AN INFORMATION THEORETIC FRAMEWORK FOR IMAGE SEGMENTATION

J. Rigau, M. Feixas, and M. Sbert

Institut d'Informàtica i Aplicacions, Universitat de Girona, 17071-Girona, Spain
{jaume.rigau,miquel.feixas,mateu.sbert}@udg.es

ABSTRACT

In this paper, an information theoretic framework for image segmentation is presented. This approach is based on the information channel that goes from the image intensity histogram to the regions of the partitioned image. It allows us to define a new family of segmentation methods which maximize the mutual information of the channel. Firstly, a greedy top-down algorithm which partitions an image into homogeneous regions is introduced. Secondly, a histogram quantization algorithm which clusters color bins in a greedy bottom-up way is defined. Finally, the resulting regions in the partitioning algorithm can optionally be merged using the quantized histogram.

1. INTRODUCTION

In image processing, grouping parts of an image into units that are homogeneous with respect to one or more characteristics (or features) results in a segmented image. Thus, we expect that segmentation subdivides an image on its constituent regions or objects. Segmentation of non trivial images is one of the most difficult tasks in image processing. Image segmentation algorithms are generally based on one of two basic properties of intensity values: discontinuity and similarity. In the first category, the approach is to partition the image based on abrupt changes in intensity, such as edges in an image. The principal approaches in the second category are based on partitioning an image into regions that are similar according to a set of predefined criteria. Thresholding, region growing, and region splitting and merging are examples of methods of this category [1, 2].

In this paper, we introduce a new information theoretic framework for image segmentation, built on the information channel between the two most basic pixel characteristics: its intensity and its spatial position into the image. Using this channel, we present two algorithms based on the maximization of the mutual information (MI). The first algorithm partitions an image into relatively homogeneous regions using a binary space partition (BSP). The second segments an image from the clustering of the histogram bins. The resulting regions in the first algorithm can be merged using the quantized histogram obtained in the second one.

2. INFORMATION THEORY TOOLS

The following information theoretic definitions and inequalities [3] are fundamental to develop the most basic ideas of this paper.

The *Shannon entropy* $H(X)$ of a discrete random variable X with values in the set $\mathcal{X} = \{x_1, \ldots, x_n\}$ is defined as

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i, \tag{1}$$

where $n = |\mathcal{X}|$ and $p_i = Pr[X = x_i]$. The logarithms are taken in base 2 and entropy is expressed in bits. If we consider another random variable Y with values in the set $\mathcal{Y} = \{y_1, \ldots, y_m\}$ and $q_j = Pr[Y = y_j]$, the *conditional entropy* is defined as

$$H(X|Y) = -\sum_{j=1}^{m} q_j \sum_{i=1}^{n} p_{i|j} \log p_{i|j} \tag{2}$$

where $m = |\mathcal{Y}|$ and $p_{i|j} = Pr[X = x_i|Y = y_j]$ is the conditional probability. $H(X|Y)$ corresponds to the uncertainty in the *information channel* input X from the point of view of receiver Y, and vice versa for $H(Y|X)$.

The *mutual information* between X and Y is defined as

$$I(X,Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \log \frac{p_{ij}}{p_i q_j}, \tag{3}$$

where $p_{ij} = Pr[X = x_i, Y = y_j]$ is the joint probability. It can also be expressed by $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ and is a measure of the *shared information* between X and Y.

Next, we give two basic inequalities:
Data processing inequality. If $X \to Y \to Z$ is a Markov chain, i.e., $p(x,y,z) = p(x)p(y|x)p(z|y)$, then

$$I(X,Y) \geq I(X,Z). \tag{4}$$

This result demonstrates that no processing of Y, deterministic or random, can increase the information that Y contains about X.

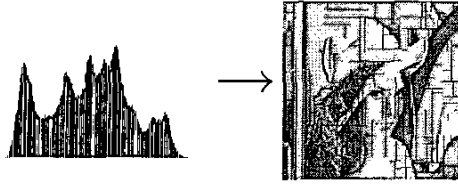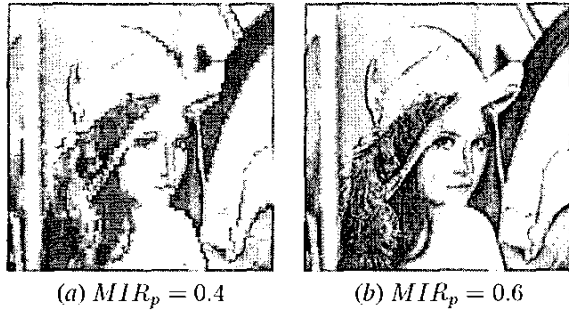**Fig. 1.** Input and output distributions for the information channel.

(a) $MIR_p = 0.4$ (b) $MIR_p = 0.6$

Fig. 2. Two partitions of the *Lena* image (512×512), over luminance channel Y_{709} obtained with the given MIR_p values. The number of regions R is (a) 1553 and (b) 15316. RMSE and PSNR values are respectively (a) (16.232, 22.681) and (b) (9.710, 27.490).

Fano's inequality. Suppose we have two correlated random variables X and Y and we wish to measure the probability of error in guessing X from the knowledge of Y. Fano's inequality gives us a tight lower bound on this error probability in terms of the conditional entropy $H(X|Y)$. From Y we calculate a function $g(Y) = \tilde{X}$ which is an estimate of X. The probability of error is defined by $P_e = Pr[\tilde{X} \neq X]$ and the Fano's inequality is given by $H(X|Y) \leq H(P_e) + P_e \log(n - 1)$ or, equivalently, by

$$I(X,Y) \geq H(X) - H(P_e) - P_e \log(n - 1), \quad (5)$$

where $H(P_e)$ is the binary entropy from $\{P_e, 1 - P_e\}$. Thus, Fano's inequality bounds the probability that $\tilde{X} \neq X$.

3. IMAGE PARTITION

Given an image with N pixels and an intensity histogram with n_i pixels in bin i, we define a discrete information channel where input X represents the bins of the histogram, with probability distribution $\{p_i\} = \{\frac{n_i}{N}\}$, output Y the pixel-to-pixel image partition, with distribution $\{q_j\} = \{\frac{1}{N}\}$ over the N pixels, and the conditional probability distribution $\{p_{j|i}\}$ is the transition probability from bin i of the histogram to pixel j of the image. This *information channel*

can be represented by

$$X \longrightarrow Y \quad (6)$$
$$\{p_i\} \xrightarrow{\{p_{j|i}\}} \{q_j\}$$

In this channel, it can be seen that, given a pixel, there is no uncertainty about the corresponding bin of the histogram (consequently, $I(X,Y) = H(X)$). From the data processing inequality (4), we know that any clustering or quantization over X or Y will reduce the shared information $I(X,Y)$. The information channel $X \rightarrow Y$ can be defined for each color component of an image. Thus, all the algorithms presented in this paper can be applied to any component of a color system.

In this section, we present a greedy algorithm which partitions an image in quasi-homogeneous regions. The optimal partitioning algorithm is NP-complete. To do this partition, a natural approach could consider the above channel (6) as the starting point for the image partitioning, designing a pixel clustering algorithm which minimizes the loss of MI. This process can be described by a Markov chain, $X \rightarrow Y \rightarrow \hat{Y}$, where $\hat{Y} = f(Y)$ represents a clustering of Y.

However, due to the computational cost of this algorithm, a completely opposite strategy has been adopted: a top-down splitting algorithm takes the full image as the unique initial partition and progressively subdivides it with vertical or horizontal lines (BSP) chosen according to the maximum MI gain for each partitioning step. Note that other types of lines could be used, obtaining a varied polygonal subdivision. Our splitting process is represented over the channel (see Fig. 1)

$$X \longrightarrow \hat{Y}. \quad (7)$$

The channel varies at each partition step because the number of regions is increased and, consequently, the marginal probabilities of $\hat{Y}$ and the conditional probabilities of $\hat{Y}$ over X also change. This process can be interpreted in the following way: the choice of the partition which maximizes the MI increases the chances of guessing the intensity of a pixel chosen randomly from the knowledge of the region it pertains to. Similar algorithms were introduced in the context of pattern recognition [4], learning [5], DNA segmentation [6], and document clustering [7].

Our partitioning algorithm can be represented by a binary tree where each node corresponds to an image region. At each partitioning step, the tree acquires information from the original image such that each internal node i contains the mutual information I_i gained with its corresponding splitting. The total $I(X, \hat{Y})$ captured by the tree [4] can be obtained adding up the MI available at the internal nodes of the tree weighted by the relative area $q_i = \frac{N_i}{N}$ of the region i, i.e., the relative number of pixels corresponding to each

node. Thus, the total MI acquired in the process is given by

$$I(X,\widehat{Y}) = \sum_{i=1}^{T} \frac{N_i}{N} I_i, \qquad (8)$$

where T is the number of internal nodes. It is important to stress that this process of extracting information enables us to decide locally which is the best partition.

This partitioning procedure can be stopped using different criteria:

- Given the error probability P_e allowed in partitioning, Fano's inequality (5) provides us with a lower bound for the gain of MI. Taking the equality in (5), we obtain the minimum value of MI needed in the partitioning algorithm:

$$I_{min}(X,Y) = H(X) - H(P_e) - P_e \log(B-1), \qquad (9)$$

where B is the number of bins of the histogram. The process stops when $I(X,\widehat{Y}) \geq I_{min}(X,Y)$. Note that $I_{min}(X,Y)$ is calculated from the initial channel (6).

- The ratio $MIR_p = \frac{I(X,\widehat{Y})}{I(X,Y)}$ is greater than a given threshold. From it we can also determine the error probability in partitioning using (9), and vice versa.

- A predefined number of regions R.

This process can also be visualized from equation $H(X) = I(X,\widehat{Y}) + H(X|\widehat{Y})$, where the acquisition of information increases $I(X,\widehat{Y})$ and decreases $H(X|\widehat{Y})$, producing a reduction of uncertainty due to the fact that the regions become more and more homogeneous. Observe that the maximum MI that can be achieved is $H(X)$.

Two partitions of the *Lena* image over luminance channel Y_{709} in Fig. 2 illustrate the behavior of the partitioning algorithm. They have been obtained using the MIR_p criterion. Number of regions R, root mean square error (RMSE), and peak signal-to-noise ratio (PSNR) are given. The regions in the partitioned images are shown with their average intensity.
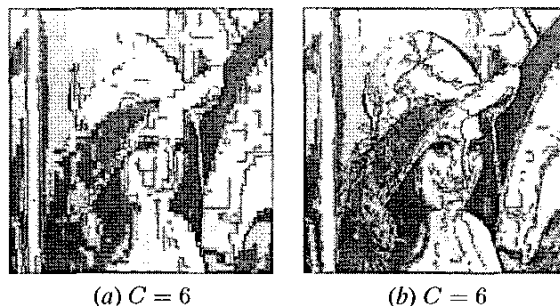
4. HISTOGRAM QUANTIZATION

In this section, a greedy bottom-up segmentation algorithm based on the minimization of the loss of MI is introduced. This algorithm produces a clustering of the histogram bins.

Now, the reverse of the channel (7) is the starting point for the histogram quantization. Thus, the histogram clustering is carried out from a given partition of an image. This process can also be described by a Markov chain, $\widehat{Y} \rightarrow X \rightarrow \widehat{X}$, where $\widehat{X} = f(X)$ represents a clustering of the


(a) $MIR_q = 0.45$ (b) $MIR_q = 0.65$
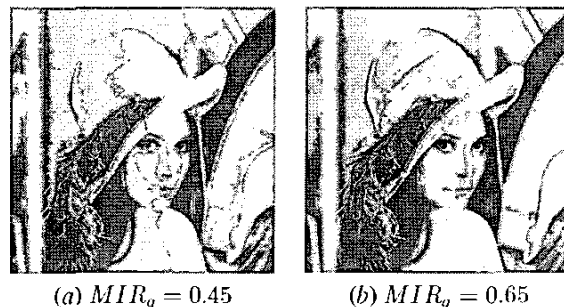
Fig. 3. Two segmentations of the *Lena* image over luminance channel Y_{709}, obtained from the partitioned image of Fig. 2.(a) using the histogram quantization algorithm with the given MIR_q values. The number of colors C is (a) 3 and (b) 6. RMSE and PSNR values are respectively (a) (19.068, 22.212) and (b) (10.683, 27.245).

(a) $C = 6$ (b) $C = 6$

Fig. 4. Two contour segmentations of the *Lena* image, over luminance channel Y_{709} obtained by merging the regions of the corresponding partitioned images of Fig. 2 from the quantized histogram of six colors of Fig. 3.(b). RMSE and PSNR values are respectively (a) (18.961, 22.261) and (b) (14.297, 24.714).

histogram. The optimal quantization algorithm is also NP-complete.

The basic idea underlying our segmentation process is to preserve the maximum information of the image with the minimum number of colors (histogram bins). The clustering of the histogram is obtained efficiently by merging two neighbor bins such that the loss of MI is minimum. The stopping criterion is given, as in the previous section, by an error probability P_e or a MI ratio $MIR_q = \frac{I(\widehat{X},\widehat{Y})}{I(X,\widehat{Y})}$. Optionally, a predefined number of colors C can also be given. An alternative to this algorithm would be to take a top-down approach, like the partitioning algorithm of the previous section. Thus, we could start from the full histogram and successively apply the binary partition which

1195

(a) R   (b) G

(c) B   (d) RGB
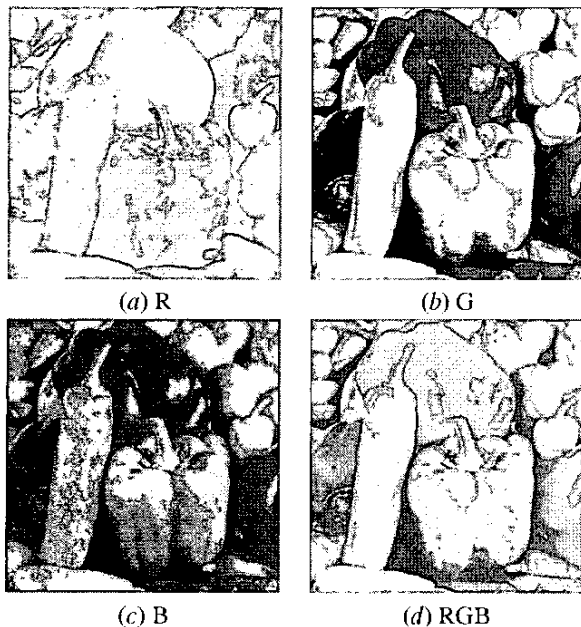
Fig. 5. Subfigures (a-c) show the segmentations of the *Peppers* image over the RGB components with $MIR_q = 0.65$ from the respective partitioned images with $MIR_p = 0.3$. Four colors have been obtained for each component. (d) shows the result of merging (a-c). RMSE and PSNR values in (d) are 19.407 and 21.606, respectively.

maximizes the MI. However, this algorithm is less accurate and more costly than the clustering one. Our clustering process is represented over the channel

$$\widehat{Y} \longrightarrow \widehat{X}. \tag{10}$$

Observe that one particular case of this channel is $Y \to \widehat{X}$. Note also that (10) changes at each clustering step because the number of bins is reduced. The choice of the clustering of the histogram which minimizes the loss of MI increases the chances of guessing the region of a randomly chosen pixel from the knowledge of its intensity. At the end of the quantization process, the MI of the channel is $I(\widehat{X}, \widehat{Y})$, and the following inequality is fulfilled: $I(X,Y) \geq I(X,\widehat{Y}) \geq I(\widehat{X},\widehat{Y})$.

The behavior of our histogram quantization algorithm is shown in Figures 3-5. In Fig. 3, two segmentations of the *Lena* image over luminance channel Y_{709} are shown. They have been obtained using the channel $\widehat{Y} \to \widehat{X}$ with the MIR_q criterion. Number of colors C, RMSE, and PSNR are given. In Fig. 4, the regions obtained in the partitions of Fig. 2 are merged using a quantized histogram of six colors. Finally, Figures 5.(a-c) illustrate the result of quantizing the three color components of the *Peppers* image. For

each component, four colors have been obtained. Fig. 5.(d) shows the result of merging Figures 5.(a-c).

5. CONCLUSIONS AND FUTURE WORK

We have presented an information theoretic framework for image segmentation, based on the information channel between the image intensity histogram and the regions of the partitioned image. Two greedy algorithms, which respectively split the image into homogeneous regions and cluster the bins of the histogram, have been introduced. Mutual information drives respectively the image partitioning [histogram quantization] so that the next image splitting [histogram clustering] is chosen to maximize [minimize] the gain [loss] in mutual information. Our approach has been validated with several experiments on standard test images. In our future work, we will study the compositional complexity of an image following the segmentation framework presented in this paper, as well as the applicability to image compression.

6. REFERENCES

[1] Dana H. Ballard and Christopher M. Brown, *Computer Vision*, Prentice Hall, Englewood Cliffs (NJ), USA, 1982.

[2] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing*, Prentice Hall, Upper Saddle River (NJ), USA, 2002.

[3] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, 1991.

[4] Ishwar K. Sethi and G.P.R. Sarvarayudu, "Hierarchical classifier design using mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, no. 4, pp. 441–445, July 1982.

[5] Sanjeev R. Kulkarni, Gábor Lugosi, and Santosh S. Venkatesh, "Learning pattern classification – a survey," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2178–2206, 1998.

[6] Pedro Bernaola, José L. Oliver, and Ramón Román, "Decomposition of DNA sequence complexity," *Physical Review Letters*, vol. 83, no. 16, pp. 3336–3339, October 1999.

[7] Noam Slonim and Naftali Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2000, pp. 208–215, ACM Press, Held in Athens, Greece.

1196