A framework of salient object detection for images and videos

Jimmy Lin

supervised by Dr. Stephen Gould
College of Engineering and Computer Science
Australian National University

April 15, 2013

Problem Motivation

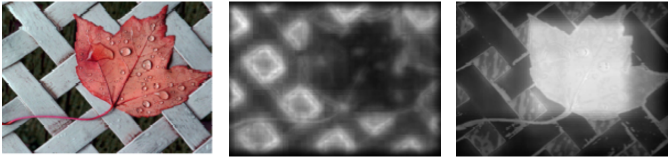
- What is salient object?

Fig. images from MSRA datasets B

  - high contrast in object's boundary
  - intensive color spatial distribution
  - spatial / temporal continuity
- Why to detect salient objects?
  - Adaptive image display on small devices
  - Automatic image cropping
  - Image/video compression
  - ...

Related Works

- Salient-based Model (SM,1998)
  - Itti, Laurent, Christof Koch, and Ernst Niebur. "A model of saliency-based visual attention for rapid scene analysis." *Pattern Analysis and Machine Intelligence, IEEE Transactions on 20.11 (1998): 1254-1259.*
- Fuzzy Growing Method (FG,2003)
  - Ma, Yu-Fei, and Hong-Jiang Zhang. "Contrast-based image attention analysis by using fuzzy growing." *Proceedings of the eleventh ACM international conference on Multimedia. ACM, 2003.*
- CRF-based Model (CRFM,2007)
  - Liu, Tie, et al. "Learning to detect a salient object." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*
  - iu, Tie, et al. "Learning to detect a salient object." *Pattern Analysis and Machine Intelligence, IEEE Transactions on 33.2 (2011): 353-367.*



(a) input image (b) feature map of SM (c) feature map of CRFM

Problem Formulation

- Binary Labelling Task,
  For each pixel x, $a_x \in \{0, 1\}$ indicate whether pixel x belongs to salient object.
- For image,
  One single image I,
  Corresponding Binary Mask A,
  Probabilistic model $P(A|I) = \frac{1}{Z} exp(-E(A|I))$.
- For video,
  A sequence of image $I_1, I_2 \dots I_N$,
  Corresponding sequence of Binary Mask $A_1, A_2 \dots A_N$,
  PM $P(A_{1,\dots,N}|I_{1,\dots,N}) = \frac{1}{Z} exp(-E(A_{1,\dots,N}|I_{1,\dots,N}))$.

$$E(A_{1,\dots,N}|I_{1,\dots,N}) = \sum_{t=1}^{N} E(A_t|I_{1,\dots,N}) = \sum_{t=1}^{N} E(A_t|I_{t-1}, I_t)$$

- Formulating Energy Function
- Learning and Inference for CRF model
- Evaluating the result of model

Formulation in a single image

Energy function is formualted as

$$E(A|I) = \sum_x \sum_{k=1}^{K} \lambda_k F_k(a_x, I) + \sum_{x,x'} S(a_x, a_{x'}, I)$$

λ_k: weight of kth feature, x, x': two adjacent pixels.

Static salient feature. $F_k(a_x, I)$ is formulated from a normalized feature map $f_k(x, I) \in [0, 1]$ for every pixel, written as:

$$F_k(a_x, I) = \begin{cases} f_k(x, I), & a_x = 0 \\ 1 - f_k(x, I), & a_x = 1 \end{cases}$$

Pairwise feature. $S(a_x, a_{x'}, I)$ exploits the spatial relationship between two adjacent pixels and can be viewed as a penalty to adjacent pixels that are assigned with different labels.

$$S(a_x, a_{x'}, I) = |a_x - a_{x'}| \cdot exp(-\beta d_{x,x'})$$

where $d_{x,x'} = ||I_x - I_{x'}||_2$ is the L2-norm of color difference, and $\beta = (2\langle||I_x - I_{x'}||^2\rangle)^{-1}$ is robust parameter weighting the color contrast.

Liu, Tie, et al. "Learning to detect a salient object." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*

Formulation in sequential images

Energy function is formulated as

$$E(A_t|I_t, I_{t-1}) = \sum_x \Big(\sum_{k=1}^{K} \lambda_k F_k(a_x, I_t) + \sum_{k=K+1}^{K+L} \lambda_k F_k(a_x, M_t) + \lambda_0 F(a_x, I_{t-1}, I_t) \Big) + S(a_x, a_{x'}, I_t),$$

Motion salient features. Processing is similar to Static salient features, but based on motion field M_t of the image I_t. The motion field is obtained by using the SIFT flow technique.

Fig. (a) Original Image (b) Motion Field

Appearance coherent feature. This feature $f(x, I_{t-1}, I_t)$ penalizes the pixels that are identified to be in the salient object, but with a large color difference between the surrounding regions from two adjacent frames.

Liu, Tie, et al. "Learning to detect a salient object." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*

Liu, Ce, et al. "SIFT flow: dense correspondence across different scenes." *Computer VisionECCV 2008. Springer Berlin Heidelberg, 2008. 28-42.*

Local Feature: Multiscale Contrast

Contrast is the most commonly used local feature for attention detection because it simulates the human visual receptive fields.

(a) Input image (b) Contrast maps at multiple scales (c) feature map

Multiscale contrast feature $f_c(x, I)$ is defined as a linear combination of contrasts in the Gaussian image pyramid:

$$f_c(x, I) = \sum_{q=1}^{Q} \sum_{x' \in N(x)} ||I^q(x) - I^q(x')||^2$$

where $N(x)$ is 9×9 windows, q is the index for the scales in pyramid.

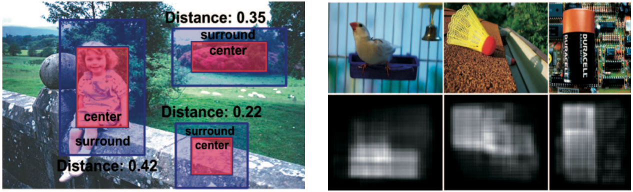
Liu, Tie, et al. "Learning to detect a salient object." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*

Regional Feature: Center-Surround Histogram

Find the most distinct rectangle, $R^*(x)$, centered at each pixel x by varying the size and aspect ratio:

$$R^*(x) = \arg\max_{R(x)} \chi^2(R(x), R_S(x))$$

Five templates of aspect ratios $\{0.5, 0.75, 1.0, 1.5, 2.0\}$.
Size range of rectangle $R(x)$ can be set as $[0.1, 0.7] \times min(w, h)$.



(a) Center-Surround Histogram (b)Feature Maps of Center-Surround Histogram

The center-surround histogram feature $f_h(x, I)$ is defined as

$$f_h(x, I) \propto \sum_{\{x' | x \in R^*(x')\}} w_{xx'} \chi^2(R^*(x'), R_S^*(x'))$$

Liu, Tie, et al. "Learning to detect a salient object." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*

Global Feature: Color Spatial Distribution

We use Gaussian Mixture Models (GMMs) $\{w_c, \mu_c, \Sigma_c\}_{c=1}^{C}$ to represent all colors in the image. Each pixel is assigned to a color component with the probability

$$P(c|I_x) = \frac{w_c \mathcal{N}(I_x|\mu_c, \sigma_c)}{\sum_c w_c \mathcal{N}(I_x|\mu_c, \sum_c)}$$

Fig. three examples making use of global feature (a) grass (b) sky (c) soil

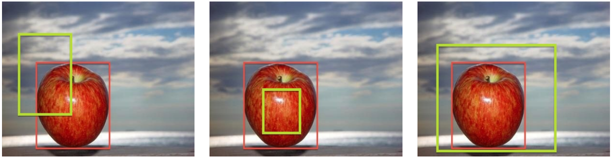
Then we compute composite variance and normalize it to $[0, 1]$. Finally define the color spatial-distribution feature as,

$$f_s(x, I) \propto \sum_c p(c|I_x) \cdot (1 - V(c))$$

Liu, Tie, et al. "Learning to detect a salient object." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*

Evaluation Criteria

- Region-based measurement



(a) arbitrary labelling (b) large prec but low recall (c) large recall but low prec

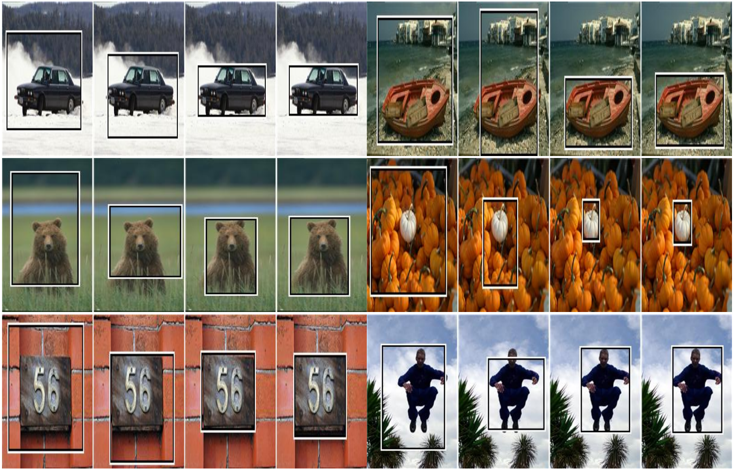
  - Ratio of Precision to Recall
    Precision: % of pixels that are correctly detected in ground truth
    Recall: % of pixels that are correctly detected in resulted detection
  - F-Measure

$$F_\alpha = \frac{(1+\alpha) \times Precision \times Recall}{\alpha \times Precision + Recall}$$

- Boundary-based measurement
  - Boundary Displacement Error (BDE)
    Measures the average of positional difference of ground truth and resulted detection.

Liu, Tie, et al. "Learning to detect a salient object." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*

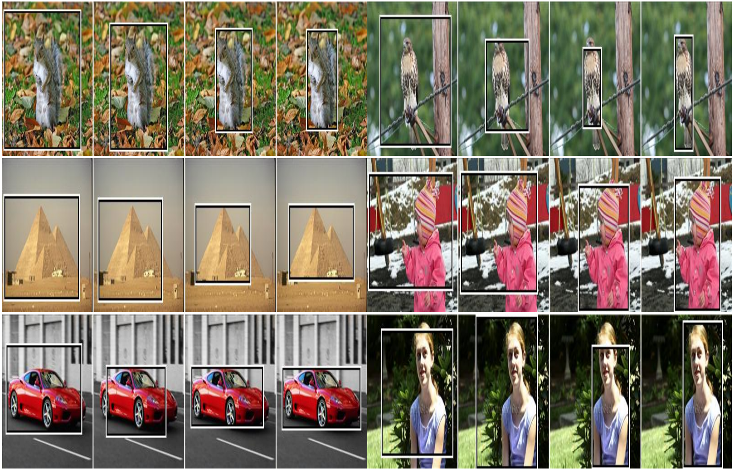
Result Comparisons: perfect detection



(a) FG(Ma,2003) (b) SM(Itti,1998) (c) CRFM(Liu,2007) (d) Ground truth

Liu, Tie, et al. "Learning to detect a salient object." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*

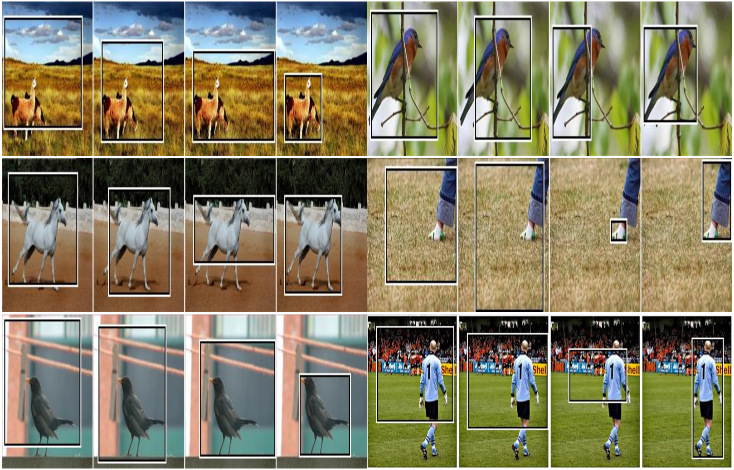
Result Comparisons: decent detection



(a) FG(Ma,2003) (b) SM(Itti,1998) (c) CRFM(Liu,2007) (d) Ground truth

Liu, Tie, et al. "Learning to detect a salient object." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*

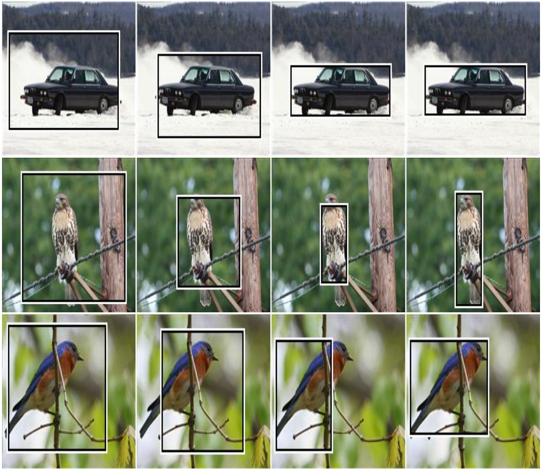
Result Comparisons: terrible detection



(a) FG(Ma,2003) (b) SM(Itti,1998) (c) CRFM(Liu,2007) (d) Ground truth

Liu, Tie, et al. "Learning to detect a salient object." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*

Thank You! Suggestions and Questions Please.



Liu, Tie, et al. "Learning to detect a salient object." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*