CONTEXT SALIENCY BASED IMAGE SUMMARIZATION

Liang Shi[1], *Jinqiao Wang*[2], *Lei Xu*[3], *Hanqing Lu*[2], *Changsheng Xu*[2]

[1]Beijing University of Posts and Telecommunications, Beijing 100876, China
{freybupt@gmail.com}
[2]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{jqwang, luhq, csxu@nlpr.ia.ac.cn}
[3]Nokia Research Center, No.5 Dong Huan Zhouglu, BDA, Beijing, 100176, China
{ext-lei.2.xu@nokia.com}

ABSTRACT

Image summarization is to determine a smaller but faithful representation of the original visual content. In this paper, we propose a *context saliency* based image summarization approach, incorporating statistical saliency and geometric information as the importance measurement instead of visual saliency. To ensure image summaries to be adaptive to target device under perception constraint, we present a grid-based piecewise linear image warping scaleplate, and adopt the *sweet spot* evaluation to generate a flexible model combining the cropping and warping methods. Additionally, we explore potential extensions on image retargeting, thumbnail generation, digital matting and photo browsing. Experimental results show comparable performance compared to the-state-of-art on common data sets.

Index Terms— Image summarization, context saliency, geometric segmentation

1. INTRODUCTION

Instead of reading plain text, users are more willing to interact with visual content, which is much more attractive and informative. Along with the booming of cellphone cameras and picture sharing communities, there is an ever-increasing variety of viewing options available: cell-phones, digital cameras, ipods, laptops, PDA, PSP, and so on. These devices play an increasingly important role in daily life of millions of people world wide, as Paul Levinson calls in his book to be "*the media-in-motion business*"[8]. Meanwhile, this trend brings us a challenging question: How could ONE source image be adapted to ALL displays with equally satisfying viewing experience? There should be better answers than current industrial methods of squeezing, center cropping and black padding, and the representation should be more adaptive to apply to both HVS(Human Vision System) and heterogeneity of devices. Given a source image, summarization(retargeting) techniques aim at generating a more compacted yet pleasing version as *visual summary*, which has widespread applications in thumbnail generation, photo browsing, digital matting, net meeting, E-commerce, etc. In [11], visual summary results are evaluated by Completeness and Coherence, which requires more possible patches from the input data and fewer visual artifacts that were not in the input data.

To be more specific, two major concerns are involved in image summarization: (1) Information Maximization. The image summaries should contain enough information with equally satisfying viewing expectancy, ensuring important objects to be large enough to identify on smaller displays; (2) Deformation Minimization. It should introduce less distortion on salient regions than viewer's tolerance, which might be caused by aspect ratio change like mapping $4:3$ videos to $16:9$ wide screen of a HDTV set or iPhone.

The-state-of-art methods mainly adopted two types of features: visual saliency and image entropy. Setlur first introduced bi-layer segmentation for respective scaling of the filled-in background and removed objects [10], but his work is over-dependant on the segmentation results. Non-homogeneous pixel mapping is adopted by [13] which is expensive in calculation. Hou and Zhang [4] introduced a code length method to crop important patch as a thumbnail. It is limited with two separate objects in an image. Comparatively, Avidan and Shamir [1] took a discrete approach to add or remove "seams" of least energy backward or forward to draw a centralized effect, yet will cause noticeable deformation on smooth objects. Recently, Wang et.al.[12] presented an image resizing method with a scale-and-stretch mesh, distorting the structure of complex background and the relative proportion between object components. To sum up, bottom-up visual saliency and image entropy are unstable and irrelevant to scene layout, thus is insufficient for summarization problems.

Though the definition of "important regions" is controversial, due to the human-centered nature of summarization, object-oriented context saliency is considered as a more accurate description from physiology behavior analysis. Besides

Kadir and Brady's rarity assumptions for saliency [6], we argue that there should be three considerations: Firstly, rarity should based on the global statistical distribution of features among a group of similar images rather than local contrast; In addition, geometric constraints are essential in determining saliency as context information; Finally, the summarizing strategy has close relevance with not only the image content, but also the viewing distance and the real size of terminal displays.

In this paper, we proposed a *context saliency* based image summarization approach. After global redundancy, contrast and geometric analysis, we integrate them into the *context saliency* within a naive Bayesian framework by posteriori expectation-maximization. In addition, we bring up an effective grid framework for piecewise linear image warping to keep the proportion of context salient regions. This approach could elegantly merge the crop-based and warp-based methods and is adaptable to the scale and aspect ratio changes by *sweet spot* perception analysis.

2. CONTEXT SALIENCY

As the preprocessing step in many applications, how to select salient regions is a classical problem in computer vision. Besides the traditional physiological salient studies, we aim to draw a prediction on context importance as a more accurate description for the image content. Two assumptions are made about *context saliency*:

- The salient region is more likely to be foreground rather than background;
- The salient region is statistically rare, not belonging to common objects like trees, grass and mountains.

In other words, the salient region should be foreground, but being foreground does not always imply importance, such as trees in the foreground. By combining statistical saliency from their rare occurrence and geometric constraints from multiple visual features, we present a *context saliency* to integrate both visual attractiveness and semantic expectancy in a unified framework, as shown in Fig 1.

2.1. Statistical Saliency

As illustrated in Fig 1, the statistical saliency includes constrast saliency and redundant analysis. From the physiological perspective, visual attractiveness is essential, so we first compute the multiscale contrast feature $S(x, P)$ as a linear combination of contrasts in the Gaussian image pyramid [9]:

$$S(x, P) = \sum_{l=1}^{L} \sum_{x' \in r} \|I^l(x) - I^l(x')\|^2 \qquad (1)$$

where I^l is the l-th level image P in the pyramid. The number of pyramid levels L is 3 with a 9×9 window. r is the position

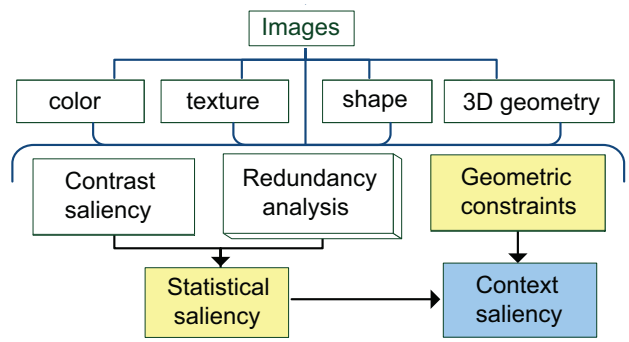

Fig. 1. Context Saliency.

of (i, j), and $S(x, P)$ is normalized to a fixed range of $[0, 1]$. A face detection module is added to improve the saliency map with the tree-structured multiview face detector(MVFD) [5] that $S'(x, P) = S(x, P) + \sum_{k=1}^{N} \pi_k N(p_k, v_k)$. p_k is the center of detected face while v_k is corresponding variances.

Then we adopt redundancy analysis to modify the visual attention into the statistical saliency. We use 50 outdoor and indoor images respectively for training, which are labeled with segmentation of foreground and background. Based on the contrast saliency map of each image, 100 feature points per image are selected under the principle that the less the saliency, the more possible it could be selected. As a result, the less important background is more likely to be selected. Next, a feature descriptor is computed by color histogram and gray-level co-occurrence matrix as texture-analysis in 9px $\times$ 9px non-overlapping patch. A simple K-means clustering is applied to get 7 groups of "sample patches" representing the sky, cloud, water, sand, ground, grass and tree, respectively. Afterward, the saliency map of each test image is modified by a sliding 9px $\times$ 9px window to "shrink" the saliency by contextual clues. We compute the information density considering both the image layout and the distance to "sample patches" based on the contrast saliency priori:

$$S_r = -log_2 \frac{S_{r-1}}{\min\{Dis(h_r, h_s)\}} \times S'_r(x, P) \qquad (2)$$

where S_{r-1} denotes the saliency of the last patch by the sliding window. $Dis(h_r, h_s)$ is the set of color histogram distances between current patch and sample patches. S_r is then normalized by $S'_r = \frac{S_r}{\max(S_r)}$. This method is more robust to local redundancies and blob noises, which also consider the global scene layout between foreground and background within the image. An example is shown in Fig 2.b.

2.2. Geometric Constraint

To extract geometric information, we adopt geometric context from a single image[3]. It estimates the coarse geometric properties of a scene by learning appearance-based models

of geometric classes with a multiple hypothesis framework. The superpixel label confidences, weighted by the homogeneity likelihood, are determined by averaging the confidence in each geometric label of the corresponding regions:

$$G(y_i = v \mid x) = \sum_{j}^{n_h} P(y_j = v \mid x, h_{ji}) P(h_{ji} \mid x) \quad (3)$$

where G is the label confidence, y_i is the superpixel label, v is a possible label value, x is the image data, n_h is the number of hypotheses and h_{ji} defines the region containing the ith superpixel for the jth hypothesis with the region label y_j. (See Fig 2.c)

2.3. Naive Bayesian Incorporation

With observed statistical saliency $S(x, y)$ and geometric constraint $G(x, y)$, the relation to the unknown desired feature $C(s)$ of context saliency could be modeled by a well-defined Bayesian framework using the maximum a posteriori (MAP) critierion. We try to find the most likely estimates for C, given S and G, which could be expressed as a maximization over a probability distribution P over a sum of $\log$ likelihood.

$$\begin{aligned} arg\max_{C} P(C \mid S, G) &= arg\max_{C} L(S \mid C) \\ &\quad + \log P(G \mid C) P(C) \quad (4) \end{aligned}$$

where $P(C)$ is set to 1 for simplicity. The final binary results of $C(x, y) \in (0, 1]$ are normalized as *context saliency map*(CSM) shown in Fig 2.d.

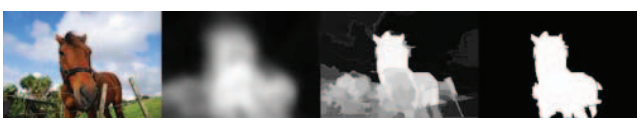

Fig. 2. From left to right: (a)Original Image; (b)Statistical Saliency; (c)Geometric Constraint; (d)Context Saliency Map

3. SWEET SPOT BASED IMAGE SUMMARIZATION

To adapt image summaries with different target devices, we build a grid scaleplate for non-homogeneous image warping. Differed from [12], we use rectilinear grids to reduce the structural deformation with less parameters. Because of the subjectiveness of summarization results, we take advantage of a multimedia perception study called "*sweet spot*"[7], which provides a quantified preference from end users.

3.1. Physiological Sweet Spot Evaluation

The experiment in [7] showed that extreme long shots were best when depicted actors were at least $0.7°$ high. We utilize this results in our warping methods by adjusting the weight discrepancy of CSM. B_h is the height of the bounding box on the focuses, initialized from the image center with the original aspect ratio. T_d is the diagonal length of the display and d is the distance between viewers' eyes and the screen. Generally we have $d = 3 \times T_d$, and according to *sweet spot* that $\frac{B_h}{d} > \tan 0.7°$. So given the specification of target device, we modify the CSM by declining the saliency outside the bounding box.

3.2. Image summarization

Each image is represented as a $2D$ grid $g = (M, Q)$, in which marks M are the coordinates and Q is a set of quads on the diagonal from lower left to upper right. $M = \{m_0^T, m_1^T, ..., m_n^2\}$ and n is the total quads number, which is initialized equally along the width and height of the output image. Based on CSM and grid line constraint, M is adjusted through minimizing the deformation energy, which is computed in a global optimization with gradient descent method in real-time. We measure the distortion by the weighted summation according to the slant angle differences between original and optimal quads on the diagonal:

$$D_s(M) = \sum_{m \in M(q)} \overline{c}_q (1 - \frac{m_q \cdot m_q^{'}}{\|m_q\| \times \|m_q^{'}\|})^2 \quad (5)$$

The weight $\overline{c}_q \in [0, 1]$ from CSM ensures that the deformation of importance regions contributes more to D_s. As a result, the change in the diagonal slant angles of these quads is minimized, meaning corresponding aspect ratio is constant while other parts absorb the distortion. After differentiating to $M(q)$, each vertical and horizontal mark is updated if ${D_s}^{k+1} < {D_s}^{k}$. It is repeated until k equals the maximal times of iteration, which depends on the image size and grid granularity. (e.g. 10 in the 448×336 frames by 20×20 rectilinear grid).

4. EXPERIMENTS

Since perceptual satisfaction is foremost in the task of summarization, we conducted user studies to assess the viewers' reactions. The experiment involved 300 images from VOC07 [2] to be downsized to 60% vertically and horizontally respectively. We randomly select 20 groups of results to be evaluated by viewers. Each subject was shown the original images compared to a random sequence of other results by seam carving [1], optimal scale-and-stretch mesh method [12] and our algorithm, who were asked to point out the most accepted one according to object clarity, photography invariance and overall perceptual quality. Results are given in Tab. 1. It can be seen that our results are more enjoyed by viewers. The optimized grid ensures local aspect ratio to be perceptibly acceptable. Typical results are shown in Fig 3(upper two rows).

	Vertically 60%	Horizontally 60%
Seam Carving [1]	10.5%	17.3%
Optimal mesh [12]	26.1%	18.2%
Our results	63.4%	64.5%

Table 1. User study for image summarization.

In addition to aspect ratio changes, we apply the method to thumbnail generation(middle two rows) and assisted image editing(lower two rows). We preserve both object proportion and clarity in the thumbnail compared to uniformly squeezing, while the CSM could serve as an region filtering technique that extract a group of important objects for further processing. Considering that the context saliency is a better representation for foreground and background, this framework is potentially available for image matting and collaging. It could also be easy to expand to video retargeting and dynamic image browsing with temporal constraint in Human-Computer-Interaction applications. More results can be seen at http://nlpr-web.ia.ac.cn/english/iva/homepage/jqwang/Demos.htm

5. CONCLUSIONS

We proposed a novel method for image summarization by *context saliency*, which combines improved contrast saliency and geometric information to formulate a more accurate description for semantic importance under Bayesian framework. Practically, our algorithm is based on perception preference and could be adapted to diversified displays flexibly with satisfying viewer experience. We build a computationally efficient and effective grid framework for non-homogeneous image warping by a global optimization to minimize the deformation. Results are evaluated by user studies on aspect ratio changes, thumbnail generation and assisted image editing, suggesting its potential on different applications.

6. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China No. 60833006, Natural Science Foundation of Beijing No. 4072025, and the 863 program No. 6083300660121302.

Fig. 3. Results of aspect ratio change compared to [1] and [12](Upper); Thumbnail generation by squeezing and our method(middle); Assisted image editing(lower two rows).

7. REFERENCES

[1] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. In *ACM Trans. Graph*, 2007.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[3] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, 2005.

[4] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. In *NIPS*, 2008.

[5] C. Huang, H. Ai, Y. Li, and S. Lao. Vector boosting for rotation invariant multi-view face detection. In *ICCV*, 2005.

[6] T. Kadir and M. Brady. Saliency, scale and image description. *IJCV*, V45(2):83–105, November 2001.

[7] H. O. Knoche and M. A. Sasse. The sweet spot: How people trade off size and definition on mobile devices. In *MM*, 2008.

[8] P. Levinson. Cellphone. *Palgrave/St. Martins, New York*, 2004.

[9] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. 2007.

[10] V. Setlur and S. Takagi. Automatic image retargetting. In *MUM*, 2005.

[11] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *CVPR*, 2008.

[12] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee. Optimized scale-and-stretch for image resizing. In *ACM Trans. Graph*, 2008.

[13] L. Wolf, M. Guttmann, and D. Cohen-Or. Non-homogeneous content-driven video retargeting. In *ICCV*, 2007.