Application of Conditional Random Fields in Pixel-Level Salient Object Detection within an Image using Local, Regional, and Global Features

Jimmy Lin
Chris Claoue Long

Dr. Stephen Gould

College of Engineering and Computer Science
Australian National University

May 22, 2013

Introduction and Motivation

Fig.1 Images from MSRA dataset B

Saliency is the prominence of an object in an image.

Salient object detection is useful in numerous areas, for instance, in siimulating human vision by robots, augmented Reality, 3D surface reconstruction and more.

Often detected by its **high contrast boundary** to its near neighbours, **distinction from its surrounds**, **intensive colour distribution** compared to all other color component in candidate image and **space continuity of saliency**.

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)

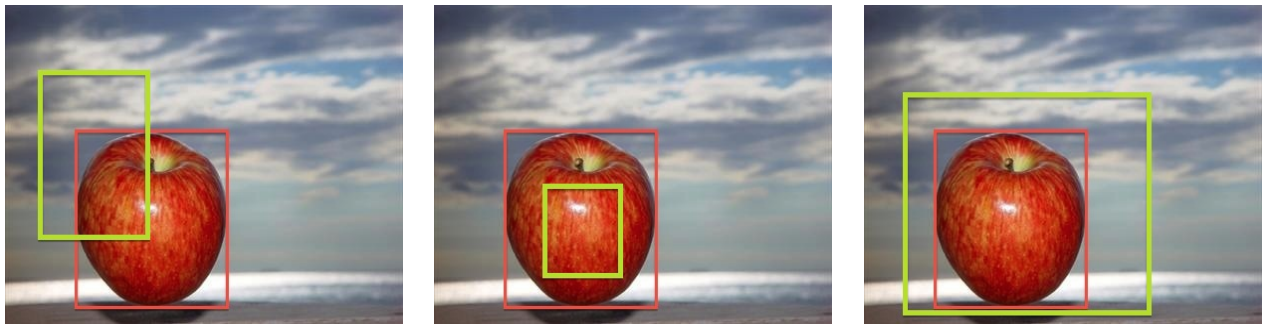
  Liu, Tie, et al. "Learning to detect a salient object." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*

  iu, Tie, et al. "Learning to detect a salient object." *Pattern Analysis and Machine Intelligence, IEEE Transactions on 33.2 (2011): 353-367.*

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.*

Comparisons between Existing Approaches

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

Formulation

Given an image I, we want to compute the location of a salient object.

Binary labelling task – for each pixel x, indicate whether it belongs to the salient object (1) or not (0). Thus, our objective is to have corresponding map A, indicating binary saliency of one pixel.

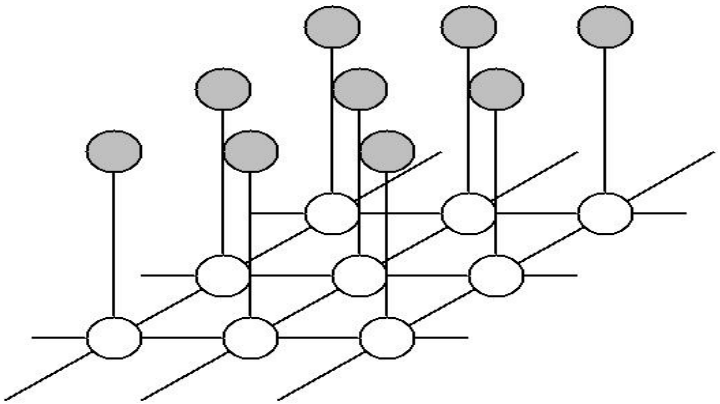


Fig.2 graph for Conditional Random Field

Build up a probabilistic model $P(A|I) = \frac{1}{Z}e^{-E(A|I)}$, where $\frac{1}{Z}$ is the normalising factor, and $E(A|I)$ is the energy function incorporating both unary and pairwise potentials between pixels.

Pairwise Feature

Formally, the energy function can be represented as

$$E(A|I) = \sum_{x} S_{unary}(a_x, I) + \lambda_0 \sum_{x,x'} S_{pair}(a_x, a_{x'}, I)$$

where λ is the relative weight between the summary of multiple unary features and pairwise features.

The pairwise feature $S(a_x, a_{x'}, I)$ exploits the spatial relationship between two adjacent pixels. It can be viewed as a "penalty" for labelling adjacent pixels the same or differently.

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

where x, x' represent two adjacent pixels, $d_{x,x'}$ is the L2-norm (standard norm) representing the colour difference between the two pixels, and $\beta = (2\langle ||I_x - I_{x'}||^2 \rangle)^{-1}$ is a robust parameter to weight the colour contrast.

Unary Features Combination

The unary potential for combination of three features is specified as

$$S_{unary}(a_x, I) = \sum_{k=1}^{K} \lambda_k \cdot F_k(a_x, I)$$

where λ_k is the weight of the k^{th} feature conforming to $\sum_{k=1}^{K} \lambda_k = 1$.

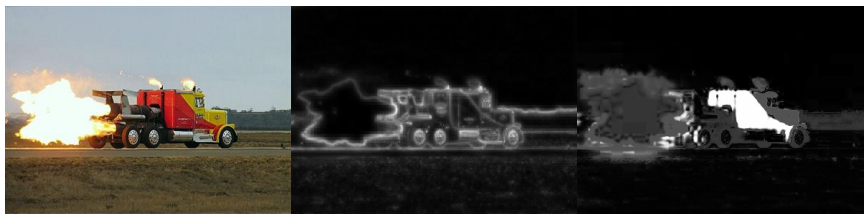


Fig.3 Preview of feature maps

The value of each feature $F_k(a_x, I)$ comes from a normalised feature map $f_k(x, I) \in [0, 1]$, and for each pixel:

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

Learning

It is technically difficult to directly compute the optimisation of the following,

$$E(A|I) = \sum_{x} \sum_{k=1}^{K} \lambda_k \cdot F_k(a_x, I) + \lambda_0 \sum_{x,x'} S_{pair}(a_x, a_{x'}, I)$$

Thus, we separate the different types of potentials to derive an approximate solution

$$S_{unary}(a_x, I) = \sum_{k=1}^{K} \lambda_k \cdot F_k(a_x, I)$$

This can be achieved by the logistic regression.

Then we optimise λ_0 as an ordinary CRF learning task.

$$E(A|I) = \sum_{x} S_{unary}(a_x, I) + \lambda_0 \sum_{x,x'} S_{pair}(a_x, a_{x'}, I)$$

It is somewhat similar to the Coordinate Descent Method with one iteration.

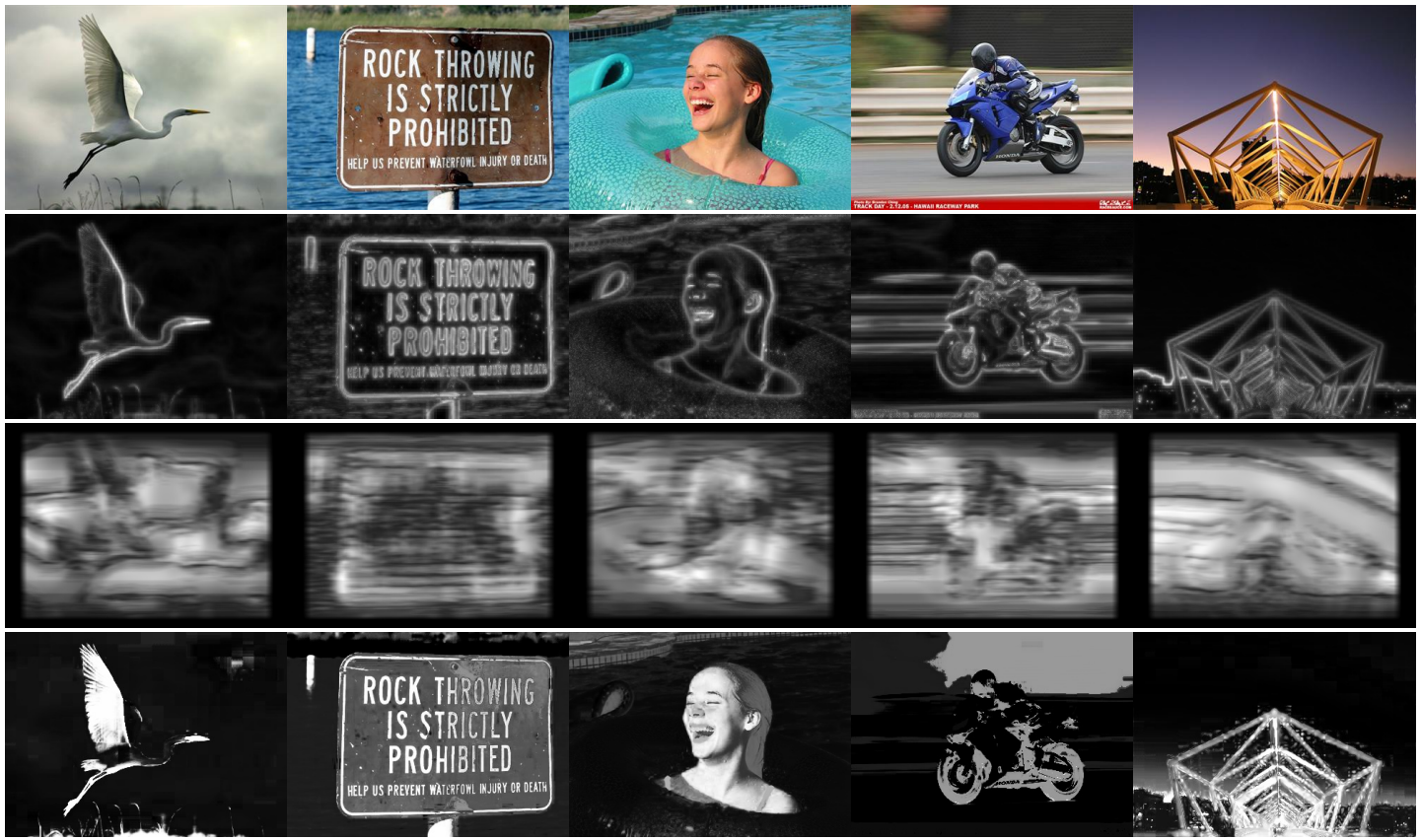
Inference

After the linear coefficient λ is determined, we can infer the binary saliency mask of the image.

CRF inference is used to determine the assignment of binary saliency of each pixel a_x, which minimises the energy function

$$E(A|I) = \sum_x \sum_{k=1}^{K} \lambda_k \cdot F_k(a_x, I) + \lambda_0 \sum_{x,x'} S_{pair}(a_x, a_{x'}, I)$$

Feature Extraction



From Top to Bottom Row: (1) Original Image (2) Local: MultiScale Contrast (3) Regional: Center-Surround Histogram (4) Global: Color Spatial Distribution

Local: Multiscale Contrast

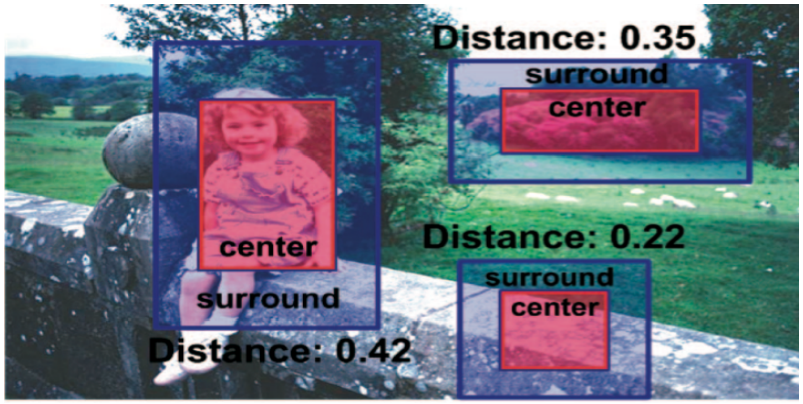
Create a contrast map from the linear combination of image contrast at all levels of an N-level gaussian image pyramid, using the pixels x in the image I:

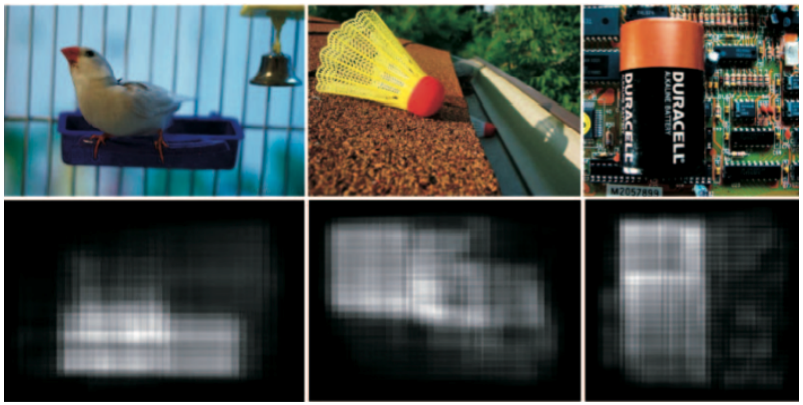
$$f_c(x, I) = \sum_{n=1}^{N} \sum_{x' \in W(x)} ||I^n(x) - I^n(x')||^2$$

where W(x) is a window that delineates which area to consider for neighbouring pixels to compare contrast values.

Regional: Center-Surround Histogram

Given a rectangle $R_s(x)$around a salient region, create a frame $R(x)$ around it so that the area of the frame is equal to that of the rectangle (this is displaced as needed to fit into the image dimensions), at a suitable aspect ratio.





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

Regional: Center-Surround Histogram

Create a colour RGB histogram for both the rectangle and the surrounding frame with a certain resolution (number of "bins" for each colour)

Calculate the χ^2 value between the two histograms to obtain the differences between the rectangle and the surrounding frame. Do this for multiple aspect ratios, and keep the largest χ^2 value:

$$R(x) = \arg\max_{R(x)} \chi^2(R(x), R_s(x)) = \arg\max_{R(x)} \frac{1}{2} \cdot \sum_{i \in bins} \frac{(hist_{R(x)_i} - hist_{R_s(x)_i})^2}{hist_{R(x)_i} + hist_{R_s(x)_i}}$$

The center-surround histogram feature is finally calculated as:

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

Global: Colour Spatial Distribution

Create a Gaussian Mixture Model to compute the spatial variance and continuity of colour in an image.

The component model is created from only a subset of the pixels in the image, and the maximum number of iterations is limited in order to reduce the time taken to compute this feature without sacrificing too much accuracy.

Global: Colour Spatial Distribution

Each pixel is associated to a colour component with the probability

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

where ω_c is the weight, μ_c is the mean colour, σ_c is the covariance, and $\mathcal{N}(I_x|\mu_c, \sigma_c)$ is the multivariate normal distribution of the c^{th} *component*

The final colour spatial distribution feature is defined as a weighted sum:

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

where $V(c)$ is the normalised covariance (horizontal and vertical variances) of the c^{th} component, contained between 0 and 1.

Our variations to previous works

Separate learning of parameters in the CRF model

Improved speed and performance of the given feature extraction processes

- Reducing the resolution of the image before calculating its colour spatial distribution, and limiting the complexity of the gaussian model fitting without reducing overall quality of the feature map:
- Accounting for center-surround histograms that would otherwise exceed the boundary of the image, especially in corners
- Improved calculation of the aspect ratio to use when calculating the center-surround histogram frame

Our Project Progress

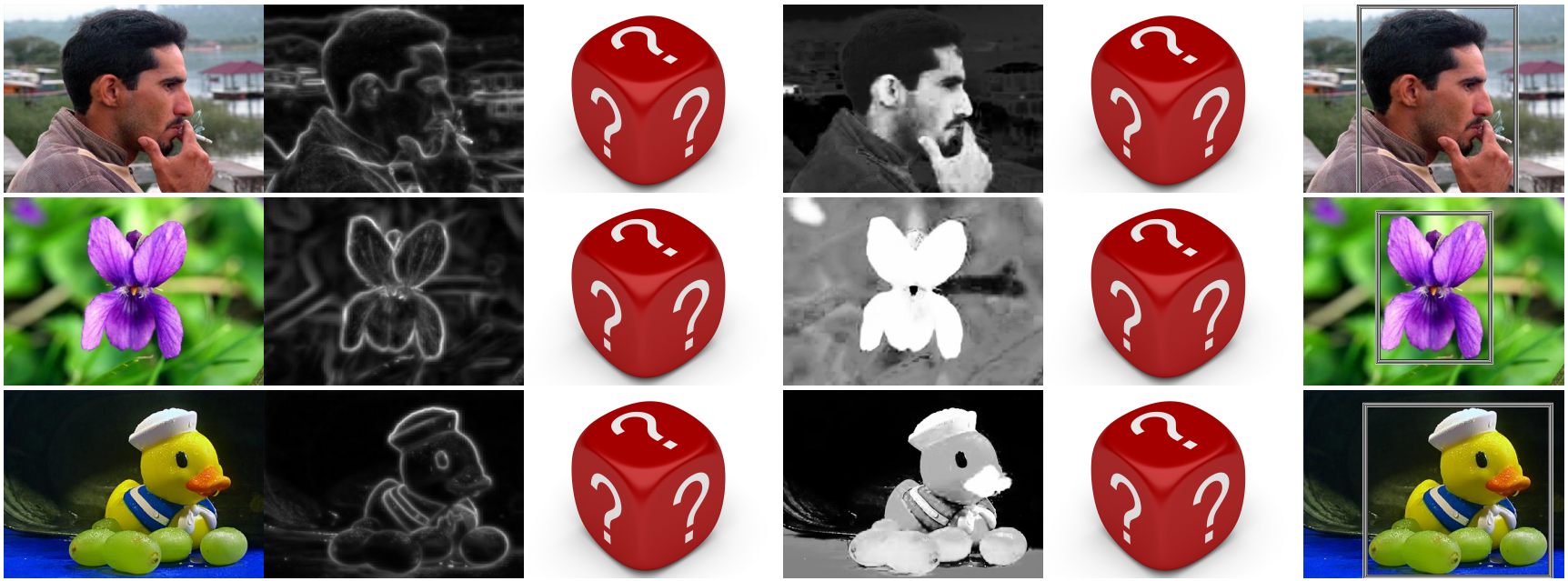


Fig.5 Our implementation

Thank you! Questions?

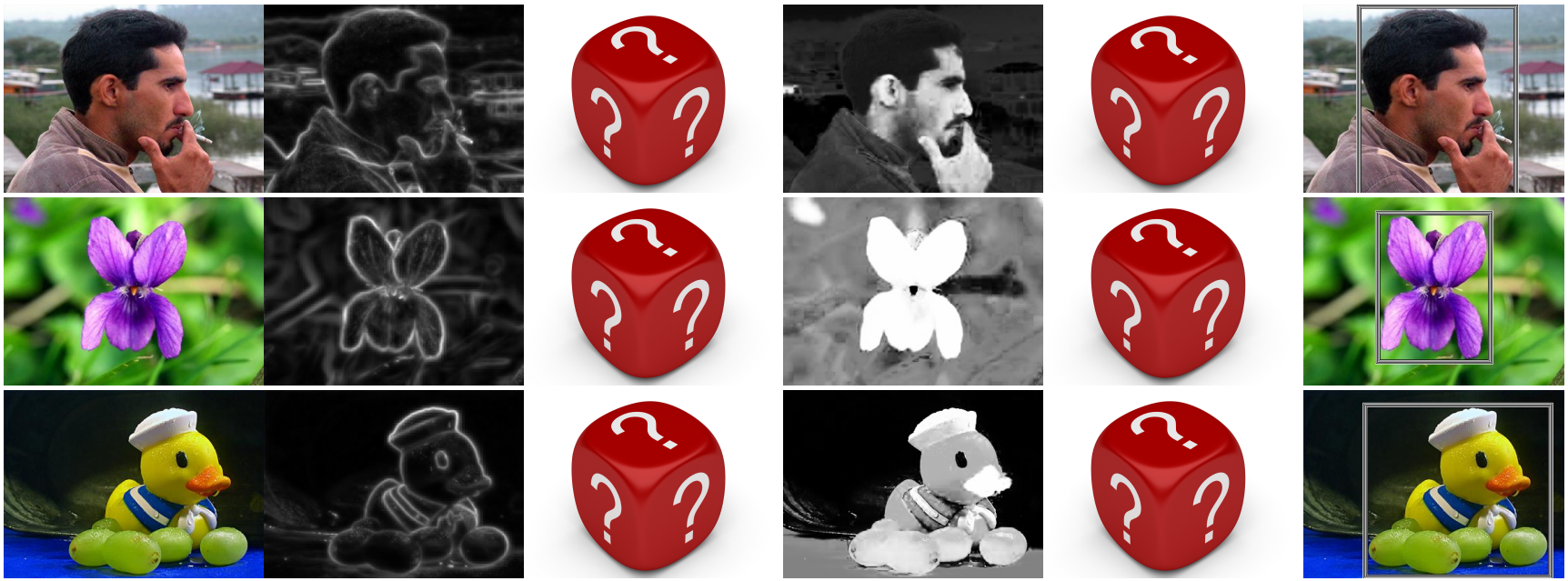


Fig.5 Our implementation

References


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


Liu, Tie, et al. "Learning to detect a salient object." *Pattern Analysis and Machine Intelligence, IEEE Transactions on 33.2 (2011): 353-367.*


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.*


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.*


Stephen Gould, "DARWIN: A Framework for Machine Learning and Computer Vision Research and Development", *Journal of Machine Learning Research (JMLR), 13(Dec):35333537, 2012.*