TouchCut: Fast Image and Video Segmentation using Single-Touch Interaction

Tinghuai Wang[a,*], Bo Han[b], John Collomosse[a]

[a] *Centre for Vision, Speech and Signal Processing, University of Surrey, UK.*
[b] *Sony China Research Laboratory, Beijing, China.*

Abstract

We present TouchCut; a robust and efficient algorithm for segmenting image and video sequences with minimal user interaction. Our algorithm requires only a single finger touch to identify the object of interest in the image or first frame of video. Our approach is based on a level set framework, with an appearance model fusing edge, region texture and geometric information sampled local to the touched point. We first present our image segmentation solution, then extend this framework to progressive (per-frame) video segmentation, encouraging temporal coherence by incorporating motion estimation and a shape prior learned from previous frames. This new approach to visual object cut-out provides a practical solution for image and video segmentation on compact touch screen devices, facilitating spatially localized media manipulation. We describe such a case study, enabling users to selectively stylize video objects to create a hand-painted effect. We demonstrate the advantages of TouchCut by quantitatively comparing against the state of the art both in terms of accuracy, and run-time performance.

Keywords: Object cut-out, Touch interaction, Image and video segmentation, Level set methods

1. Introduction

The segmentation of objects from cluttered natural images remains a fundamental and inherently challenging Computer Vision problem. The task

*Corresponding author.
Email addresses: tinghuai.wang@surrey.ac.uk (Tinghuai Wang), han_bo@outlook.com (Bo Han), j.collomosse@surrey.ac.uk (John Collomosse)

is generally regarded as under-constrained since, in the absence of high level scene understanding, there can be more than one interpretation of pixels comprising the desired object of interest or 'foreground' object. The past decade has seen a trend toward better constraining the segmentation task through: i) the development and increased reliance on global optimization methods; and ii) the combination of high-level prior scene understanding via *user interaction* with low-level cues such as color and edges observed in the image.

A key challenge of interactive segmentation is to maximize user provided prior information whilst minimizing user intervention. Although recent years have delivered significant advances, a considerable amount of user intervention is still required to achieve a satisfactory segmentation. Typically this involves the user indicating positive and negative class examples of pixels or regions. Often such indications requires correction either automatically to boost discrimination between by the positive or negative classes [55], or by the user iteratively working with the system to supply additional constraints [48]. Regardless of the interaction modality, the goal of any interactive image segmentation is to minimize the amount of effort to cut out a desired object while accurately selecting objects of interest.

To address this problem, this paper contributes an efficient algorithm for object segmentation driven by minimal user effort — a single touch. In contrast to previous interaction modes, ranging from roughly marking the desired boundary [29, 7, 65] to loosely drawing scribbles labeling the desired object and the background [8, 30, 4], to placing a bounding box around the desired object [57, 40], our system requires only a single (x, y) coordinate from the user offering an intuitive and maximally "economical" interaction method. Our method has clear applications on emerging touch-screen tablet, mobile and pervasive devices, the form factor of which devices may be inconvenient for fine-motor interaction. Further, in some use cases (e.g. deployment in high throughput, or multi-tasking situations such as driving) the high cognitive load required to trace outlines or regions may also be undesirable.

Our core technical contribution is a new model for object segmentation that fuses edge, region, and geometric cues within a *level set* [71, 52] framework. In contrast to previous expanding contour approaches relying on intensity gradient, our proposed model incorporates a probabilistic estimate of edge location derived from a novel dominant color extraction scheme. This scheme offers improved robustness when filling color or texture coherent regions, leading to more accurate localization of the desired object's boundary.

We also fuse this boundary information with a region-based maximum *a posteriori* (MAP) criterion designed to promote color similarity with pixels local to the touched foreground point. The robustness of this region-based criterion is further enhanced by a consistency constraint enforcing uniformity of deviation from the foreground color model, learned from the touched foreground region. This approach to color consistency, combined with a novel per-pixel adaptive weighting scheme, mitigates the tendency for contour expansion to skip the real boundary when color models of the foreground and background are indistinct. Finally, our proposed model also utilizes the geometric cue implied in single-touch input, that users typically touch image in close vicinity to the geometric center of the desired object.

All together, our edge-region-geometry model provides a robust and flexible description of the interactive object segmentation problem, leveraging the flexibility of level set methods to promote accurate boundary placement and strong region connectivity while requiring minimum user interaction. Using an incrementally built foreground color model, our framework also extends to address the temporally coherent video object segmentation problem, creating regions whose shape and neighborhood topology evolve smoothly over time whilst tracking the underlying video content. A motion estimation enabled shape prior is further introduced into the video adaptation to preserve temporal coherence when the foreground and background color distributions are indistinct.

Following a literature review, we briefly revisit level set methods in Sec. 3. We then describe the proposed framework for interactive object segmentation on still images (Sec. 4), explaining each of the energy terms comprising the proposed energy functional. We extend our system on still images to video sequences in Sec. 5, presenting an application of our proposed algorithm to tablet-based video manipulation. We present a comparative evaluation with previous work in Sec. 6 on both a qualitative and quantitative basis, concluding in Sec. 7.

2. Related Work

Image segmentation underpins many visual analysis tasks and is a long-standing research topic in Computer Vision. Over the past decade a number of successful interactive approaches have emerged, enabling the user to *seed* or '*scribble*' on part of the desired object and background to initialize the segmentation [8, 45, 30, 4, 40, 61, 37]. This approach is intuitive and generally tolerant of low accuracy user input, though requires the user to trace

a contour contacting multiple points in the image. Drawing a bounding box [57, 40] to constrain the spatial extent of object is simpler in many cases, taking two mouse clicks to specify the box. Yet scribble-based corrections are often needed to refine the results as the bounding box may not provide sufficiently tight capture for some object shapes.

Many methods [4, 56] driven by scribbles selectively fill the desired region by expanding from the interior of the selected object outwards and do not explicitly consider the object boundary. This make it advantageous for segmenting objects with complex topologies, whilst it may suffer from a bias that favors shorter paths from the seeds. Another drawback of these methods is that it may fail to accurately identify the real object boundaries due to the lack of an explicit presentation of edge contrast.

An important class of seeded segmentation algorithm are those performing a Graph Cut optimization, after Boykov and Jolly [8] who address object segmentation in images via max-flow/min-cut energy minimization. Typical energy functionals balance the probability of pixels belonging to the foreground/background with spatial coherence constraints expressed via edge contrast. The user-specified scribbles serve as hard constraints and also provide statistical information. This region-edge combination is very effective in improving segmentations based on edge or region alone. However, there is an inherent bias of graph cut towards shorter paths, i.e. small segments as the optimization sums over the boundaries of segmented regions. By contrast, level set based methods include a length-based "ballooning" term which encourages a larger object segment (Fig. 10).

Our work is most closely related to prior image segmentation approaches using level set methods [71], which neatly enable the minimization of energy functionals such as those proposed by Mumford-Shah [49] or Zhu-Yuille [72].

An early application of level sets to image segmentation was the edge-based active contour model [13, 35]. This approach depends primarily upon image gradient and therefore is sensitive to clutter and image noise, which can cause convergence to local minima that do not well describe the intended region shape. More robust approaches that encode region information were proposed later by Paragios *et al.* [51], who also tackle video segmentation, by modelling inter-frame difference using Laplacian or Gaussian distributions. Chan and Vese [14] propose the approach of active contours also harness Gaussian distributions, partitioning objects via intensity distributions modelled with different variances. Heiler *et al.* [32] also adopt Laplacan distributions for natural image segmentation.

Non-parametric methods have also been proposed to model the intensity

distribution of image regions [33, 36, 11]. Although the method of modelling the distribution varies (e.g. histograms, compact encodings) all assume that pixels belonging to one region all share the same probability distribution and thus can not handle intensity inhomogeneity in the sought regions. More recent work [10, 39, 41] have been proposed to incorporate local intensity statistics instead of modeling the intensity distribution globally for each region.

Higher level prior knowledge such as geometric shape priors have been introduced to level set framework [59, 53, 21, 9, 62, 58, 18, 20]. Variational integration of the shape prior based on the assumption of a Gaussian distribution were proposed by Rousson and Paragios [59]. Implicit representations and distance transforms were considered to represent shape in a higher dimensional framework for optimization under level sets by Paragios *et al.* [53]. Tsai *et al.* [62] proposed a very efficient implementation of shape-driven level set segmentation by directly optimizing in the linear subspace spanned by the principal components. The use of nonparametric density estimation to model larger classes of level set based shape distributions was proposed by Cremers *et al.* [18] and Rousson and Cremers [58]. Paragios [50] proposes to integrate user interactive constraints to level set framework by taking into account distance from seeds, which is similar in aim to the geometric prior proposed in our work. Yezzi *et al.* have explored the fuion of multiple cues, including geometry, in their segmentation frameworks [22, 34] and so share our goal. A more complete survey of previous works integrating multiple priors to level set framework, the reader is refereed to [19]. The incorporation of shape priors has also been investigated in non-level set based segmentation techniques. In particular, geometric priors based on distance from object centre have been incorporated into Graph Cut approaches [64, 31]. We later compare our proposed technique against GrabCut modified to incorporate such a shape prior [64].

One of the appealing advantages of level set methods is that they can neatly enable flexible forms of energy functionals. However, they are prone to getting stuck in a local minimum frequently caused by the sensitive edge-based term. Early approaches to edge detection aim at identifying the presence of a boundary through local measurements, such as Sobel operator [27] and Canny detector [12]. Recent local approaches incorporate color and texture information, either taking advantage of learning techniques for cue combination [47, 26] or observing the local distribution of quantized color class labels without estimating a specific model for a texture region [24]. The latter approach provides a more efficient means to calculate an implicit

edge indicator and is adapted in this paper.

The implicit edge detector proposed in [24] employs quantized color label as the basis of the color-texture property analysis. However, this can lead to an incorrect observation as color homogeneous regions with minor color or luminance variance might be treated non-homogeneous due to the sensitivity of color quantization. To circumvent this issue, we develop an efficient and robust dominant color extraction algorithm to facilitate the acquisition of color class label in [24]. A dominant color (DC) is defined as a set of similar colors, of which the corresponding pixels occupy a relatively large proportion in (a specific region of) an image. There have been many approaches proposed to address the DC extraction problem. Lin and Zhang [44] extract coarse DCs by considering each local maximum and its neighborhood within a diameter-fixed sphere in the HSV color space as a possible DC. Wang *et al.* [67] adopt EM algorithm to estimate the GMMs of the input colors. However, it is difficult to properly set the number of the Gaussians. The Generalized Lloyd Algorithm (GLA) is adopted to divide the input colors into clusters in [23, 25]. Because the GLA aims at minimizing the global quantization distortion, the color ranges with high frequency are apt to be over-divided, and those with low frequency are apt to be under-divided. The Mean Shift algorithm is adopted to identify the dominant colors in [17]. However, it suffers from scale problem which makes it difficult to adaptively make a good trade-off between precision, robustness and roughness in the color histogram. To address these problems, we propose a novel non-parametric DC extraction algorithm which considers both color distribution and color similarity, to better explore the inherent characteristics of DC.

We additionally extend the proposed TouchCut segmentation system from still images to a video sequence, so facilitating interactive video object segmentation. Interactive video object segmentation systems have been proposed in recent years. Various directions have been investigated such as tracking region boundaries over time [2, 54], extending 2D segmentation to 3D video volumes [43, 66, 3, 4], applying graph cut segmentation on successive frames driven by motion flow [6, 5, 70, 68]. Our algorithm propagates the foreground mask forward which initializes the segmentation on the new frame and provides a shape prior taking account of the inherent error in optical flow.

3. Level Set Implementation of Active Contours

The basic idea of active contour models implemented via level set methods is that a contour $\mathcal{C}$ in a domain Ω can be represented by the zero level set

of a higher level embedding function ϕ: $\Omega \rightarrow \Re$. Evolving the contour $\mathcal{C}$ is achieved by evolving the embedding function ϕ which is defined as the signed distance function with $\phi > 0$ inside the contour, $\phi < 0$ outside the contour and $|\nabla\phi| = 1$ almost everywhere.

The evolution of the level set function ϕ is governed by a partial differential equation (PDE). One can directly derive the PDE from a certain energy functional $E(\phi)$ on the space of level set functions. Subsequently one can derive the Euler-Lagrange equation which minimizes $E(\phi)$:

$$\frac{\partial\phi}{\partial t} = -\frac{\partial E(\phi)}{\partial\phi}$$

These methods are known as variational level set methods [71]. Thus the segmentation boundary $\mathcal{C}$ is derived by obtaining the best ϕ at the zero level as

$$\mathcal{C} = \{x \in \Omega \mid \phi(x) = 0\}$$

The level set function is not uniquely defined. Depending on the chosen embedding, one can obtain slightly different evolution equations for ϕ. This formulation enables direct incorporation of edge, shape and statistical prior information into the design of $E(\phi)$ in the segmentation framework. For instance image gradient is often incorporated to the energy function to align the contour with the boundaries of object, driven by an additional balloon force which leads to either a shrinking or an expansion of contours. It is also amenable to the introduction of shape information which was shown to significantly improve the segmentation of known objects in a given image [20].

4. Segmentation Framework of Still Images

In the level set paradigm, we propose a new energy functional taking account of probabilistic edge map, color distribution of foreground and background in an adaptive manner as well as the geometric cue implied by user touch:

$$E(\phi) = [\omega_e\ \omega_a\ \omega_b\ \omega_u\ \omega_g][E_e(\phi)\ E_a(\phi)\ E_b(\phi)\ E_u(\phi)\ E_g(\phi)]^T + \mathcal{R}(\phi) \quad (1)$$

where

- $E_e(\phi)$ is the edge probability term, which is minimized when the zero level contour of ϕ is located at the object boundaries; it is essential for the contour evolution to stop at the desired object boundary

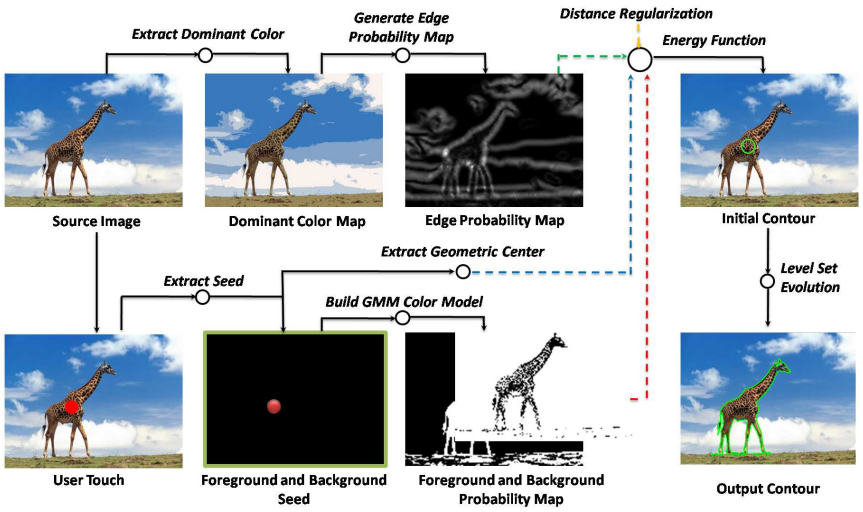


Figure 1: System overview. Dominant color extraction is performed on the input image for calculating the edge probability map (first row). Foreground/background color model is estimated based on user input and the image border respectively (second row). The energy function incorporates the various energy terms. The evolution of the embedding function ϕ is specified by the energy function (right column). The zero level contour converges to the object boundary to generate the segmentation (bottom right).

- $E_a(\phi)$ is the ballooning term, which is introduced to speed up the motion of the zero level contour in the level set evolution process, necessary when the initial contour is placed far away from the desired object boundaries

- $E_u(\phi)$ is the foreground consistency term is tailored for single-touch interaction which preserves the consistency of appearance within zero level contour; the foreground appearance model has higher confidence than the background appearance model, especially when the desired object intersects the border of the image

- $E_b(\phi)$ is the Bayes statistical error term based on color distributions which measures the *a posterior* probability of the regions inside and outside the contour under the given prior, warranting a better separation of foreground and background

- $E_g(\phi)$ is the geometric cue term which incorporates a higher level semantic cue from single-touch interaction into the object segmentation framework which imposes a weak constraint that the user-input point is

Table 1: Parameters settings in the energy function

ω_e	$\overline{\omega_b}$	ω_u	ω_g	ω_a	σ	ϵ	K
6	1	10	5	-4	-24	-1.5	400

in the close vicinity of the geometry center of object; this allows incorporation of regions close to the touched location that significantly differ in appearance from the touched location which justified the importance of this term

- $\mathcal{R}(\phi)$ indicates the distance regularization term to ensure the stable evolution of the level set function by penalizing the deviation of the level set function from a signed distance function.

These terms can be categorized as: edge based energy, statistical prior energy, geometry energy and distance regularization. $[\omega_e\ \omega_a\ \omega_b\ \omega_u\ \omega_g]$ is a set of coefficients for each energy term; $[\omega_e\ \omega_a\ \omega_u\ \omega_g]$ are specified in Table 1, and ω_b is an adaptive parameter defined in Subsec. 4.5. Fig. 1 presents an overview of the proposed system, where the dashed lines indicate these four energy categories. Each individual energy term is detailed in the following subsections, and we also illustrate the importance of each by disabling various terms to qualitatively demonstrate their contribution to segmentation performance.

4.1. Edge Based Energy

Classical snakes and active contour models [13] typically use an edge detector to halt the evolution of the curve on the boundary of the desired object. The gradient based edge detector inherently captures high frequency information but not necessarily the real boundary of the desired object. Moreover, it is also sensitive to noise. The edge-based active contour model is thus not applicable to most natural images especially texture rich or noisy data.

In order to describe the edge probability of color-texture homogeneous region in natural image, we propose an approach inspired by JSEG [25], which calculates an edge indicator by observing the local distribution of color class labels without estimating a specific model for a texture region. In our proposed method, the color class labels are generated by extracting the dominant color modes and assigning each pixel with the label of according color mode.

4.1.1. Extracting Dominant Color

The proposed algorithm is performed on the histogram in a CIE Lab color space due it being a perceptually uniform space, i.e. over which the differene of perceived color is approximated by Euclidean distance. The key procedure of this algorithm is shown in Fig. 2.

The first step finds all the local maximums in the histogram and assigns a unique label to each of them. A histogram bin x (a vector), is a local maximum if the following condition, where $H(\cdot)$ denotes the histogram value, and $N(\cdot)$ denotes the neighborhood, is satisfied,

$$H(x) \geq H(y), \forall y \in N(x).$$

In case that neighboring peak bins have different labels, their labels are unified.

In the second step, the input colors are clustered via iteratively spreading the labels of all the peaks and regarding the bins with one same label as one cluster. The label spreading process is iteratively performed until every bin x with $H(x) > 0$ is labeled. In each iteration, bin x, which has not been labeled, may inherit the label of bin y, if $H(x) \leq H(y), y \in N(x)$. As the labels of different peaks are spread simultaneously, label of the peak that is closer to the bin is likely to arrive first. This scheme seeks the shortest ascending route to a local maximum. If multiple labels arrive at x in the same iteration, then x is labeled as the same as the neighboring bin with the larger histogram value.

A bin x is defined as a joint $J_t(\cdot, \cdot)$ of two adjacent clusters Δ and Σ if the following is satisfied

$$x \in \Delta, \exists y \mid (y \in \Sigma) \wedge [y \in N(x)] \wedge [H(y) \geq H(x)],$$

and thus $x \in J_t(\Delta, \Sigma)$. The connection value $v_c(\cdot, \cdot)$ of these two adjacent clusters can be defined as

$$V_c(\Delta, \Sigma) = \max[H(x) \mid x \in J_t(\Delta, \Sigma)]$$

After the second step, all the colors are clustered. However, color histograms, especially the high-resolution ones, are not smooth which normally leads to many local peaks. To make the algorithm robust to roughness of the histogram, some of the adjacent clusters should be properly merged. Considering the peaks as islets in a lake, some of small islets will be connected to form larger islets as the water level in the lake decreases. To this end, all the histogram values are first sorted in descending order. Then we scan

the sorted values one by one to simulate the water level decreasing. We only consider merging the connected clusters during the scanning. The dominant mean color $C_m(\cdot)$ of each cluster Δ is updated as the water level h decreases as

$$C_m(\Delta, h) = \frac{\sum x \cdot H(x)}{\sum H(x)} \mid (x \in \Delta) \wedge [H(x) \geq h]$$

where x indicates both the bin and the color vector. Only the bins which are above the water level h contribute to the dominant mean color of their cluster. When the water level h reaches the joint of two adjacent clusters, they can be merged if the following two conditions are satisfied

$$||C_m(\Delta, h) - C_m(\Sigma, h)|| \leq T_d,$$

$$V_c(\Delta, \Sigma) \geq T_p \cdot \min\{\max[H(x) \mid x \in \Delta], \max[H(y) \mid y \in \Sigma]\},$$

where the threshold T_d indicates a color difference that is distinctly visible to human eyes (suggested as 0.07) and $T_p \in [0.5, 0.75]$. The conditions constrain that two clusters can be merged only if their dominant mean colors are similar enough and their connection value is not too small compared to their peak values.

Compared with the agglomerative algorithm used in [23, 25], which only considers color similarity, the proposed cluster merging scheme also considers color distribution. When two clusters are merged, their connection relationships with other clusters will be inherited by the new cluster, so that the new cluster may be further merged with the adjacent clusters. As all the connection values are scanned, all the adjacent clusters will be considered for merging. Thus this step is finished when the iteration is over.

The average processing time on VGA (640×480) image is less than 60 ms (Intel Core2 CPU 2.1 GHz, single thread) which makes it a very efficient and robust algorithm for our application.

4.1.2. Edge Indicator and Energy

Suppose Z is the set of all N pixels in a dominant color mode map. Let $z = (x, y), z \in Z$. Z is classified into C DC modes. The means of Z and class Z_i ($i \in C$) in Z are

$$\begin{aligned} m &= \frac{1}{N} \sum_{z \in Z} z. \\ m_i &= \frac{1}{N_i} \sum_{z \in Z_i} z. \end{aligned}$$

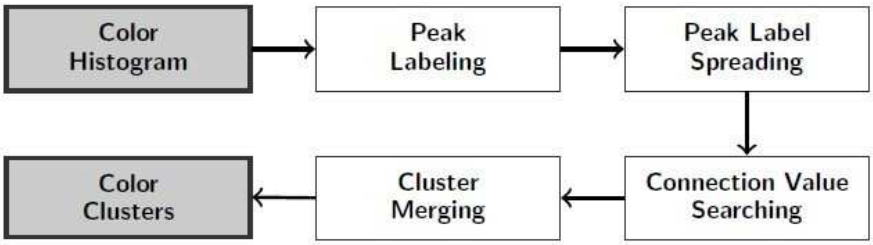


Figure 2: Key procedures of the proposed dominant color extraction algorithm

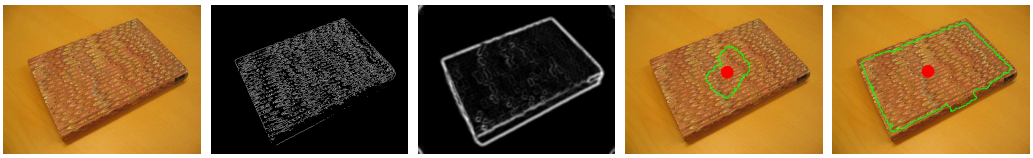


Figure 3: Comparisons of edge map by Sobel operator and the proposed edge probability: (1) Source image (2) Edge map by Sobel operator (3) Edge probability map by our approach (4) Segmentation with only the edge energy based on Sobel operator (5) Segmentation with only the edge energy based on proposed edge probability.

respectively. Let

$$\begin{aligned} S_T &= \sum_{z \in Z} ||z - m||^2 \\ S_W &= \sum_{i=1}^{C} S_i = \sum_{i=1}^{C} \sum_{z \in Z_i} ||z - m_i||^2 \end{aligned}$$

be the variance of pixels in Z and the total variance of pixels belonging to the same DC mode. The edge indicator is defined as

$$J = (S_T - S_W)/S_W.$$

The value of J is large near the boundaries of color-texture homogeneous region and small in region interiors, and thus can serve as edge "probability" while suppressing high frequency information and noise as opposed to traditional edge detectors.

E_e incorporates the edge indicator J and is defined similarly as the geodesic model [13], which is minimized when the zero level contour of ϕ is located at the object boundaries

$$E_e = \int_{\Omega} g\delta(\phi)|\nabla\phi| d\mathbf{x} \tag{2}$$

where $g = \frac{1}{1+cJ}$, c is a constant, H is the Heaviside function and δ is the Dirac delta function.

We define the ballooning term as

$$E_a = \int_\Omega gH(\phi)d\mathbf{x} \tag{3}$$

which computes a weighted area of the region $\Omega_\phi^+ \triangleq \{\mathbf{x} : \phi(\mathbf{x}) > 0\}$. This energy is introduced to speed up the motion of the zero level contour in the evolution process when the initial contour is not placed in the vicinity of the desired object boundary. The ballooning of the zero level contour is inhabited near the boundaries where J takes larger values.

A comparison of the Sobel edge map and the proposed edge probability is shown in Fig. 3. The awareness of local color distribution avoids the converge of zero level contour stuck in a local minimum frequently caused by the traditional local edge detectors and facilitates the segmentation of natural image.

4.2. Statistical Prior Energy

An optimal partition $\mathcal{P}(\Omega)$ of the image plane Ω can be computed by maximizing the *a posterior* probability $p(\mathcal{P}(\Omega)|I)$ for the given image I [52]. Applying Bayes' rule, it can be expressed as

$$p(\mathcal{P}(\Omega)|I) \propto p(I|\mathcal{P}(\Omega))p(\mathcal{P}(\Omega)).$$

$p(\mathcal{P}(\Omega))$ allows to introduce prior knowledge such as geometric priors to cope with missing low-level information. Under the given prior, optimal two-region partition is achieved by maximizing

$$p(I|\mathcal{P}(\Omega)) = p(I|\Omega^+)p(I|\Omega^-). \tag{4}$$

where Ω^+ and Ω^- represent the regions inside and outside the contour respectively. Maximization of (4) is equivalent to minimizing its negative logarithm, we define $E_b(\phi)$ as

$$E_b(\phi) = \log p(I|\Omega^+) + \log p(I|\Omega^-). \tag{5}$$

We assume that the image I in each region is characterized by the individual pixel values at different locations $\mathbf{x}$ and the pixel values are i.i.d. Let $\phi(\mathbf{x}) > 0$ if $\mathbf{x} \in \Omega^+$ and $\phi(\mathbf{x}) < 0$ if $\mathbf{x} \in \Omega^-$. We reduce (5) to

$$\begin{aligned} E_b(\phi) = -\int_\Omega (H(\phi) \log p(I(\mathbf{x})|\theta^+) + \\ (1 - H(\phi)) \log p(I(\mathbf{x})|\theta^-))d\mathbf{x}. \end{aligned} \tag{6}$$

where θ^+ and θ^- represent the foreground and background color model respectively.

The foreground and background color model are represented by Gaussian Mixture Model (GMM) as

$$p(I(\mathbf{x})|\theta_i) = \sum_{k=1}^{K_i} w_{ik}\mathcal{N}(I(\mathbf{x}); \omega_{ik}, \Sigma_{ik}),$$

with parameters w_{ik}, ω_{ik} and Σ_{ik} representing the weight, the mean and the covariance of the k^{th} component. The parameters of all GMMs ($\theta_i = \{w_{ik}, \omega_{ik}, \Sigma_{ik}, i = 1, \ldots, L, k = 1, \ldots, K_i\}$) are learned from observations of pixels; specifically the pixels in the user-specified area are assumed to be foreground and the border of the image is assumed to be the background. The second row in Fig. 1 visualizes the process of estimating the foreground and background color model.

The user-touched area is usually a part of the desired object, and thus the foreground color model has higher confidence than the background color model, especially when the desired object intersects the border of the image. We propose a foreground consistency term to enforce the minimization of foreground statistical error as

$$E_u(\phi) = \frac{\int_\Omega H(\phi)(1 - p(I(\mathbf{x})|\theta^+))d\mathbf{x}}{\int_\Omega H(\phi)d\mathbf{x}}. \quad (7)$$

This energy term computes the averaged classification error $1-p(I(\mathbf{x})|\theta^+)$ inside the zero level contour regardless the accuracy of background color model. Fig. 5 gives examples where the background color model is confused with the foreground whilst foreground consistence term imposes the contour evolution proceeds as long as the foreground statistical error inside the contour is minimized.

Fig. 4 presents segmentation results by disabling statistical prior energy and edge based energy respectively. Based on edge information alone, the system fails to converge to the correct object boundary without considering the global color distribution. Disabling edge based energy leads to unsmooth segmentation caused by the statistical prior error. Combining these two achieves robust segmentation in the presence of inaccurate edge information or color modeling.

4.3. Geometry Energy

People tend to select the geometrical center when they are indicating the object of interest. Although not a precise measurement, such a geometrical

Figure 4: Contribution of edge based energy and statistical prior energy: (1) Source image (2) Segmentation by full system (3) Segmentation by disabling statistical prior energy (4) Segmentation by disabling edge based energy.

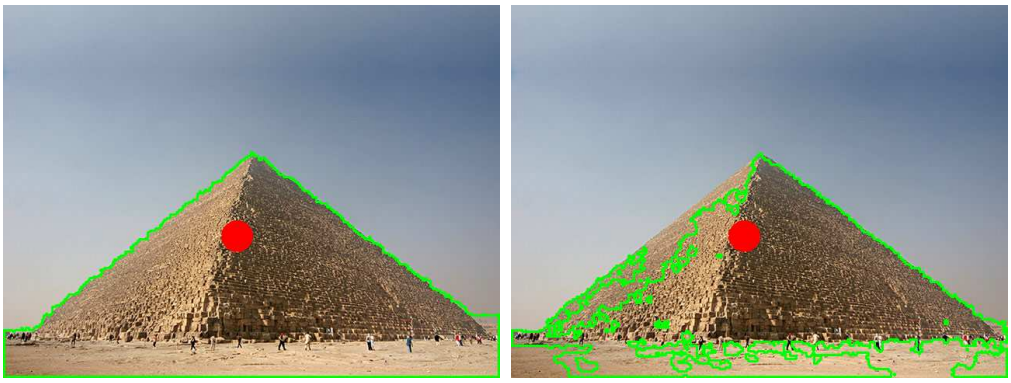


Figure 5: Contribution of foreground consistency term: (1) Segmentation by full system (2) Segmentation by disabling foreground consistency term.

constraint provides a weak cue for the contour evolution process. We propose a central symmetry term to reflect this geometrical constraint, by computing the spatial deviation of the geometrical center of zero level contour from the user-input point as

$$E_g(\phi) = |\frac{\int_\Omega H(\phi)(\mathbf{x} - \overline{\mathbf{x}})d\mathbf{x}}{\int_\Omega H(\phi)d\mathbf{x}}| \tag{8}$$

where $\overline{\mathbf{x}}$ represents the user-input point. As the desired object could have very complex shape, this term is regarded as a relatively weak indication of the desired object's geometry. The utility of the geometric term is to allow incorporation of regions close to the touched location that significantly differ

in appearance (e.g. color) from the touched location. This is well illustrated in Fig. 8; in the first example, since the touched area (white) is significantly different in color but still spatially local to the region of the blue trousers; the second example further demonstrates a situation where this higher level knowledge is explored to guide the contour evolution to segment the whole object of interest which can not be achieved by low level color and edge information.

4.4. Distance Regularization

The proposed model incorporates the distance regularization term present in [42] to ensure stable evolution of the level set function, by penalizing the deviation of the level set function from a signed distance function. This deviation is characterized by the following integral

$$\mathcal{R}(\phi) = \mu \int_{\Omega} \mathcal{P}(|\nabla\phi|) d\mathbf{x} \tag{9}$$

where μ relates to the numerical stability condition detailed in Subsec. 4.7 and $\mathcal{P}(s)$ is a double-well potential function defined as

$$\mathcal{P}(s) = \begin{cases} \frac{1}{(2\pi)^2}(1-\cos(2\pi s)), & \text{if} \quad s \leq 1 \\ \frac{1}{2}(s-1)^2, & \text{if} \quad s \geq 1. \end{cases}$$

This potential function maintains the signed distance property $|\nabla\phi| = 1$ only in a vicinity of the zero level contour, while keeps the embedding function ϕ as a constant with $|\nabla\phi| = 0$ at locations far away from the zero level contour.

4.5. Adaptive Weighting

Minimizing the proposed energy functional (1) with constant coefficients usually gives good segmentations. However, when the foreground and background distribution is not distinct, the Bayes error term would be non-discriminative and the contour evolution process would not converge to the desired object boundaries. In this case, the weight of Bayes' error term should be relatively small to increase the influence of other reliable terms. We expect it to be adaptively tuned based on the color modeling error on a per image basis. To this end, we estimate the misclassifying error in foreground/background seeds based on the posterior probability

$$\eta = \frac{1}{|\Omega^+|} \sum_{\mathbf{x}\in\Omega^+} p(I(\mathbf{x})|\theta^-) + \frac{1}{|\Omega^-|} \sum_{\mathbf{x}\in\Omega^-} p(I(\mathbf{x})|\theta^+)$$

and define coefficient $\omega_b = \max\{\overline{\omega_b}(1-\eta), 0\}$. When the misclassifying error η is close to zero, the weight approaches $\overline{\omega_b}$. When the color models are indistinct, ω_b approaches 0.

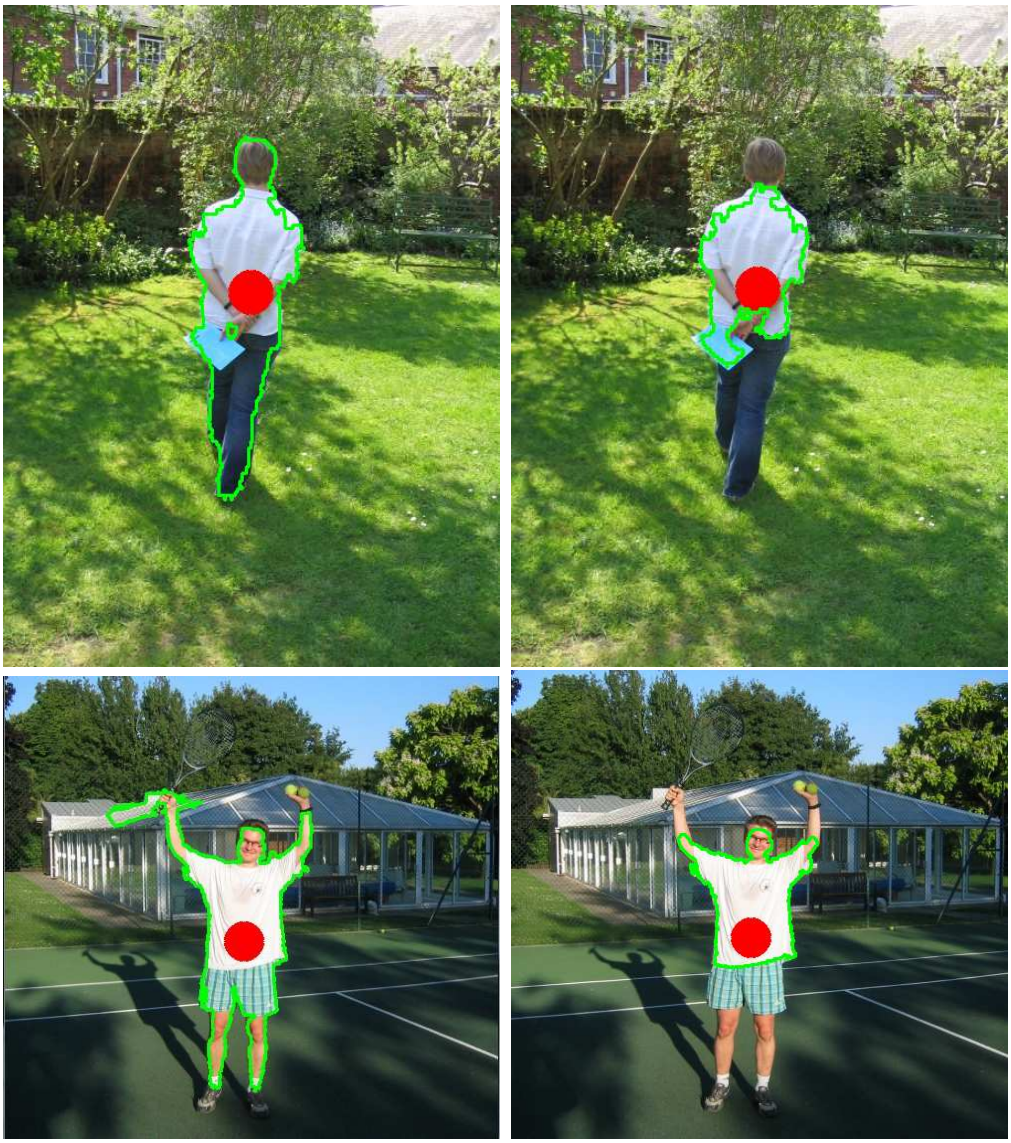


Figure 6: Contribution of geometry energy: (left) - Segmentation by full system (right) Segmentation by disabling geometry energy.

4.6. Impact of Varying Weights

We empirically choose the weights listed in Table 1 by evaluating the segmentation quality resulted by a large range of weights. However note the system is not highly sensitive to weight choices and a single pre-configured

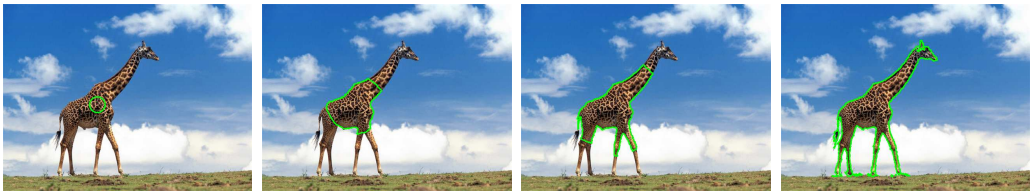


Figure 7: Contour evolution process: The initial contour and the zero level contours (green curve) after 100 ($\sim 65\ ms$), 200 ($\sim 132\ ms$), 400 ($\sim 255\ ms$) iterations respectively

system using the same weightings is applicable to a broad class of imagery tested. We owe this to the balance of energies in our proposed object segmentation model, as well as the adaptive weighting scheme we adopted described in Subsec. 4.5. Fig. 8 demonstrates the effect on the segmentation result by significantly raising or lowering the weights ω_e, ω_a, $\overline{\omega_b}$ and ω_u, compared to the result (row 1 right) using our chosen weights listed in Table 1. Lower ω_e (row 2 left) causes mis-segmentation dominated by statistic prior energy when the color distributions of foreground and background are not distinct, whilst higher ω_e (row 2 right) stops the contour evolution at strong local edges. Lower ω_a (row 3 left) causes prematured convergence whilst higher ω_a (row 3 right) forces the contour to exceed the real boundary. Adjusting $\overline{\omega_b}$ (row 4) around our chosen weight results in similar observations as adjusting ω_e inversely. Higher ω_u (row 5 right) strongly enforces the foreground consistency which results in under-segmentation compared to lower ω_u (row 5 left). The weight of geometry energy ω_g is relatively insensitive, though its presence incorporates the strong indication of higher-level semantic (Fig. 6).

4.7. Numerical Approximation and Implementation

We use the standard gradient descent method to minimize the energy functional (1)

$$\begin{aligned}\frac{\partial \phi}{\partial t} = & -\frac{\partial E_e(\phi)}{\partial \phi} - \frac{\partial E_b(\phi)}{\partial \phi} - \frac{\partial E_u(\phi)}{\partial \phi} \\ & -\frac{\partial E_g(\phi)}{\partial \phi} - \frac{\partial E_a(\phi)}{\partial \phi} - \frac{\partial \mathcal{R}(\phi)}{\partial \phi}\end{aligned} \tag{10}$$

where the gradient flows are deducted as follows:

$$\begin{aligned}
\frac{\partial E_e(\phi)}{\partial \phi} &= \omega_e \delta(\phi) \mathrm{div}(g \frac{\nabla \phi}{|\nabla \phi|}) \\
\frac{\partial E_b(\phi)}{\partial \phi} &= \omega_b \delta(\phi) \log \frac{p(I(\mathbf{x})|\theta^+)}{p(I(\mathbf{x})|\theta^-)} \\
\frac{\partial E_u(\phi)}{\partial \phi} &= \omega_u \delta(\phi) [\frac{(1 - p(I(\mathbf{x})|\theta^+))}{(\int_\Omega H(\phi) d\mathbf{x})^2} \\
&\quad - \frac{\int_\Omega (1 - p(I(\mathbf{x})|\theta^+)) H(\phi) d\mathbf{x}}{(\int_\Omega H(\phi) d\mathbf{x})^2}] \\
\frac{\partial E_g(\phi)}{\partial \phi} &= \omega_g \delta(\phi) \frac{|(\mathbf{x} - \overline{\mathbf{x}}) - \int_\Omega (\mathbf{x} - \overline{\mathbf{x}}) H(\phi) d\mathbf{x}|}{(\int_\Omega H(\phi) d\mathbf{x})^2} \\
\frac{\partial E_a(\phi)}{\partial \phi} &= \omega_a g \delta(\phi) \\
\frac{\partial \mathcal{R}(\phi)}{\partial \phi} &= \mu \mathrm{div}(\frac{\mathcal{P}'(|\nabla \phi|)}{|\nabla \phi|} \nabla \phi)
\end{aligned}$$

To discretize the equations, we use a finite differences scheme. Considering the 2D case with a time dependent embedding function $\phi(x, y, t)$, the spatial derivatives $\partial\phi/\partial x$ and $\partial\phi/\partial y$ are approximated by the central difference, where the space steps are fixed as $\Delta x = \Delta y = 1$. The temporal partial derivative $\partial\phi/\partial t$ is approximated by the forward difference.

We discretize embedding function $\phi(x, y, t)$ as $\phi_{i,j}^k$, where (i, j) is the spatial index and k is the temporal index. The level set evolution equation (1) is discretized as $(\phi_{i,j}^{k+1} - \phi_{i,j}^k)/\Delta t = F(\phi_{i,j}^k)$ where $F(\phi_{i,j}^k)$ approximates the right hand side in (1). The level set evolution is then expressed as an iteration process

$$\phi_{i,j}^{k+1} = \phi_{i,j}^k + \Delta t F(\phi_{i,j}^k), k = 0, 1, 2, \ldots \tag{11}$$

In the implementation, the Heaviside function H is approximated by a smooth function defined by

$$H_\epsilon(x) = \begin{cases} \frac{1}{2}(1 + \frac{x}{\epsilon} + \frac{1}{\pi} \sin(\frac{\pi x}{\epsilon}), & |x| \leq \epsilon \\ 1, & x > \epsilon \\ 0, & x < -\epsilon. \end{cases}$$

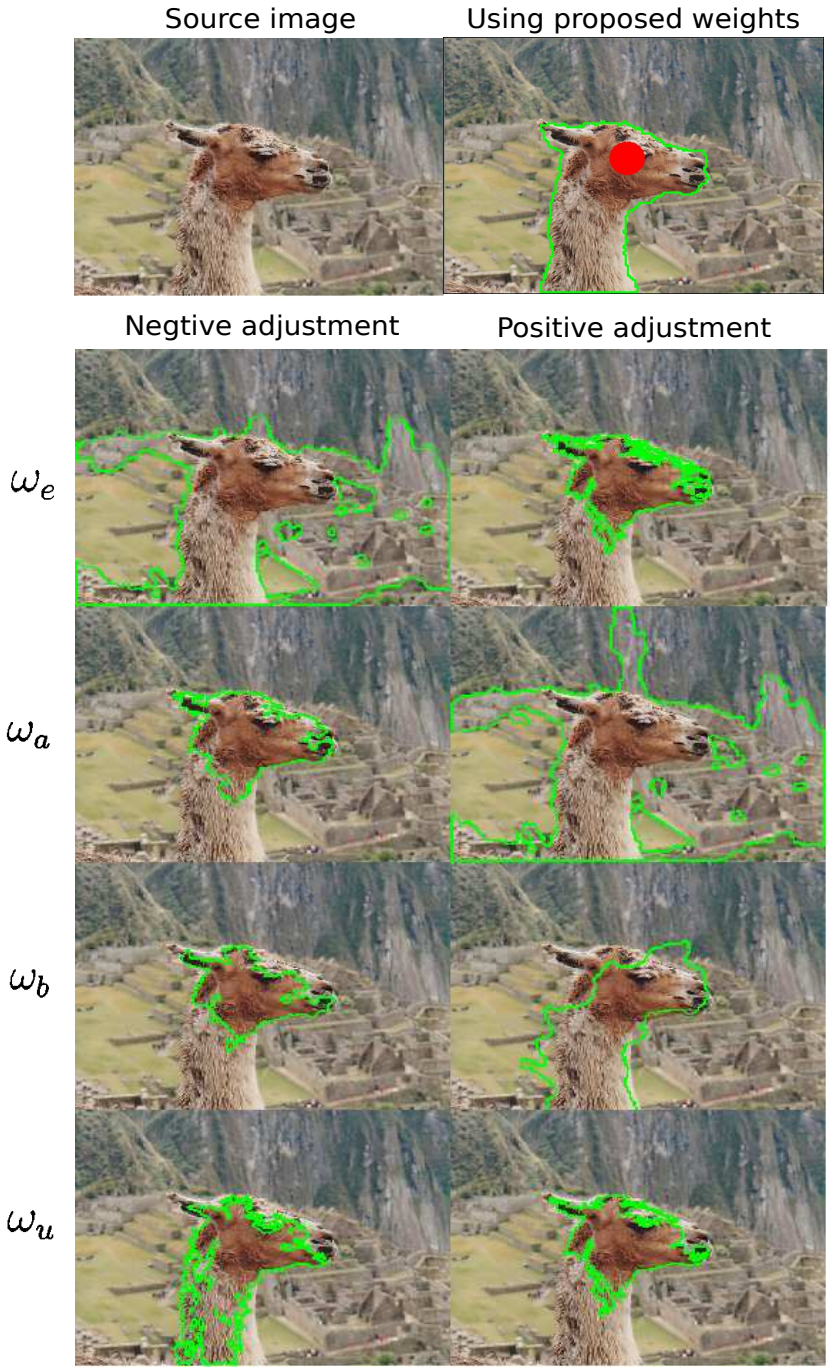


Figure 8: Segmentation results by significantly lowering (row 2-5 left column) or raising (row 2-5 right column) the weights ω_e, ω_a, $\overline{\omega_b}$ and ω_u respectively. The source image and the segmentation by our pre-configured system is shown in row 1.

and the Dirac delta function δ is approximated by

$$\delta_\epsilon(x) = \begin{cases} \frac{1}{2\epsilon}(1+\cos(\frac{\pi x}{\epsilon})), & |x| \leq \epsilon \\ 0, & |x| > \epsilon. \end{cases}$$

As the Dirac delta function and the Heaviside function multiply the entire image plane in (10), only the ϕ values in the vicinity of the zero crossing points need to be updated. This is the central idea of the narrow band methods [1]. The computational cost of a level set method can be substantially reduced by confining the computation to a narrow band around the zero level set contour. For our proposed formulation, as re-initialization is not needed due to the incorporation of distance regularization term E_d [42], the narrow band implementation is simple and straightforward. Our narrow band implementation allows the use of a large time step in the finite difference scheme to greatly reduce the iterations as long as the choice of the time step $\triangle t$ satisfies the Courant-Friedrichs-Lewy (CFL) condition $\mu\triangle t < (1/4)$ for numerical stability.

We adopt the narrow band method [1] to substantially reduce the computational cost of level set method by confining the computation to a narrow band around the zero level set contour. The narrow band scheme is implemented in the following major steps:

1. Compute the narrow band $B_k = \bigcup_{(i,j)\in C_0} N_{i,j}$, where C_k is the set of zero crossing points of ϕ_k and $N_{i,j}$ is a 3×3 neighborhood system centered at each point (i, j). If either $\phi_{i-1,j}\phi_{i+1,j} \leq 0$ or $\phi_{i,j-1}\phi_{i,j+1} \leq 0$, point (i, j) is regarded as a zero crossing point. $\forall(i, j) \in B_k$ and $(i, j) \notin B_{k-1}$, set $\phi^k_{i,j} = 2$ if $\phi^{k-1}_{i,j} > 0$, or else set $\phi^k_{i,j} = -2$ if $\phi^{k-1}_{i,j} < 0$.
2. Update the embedding function $\phi^{k+1}_{i,j} = \phi^k_{i,j} + \Delta t F(\phi^k_{i,j})$ on the narrow band B_k.
3. If k exceeds a predefined maximum number of iterations K, the evolution process is halted. Otherwise, go to (1).

In our prototype, we use a mouse click and a fixed brush size σ to simulate the finger touch of the user. The embedding function ϕ is initialized by extracting the contour of the user input brush stroke. The embedding function is assigned as 2 inside the contour and -2 outside the contour. We empirically choose the parameters in the formulation which are listed in Table 1. Fig. 7 illustrates the evolution process of the zero level contour which clearly indicates the speed of convergence.

5. Extension to Video Object Segmentation

The proposed TouchCut framework enables fast object segmentation with accurate boundary placement and strong region connectivity on still images. In this section, we extend TouchCut framework to video object segmentation. As one of the potential applications enabled by the proposed system, we stylize video objects or background into paintings based on the temporally coherent object/background region map.

After acquiring the object segmentation on the initial frame, TouchCut is performed on successive video frames using both photometric properties of the current frame and prior information propagated forward from previous frames. This information consists of:

i. an incrementally built GMM encoding the color distribution of foreground/background over past frames;
ii. an initial contour for level set evolution;
iii. an estimated foreground object mask.

The image data labeled by the binary segmentation mask on previously segmented frames underpins accurate color distribution of foreground and background region respectively when the luminance variation on successive frames is minor. In practice, to cope with variations in luminance often present in the sequence and cumulative segmentation error near boundary, the proportion of samples $S_{l,t-t_d} \in [0,1]$ $(t_d > 0), l \in l_f, l_b$ drawn from all foreground (l_f) and background (l_b) pixels from historical frame I_{t-t_d} decreases exponentially as the temporal distance t_d from the current frame I_t increases where $\sigma^2_{t_d}$ is determined by the level of luminance variance. Smaller $\sigma^2_{t_d}$ is selected when luminance variance is large, contributing more recent data to the GMM, otherwise the historical data contributes more to increase robustness.

$$S_{l,t-t_d} \propto e^{-t_d^2/\sigma^2_{t_d}}, \tag{12}$$

We employ optical flow to create a per-pixel propagation of the foreground mask from frame I_{t-1} to create an estimated mask on frame I_t which is used as the shape prior $\tilde{\phi}$ which takes the value of the initial embedding function. We propose a shape energy term measuring the shape dissimilarity of two shapes represented by the embedding functions ϕ and $\tilde{\phi}$, which is commonly computing the area of the set symmetric difference

$$E_s(\phi) = \omega_s \int_\Omega (H(\phi) - H(\tilde{\phi}))^2 d\mathbf{x} \tag{13}$$

ω_s is inversely proportional to the alignment error in the scope of the foreground object Ω_f

$$\omega_s \propto 1/\sqrt{\frac{1}{|\Omega_f|}\sum_{\mathbf{x}\in\Omega_f} ||I_{t-1}(\mathbf{x}) - I_t'(\mathbf{x})||^2}. \tag{14}$$

where I_t' is the warped color image from frame I_{t-1} to I_t by the optical flow. Accurate alignment generally indicates reliable motion estimation and such shape priors thus contribute more to the contour evolution.

Applying the standard gradient descent method to minimize the shape energy term, we deduct the gradient flow of shape energy as

$$\frac{\partial E_s(\phi)}{\partial \phi} = 2\omega_s \delta(\phi)(H(\phi) - H(\tilde{\phi})). \tag{15}$$

The initial contour of TouchCut on current frame I_t is acquired by computing the deviation of the initial contour ϕ_0^0 on the first frame from the geometrical center g_0^0. Let the center of ϕ_0^0 be b_0, and the center b_t of the initial contour on I_t is estimated as $b_t = g_0^t + b_0 - g_0^0$. The geometrical center on frame I_t, g_0^t is estimated from the estimated foreground mask. We use a circle centered at b_t with a radius r_c as the initial contour on frame I_t. r_c is two times as large as the maximum distance from the contour to the touch point on the initial frame.

Due to the inherent error of optical flow, the new initial contour might be slightly drifting from the desired object, i.e. part of the area inside the initial contour might be the background. The robust formulation based on color and shape priors push the contour evolution, minimizing the pixel classification error inside and outside the contour, while satisfying other criteria defined in the energy function, to achieve accurate and temporally coherent segmentation.

6. Experiments and Comparisons

We have applied the proposed algorithm on a dataset consisting of the combined Berkeley $BSDS300$ dataset [46], and image dataset accompanying GrabCut [57]. We also demonstrate the application of TouchCut to video sequences exhibiting clutter and agile motion. We assess segmentation performance on both qualitatively through visual comparison to prior work, and quantitatively based on a manual ground truth segmentation. We indicate relative performance to the state of the art for both image and video comparisons.

	TouchCut	Graph Cut	GrabCut	Star shape	Geodesic star
Average	**0.8648**	0.6507	0.7977	0.7977	0.7016
Median	**0.8989**	0.6755	0.8566	0.7423	0.7243
Std	**0.1330**	0.2365	0.2054	0.2173	0.2102

Table 2: Statistics of segmentation overlap scores from our objective comparison on the dataset

6.1. Segmentation of Still Images

Fig. 10 presents a qualitative comparison of the proposed method with standard graph cut (column 2) [8], GrabCut (column 3) [57], star-convexity prior graph cut (column 4) [64], and its variant using geodesic distance (column 5) [31]. In all cases, other than Graph cut, we adapted such the modeling of color distributions to exactly that of the proposed approach to make a fair comparison — i.e. to solely evaluate performance of the single touch segmentation paradigm. Specifically, the foreground color was modeled from the pixels in user-touch area while the background color was modeled by taking pixels from the border of the image. With significantly less user input, our method gives satisfactory segmentation even when the foreground and background colors lack distinction (first row) or regions exhibit complex topology (second row). Graph cut fails to separate the objects exhibiting a similar color to the desired object, whilst our approach fills the desired region by expanding from the interior of the selected object outwards and explicitly considers the object boundary and geometric properties. GrabCut presents better spatial constraints than graph cut, benefiting from the bounding box while failed to exclude noisy extraneous regions (e.g. the varying levels of luminance underneath the tiger) which do not appear outside the bounding box. The latter method also suffers from "short-cutting" regions (e.g. the elephant's legs and trunk). Star-convexity prior graph cut appears to alleviate the bias of a graph cut towards shorter segmentation boundaries and also spatially confine the segmentation to the proximity of foreground (row 4 and 5). However, the strong assumption of star convexity does not apply on complex object which restricts its applications to natural image (e.g. the elephant's legs and trunk, and the head of the jockey), although geodesic distance is also adopted in its variant.

For our objective comparison, we adopt the measure used to evaluate segmentation quality in the VOC segmentation challenge [28] to quantify segmentation accuracy against a manual specified ground-truth, which is given by $\frac{y \cap y_{gt}}{y \cup y_{gt}}$ (with y denoting the output segmentation and y_{gt} denoting ground

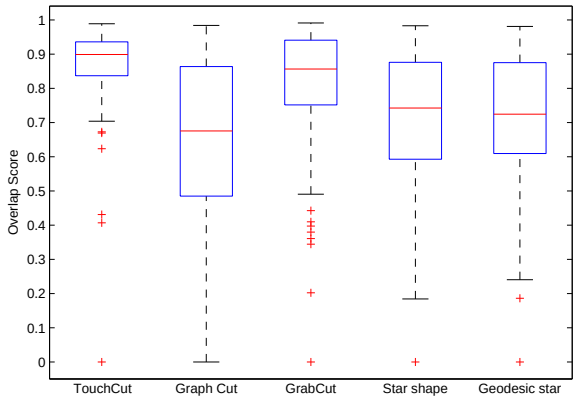


Figure 9: Box-plot of segmentation overlap scores of the proposed TouchCut, standard graph cut, GrabCut, star-convexity prior graph cut, and geodesic star convexity graph cut respectively.

truth). Table 6.1 and Fig. **??** summarise the results of this objective comparison, confirming that the proposed method outperforms state-of-the-art methods adapted to accept a single touch, with the highest average overlap score 0.8648 across the dataset. All graph cut variants enhanced by spatial or shape constraint outperform the standard graph cut which reaches a low average overlap score 0.6607. GrabCut outperforms (average overlap score 0.7977) the star-convexity prior graph cut methods (average overlap score 0.7097 and 0.7016 respectively) owing to its robust foreground and background prior. The statistics of the segmentation overlap score in Table 6.1 further confirms the robustness of the proposed method across various natural images in the dataset, with a much tighter standard deviation (0.133). The box-plot of Fig. **??** in demonstrates higher reliability than methods, such as regular Graph Cut, which have both a broader standard deivation and in sporadic poor performance (indicated by box limits). All of the above analysis matches our intuition and theory that our proposed model gives more robust representation of the object segmentation problem given very limited input.

Figure 10: Comparison of proposed method (left) with graph cut (column 2) [8], GrabCut (column 3) [57], star-convexity graph cut (column 4) [64], and its variant using geodesic distance (column 5) [31]. The contour of segmented object is shown in green.

Figure 11: Representative segmentation results from our dataset, discussed within Subsec.6.1

Fig. 11 presents further segmentation results [1]. The first row shows the results on highly-textured images. The edge probability map enables the contour evolution over color-texture homogeneous regions without being stopped at local minimum. The second row shows the segmentation results of images with indistinct foreground and background colors. In this case, the color modeling error is large which adaptively results in a small weight on color based term E_b. On the other hand, the foreground consistency term E_u enforces the region inside the zero level contour to be coherent in the sense of color distribution regardless the background color distribution. Such a constraint significantly imposes the stability of the contour evolution process in the case of indistinct color distributions. The third row contains segmentation results to deal with objects with complex shape. By leverag-

[1]More results can be viewed online at: `http://personal.ee.surrey.ac.uk/Personal/Tinghuai.Wang/CVIU2012`

ing the strength of the implicit contour representation in level set methods, our system is robust in coping with complex topologies without exhibiting short-cutting problem which is common in graph-cut based systems. The system is able to cope with weak boundaries and complex foregrounds and backgrounds, to extract meaningful objects in most cases.

For all image results the running time on a Core2 2.1 GHz PC is constant ~ 0.4 second per VGA image (640×480). More representative segmentation results are presented in Fig. 12.

6.2. Application to Stroke-Based Painterly Rendering

Object segmentation provides an object level parsing for a scene, where special effects can be applied on either the object or background to create novel compositions in photography or media production. Our proposed single-touch object segmentation algorithm can deliver intuitive, efficient and precise object segmentation. We demonstrate an application of TouchCut to create oil painting effects by compositing virtual brush strokes, which is also known as stroke-based painterly rendering (SBR). We apply an automatic SBR algorithm by Shugrina *et al.* [60] to create novel oil painting and photo composition as shown in Fig. 13, where the background scene is stylized as oil painting effects.

6.3. Segmentation of Video Sequences

We quantitatively test TouchCut on three videos and ground-truth for the primary foreground object present in [15] and [63]. The videos tested exhibit various challenging conditions such as foreground and background color overlap, luminance variation, shape deformation and camera motion. We compare against two state-of-the-art approaches: another level set based tracking approach by Chockalingham *et al.* [15] and the 'motion coherent segmentation' method of Tsai *et al.* [63]. These methods require human labeling of the object boundary (contour) in the first frame, whilst TouchCut requires minimal user intervention to guide the segmentation of whole video via a single touch on the first frame. The segmentation accuracy is quantified as the average per-frame pixel error rate, $\epsilon(S) = \frac{|XOR(S,GT)|}{F}$, where S is the segmentation, GT is the ground-truth segmentation and F is the total number of frames. As shown in Table 6.3, our method outperforms the approaches present in [15] and [63] on two of the three videos (*PARACHUTE, BIRDFALL*), and produces the second best result on the *GIRL* video. Our higher error rate on $GIRL$ is caused by the inaccurate optical flow motion estimations and indistinct color of foreground and background, which is

Figure 12: Additional representative segmentation results from our combined *BSDS*300 and GrabCut dataset.

reasonable since the object exhibits large appearance variation, considering TouchCut does not either explicitly employ global optimizations to enforce motion coherence across frames as in [63] and or manually bootstrapping the

Figure 13: Segmentation applying TouchCut to image and background oil painting effects (Segmentation on the left and SBR effects on the right).

	TouchCut	[63]	[15]	Running Time
BIRDFALL	**0.003 (248)**	0.003 (252)	0.005 (454)	20 s
PARACHUTE	**0.002 (228)**	0.002 (235)	0.004 (502)	35 s
GIRL	0.012 (1691)	**0.009 (1304)**	0.012 (1755)	16 s

Table 3: Video segmentation error expressed as the average fraction of mis-segmented pixels (false positive plus false negative) per frame. Absolute number of mis-segmented pixels in parentheses (averaged per frame). Execution times for each sequence using TouchCut.

first frame. Further, TouchCut requires only a single touch to bootstrap the entire video segmentation — and we believe these comparative results to be very encouraging. Further qualitative segmentation results with comparisons to [63] are shown in Fig. 14.

We study the impact of the proposed shape prior underpinning our video segmentation, comparing against a baseline implementation that does not incorporate the shape prior, but otherwise follows the same pipeline as our full method. Fig. 15 shows the results. The shape energy significantly improves segmentation accuracy, quantified against a manual ground truth.

6.4. Application to Video Stylization

Temporally coherent segmentation of video forms a stable representation of visual structure in the scene which enables other computer vision and graphics applications. We demonstrate an application of TouchCut system on videos to create stylized region-based effects such as painterly rendering. We incorporate a framework of automatic non-photorealistic rendering by Kyprianidis and Döllner [38] to facilitate the domestic user to create artistic stylizations on either the desired object or the background scene.

Figure 14: Qualitative segmentation results on *BIRDFALL* (row 1), *PARACHUTE* (row 3) and *GIRL* (row 5) respectively, compared to the corresponding segmentation results (row 2, 4, 6 respectively) from [63].

As qualitative evaluation, we apply our segmentation algorithm to several video sequences exhibiting both slow moving and agile motion — summarized in Table 4. Fig. 17 presents the segmentation results applying TouchCut to these five video sequences and the foreground object or background painterly stylization effects. Our segmentation algorithm ensures the foreground and background regions deform in a coherent manner.

In Fig. 17(a) there is significant agile motion in "YUNAKIM" – Yuna swings and suffers frequent inter-occlusion over duration of the clip. Despite the adoption of a forward propagation (2D+t) strategy over several hundred

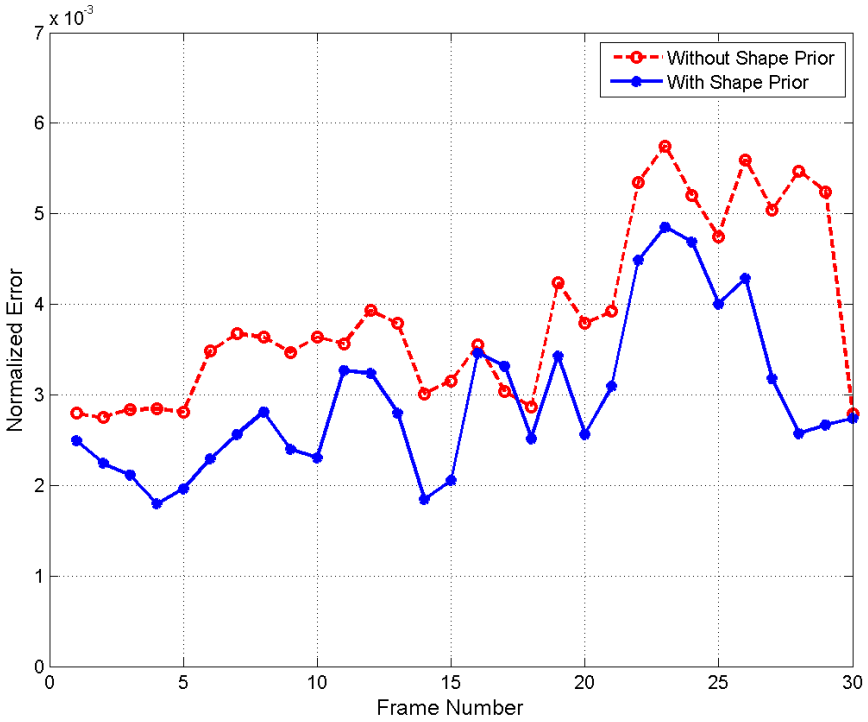


Figure 15: Illustrating the importance of the shape prior for Video TouchCut. Comparison of the full TouchCut with a baseline that does not incorporate a shape prior. Shape energy improves segmentation quality (normalized pixel error rate on *BIRDFALL*).

frames of video there is no significant degradation. With an incrementally learned GMM color model, TouchCut is able to deal with the strong luminance variation on the boy's face ("BEACH") and produce stable segmentations with temporal coherence present in Fig. 18(a). Similar situation can

Sequence	**Motion**	**# of Frames**
YUNAKIM (Fig. 17(a))	Agile	225
BOY (Fig. 17(b))	Slow	190
BEACH (Fig. 18(a))	Medium	300
LION (Fig. 18(b))	Slow	201
WALK (Fig. 18(c))	Medium	200

Table 4: Summary of video sequences used in our qualitative evaluation, annotated as to motion and number of frames present.

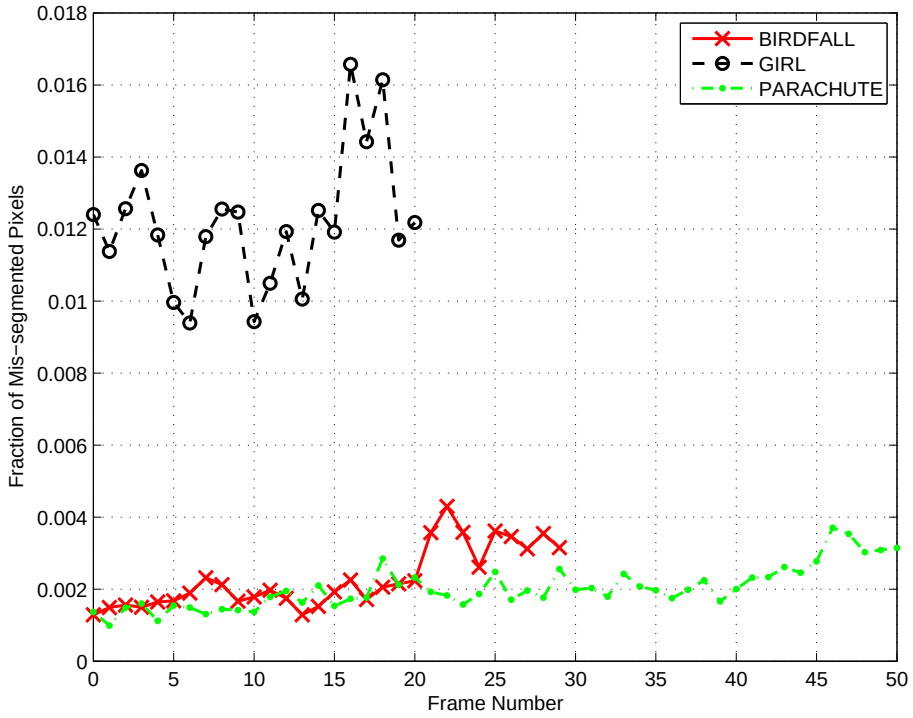
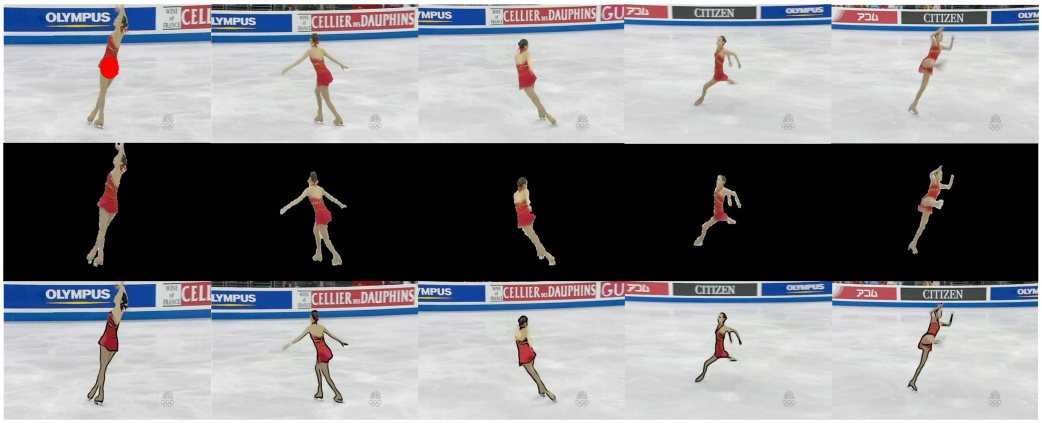
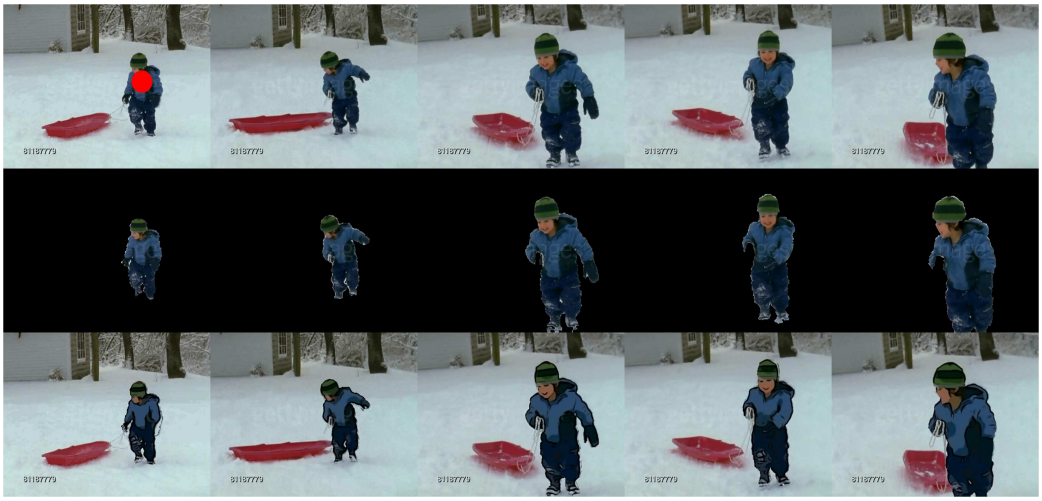


Figure 16: Fraction of mis-segmented pixels per frame for all frames of the sequences *BIRDFALL*, *PARACHUTE* and *GIRL* as used in [15] and [63].

be observed in "WALK" sequence, where both strong luminance variation and agile motion present. As an application of TouchCut, the domestic user can choose either the foreground or background to create special effects, i.e. painterly stylization, with single finger touch on the first frame as shown in Fig. 17 (foreground object stylized) and Fig. 18(c) (background stylized). Drawing upon the coherent video segmentation by TouchCut, the stylized videos create either a painting object in a realistic scene or a realistic object in an unrealistic scene with painting style. Other special effects would be creating a new movie with the selected object in a totally different scene or emphasizing the desired object by blurring the background.



(a) Representative frames from "YUNAKIM" sequence



(b) Representative frames from "BOY" sequence

Figure 17: Segmentation results applying TouchCut to video sequences and foreground object stylization effects (source in top row, foreground object cut-out in middle row, painterly rendering on foreground object in bottom row). Please refer to **http://personal.ee.surrey.ac.uk/Personal/Tinghuai.Wang/CVIU2012** for these and further results.

7. Conclusion and Future Work

We have presented a single-touch object segmentation system using level set methods. The core contribution is an edge-region-geometry based segmentation model to robustly tackle the interactive object segmentation problem — encoding boundary probabilities of color-texture homogeneous re-

gions, and the statistical and geometric priors inferred from the user input. Our edge model gives a robust description of the coherent color-texture region, which mitigates against the contour becoming stuck in local minima in the presence of noisy data. This frequently occurs in prior approaches, where traditional intensity gradient-based edge maps are used. Edge information alone only provides local information to drive contour evolution towards potential object boundary. Augmenting this model with color information from user input introduced a global term, balancing the *a posterior* probabilities of region models inside and outside the putative object contour.

By leveraging the flexibility of level set methods in energy minimization, our system achieved promising results in various natural images with complex scenes and objects. We also demonstrated that TouchCut can be extended to segment video sequences into temporally coherent foreground and background region maps. This gives rise to potential applications to video special effects (e.g. artistic stylization) with minimal user intervention, that may be suited to consumer touch-screen video cameras. Coherence was promoted through an incrementally learned color model, providing robustness against drift of the contour otherwise caused by motion estimation error. The introduction of a shape prior into the motion estimation framework was shown to deliver a further significant enhancement to coherence, especially when the foreground and background color distribution became indistinct.

Nevertheless, TouchCut can still experiences difficulties in separating the desired object from the adjacent background in the presence of highly similar colors. This remains an open question in the image segmentation community in the absence of other higher level semantic priors, e.g. shape, or other forms of global measurement. One interesting direction for future work would be to improve the background color modeling by measuring the salience of different dominant color modes. Another direction of future work with respect to the video extension might include detecting occlusion boundaries discovered from motion disparity in the scene, and using these to compensate for any ambiguity in appearance between the foreground and background.

Future applications of TouchCut fall within our original project motivation, to develop an image and video object segmentation algorithm with minimal user intervention suitable for emerging tablet and touch-screen devices. These applications could span embedded object extraction and tracking, intelligent focus, and video stylization [16, 69] on these devices.

[1] Adalsteinsson, D., Sethian, J., 1995. A fast level set method for propagating interfaces. Journal of Computational Physics, 269–277.

[2] Agarwala, A., Hertzmann, A., Salesin, D., Seitz, S., 2004. Keyframe-based tracking for rotoscoping and animation. In: Proc. SIGGRAPH. pp. 584–591.

[3] Armstrong, C., Price, B., Barrett, W., 2007. Interactive segmentation of image volumes with live surface. Computers & Graphics 31 (2), 212–229.

[4] Bai, X., Sapiro, G., 2007. A geodesic framework for fast interactive image and video segmentation and matting. In: Proc. ICCV. pp. 1–8.

[5] Bai, X., Wang, J., Sapiro, G., 2010. Dynamic color flow: A motion-adaptive color model for object segmentation in video. In: Proc. ECCV. pp. 617–630.

[6] Bai, X., Wang, J., Simons, D., Sapiro, G., 2009. Video snapcut: robust video object cutout using localized classifiers. In: Proc. SIGGRAPH. pp. 1–11.

[7] Blake, A., Rother, C., Brown, M., Perez, P., Torr, P., 2004. Interactive image segmentation using an adaptive gmmrf model. In: Proc. ECCV. pp. 428–441.

[8] Boykov, Y., Jolly, M.-P., 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: Proc. ICCV. pp. 105–112.

[9] Bresson, X., Vandergheynst, P., Thiran, J., 2003. A priori information in image segmentation: Energy functional based on shape statistical model and image information. In: Proc. ICIP. pp. 425–428.

[10] Brox, T., Cremers, D., 2007. On the statistical interpretation of the piecewise smooth mumford-shah functional. In: Proc. SSVM. pp. 203–213.

[11] Brox, T., Weickert, J., 2006. Level set segmentation with multiple regions. IEEE Trans. on Image Process. 15 (10), 3213–3218.

[12] Canny, J., 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. 8 (6), 679–698.

[13] Caselles, V., Kimmel, R., Sapiro, G., 1995. Geodesic active contours. In: Proc. ICCV. pp. 694–699.

[14] Chan, T., Vese, L., 2001. Active contours without edges. IEEE Trans. Image Process., 266–277.

[15] Chockalingam, P., Pradeep, S. N., Birchfield, S., 2009. Adaptive fragments-based tracking of non-rigid objects using level sets. In: Proc. ICCV. pp. 1530–1537.

[16] Collomosse, J., Rowntree, D., Hall, P. M., 2005. Stroke surfaces: Temporally coherent artistic animations from video. IEEE Trans. Vis. Comput. Graph. 11 (5), 540–549.

[17] Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24, 603–619.

[18] Cremers, D., 2006. Dynamical statistical shape priors for level set-based tracking. IEEE Trans. Pattern Anal. Mach. Intell. 28 (8), 1262–1273.

[19] Cremers, D., Rousson, M., Deriche, R., 2007. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. International Journal of Computer Vision 72 (2), 195–215.

[20] Cremers, D., Schmidt, F. R., Barthel, F., 2008. Shape priors in variational image segmentation: Convexity, lipschitz continuity and globally optimal solutions. In: Proc. CVPR.

[21] Cremers, D., Sochen, N. A., Schnörr, C., 2003. Towards recognition-based variational segmentation using shape priors and dynamic labeling. In: Scale-Space. pp. 388–400.

[22] Dambreville, S., Niethammer, M., Yezzi, A., Tannenbaum, A., 2007. A variational framework combining level-sets and thresholding. In: Proc. British Machine Vision Conference (BMVC). pp. 1–10.

[23] Deng, Y., Kenney, C., Moore, M., Manjunath, B. S., 1999. Peer group filtering and perceptual color image quantization. In: Proc. ISCAS. pp. 21–24.

[24] Deng, Y., Manjunath, B. S., 2001. Unsupervised segmentation of color-texture regions in images and video. IEEE Trans. Pattern Anal. Mach. Intell., 1139–1145.

[25] Deng, Y., Manjunath, B. S., 2001. Unsupervised segmentation of color-texture regions in images and video. IEEE Trans. Pattern Anal. Mach. Intell. 23 (8), 800–810.

[26] Dollar, P., 2006. Supervised learning of edges and object boundaries. In: Proc. CVPR. pp. 1964–1971.

[27] Duda, R., Hart, P., 1973. Pattern Classification and Scene Analysis. Wiley.

[28] Everingham, M., Gool, L. J. V., Williams, C. K. I., Winn, J. M., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88 (2), 303–338.

[29] Gleicher, M., 1995. Image snapping. In: Proc. SIGGRAPH. ACM, pp. 183–190.

[30] Grady, L., 2006. Random walks for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 1768–1783.

[31] Gulshan, V., Rother, C., Criminisi, A., Blake, A., Zisserman, A., 2010. Geodesic star convexity for interactive image segmentation. In: Proc. CVPR. pp. 3129–3136.

[32] Heiler, M., Schnoerr, C., 2003. Natural image statistics for natural image segmentation. In: IJCV. pp. 1259–1266.

[33] Herbulot, A., Jehan-Besson, S., Barlaud, M., Aubert, G., 2004. Shape gradient for image segmentation using information theory. In: Proc. ICASSP. pp. 17–21.

[34] Jackson, J., Yezzi, A., Soatto, S., 2007. Joint priors for variational shape and appearance modeling. In: Proc. CVPR. pp. 1–10.

[35] Kichenassamy, S., Kumar, A., Olver, P., Tannenbaum, A., Yezzi, A., 1995. Gradient flows and geometric active contour models. In: Proc. ICCV. pp. 810–815.

[36] Kim, J., Fisher, J. W., Yezzi, A., Cetin, M., Willsky, A. S., 2005. A non-parametric statistical method for image segmentation using information theory and curve evolution. IEEE Trans. Image Process. 14, 1486–1502.

[37] Kim, T. H., Lee, K. M., Lee, S. U., 2010. Nonparametric higher-order learning for interactive segmentation. In: Proc. CVPR. pp. 3201–3208.

[38] Kyprianidis, J.-E., Döllner, J., 2008. Image abstraction by structure adaptive filtering. In: Proc. EG UK Theory and Practice of Computer Graphics. pp. 51C–58.

[39] Lankton, S., Tannenbaum, A., 2008. Localizing region-based active contours. IEEE Trans. on Image Process., 2029–2039.

[40] Lempitsky, V., Kohli, P., Rother, C., Sharp, T., 2009. Image segmentation with a bounding box prior. In: Proc. ICCV. pp. 277–284.

[41] Li, C., Kao, C., Gore, J. C., Ding, Z., October 2008. Minimization of region-scalable fitting energy for image segmentation. IEEE Trans. Image Process. 17 (10), 1940–1949.

[42] Li, C., Xu, C., Gui, C., Fox, M. D., 2005. Level set evolution without re-initialization: A new variational formulation. In: Proc. CVPR. IEEE, pp. 430–436.

[43] Li, Y., Sun, J., Shum, H.-Y., 2005. Video object cut and paste. In: Proc. SIGGRAPH. pp. 595–600.

[44] Lin, T., 2000. Automatic video scene extraction by shot grouping. In: Proc. ICPR. pp. 39–42.

[45] Lombaert, H., Sun, Y., Grady, L., Xu, C., 2005. A multilevel banded graph cuts method for fast image segmentation. In: Proc. ICCV. pp. 259–265.

[46] Martin, D., Fowlkes, C., Tal, D., Malik, J., 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. ICCV. pp. 416–423.

[47] Martin, D. R., Fowlkes, C., Malik, J., 2004. Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Trans. Pattern Anal. Mach. Intell. 26 (5), 530–549.

[48] Mehrani, P., Veksler, O., 2010. Saliency segmentation based on learning and graph cut refinement. In: Proc. BMVC. pp. 110.1–12.

[49] Mumford, D., Shah, J., 1985. Boundary detection by minimizing functionals. In: Proc. CVPR. IEEE, pp. 22–26.

[50] Paragios, N., 2003. User-aided boundary delineation through the propagation of implicit representations. In: MICCAI (2). pp. 678–686.

[51] Paragios, N., Deriche, R., 1998. A pde-based level-set approach for detection and tracking of moving objects. In: Proc. ICCV. pp. 1139–1145.

[52] Paragios, N., Deriche, R., 2002. Geodesic active regions: a new paradigm to deal with frame partition problems in computer vision. Journal of Visual Communication and Image Representation, 249–268.

[53] Paragios, N., Rousson, M., Ramesh, V., 2002. Matching distance functions: A shape-to-area variational approach for global-to-local registration. In: ECCV. pp. 775–789.

[54] Price, B., Morse, B., Cohen, S., 2009. Learning-based interactive video segmentation by evaluation of multiple propogated cues. In: Proc. ICCV.

[55] Primo, C., Hernandez, A., Escalera, S., 2011. Automatic user interaction correciton via multi-label graph cuts. In: Proc. ICCV Workshop on HCI in Computer Vision.

[56] Protiere, A., Sapiro, G., 2007. Interactive image segmentation via adaptive weighted distances. IEEE Trans. Image Process. 16.

[57] Rother, C., Kolmogorov, V., Blake, A., 2004. Grabcut - interactive foreground extraction using iterated graph cuts. In: Proc. SIGGRAPH. ACM.

[58] Rousson, M., Cremers, D., 2005. Efficient kernel density estimation of shape and intensity priors for level set segmentation. In: MICCAI. pp. 757–764.

[59] Rousson, M., Paragios, N., 2002. Shape priors for level set representations. In: Proc. ECCV. Springer, pp. 78–92.

[60] Shugrina, M., Betke, M., Collomosse, J., 2006. Empathic painting: Interactive stylization using observed emotional state. In: Proc. NPAR. pp. 87–96.

[61] Singaraju, D., Grady, L., Vidal, R., 2009. P-brush: Continuous valued mrfs with normed pairwise distributions for image segmentation. In: Proc. CVPR. pp. 1303–1310.

[62] Tsai, A., Yezzi, A. J., III, W. M. W., Tempany, C. M., Tucker, D., Fan, A. C., Grimson, W. E., Willsky, A. S., 2003. A shape-based approach to the segmentation of medical imagery using level sets. IEEE Trans. Med. Imaging 22 (2), 137–154.

[63] Tsai, D., Flagg, M., Rehg, J., 2010. Motion coherent tracking with multi-label mrf optimization. In: Proc. BMVC. pp. 56.1–11.

[64] Veksler, O., 2008. Star shape prior for graph-cut image segmentation. In: ECCV. ECCV '08. Springer-Verlag, Berlin, Heidelberg, pp. 454–467. URL http://dx.doi.org/10.1007/978-3-540-88690-7_34

[65] Wang, J., Agrawala, M., Cohen, M. F., 2007. Soft scissors: an interactive tool for realtime high quality matting. In: Proc. SIGGRAPH. ACM, pp. 585–594.

[66] Wang, J., Bhat, P., Colburn, A., Agrawala, M., Cohen, M. F., 2005. Interactive video cutout. In: Proc. SIGGRAPH. pp. 585–594.

[67] Wang, L., Zeng, B., Lin, S., Xu, G., Shum, H.-Y., 2004. Automatic extraction of semantic colors in sports video. In: Proc. ICASSP. pp. 617–620.

[68] Wang, T., Collomosse, J. P., Hu, R., Slatter, D., Greig, D., Cheatle, P., 2011. Stylized ambient displays of digital media collections. Computers & Graphics 35 (1), 54–66.

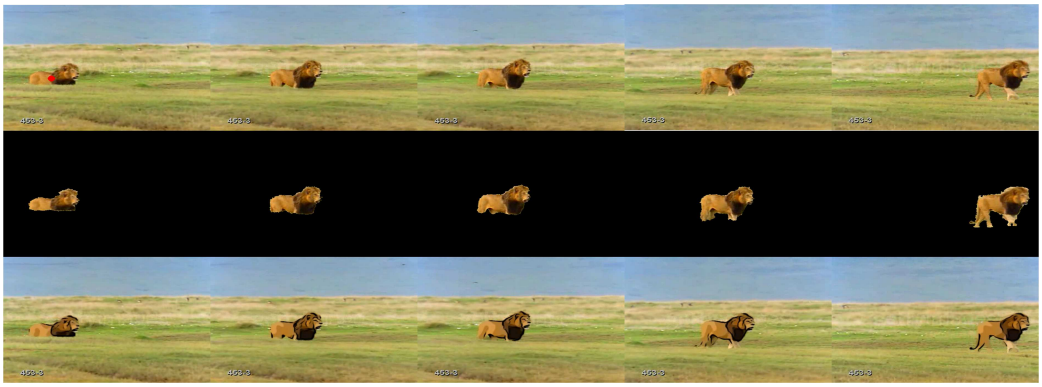
[69] Wang, T., Collomosse, J. P., Slatter, D., Cheatle, P., Greig, D., 2010. Video stylization for digital ambient displays of home movies. In: Proc. NPAR. pp. 137–146.

[70] Wang, T., Guillemaut, J.-Y., Collomosse, J. P., 2010. Multi-label propagation for coherent video segmentation and artistic stylization. In: Proc. ICIP. pp. 3005–3008.

[71] Zhao, H., Chan, T., Merriman, B., Osher, S., 1996. A variational level set approach to multiphase motion. Journal of Computational Physics, 179–195.

[72] Zhu, S., Yuille, A., 1996. Region competition: unifying snakes, region growing, and bayes/mdl for multiband image segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 884–900.

(a) Representative frames from “BEACH” sequence



(b) Representative frames from “LION” sequence

(c) Representative frames from “WALK” sequence

Figure 18: Additional segmentation results applying TouchCut to video sequences and using the matte to create foreground (a and b) or background (c) object stylization effects (source in top row, foreground object cut-out in middle row, painterly rendering on foreground/background object in bottom row).