Self-Adaptive Image Cropping for Small Displays

Gianluigi Ciocca, Claudio Cusano, Francesca Gasparini, Raimondo Schettini

Abstract — *We propose a new self-adaptive image cropping algorithm where the processing steps are driven by the classification of the images into semantic classes. The algorithm exploits both visual and semantic information. Visual information is obtained by a visual attention model, while semantic information relates to the automatically assigned image genre and to the detection of face and skin regions.*

Index Terms — Adaptive image cropping, small displays, image rendering, image classification, face detection, skin detection.

I. INTRODUCTION

Images are playing a more and more important role in sharing, expressing and exchanging information in our daily lives. Accompanying this revolution, mobile handheld devices with different capabilities are undergoing considerable progress. Now we all can easily capture and share personal photos on these small-form-factor devices anywhere and anytime. However, many hurdles still need to be crossed. Among them, major crucial challenges include the limited accessing bandwidth and small display sizes. Some of the efforts that have been put on image adaptation are related to the ROI coding scheme introduced in JPEG 2000 [1]. Most of the approaches for adapting images only focused on compressing the whole image in order to reduce the data transmitted. Few other methods use an auto-cropping technique to reduce the size of the image transmitted [2][3]. These methods decompose the image into a set of spatial information elements (saliency regions) which are then displayed serially to help users' browsing or searching through the whole image. These methods are heavily based on a visual attention model technique that is used to identify the saliency regions to be cropped [4].

In this paper we show how both visual and semantic information can be exploited to design a self-adaptive image cropping algorithm for small displays.

II. SELF-ADAPTIVE IMAGE CROPPING

Fig. 1 shows the flow diagram of the proposed algorithm. The images are first classified into three broad classes, that is, "landscape", "close-up", and "other". The classification is based on the use of ensembles of decision trees, called decision forests. The trees of the forests are constructed according to CART methodology [5]. The features used in this classification process are related to color, texture, edge and composition of the image [6][7]. Then, an ad-hoc cropping strategy is applied for each image class. A landscape image, for example, due to its lack of specific focus elements, is not processed at all: the image is regarded as being wholly relevant. A close up image, generally, shows only a single object or subject in the foreground, and thus, the cropping region should take into account only this discarding any region that can be considered as background. In case of an image belonging to the other class, we are concerned in whether it contains people or not. In the first case, the cropping strategy should prioritize the selection of regions containing most of the people.

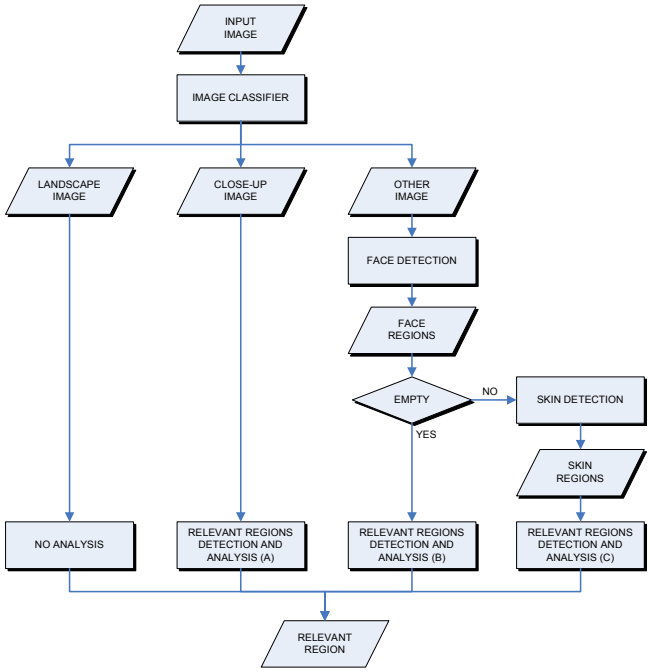

Fig. 1. The flow diagram of the proposed algorithm.

Landscape images. In the case of landscape images, no cropping is performed and the image is only adapted to fit the display dimensions. We adopt this strategy because landscape images usually do not present a specific subject to be focalized.

Close-up images. For close-up images here we define a new procedure which we called "Relevant regions detection and analysis (A)" in Figure 1:

a. A saliency map is generated based on the Itti-Koch visual attention model [8]. Visual attention facilitates the processing of the portion of the input associated with the

G. Ciocca, C. Cusano, F. Gasparini, and R. Schettini are with the Department of IT, Systems and Communications (DISCo) at the University of Milano-Bicocca, Milan, Italy (email: ciocca@disco.unimib.it). E-Mails: [ciocca, cusano, gasparini, schettini]@disco.unimib.it

relevant information, suppressing the remaining information.

b. The saliency map is automatically binarized in order to identify saliency regions. The regions with areas smaller than a threshold which are a function of the area of the larger region are discarded.
c. A single relevant region is obtained, considering the bounding box that includes all the saliency regions previously identified.
d. The image is then cropped and adapted with respect to this region.

Other images. A face detector inspired by the Viola and Jones one [9] is applied to distinguish between images with and without faces. To perform this, we first analyze the images in order to detect candidate face regions. The detector is composed of a chain of weak classifiers trained by the Ada-boost algorithm. These regions are then validated by checking the amount of skin present and those that have a given amount of skin are selected as true face regions.

For images without faces we designed the "relevant regions detection and analysis (B)":

a. Same as point a. of the close-up case.
b. Same as point b. of the close-up case.
c. The most salient region individuated in point b. is now considered as the relevant region.
d. The image is then cropped and adapted with respect to this region.

For images with faces, we designed the "relevant regions detection and analysis (C)":

a. Same as point a. of the close-up case.
b. A new skin color map is also computed.
c. The saliency map, the skin color map and the face regions are then combined together to form a global map, used to locate the most relevant region.
d. The image is then cropped and adapted with respect to this region.

For both close-up and other images, the borders of the final cropped region are enlarged to better include the relevant area.

The last phase in the adaptive image cropping aims at adjusting the cropping region such that the overall information is maximized and all the space in the output display is used. The region is rotated if the orientation of the region differs from the normal orientation of the display. Further, the region is enlarged and/or trimmed in order to retain the display's aspect ratio maintaining the cropped region centered. Finally, the cropped image is resized to the size of the display.

III. ALGORITHM'S BUILDING BLOCKS

A. Image Classification

Automatic image classification allows for the application of the most appropriate enhancement strategy according to the content of the photograph. To this end we designed an image classification strategy [6] based on the analysis of low-level features that can be automatically computed without any prior knowledge of the content of the image.

The landscape class includes photos with no evident focus elements and the images belonging to it are mostly related to natural panoramas (Fig. 2).

Fig. 2. Examples of "Landscape" images.

The close-up class includes portraits and photos of people and objects in which the context provides little or no information in regards to where the photo was taken (Fig. 3). All the images not classified as landscapes or close-ups are put in the other image class (Fig. 4).

Fig. 3. Examples of "Close-up" images.

Fig. 4. Examples of "Other" image

The features we used are related to color (moments of inertia of the color channels in the HSV color space, and skin color distribution), texture and edge (statistics on wavelets decomposition and on edge and texture distributions), and composition of the image (in terms of fragmentation and symmetry).

We used the CART methodology [5] to train tree classifiers. Briefly, tree classifiers are produced by recursively partitioning the space of the features which describe the

images. Each split is formed by conditions related to the values of the features. Once a tree has been constructed, a class is assigned to each of the terminal nodes, and when a new case is processed by the tree, its predicted class is the class associated with the terminal node into which the case finally moves on the basis of the values of the features. The construction process is based on a training set of cases of known class. A function of impurity of the nodes, *i(t)*, is introduced, and the decrease in its value produced by a split is taken as a measure of the goodness of the split itself. For each node, all the possible splits on all the features are considered and the split which minimizes the average impurity of the two sub nodes is selected. The function of node impurity we have used is the Gini diversity index:

$$i(t) = 1 - \sum_{c=1}^{C} p(c \mid t)^2 \tag{1}$$

where $p(c|t)$ is the resubstitution estimate of the conditional probability of class c (c = 1, …, C) in node t, that is, the probability that a case found in node t is a case of class c. When the difference in impurity between a node and best subnodes is below a threshold, the node is considered as terminal. The class assigned to a terminal node t is the class c* for which:

$$p(c^* \mid t) = \max_{c=1,\ldots,C} p(c \mid t) \tag{2}$$

In CART methodology the size of a tree is treated as a tuning parameter, and the optimal size is adaptively chosen from the data. A very large tree is grown and then pruned, using a cost-complexity criterion which governs the tradeoff between size and accuracy. Although the pruning process prevents the danger of trees too tailored to the training data, there is still overfitting due to instability (a small change in data may result in a very different tree). Decision forests can be used to overcome this problem improving, at the same time, generalization accuracy. The trees of a decision forest are generated by running the training process on boostrap replicates of the training set. The classification results produced by the single trees are combined applying a majority vote. Each decision tree has been trained on bootstrap replicates of a training set composed of about 4500 photographs manually annotated with the correct class.

B. Face Detection

Face detection in a single image is a challenging task because the overall appearance of faces ranges widely in scale, location, orientation and pose, as well as in facial expressions and lighting conditions [10] and [11]. In this work we have adopted a variation of the face detector proposed by Viola and Jones [9]. The face detector algorithm uses a multi-scale, multi-stage classifier, which operates on image intensity information. It uses an over-complete set of Haar-like features.

In order to compute these features very rapidly at many scales they introduced an integral image representation. The integral image can be computed from an image using a few operations per pixel. Using the integral image, the Haar-like features can be computed at any scale or location in constant time. Within any image sub-window, the total number of Haar-like features is very large. In order to ensure fast classification, the learning process must exclude a large majority of the available features, and focus on a small set of critical ones. Feature selection is achieved using the AdaBoost learning algorithm by constraining each weak classifier to depend on only a single feature. The boosting algorithm is used to train successively more complex classifiers in a cascade structure which increases the speed of the detector by quickly discarding background regions while spending more computation on promising face-like regions.

C. Saliency Map Detection

Visual attention models simulate the human vision system to process and analyze images in order to identify conspicuous regions within the images themselves. These regions (ROIs) can be used to guide the analysis of the images. In the follows we describe the Itti and Koch model [8] that is inspired by the behavior and the neuronal architecture of the early primate visual system.

Input is provided in the form of a color image. Nine images with descending spatial scale are created from the original image using dyadic Gaussian pyramids which progressively low-pass filter and sub-sample the input images. Each feature to be extracted is computed with a set of linear "center-surround" operations akin to visual receptive fields, where visual neurons are most sensitive in a small region of the visual space (center), while stimuli presented in a broader region concentric to the center (surround) inhibit the neuronal response. The difference between images at fine (center) and coarse (surround) scales is used to mimic this behavior. A set of feature maps are obtained by varying the scale. Three visual features are exploited: color, intensity and orientation. In the follows Θ indicates the center-surround difference operation between a center image c and a surround image s. A feature map for the intensity feature is computed as:

$$I(c,s) = \left| I(c) \Theta I(s) \right| \tag{3}$$

The second set of maps is constructed from the color channels which are represented using the so-called "color double-opponent" system. The maps are computed for the two channel pairs Red/Green and Blue/Yellow as follow:

$$RG(c,s) = \left| (R(c) - G(c)) \Theta (G(s) - R(s)) \right| \tag{4}$$

$$BY(c,s) = \left| (B(c) - Y(c)) \Theta (B(s) - Y(s)) \right| \tag{5}$$

Gabor pyramids $O(\sigma,\theta)$, with σ representing the scale and θ representing the orientation, are used to compute the orientation maps:

$$O(c,s,\theta) = \left|O(c,\theta)\Theta O(s,\theta)\right| \quad (6)$$

In total 42 feature maps are computed: 6 for the intensity, 12 for color and 24 for orientation. Before computing the final saliency map, the feature maps are normalized in order to globally promote maps in which a small number of strong peaks of activity (conspicuous locations) are present, while globally suppress maps which contain numerous comparable peak responses. After the normalization process, all the sets of the normalized maps are combined by average into three conspicuity maps, $\bar{I}$, $\overline{C}$, and $\overline{O}$, one for each feature. The three conspicuity maps are normalized, and linearly combined together into the final saliency map S:

$$S = \tfrac{1}{3}\left(N(\bar{I}) + N(\overline{C}) + N(\overline{O})\right) \quad (7)$$

An example of saliency map is shown in Fig. 5.

Fig. 5. The saliency map computed with the Itti-Koch visual attention model.

Although the visual attention model is used in all the different cropping strategies of our algorithm, it requires a tuning of several parameters that influence the results depending on the image content. The image classification phase allowed us to heuristically tune these parameters at the best.

D.Skin Detector

Many different methods for discriminating between skin pixels and non-skin pixels are available. The simplest and most often applied method is to build an "explicit skin cluster" classifier which expressly defines the boundaries of the skin cluster in certain color spaces. The underlying hypothesis of methods based on explicit skin clustering is that skin pixels exhibit similar color coordinates in a properly chosen color space. This type of binary method is very popular since it is easy to implement and does not require a training phase.

For this application we have adopted a method based on the YCbCr color space, developed by Chai and Ngan [12]. A skin color map is derived and used on the chrominance components of the input image to detect pixels that appear to be skin. The algorithm then employs a set of regularization processes to reinforce those regions of skin-color pixels most likely to belong to facial regions. We consider only their color segmentation step here. Working in the YCbCr space the authors find that the ranges of Cb and Cr most representative for the skin-color reference map are: $77 \le Cb \le 127$ and $133 \le Cr \le 173$**.** This range of chrominance values is obtained heuristically from many different facial images, European, Asian and African, to derive a map that models the facial color of all human races.

Considering that a good classifier should have high recall and high precision, but typically, as recall increases, precision decreases, we have applied a genetic algorithm to determine the boundaries of the skin clusters in the YCbCr color space, to favor either high recall or high precision, or to satisfy a reasonable tradeoff between the two, depending on application demands [13]. Table I shows the three different skin boundaries obtained by the genetic algorithm for the Cb and Cr color channels.

TABLE I
BOUNDARIES OF THE COLOR SKIN CLUSTERS

Application Demand	Cb Min	Cb Max	Cr Min	Cr Max
Precision	81	112	138	176
Recall	74	124	134	185
Tradeoff	88	115	139	183

For the scope of this work the boundaries of the color skin cluster were chosen to offer high recall in pixel classification.

IV. EXPERIMENTAL RESULTS

To verify the reliability of our method we used a data set composed of 532 images. These images were downloaded from personal web-pages, acquired by various digital cameras and taken from photo collection CDs, and had different sizes (from 256×384 to 2240×1488 pixels), resolutions and quality (in terms of jpeg compression, noise, dynamic range, etc.). To evaluate the goodness of the proposed image cropping strategy, we collected the judgments of a panel of five non-professional photographers which can be summarized as follows: about 7% of the output images were judged worse than the original image scaled to fit the small screen display; about 40% were judged equivalent (the great majority of landscape images are correctly classified and since they are not cropped according to our procedure they are equivalent to the resized original), while the remaining 53% of the images were judged better. Without considering landscapes, the percentage of images judged better increases. The worst results are mostly due to the errors introduced by the face detector. Most of the misclassified images by the CART algorithm are still cropped acceptably. With more insight into the semantic contents of the images, we expect the whole procedure to perform better and more reliably.

Fig. 6, 7, and 8 show some results of the adaptive cropping algorithm. The bounding box superimposed on the images represents the relevant region detected by the algorithm without any adaptation attempt. Fig. 9 and 10 show some cropping results using a simulated small display. It can be seen how the adaptation stage fills the display surface while maintaining the focus on the relevant area.

Fig. 6. Relevant regions selected within some of the "close-up" images.

Fig. 7. Relevant regions selected within some of the "Other" images. No faces are present or detected.

Fig. 8. Relevant regions selected within some of the "Other" images containing faces.

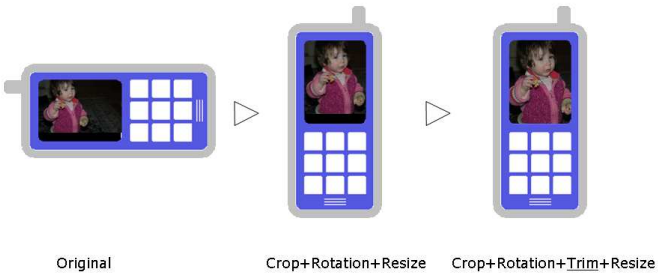

Fig. 9. Image cropping results in a simulated display. Two different adaptation strategies are used to demonstrate the effect of the region trimming operation. "Original" refers to the image fitted to the display without any processing.

V.CONCLUSIONS

We designed a self-adaptive image cropping algorithm exploiting both visual and semantic information. Visual information is obtained by a visual attention model, while semantic information relates to the automatically assigned image genre and to the detection of face and skin regions. The classification of the images to be cropped allowed us to build a cropping strategy specific for each image class. The image classification phase made it possible to also tune the saliency map parameters at the best. The efficacy of these strategies is further improved prioritizing the visualization of some objects in the image instead of others. Here we checked the presence of people in the images belonging to the other class, but the analysis can be further specialized for the identification of other objects.

Fig. 10. More results in a simulated display. The image at the left is not cropped since it is classified as landscape. The image on the left(top) refers to the image fitted to the display without any processing.

REFERENCES

[1] C., Christopoulos, A., Skodras, T., Ebrahimi, "The JPEG2000 still image coding system: an overview", IEEE Trans Cons Elect. vol. 46, no. 4, pp. 1103–1127, 2000 .

[2] L., Chen, X., Xie, X., Fan, W., Ma, H.J., Zhang, H.Q., Zhou, "A visual attention model for adapting images on small displays", Multimedia Systems, vol. 9, pp. 353–364, 2003.

[3] B., Suh, H., Ling, B.B., Bederson, D.W., Jacobs, "Automatic Thumbnail Cropping and its Effectiveness", In Proc. UIST'03, pp. 95-104, 2003.

[4] M., Kimura, M.A., Yamauchi, "A method for extracting region of interest based on attractiveness", IEEE Trans. On Consumer Electronics, Vol. 52, No. 2, pp. 312-316, 2006.

[5] L., Breiman, J.H., Friedman, R.A., Olshen, C.J., Stone, "Classification and Regression Trees", (Wadsworth and Brooks/Cole), 1984.

[6] R., Schettini, C., Brambilla, C., Cusano, G., Ciocca, "Automatic classification of digital photographs based on decision forests", *IJPRAI*, vol. 18(5), pp. 819-846, 2004.

[7] M., De Ponti, R., Schettini, C., Brambilla, A., Valsasna, G., Ciocca, "Content-Based Digital-Image Classification Method", European Patent no. EP1102180, 2001.

[8] L., Itti, C., Koch, "A model of saliency based visual attention of rapid scene analysis", *IEEE Trans. on PAMI*, vol. 20, pp.1254-1259, 1998.

[9] P., Viola, M.J., Jones, "Robust real-time face detection", International Journal of Computer Vision, vol. 57, pp. 137-154, 2004.

[10] H., Rowley, S., Baluja, T., Kanade, "Neural Network-Based Face Detection", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20,(1), (1998) 23-28.

[11] M.H., Yang, D.J., Kriegman, N., Ahuja, "Detecting Faces in Images: A Survey", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24(1), (2002) 34-58.

[12] D. Chai and K. N. Ngan, "Face segmentation using skin colour map in videophone applications", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No 4, pp. 551-564, 1999.

[13] F., Gasparini, R., Schettini, "Skin segmentation using multiple thresholding", Proc. Internet Imaging VII, Vol. 6061 (S. Santini, R. Schettini, T. Gevers eds.), pp. 1-8, 2006.

Gianluigi Ciocca received his degree (Laurea) in Computer Science at the University of Milan in 1998, and since then he has been a fellow at the Institute of Multimedia Information Technologies and at the Imaging and Vision Laboratory in the ITC Institute of the Italian National Research Council, where his research has focused on the development of systems for the management of image and video databases and the development of new methodologies and algorithms for automatic indexing. He is currently a researcher in computer science at the Department of Information Science, Systems Theory, and Communication (DISCo) of the University of Milano-Bicocca, working on video analysis and video abstraction.

Claudio Cusano is a post-doc student at DISCo, (Department of Information Science, Systems Theory, and Communication), of the University of Milano-Bicocca, where he took his PhD in Computer Science. Since April 2001 he has been a fellow of the the ITC Institute of the Italian National Research Council. The main topics of his current research concern 2D and 3D imaging, with a particular focus on image analysis and classification, and on face recognition.

Francesca Gasparini took her degree in Nuclear Engineering at the Polytechnic of Milan in 1997 and her Ph.D in Science and Technology in Nuclear Power Plants at the Polytechnic of Milan in 2000. Since January 2001 she has been a fellow at the ITC Imaging and Vision Laboratory, of the Italian National Research Council, located in Milan, where her research has focused on image enhancement, cast detection and cast removal. She is currently a teaching assistant in computer science at DISCo (Dipartimento di Informatica, Sistemistica e Comunicazione) of the University of Milano-Bicocca, working on image processing.

Raimondo Schettini is an associate professor at DISCo, University of Milano Bicocca where he is in charge of the Imaging and Vision Lab. He has been associated with Italian National Research Council (CNR) since 1987. He has been team leader in several research projects and published more than 160 refereed papers on image processing, analysis and reproduction, and on image content-based indexing and retrieval. He is an associated editor of the Pattern Recognition Journal. He was a co-guest editor of three special issues about Internet Imaging (Journal of Electronic Imaging, 2002), Color Image Processing and Analysis (Pattern Recognition Letters, 2003), and Color for Image Indexing and Retrieval (Computer Vision and Image Understanding, 2004). He was General Co-Chairman of the 1st Workshop on Image and Video Content-based Retrieval (1998), of the First European Conference on Color in Graphics, Imaging and Vision (2002), and of the EI Internet Imaging Conferences (2000-2006).