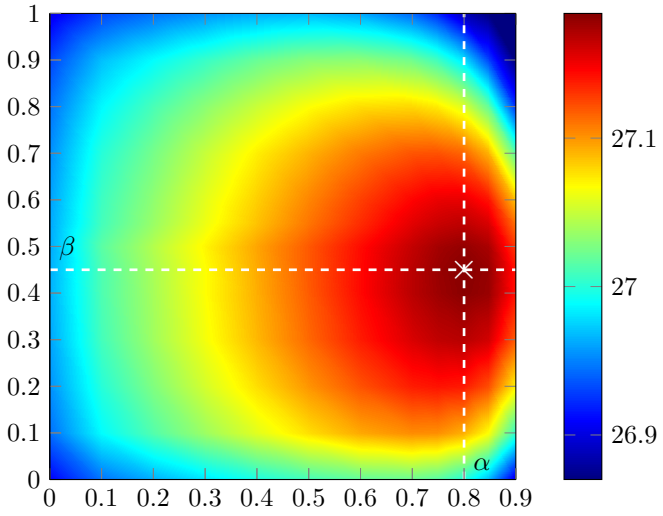

Fig. 2. Dependence of generated saliency-map quality on temporal-propagation parameters α and β for eight observers. The saliency maps come from the proposed algorithm (Section III) and are postprocessed according to the tuning procedure (Section IV-B). Estimates for the quality of the final saliency maps are based on the PSNR measure.

saliency maps predicted by various visual-attention models can be significantly improved using sequences of simple postprocessing transformations: brightness correction, blending with the center-prior image and blurring, as recent comparisons [7], [8] show. Consequently, such transformations contribute the most to the performance of modern saliency models. Practically, the choice of these transformations determines the model's advantage on a test video.

Good-enough transformations, however, are simply selected manually for each kind of video with known ground-truth saliency. Ultimately, simple transformations provide no additional benefit to visual-attention research, so straightforward model comparison using measures such as similarity score [16] is irrelevant. We believe the results of a fair comparison should be unaffected by these simple transformations.

An obvious way to achieve such behavior is to compute the optimal transformations for each model, maximizing its quality for a training data set. The transformed models can then be compared using common techniques, since no competitor employs the simple transforms better than any other. Note that the described procedure can also improve the quality of existing models and tune them for particular applications. It's especially useful for video compression in which each kind of artistic content has its own distinctive "saliency pattern."

The simple transformations discussed above can be represented by a vector of their parameters. The blending with the center-prior image is described by the blending weight and the blending image, depicted as a two-dimensional normal distribution. The brightness correction is determined by a monotonic function in gray scale, and blurring is determined by a Gaussian-filter kernel. Choosing the optimal parameters for a model, however, is a nontrivial global-optimization problem, because its cost function depends on an enormous amount of data (thousands of maps of predicted and ground-truth saliency). Therefore, it usually involves a computationally complex function with many local minima. Also, the decision space for this task has numerous dimensions: encoding just the brightness-correction function for 8-bit saliency maps requires at least 256 numbers. Some approaches to solving this problem have been proposed, however.

The authors of [7] reduce the parameter space to six numbers: five describe the brightness-correction function, and the sixth is the blending weight of the precomputed center-prior image. They use a gradient-descent algorithm to get suboptimal values for these parameters. Since the cost function has many local minima, they repeat the optimization procedure multiple times from different initial points in an attempt to find the global minimum. The evaluation methodology proposed in [8] employs the histogram-matching algorithm to estimate

brightness correction; it determines the radius of the Gaussian blurring filter and the weight of the center-prior image by performing an independent exhaustive search in a certain range. Neither approach can guarantee globally optimal parameters, however. Moreover, they have high computational complexity, because they each employ some kind of exhaustive search for the complex cost function.

For our evaluation we developed a method that for any saliency model can find the exact globally optimal blending weight and brightness-correction function simultaneously in terms of MSE.

A. Proposed evaluation method

If we consider a method of choosing optimal transformations [7] and substitute the similarity-score (SS) measure [16] in its cost function along with the MSE, the saliency maps of the transformed models will remain visually similar. Since MSE is a quadratic function, this substitution simplifies the structure of the cost function, but it still contains strong nonlinear dependencies resulting from the complex parameterization of the brightness-correction function. Fortunately, the naïve parameterization that maps 256 values (most saliency maps are stored with 8-bit depth) allows us to eliminate these nonlinearities and reduce the optimization task to the quadratic programming problem. Moreover, increasing the number of introduced parameters improves generalization performance of the transformation.

We omit the blurring step because it significantly complicates the cost function and provides an insignificant gain [8].

B. Formal method description

Let $\mathbf{SM}^i$ and $\mathbf{GT}^i$ be the respective predicted and ground-truth saliency maps for the i^{th} frame, and let $\mathbf{CP}$ be a precomputed center-prior image. The blending weight is denoted by β, and $p \in \mathbb{R}^2$ is the pixel position. The cost function is then

$$C(\beta, M) = \sum_{i,p} \left(M(\mathbf{SM}^i_p) + \beta \mathbf{CP}_p - \mathbf{GT}^i_p \right)^2, \qquad (3)$$

where $M\colon \mathbb{R}^+ \to \mathbb{R}^+$ is the brightness-correction function. Since we store saliency maps as 8-bit images, we can represent this function by the vector $\mathbf{m} \in \mathbb{R}^N$, $N = 256$, such that $M(s) = \mathbf{m}_s \in \mathbb{R}^+$ for any saliency value $s \in \mathbb{N}^+$, with $s < N$. Then we can consider C to be the quadratic function of real-valued arguments β and $\mathbf{m}$, so the solution of the following quadratic programming problem should yield the optimal parameter values:

$$(\beta, \mathbf{m}) = \underset{\substack{\beta>0 \\ 0<\mathbf{m}_i<\mathbf{m}_{i+1}}}{\arg\min} \sum_{i,p} \left(\mathbf{m}_{\mathbf{SM}^i_p} + \beta \mathbf{CP}_p - \mathbf{GT}^i_p \right)^2. \qquad (4)$$

The constraints guarantee that the brightness-correction function $\mathbf{m} \in \mathbb{R}^N$ is monotone. This task has a canonical matrix form:

$$(\beta, \mathbf{m}) = \underset{\substack{\mathbf{x}_1>0 \\ 0<\mathbf{x}_i<\mathbf{x}_{i+1}, i>1}}{\arg\min} \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{f}^T\mathbf{x} + c, \qquad (5)$$

where $\mathbf{x} = (\beta, \mathbf{m}) \in \mathbb{R}^{N+1}$ contains the target parameters; matrix $\mathbf{H} \in \mathbb{R}^{(N+1)\times(N+1)}$, vector $\mathbf{f} \in \mathbb{R}^{N+1}$ and scalar c derive explicitly from (3) and define the optimization task. $\mathbf{H}$ is a Hermitian sparse matrix containing nonzero elements only in the first row, first column and main diagonal. Moreover, $\mathbf{H}$ is a positive-definite matrix, so the quadratic programming task is convex and easily solved using any one of numerous approaches (we use the interior-point method).

As in [7], we precompute the center-prior image $\mathbf{CP}$. We treat $\mathbf{CP}$ as the two-dimensional normal distribution and estimate its covariance matrix and mean vector from the ground-truth distribution $\mathbf{GT}$ averaged across all frames.

C. Method summary

Our proposed evaluation method is simple and has low computational complexity. The algorithm makes only two passes over the input sequences to compute the final score. On the first pass, it reads $\mathbf{GT}^i$ to compute $\mathbf{CP}$. On the second pass, it reads $\mathbf{SM}^i$ and $\mathbf{GT}^i$ to compute $\mathbf{H}$, $\mathbf{f}$ and c from (5) and to estimate the optimal parameters. Both passes consume only $O(N)$ memory and involve no extra computations.

Solving the task described in (4) is fast and a solution is globally optimal because the expression for the cost function uses a compact matrix representation (5) and has a simple convex structure. We compute the final MSE by substituting the optimal parameters into the compact presentation of the cost function; no extra passes are required. We convert the MSE to the more intuitive PSNR and use it as the final score.

The low complexity of the transformations allows us to improve compression of saliency-aware artistic video content. Fitting the simple transformations of an underlying saliency model requires a small amount of ground-truth saliency data for the target content (a TV show or film). Afterwards, the remaining portions of the target content can be compressed more efficiently using the transformed model because of that model's greater quality.

The source code for the method, along with a detailed derivation of (5), is available at the project page.

D. Comparison results

Figure 3 shows the comparison results for 15 saliency models, including 12 automatic methods; Figure 3 shows example model predictions. We used a dataset from [7] divided in half to form a training part and a test part. Despite training the transformation parameters on isolated data, these parameters still deliver a significant improvement on the test video for all models. The substantial benefits and low complexity of our transformation-tuning algorithm motivated us to include it as a postprocessing step in our model for saliency-aware compression (Section V).

As we have shown, the results from single-observer eye tracking are significantly better than those from automatic saliency models. Note that the model proposed by Judd et al. [8] outperformed "single observer" on an image-saliency-prediction benchmark. Nevertheless, our comparison showed the opposite result for video sequences. This situation demonstrates the

power of motion features and temporal coherence in predicting saliency. The correctness of our comparison goals is confirmed by the fact that applying the transformations causes all automatic models to produce similar quality and significantly changes their ranking. Worth noting is that our proposed method (using data from a single observer) offers quality similar to that of two-observer eye tracking.

V. Saliency-aware encoding

In this section we describe a saliency-aware modification of the x264 [11] encoder that we used for our compression experiments; this modification is publicly available. We also address the question of automatic quality measurement and present both objective and subjective evaluations of our semiautomatic model's performance on video compression.

A. Video-Encoder Implementation

The main idea of saliency-based compression is clear—allocating bits in favor of salient regions—yet the exact number of bits that should be transferred to the region of interest is a controversial question. Moreover, the compression artifacts that bit reallocation introduces can change the saliency distribution [31].

Ideally, a saliency favoring bit allocation should maximize the perceptual quality of the output video. But automatic objective measures such as the widely used SSIM and PSNR are pixel or patch based and are unable to capture the uneven attention distributions that are critical for proper saliency-aware compression.

Our modification of the x264 encoder has two additional command-line parameters, p and b, controlling the amount of bit allocation in salient regions. The result is that each frame macroblock whose saliency is below the p^{th} percentile receives b percent of the bit rate. In other words, $100-p$ percent of most salient pixels in a frame receive $100-b$ percent of the bit rate. Such parameters are easy to use, yet their implementation is nontrivial given that unless the compression uses multiple passes, the encoder cannot determine the required bit rate for a given frame. We can instead, however, roughly evaluate the resulting bit rate as a sum of macroblock-size predictions based on their quantizers. Figure 4 shows the empirical estimation of a macroblock size with respect to a quantization parameter for x264's "constant quantizer" mode over 15 test videos.

Let $\mathbf{Q}\colon \mathbb{R}^2 \to \mathbb{R}^+$ be a quantizer map of the currently processed frame estimated by the unmodified x264, $\mathbf{Q}'\colon \mathbb{R}^2 \to \mathbb{R}^+$ be a saliency-aware quantizer map, $\mathbf{B}\colon \mathbb{R}^+ \to \mathbb{R}^+$ be a function of macroblock-size prediction (Figure 4), $\mathbf{S}\colon \mathbb{R}^2 \to [0;1]$ be a saliency map downscaled to the resolution of $\mathbf{Q}$, and s_p be the value of its p^{th} percentile, where $\mathbf{SP} = \max(\mathbf{S}-s_p, 0)$ and $\mathbf{SN} = \max(s_p-\mathbf{S}, 0)$. Then, applying the above definitions for p and b and considering that $\mathbf{SP}$ and $\mathbf{SN}$ are defined up to linear scaling, we obtain the following system for $\mathbf{Q}'$:

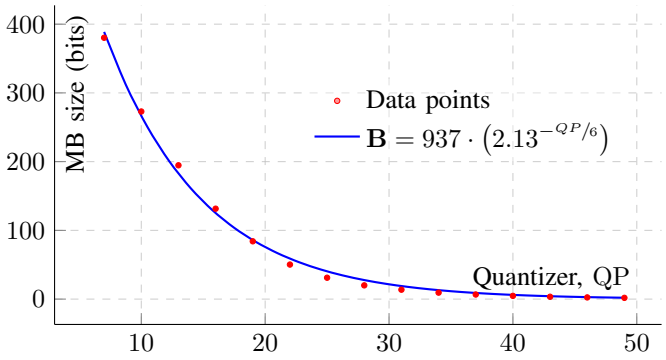

Fig. 4. Empirically estimated relationship between the quantization parameter and the encoded-macroblock (MB) size. The y-axis is a mean MB size for 15 test videos encoded in "constant quantizer" mode.

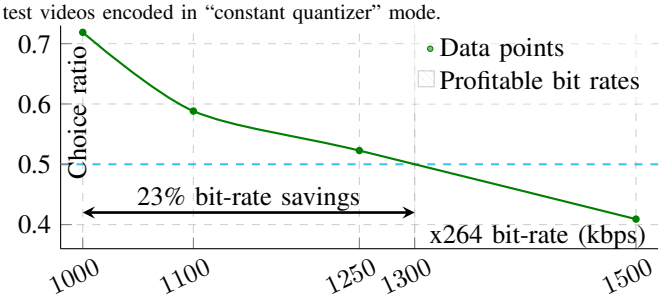

Fig. 5. Subjective comparison of saliency-aware encoding and the original x264. We asked 21 participants to select the better-quality video from the given pair. The chart reflects the ratio of choices favoring 1000 kbps saliency-aware compression over uniform-attention compression with respect to the latter bit rate. Note that the hatched bit-rate range for which saliency-aware encoding has better perceptual quality despite the lower bit rate depends sufficiently on the choice b. Here we conservatively choose $p = 80\%$ and $b = 70\%$ (i.e., 20% of the most salient pixels receive 30% of the bit rate).

$$\begin{cases} \mathbf{Q}' = \mathbf{Q} + \alpha\mathbf{SP} - \beta\mathbf{SN} \\ \displaystyle\sum_{i,j:\mathbf{SN}>0} \mathbf{B}(\mathbf{Q}'_{i,j}) = \frac{b}{100} \sum_{i,j} \mathbf{B}(\mathbf{Q}_{i,j}) \\ \displaystyle\sum_{i,j:\mathbf{SP}\geq 0} \mathbf{B}(\mathbf{Q}'_{i,j}) = \left(1 - \frac{b}{100}\right) \sum_{i,j} \mathbf{B}(\mathbf{Q}_{i,j}). \end{cases} \tag{6}$$

Now we can get an explicit expression for $\mathbf{Q}'$ by calculating α and β from the last two equations and substituting them into the first.

B. Subjective quality evaluation

Because rate-distortion curves have a logarithmic nature, only low-bit-rate encoding can benefit from attention modeling. Otherwise, the quality increase in the region of interest will be negligible compared with the distortion introduced outside that region. Selection of the highest bit rate for which saliency-aware compression still makes sense is an open research problem that we do not address in this paper.

In this section we also provide a subjective evaluation of our proposed saliency-aware encoder relative to the underlying x264 encoder; we chose a fixed-quality/variable-bit-rate strategy. Because our quality estimations are comparative rather than absolute, however, we were unable to obtain the entire rate-distortion curve for a reasonable number of comparisons. Thus,

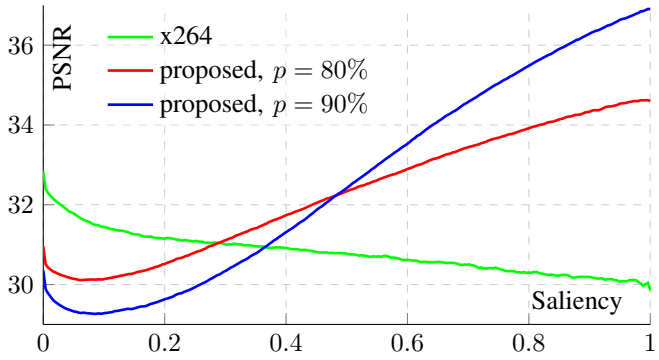

Fig. 6. Saliency–quality curve for x264 [11] and two presets of our saliency-aware encoder. Even though x264 tends to allocate bits uniformly, its curve is decreasing, since regions with complex structure and motion attract more attention. The main parts of the saliency-aware-encoder curves are increasing. The preset $p = 90\%$ allocates more bits to salient regions than to other regions, so its quality curve has a steeper slope.

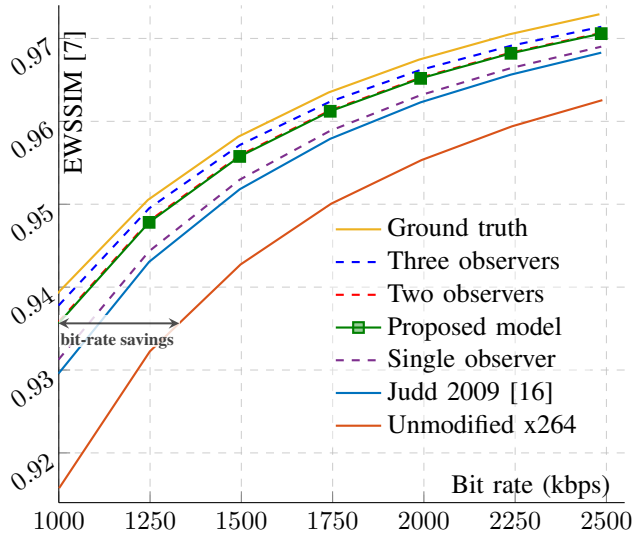

Fig. 7. Objective evaluation of our visual-attention model applied to video compression. Note that the actual bit-rate gain depends on the distortion of the non-salient region, and the optimal choice is a matter for future research. Here we conservatively chose $p = 80\%$ and $b = 70\%$ for all models.

we restricted saliency-aware encoding to a 1000 kbps bit rate and varied the original x264 bit rate until we achieved equal quality.

We showed Amazon Mechanical Turk participants a sequence of video pairs, and for each pair we asked them to choose the video with better quality or to indicate that the videos are approximately equal. We paid participants \$0.05 for 12 pairs, 2 of which were hidden quality-control comparisons between x264 videos with 1000 kbps and 2500 kbps bit rates. To accept the data from a given individual, we required correct choices for both control comparisons. We tested the encoders on 12 video sequences ranging from 16 to 20 seconds in duration and compared four x264 bit rates from 1000 kbps to 1500 kbps with saliency-aware compression, for which we used saliency maps estimated by our method and eye-tracking data from a single observer with the best ground-truth prediction.

Figure 5 shows the choice ratio for saliency-aware videos (with a fixed 1000 kbps bit rate) instead of x264 videos for each of four bit rates. Interpolating them, we denoted the equal-quality point by computing the bit rate for the x264 videos at which their subjective quality should match that of the saliency-aware videos. We acquired the above-mentioned figure during the subjective experiment in which the proposed encoder parameters were $p = 80\%$ and $b = 70\%$; its equal-quality point is at 1300 kbps, so our proposed encoder and saliency model can save 23% of the bit rate (300 kbps).

In total, we conducted 12 analogous experiments with different encoder parameters and different participants; our data corresponds to the experiment in which we obtained the maximum bit-rate gain. All experiments involved 346 participants, yielding 3460 pairwise comparisons.

C. Objective quality evaluation

To objectively evaluate the quality of our attention-based compression, we chose the simple EWSSIM metric [7] which is a weighted sum of per-pixel SSIM values using ground-truth saliency as weights. Figure 7 presents rate-distortion curves for different models. Note that the ranking of the models by EWSSIM and the similarity to ground-truth saliency (Figure 3) are alike. In particular, our model has the same quality as the average "two observers" model. As in Section IV-D, we performed the evaluation on the same test part of the data set using the same model transformations.

We chose the conservative compression-parameter values $p = 80\%$ and $b = 70\%$ and received 24% bit-rate savings according to the EWSSIM metric versus 23% according to the subjective evaluation using the same parameters (see the per-frame examples our project page). But these parameters are not EWSSIM optimal (setting $p = 85\%$ will yield a higher gain). Moreover, we found that accurate models benefit from more-aggressive presets. For instance, a preset of $p = 85\%$ and $b = 55\%$ is optimal for ground-truth saliency but one of the worst for automatic models. Unfortunately, EWSSIM-optimal presets distort non-salient regions and degrade the visual-quality perception of a sequence. A saliency-aware full-reference metric that correlates highly with subjective evaluations is necessary to automatically adjust the attention-based compression, but it is a topic of further research.

Since the proposed encoder implicitly controls quality by linearly changing macroblock quantizers, we performed a series of tests to validate the encoder and estimate the dependence of pixel distortions on pixel saliency. We fixed b and the target bit rate, then made a set of compressed videos using ground-truth saliency and various p values in the 75–99% range.

First we checked that the actual video bit rates deviate from the target bit rate by no more than 1%. Then, in each video we grouped pixels by their saliency value and computed the average PSNR for each group. We thus estimated how pixel distortions depend on saliency and p. Figure 6 shows this dependence for two different p values (and for x264), it confirms that quality increases with saliency and that lower p values produce more-uniform distortions. Despite compression uniformity, the x264 curve is strictly decreasing, since more-complex regions are

usually more salient. Also, the least-salient regions contain a lot of flat and efficiently encoded blocks, which explains why the curves for the modified encoder decrease initially. The remaining part of these curves, however, is increasing and can be considered linear in most practical cases and has a steeper slope for more-aggressive presets.

VI. Conclusion and Future Work

In this paper we introduce a novel method for saliency-map estimation using postprocessing of eye-tracking data for a single observer. Our objective comparison shows that our proposed method significantly outperforms other visual-attention models and that its quality is as good as that of the average "two observers" model.

For this research we paid special attention to the quality-measurement procedure. Having managed to improve many modern visual-attention models by applying brightness correction and "center prior," we designed the evaluation pipeline to eliminate the effect of these simple transformations.

To show the practicality of our model, we modified the x264 video encoder and added saliency-map support. The encoder is publicly available for other studies.

Also, both objective and subjective evaluations of the encoder performance were conducted in which our attention model gave 23% bitrate savings in comparison with regular x264 and showed the same performance as eye-tracking from two observers.

Also, we conducted both objective and subjective evaluations of encoder performance; for these evaluations, our attention model yielded 23% bit-rate savings compared with regular x264 and showed the same performance as eye tracking for two observers. We plan to enhance the temporal-propagation algorithm using recurrent neural networks and integrate it into modern video codecs. Also, we believe development of a full-reference metric for saliency-aware video quality derived from subjective evaluations is promising, since it would enable more-optimal bit allocation and perform more-representative objective comparisons.

References

[1] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros, "Data-driven visual similarity for cross-domain image matching," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 6, p. 154, 2011.

[2] D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, and D. Kukolj, "Salient motion features for video quality assessment," *IEEE Transactions on Image Processing (TIP)*, vol. 20, no. 4, pp. 948–958, 2011.

[3] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 10, pp. 1915–1926, 2012.

[4] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing (TIP)*, vol. 19, no. 1, pp. 185–198, 2010.

[5] Cisco Visual Networking Index, "Global mobile data traffic forecast update, 2015–2020," *Cisco white paper*, 2016.

[6] C. V. N. Index, "The zettabyte era–trends and analysis," *Cisco white paper*, 2014.

[7] Y. Gitman, M. Erofeev, D. Vatolin, B. Andrey, and F. Alexey, "Semiautomatic visual-attention modeling and its application to video compression," in *International Conference on Image Processing (ICIP)*, 2014, pp. 1105–1109.

[8] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Tech. Rep., 2012.

[9] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing (TIP)*, vol. 13, no. 10, pp. 1304–1318, 2004.

[10] S.-P. Lu and S.-H. Zhang, "Saliency-based fidelity adaptation preprocessing for video coding," *Journal of Computer Science and Technology*, vol. 26, no. 1, pp. 195–202, 2011.

[11] "x264 software video encoder," http://www.videolan.org/developers/x264.html.

[12] M. Land and B. Tatler, "How our eyes question the world," in *Looking and Acting: Vision and Eye Movements in Natural Behaviour*. Oxford University Press, 2009.

[13] R. Margolin, L. Zelnik-Manor, and A. Tal, "Saliency for image manipulation," *The Visual Computer*, pp. 1–12, 2013.

[14] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[15] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 19, 2007, pp. 545–552.

[16] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *International Conference on Computer Vision (ICCV)*, 2009, pp. 2106–2113.

[17] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1265–1274.

[18] A. L. Yarbus, *Eye movements during perception of complex objects*. Springer, 1967.

[19] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image and Vision Computing*, vol. 29, no. 1, pp. 1–14, 2011.

[20] R. Gupta, M. T. Khanna, and S. Chaudhury, "Visual saliency guided video compression algorithm," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1006–1022, 2013.

[21] H. Hadizadeh and I. Bajic, "Saliency-preserving video compression," in *International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1–6.

[22] H. Hadizadeh, "Visual saliency in video compression and transmission," Ph.D. dissertation, School of Engineering Science, Simon Fraser University, 2013.

[23] N. Riche, M. Mancas, B. Gosselin, and T. Dutoit, "RARE: A new bottom-up saliency model," in *International Conference on Image Processing (ICIP)*, 2012, pp. 641–644.

[24] H. J. Seo and P. Milanfar, "Nonparametric bottom-up saliency detection by self-resemblance," in *IEEE International Workshop on Computer Vision and Pattern Recognition*, 2009, pp. 45–52.

[25] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of Vision*, vol. 11, no. 3, 2011.

[26] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 4, pp. 996–1010, 2013.

[27] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems (NIPS)*, 2005, pp. 155–162.

[28] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 1–20, 2008.

[29] M. P. Aivar, M. M. Hayhoe, C. L. Chizk, and R. E. B. Mruczek, "Spatial memory and saccadic targeting in a natural task," *Journal of Vision*, vol. 5, no. 3, 2005.

[30] K. Simonyan, S. Grishin, D. Vatolin, and D. Popov, "Fast video super-resolution via classification," in *International Conference on Image Processing (ICIP)*, 2008, pp. 349–352.

[31] X. Min, G. Zhai, Z. Gao, and C. Hu, "Influence of compression artifacts on visual attention," in *International Conference on Multimedia and Expo (ICME)*, 2014, pp. 1–6.