Classification-Aware Distortion Metric for HEVC Intra Coding

Massimo Minervini #1, Sotirios A. Tsaftaris *#2

# IMT Institute for Advanced Studies, Lucca, Italy
1 m.minervini@imtlucca.it

* University of Edinburgh, Edinburgh, UK
2 stsaft@gmail.com

***Abstract*—Increasingly many vision applications necessitate the transmission of acquired images and video to a remote location for automated processing. When the image data are consumed by analysis algorithms and possibly never seen by a human, tailoring compression to the application is beneficial from a bit rate perspective. We inject prior knowledge of the application in the encoder to make rate-distortion decisions based on an estimate of the accuracy that will be achieved when analyzing reconstructed image data. Focusing on classification (e.g., used for image segmentation), we propose a new application-aware distortion metric based on a geometric interpretation of classification error. We devise an implementation for the High Efficiency Video Coding standard, and derive optimal model parameters for the λ-domain rate control algorithm by curve fitting procedures. We evaluate our approach on time-lapse sequences from plant phenotyping experiments and cell fluorescence microscopy encoded in intra-only mode, observing a reduction in segmentation error across bit rates.**

I. Introduction

In present days, more often than not image data are analyzed by computer vision algorithms and their transmission over channels necessitates their compression to reduce bandwidth costs. Progress in vision and automation technology is leading to increasingly many applications in which images are acquired, transmitted, and analyzed, largely without any human intervention (e.g., traffic video surveillance, industrial inspection, remote sensing, precision agriculture, robot navigation). It was shown recently that considering the application and designing data codecs appropriately not to optimize fidelity criteria (e.g., mean squared error) or psycho-visual criteria (e.g., structural similarity), but considering how would an analysis algorithm perform on compressed data, is beneficial from a bit rate perspective [1]. For example, several vision tasks (e.g., object detection, segmentation, image retrieval) can be formulated as classification problems, and early studies on the effects of lossy compression on classification can be found in [2].

In this paper, we propose to guide encoding decisions (and corresponding bit allocation) explicitly within a rate-distortion (R-D) framework, to maximize post-compression classification accuracy. We devise a distortion metric which aims to focus bit budget on points that affect most classification accuracy (Figure 1). Our work is the first to demonstrate this concept within a standard compliant context, in the recent High Efficiency Video Coding (HEVC) standard [3].

Image and video coding schemes typically split the input signal into coding units (CUs), corresponding e.g. to non-overlapping image blocks or sub-bands. While in early compression standards all CUs were encoded using the same settings, recent approaches offer the possibility to select different coding options for each CU (e.g., quantizer, CU size, prediction mode), that will result in different rate requirements and levels of distortion. In an operational R-D optimization framework, the resource allocation problem at the encoder is addressed by minimizing a distortion metric D between original $\mathbf{X}$ and reconstructed $\hat{\mathbf{X}}$ image while satisfying a bit rate constraint:

$$\underset{\Theta}{\operatorname{minimize}}\, D(\mathbf{X}, \hat{\mathbf{X}}; \Theta) \text{ subject to } R(\mathbf{X}; \Theta) \le R_{\mathrm{tot}}, \qquad (1)$$

where $R(\mathbf{X}; \Theta)$ is the output rate obtained using parameters Θ, and R_{tot} is available bit budget. In a Lagrangian formulation, optimal parameters Θ^* that minimize the so-called R-D cost are found solving the unconstrained problem:

$$\Theta^* = \arg\min_{\Theta} D(\mathbf{X}, \hat{\mathbf{X}}; \Theta) + \lambda R(\mathbf{X}; \Theta), \qquad (2)$$

where $\lambda \ge 0$ sets the trade-off between rate and distortion. To reduce perceived visual distortion while maintaining computational efficiency, distortion metrics akin to the mean squared error are routinely adopted by general purpose video encoders. To focus bits in critical for the classifier regions, here we inject prior knowledge in the R-D optimization process via a new distortion metric based on a geometric interpretation of compression error in relation to decision boundary (Figure 1).

A distortion metric based on the Kullback-Leibler divergence is used in [4] to optimize scalar quantization of synthetic signals for classification at the decoder. The joint design of vector quantization and classification has been widely investigated, e.g., combining squared error and a penalty for misclassification based on a Bayes risk term [5], 2-D hidden Markov model of image blocks [6], or Hamming distortion [7]. Chao et al. [8] modify rate control of H.264 to preserve SIFT features. Pu et al. [9] define a distortion metric based on conditional class entropy to tailor JPEG 2000 to a target detection task. However, such investigations either are aimed at quantizer design and the approaches do not comply to any compression standard, or focus on optimizing (the compression of) related features for detection and tracking tasks with potential larger computational requirements or modifications on the encoder side. On the other hand, our application-aware R-D optimization approach focuses on general classification (which can be used also as part of detection/tracking) and is compliant with the HEVC standard, adding limited computational overhead on the encoding side.

In Section II we define our proposed distortion metric, and in Section III we discuss its implementation in the λ-domain rate control algorithm [10] of HEVC. We focus on intra-only encoding mode, which ensures efficient access to any frame of a sequence. We validate our approach on a time-lapse sequence arising from plant phenotyping experiments (the task is to delineate plant objects from background). We adopt this application because due to design requirements lightweight sensors and significant compression levels are necessary [11]. We also test our approach on a sequence from fluorescence microscopy where the goal is to segment moving cells [12]. In Section IV, we observe a reduction in segmentation error across bit rates, when compared to the baseline approach based on fidelity criteria. Finally, Section V offers concluding remarks.

II. Classification-Aware Distortion Metric

Prior to defining our proposed distortion metric for classification tasks, we discuss the relation between compression and classification. For simplicity we will focus on binary classification (e.g., object segmentation) and linearly separable data, although generalizations to nonlinear and multi-class problems are possible.

Let $\delta_H : \mathcal{X} \to \{0,1\}$ be the discriminant function of a linear classifier, where $H = \{\mathbf{x} \in \mathcal{X} \,|\, \mathbf{w}^\mathsf{T}\mathbf{x} + w_0 = 0\}$ is the decision boundary defined in the pixel domain $\mathcal{X}$. In a distributed sensing and analysis framework, the classifier δ_H represents a *surrogate* of the application at the receiver, which may involve more sophisticated vision algorithms and would be difficult to use directly at the encoder (it would be computationally inefficient and also less general, since each application will require a customized solution). Thus, the $\mathbf{w}$, w_0 parameters are the *prior knowledge* of the application available at the encoder, which can be e.g. fixed prior to sensor deployment, or estimated at receiver and communicated (few bytes) to sensor.

Figure 1 shows a graphical representation of a pixel intensity (or in general feature) space $\mathcal{X}$, and several reconstructions of an original point $\mathbf{x}$ resulting from different coding options (which in turn may be associated with different rates). We aim to show that Euclidean choices to distortion optimization (e.g., mean squared error) can lead to undesired classification error. In this example, $\hat{\mathbf{x}}_2$ and $\hat{\mathbf{x}}_3$ are better (in a Euclidean sense) approximations of $\mathbf{x}$ than $\hat{\mathbf{x}}_1$, since $\|\boldsymbol{\varepsilon}_2\| < \|\boldsymbol{\varepsilon}_3\| < \|\boldsymbol{\varepsilon}_1\|$. However, $\hat{\mathbf{x}}_2$ is closer to the decision boundary, thus ambiguity of its classification at the receiver may increase, and $\hat{\mathbf{x}}_3$ is on the other side of the boundary (i.e. $\delta_H(\hat{\mathbf{x}}_3) \neq \delta_H(\mathbf{x})$), which will likely lead to a classification error on the decompressed image. We aim to define a metric to choose $\hat{\mathbf{x}}_1$, because $\hat{\mathbf{x}}_1$ moves farther from H than $\mathbf{x}$, and in an application-aware context it is preferable to $\hat{\mathbf{x}}_2$ and $\hat{\mathbf{x}}_3$. The example of Figure 1 motivates the following remarks.

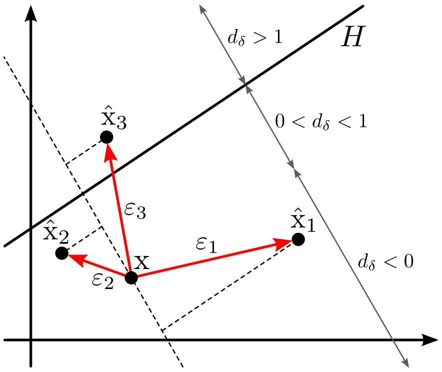


Fig. 1. Graphical example illustrating the proposed distortion metric d_δ of Eq. (3) in a 2-D feature space. Shown are: decision hyperplane H of a binary classifier; original pixel value $\mathbf{x}$; example reconstructions $\hat{\mathbf{x}}_1$, $\hat{\mathbf{x}}_2$, $\hat{\mathbf{x}}_3$ (and corresponding error vectors $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\boldsymbol{\varepsilon}_3$) after lossy compression of $\mathbf{x}$ with different parameters (each associated with different rate requirements). For post-compression classification, $\hat{\mathbf{x}}_1$ is preferable although its Euclidean error $\|\boldsymbol{\varepsilon}_1\|$ is greater than that of $\hat{\mathbf{x}}_2$ and $\hat{\mathbf{x}}_3$.

Observation 1: Distortion estimation should be inversely proportional to the distance $\Delta(\mathbf{x}, H) = (\mathbf{w}^\mathsf{T}\mathbf{x} + w_0)/\|\mathbf{w}\|$ between an original data point $\mathbf{x}$ and the decision hyperplane H, where $\Delta(\mathbf{x}, H) > 0$ if $\mathbf{x}$ lies on the same side of the plane H as the normal vector $\mathbf{w}$ and negative otherwise.

Observation 2: Distortion should be proportional to the component of the error vector (oriented from $\mathbf{x}$ to $\hat{\mathbf{x}}$) in the direction normal to the decision hyperplane $\varepsilon_{\|\mathbf{w}} = \big((\hat{\mathbf{x}} - \mathbf{x})^\mathsf{T}\mathbf{w}\big)/\|\mathbf{w}\|$.

Based on these observations, we define our proposed classification-aware distortion metric d_δ as:

$$d_\delta(\mathbf{x}, \hat{\mathbf{x}}; H) = -\operatorname{sgn}(\mathbf{w}^\mathsf{T}\mathbf{x} + w_0) \cdot \frac{\|\mathbf{w}\|}{|\mathbf{w}^\mathsf{T}\mathbf{x} + w_0|} \cdot \frac{(\hat{\mathbf{x}} - \mathbf{x})^\mathsf{T}\mathbf{w}}{\|\mathbf{w}\|} \\ = -\frac{(\hat{\mathbf{x}} - \mathbf{x})^\mathsf{T}\mathbf{w}}{\mathbf{w}^\mathsf{T}\mathbf{x} + w_0}, \tag{3}$$

where $\mathbf{w}^\mathsf{T}\mathbf{x} + w_0 \neq 0$. The first term adjusts for the sign, such that: $d_\delta < 0$ if $\hat{\mathbf{x}}$ is farther than $\mathbf{x}$ with respect to H; $0 < d_\delta < 1$ if $\hat{\mathbf{x}}$ lies between $\mathbf{x}$ and H; and $d_\delta > 1$ if $\hat{\mathbf{x}}$ crosses the decision boundary (see Figure 1). If the original point $\mathbf{x}$ lies exactly on the boundary H, the denominator of Eq. (3) is zero: prior knowledge available at the encoder is not sufficient to make an informed decision, and any displacement of $\mathbf{x}$ may result in a misclassification at the receiver, thus distortion is 'infinite' (in the implementation, to avoid singularity a small quantity is added to the denominator).

To satisfy the conditions of the generalized Lagrange multiplier method, distortion values must be nonnegative. Thus, d_δ in Eq. (3) is composed with an exponential function:

$$D_\delta(\mathbf{x}, \hat{\mathbf{x}}; H) = \exp(d_\delta(\mathbf{x}, \hat{\mathbf{x}}; H)). \tag{4}$$

If $\hat{\mathbf{x}}$ is farther than $\mathbf{x}$ with respect to H, then $0 < D_\delta < 1$; while $D_\delta > 1$ for $\hat{\mathbf{x}}$ closer to H than $\mathbf{x}$ and it grows rapidly after the decision boundary is crossed. Distortion at the CU level is obtained as the sum $D_\delta^{CU} = \sum_{i=1}^{N} D_\delta(\mathbf{x}_i, \hat{\mathbf{x}}_i; H)$, over pointwise distortion values at pixels $\mathbf{x}_1, \ldots, \mathbf{x}_N$ in a CU.

We implement the proposed distortion metric D_δ in the R-D optimization framework of the HEVC compression standard [3].

In the next section, we first outline relevant aspects of the λ-domain rate control algorithm [10] currently adopted in HEVC. Next, we discuss how we obtain suitable model parameters based on curve fitting procedures and training data.

III. Rate-Distortion Modeling in HEVC

In the HEVC standard [3], R-D optimization is employed at the encoder to decide coding parameters, and distortion is measured by the sum of squared errors (SSE) between original and reconstructed pixel values. Accurate rate control is achieved by jointly using several models, however, default model parameters are tailored to SSE as distortion metric, whereas here we adopt D_δ (Eq. (4)). Below we describe relevant models, and how we tailor them to our metric D_δ.

The JCT-VC has recently adopted the λ-domain rate control algorithm [10], which identifies the Lagrange multiplier λ (see Eq. (2)) as key to accurate bit allocation. The relationship between R and D is modeled by the hyperbolic function $D(R) = CR^{-K}$, where C and K are model parameters.

***R*-λ and λ-QP models.** By differentiating $D(R)$, the R-λ relationship $\lambda = \alpha R^{\beta}$ is obtained, where α and β are model parameters. Due to differences in the R-D characteristics of intra-coded (I-frames) and inter-coded pictures, the R-λ model for I-frames includes an image complexity measure C based on the Sum of Absolute Transformed Differences (SATD) [13]:

$$\lambda = \frac{\alpha}{256}\left(\frac{C^{\beta_1}}{R}\right)^{\beta_2}, \tag{5}$$

where $\alpha = 6.7542$, $\beta_1 = 1.2517$, and $\beta_2 = 1.7860$ are default values. To adapt to source characteristics, α and β_2 are updated with the encoding process as described in [13].

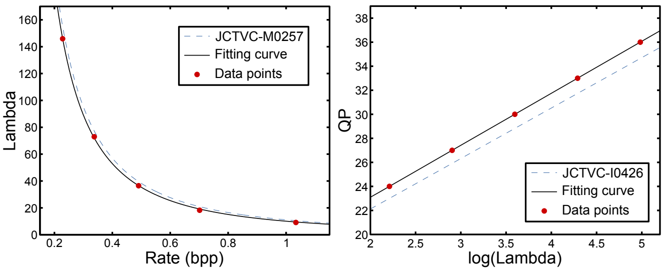
When encoding a picture or a block within a picture (so-called 'LCU', i.e. largest coding unit), target rate is estimated based on available to that point bit budget and an estimate of the bits that will be required to encode the remaining data [10]. With the target rate known, λ is calculated at the frame level and also for each LCU in an I-frame using the R-λ model of Eq. (5). Subsequently, the majority of coding parameters are determined by exhaustive search, evaluating for each option the objective function in Eq. (2) (in this work we use D_δ as distortion metric). The configuration obtaining minimum R-D cost is eventually selected.

To reduce encoding complexity, the quantization parameter (QP) is more efficiently obtained using the λ-QP model instead of exhaustive search. A linear-log relationship is used [14]:

$$QP = a \log \lambda + b\,, \tag{6}$$

where $a = 4.2005$ and $b = 13.7122$ are default values [14].

Fitting models for D_δ. To tailor R-λ and λ-QP models to our proposed distortion metric (instead of SSE), we estimate suitable model parameters based on training data and curve fitting procedures. We encode training images with different QP values in a fixed-QP strategy, using our distortion metric D_δ. Each time we record λ, QP, SATD, and the resulting rate R. To estimate parameters α and β_2 of the R-λ model for I-frames, we fit Eq. (5) to the observed image statistics using the Levenberg-Marquardt algorithm. Finally, we use least squares fitting to estimate parameters a and b of the λ-QP model in Eq. (6). We measure goodness of fit using the r^2 coefficient of determination. The values of λ and QP estimated by the models of Eq. (5) and (6) are normally restricted to a narrow range [10]. Since quality consistency among neighboring LCUs may be unnecessary in an application-aware context, we disable this clipping operation. We only ensure that $0 \leq \text{QP} \leq 51$.



(a) R-λ model (b) λ-QP model

Fig. 2. Image statistics obtained from plant image training data and fitting curves for the λ-domain models. Dashed lines denote default models in HM.

IV. Results and Discussion

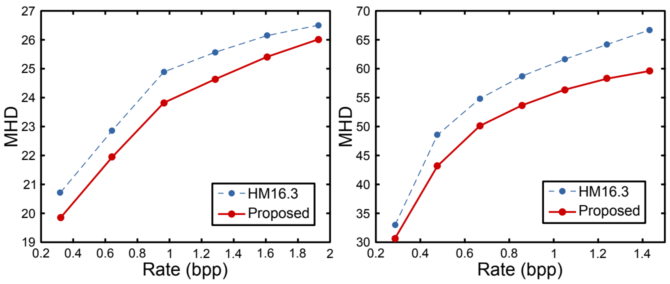
A. Experimental Settings

Image data: We evaluate our approach on a time-lapse sequence composed of 21 gray-scale images (1080×432), showing a top view on 11 growing *Arabidopsis thaliana* plant subjects [11] (see Figure 4a). We also adopt a fluorescence microscopy sequence (N2DH-GOWT1 [12]), composed of 92 frames (1024×1024) showing moving mouse stem cells.

Codec settings: We implement the proposed D_δ metric of Eq. (4) in the HM v16.3 reference encoder. We enable rate control and encode the test sequences with the *Proposed* approach at a variety of bit rates. For comparison, we also adopt the plain HM encoder, referred to as *HM16.3*, with SSE distortion metric and default λ-domain parameters (Section III). For decoding we use the HM v16.3 reference decoder.

λ-domain models: We estimate model parameters of the rate control algorithm as described in Section III. To collect training data we encode the first image of each sequence using a fixed-QP approach, with $\text{QP} \in \{24, 27, 30, 33, 36\}$. Fitting parameters are initialized to default values in HM16.3. For example, best-fit parameters ($r^2 \approx 1$, cf. Figure 2) for the plant image sequence are: $\alpha = 5.6344$, $\beta_2 = 1.8110$ for R-λ model and $a = 4.3281$, $b = 14.4329$ for λ-QP model.

Classifier: To segment the images we adopt a pixel-level classifier based on logistic regression operating on pixel intensities. Classes are defined as 'foreground' (plant or cell) and 'background'. Let $\mathbf{x} \in \mathcal{X}$ be a pixel value and y the corresponding (unknown) label. We predict the probability of y being 'foreground' as $\mathcal{P}(y = \text{foreground}|\mathbf{x}) = 1/(1 + e^{-\mathbf{w}^\mathsf{T}\mathbf{x}+w_0})$. Based on a training (uncompressed) image and corresponding ground-truth pixel labels, model parameters $\mathbf{w}$, w_0 are found using maximum likelihood estimation. Note that the same parameters are used at the encoder in the proposed distortion metric (cf. Eq. (3)). We use for training the first image in the



(a) Plants (b) Cells

Fig. 3. Compression performance with respect to segmentation accuracy (MHD, lower is better). Rate is measured in bits per pixel (bpp), averaged over the entire sequence.

sequence, which is excluded from testing. At the decoder, we decide the class of a pixel based on a threshold ϑ^* on the probability values, estimated from the training data by sweeping over a range of values in $[0, 1]$. We select $\vartheta^* = \arg\max_{\vartheta}(2|\mathbf{Y}^{gt} \cap \mathbf{Y}^c|)/(|\mathbf{Y}^{gt}| + |\mathbf{Y}^c|)$ that maximizes Dice score between ground truth $\mathbf{Y}^{gt}$ and classifier output $\mathbf{Y}^c$.

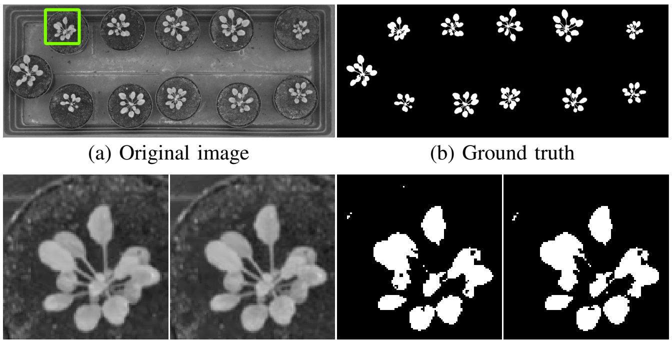
Evaluation: We evaluate our approach on post-compression classification accuracy. To compare segmentation of the image reconstructed after compression with ground-truth segmentation, we adopt the Modified Hausdorff Distance (MHD) [15].

B. Results

Figure 3 shows average R-D performance using the distortion metric D_δ of Eq. (4). It is readily seen that with the Proposed approach segmentation error (measured by MHD) is consistently lower than HM16.3 at all bit rates and for both test sequences, demonstrating that by using our classification-aware distortion metric the encoder focuses bit rate in a way to preserve classification accuracy. A reduction in image fidelity of approximately 1 dB of Peak Signal-to-Noise Ratio is observed with Proposed, which is explained by the effect discussed in Section II (classification accuracy is preferred over fidelity). With the Proposed approach, encoding time increases with respect to HM16.3 on average by 17% and 10% for plant and cell sequences, respectively.

To appreciate the benefits of our approach, Figure 4 shows the outcome of post-compression classification on an example plant image. For equivalent bit rate (1 bpp), reconstructed images in (c) and (d) appear visually almost identical. On the other hand, segmentation in (f) obtained on the image compressed with the Proposed approach is less noisy than the one in (e) obtained on the image compressed with HM16.3 (e.g., see leaf borders and holes inside some leaves).

V. Conclusion



(a) Original image (b) Ground truth

(c) HM16.3 (d) Proposed (e) HM16.3 (f) Proposed

Fig. 4. Example of post-compression classification. Shown are: (a) original image; (b) ground-truth segmentation obtained manually; detail (green box) of reconstructed image after compression at 1 bpp using (c) HM16.3, and (d) proposed approach; (e) classifier output of (c); (f) classifier output of (d).

We presented an approach to application-aware R-D optimization for image and video compression based on a new distortion metric evaluating classification errors due to lossy compression. While our metric is general and could be adapted to different coding schemes (and types of signal), we devise an implementation for the HEVC standard. Our approach involves only encoder-side optimizations and does not include free parameters tunable by the user. The resulting bit stream is standard compliant. A reduction in object segmentation (via classification) error is consistently observed when analyzing image data compressed using our methodology. We expect that our approach will have numerous practical implications in multimedia communications involving automated analyses.

Acknowledgment

This work is partially supported by a Marie Curie Action: “Reintegration Grant” (grant no. 256534) of the EU’s FP7.

References

[1] E. Soyak, S. A. Tsaftaris, and A. K. Katsaggelos, “Low-complexity tracking-aware H.264 video compression for transportation surveillance,” *IEEE TCSVT*, vol. 21, no. 10, pp. 1378–1389, 2011.

[2] J. D. Paola and R. A. Schowengerdt, “The effect of lossy image compression on image classification,” in *IGARSS*, 1995, pp. 118–120.

[3] G. J. Sullivan, J. Ohm, W. J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[4] B. M. Dogahe and M. N. Murthi, “Quantization for classification accuracy in high-rate quantizers,” in *IEEE DSP Workshop*, 2011, pp. 277–282.

[5] K. L. Oehler and R. M. Gray, “Combining image compression and classification using vector quantization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 461–473, 1995.

[6] J. Li, R. M. Gray, and R. Olshen, “Joint image compression and classification with vector quantization and a two dimensional hidden Markov model,” in *Data Compression Conference*, 1999, pp. 23–32.

[7] Y. Dong, S. Chang, and L. Carin, “Rate-distortion bound for joint compression and classification with application to multiaspect scattering,” *IEEE Sensors Journal*, vol. 5, no. 3, pp. 481–492, 2005.

[8] J. Chao, R. Huitl, E. Steinbach, and D. Schroeder, “A novel rate control framework for SIFT/SURF feature preservation in H.264/AVC video compression,” *IEEE TCSVT*, vol. 25, no. 6, pp. 958–972, 2015.

[9] L. Pu, M. W. Marcellin, A. Bilgin, and A. Ashok, “Image compression based on task-specific information,” in *ICIP*, 2014, pp. 4817–4821.

[10] B. Li, H. Li, L. Li, and J. Zhang, “λ domain rate control algorithm for High Efficiency Video Coding,” *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3841–3854, 2014.

[11] M. Minervini and S. A. Tsaftaris, “Application-aware image compression for low cost and distributed plant phenotyping,” in *DSP*, 2013, pp. 1–6.

[12] M. Maška *et al.*, “A benchmark for comparison of cell tracking algorithms,” *Bioinformatics*, vol. 30, no. 11, pp. 1609–1617, 2014.

[13] M. Karczewicz and X. Wang, “Intra frame rate control based on SATD,” in *13th Meeting of the JCT-VC*, no. JCTVC-M0257, Apr. 2013.

[14] B. Li, D. Zhang, H. Li, and J. Xu, “QP determination by lambda value,” in *9th Meeting of the JCT-VC*, no. JCTVC-I0426, May 2012.

[15] M. P. Dubuisson and A. K. Jain, “A modified Hausdorff distance for object matching,” in *Int. Conf. on Pattern Recognition*, 1994, pp. 566–568.