Pattern Recognition 43 (2010) 676–684

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

POSIT: Part-based object segmentation without intensive training

Jue Wu*, Wenchao Cai, Albert C.S. Chung

Department of Computer Science and Engineering, and Bioengineering Program, The Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Article history:
Received 6 August 2008
Received in revised form 5 May 2009
Accepted 22 July 2009

Keywords:
Object segmentation
Training
Horse and cow segmentation
Part-based model

ABSTRACT

Object segmentation is a well-known difficult problem in pattern recognition. Until now, most of the existing object segmentation methods need to go through a time-consuming training phase prior to segmentation. Both robustness and efficiency of the existing methods have room for improvement. In this work, we propose a new methodology, called POSIT, for object segmentation without intensive training process. We construct a part-based shape model to substitute the training process. In the part-based framework, we sequentially register object parts in the prior model to an image so that the searching space is largely reduced. Another advantage of the sequential matching is that, instead of predefining the weighting parameters for the terms in the matching evaluation function, we can estimate the parameters in our model on the fly. Finally, we fine-tune the previous coarse segmentation by localized graph cuts. In the experiments, POSIT has been tested on numerous natural horse and cow images and the obtained results show the accuracy, robustness and efficiency of the proposed object segmentation method.

1. Object segmentation status quo

Object segmentation is a fundamental and challenging problem in the field of pattern recognition. Its goal is to segment a whole meaningful object in a natural scene. The big challenges for object segmentation methods on real world images are the following: (1) The target object itself is complicated and may have large variance in terms of pose, intensity, color, boundary sharpness, texture, etc. See the examples shown in the 1st and 3rd columns of Fig. 1. (2) The background may be chaotic and can be confused with the foreground (object). See the examples shown in the 1st and 2nd columns of Fig. 1. (3) The images can be noise-corrupted or the object may be occluded by irrelevant objects. See an example shown in the 4th column of Fig. 1.

The conventional low-level segmentation methods usually fail to tackle these notorious obstacles. In order to meet these challenges, object segmentation methods with both top-down and bottom-up styles have received extensive interests [1–6] in the past few years. These methods consider both low-level and high-level information in images and attempt to overcome the shortcomings of the conventional approaches.

The strategy of top-down and bottom-up combo methods is to introduce a prior shape information as a high-level guidance for segmentation. The cost of introducing prior information is the extra complexity added to the methods. Until now, all the top-down and bottom-up combo methods are developed on the basis of intensive training.[1] The information about shape and appearance of the target object is needed to extract in the training phase. This phase can be time-consuming and labor-intensive because plenty of segmented images with the same object are necessary to characterize the target object. With very few exceptions, all the training data sets are segmented by manual work. This produces the most accurate segmentation ground truth but the process is quite inefficient. Among numerous training-based methods, we will review three representative top-down and bottom-up combo methods in this section.

Borenstein and his colleagues have published several papers in an effort to integrate bottom-up with top-down criteria [1,2]. This methodology, in fact, relies on low-level segments more than high-level shapes. Hence, when the low-level segments cannot separate the foreground from the background, the final segmentation will be inaccurate. Moreover, the training procedure either includes non-class training images [2] or needs to extract a large number of informative segments as templates [1]. In order to mitigate the problem of huge training burden, the authors proposed a new learning process to automatically label the unsegmented training images in

* Corresponding author.
E-mail address: woojohn911@gmail.com (J. Wu).

[1] Training is defined as the prior procedure of studying the examples of known input/output functionality. Intensive training indicates the involvement of a large number of training examples.

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.07.013

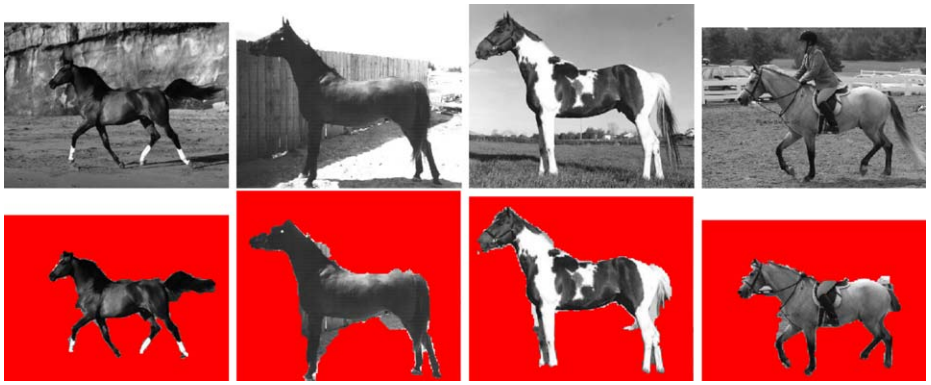

Fig. 1. Examples of horse images (1st row) and the corresponding segmentation results using the proposed method (2nd row). A brown horse may have white hoofs (1st column). The background can be easily confused with the foreground (2nd column). The skin of a horse has both slight and deep colored patches (3rd column). Part of a horse is occluded by a rider (4th column). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

another work [7]. The learning process, which combined top-down and bottom-up cues, could avoid the labor-intensive work of manual segmentation of training images. However, this method does not seem accurate enough because, when the target object has highly variable appearance, it may be in trouble due to the difficulty in matching of fragments[2] in different images.

The work of Levin and Weiss [4] formulated the bottom-up and top-down segmentation into a conditional random fields (CRF) framework. The segmentation component of the algorithm was similar to other shape-based methods but it extracted relatively fewer numbers of fragments from the training data. It was claimed in the paper that the training procedure not only considered the high-level cues but also the low-level features. This may be problematic when the target object has different low-level features (e.g. various colors or textures) in training and testing images. For example, this method may have trouble in segmenting images like the ones shown in the 1st and 3rd columns of Fig. 1. Moreover, it is unknown about how to decide the proper number of fragments. If insufficient number of fragments is chosen, the segmentation can be inaccurate.

It was shown in [3] that Obj Cut was an accurate object-category-specific segmentation method based on top-down and bottom-up cues. It combined the low-level Markov random field (MRF) model and high-level layered pictorial structure (LPS) [8] model. The accuracy of Obj Cut depends on the goodness of LPS samples because the final segmentation is the averaging result over all the samples. The requirement for the training data is demanding in Obj Cut because a number of video frames of the moving object are needed. The features of objects to be trained include both object outline and texture, which are not general enough if the object in images has various features or lacks texture patterns. Besides, although the accuracy of Obj Cut was shown to be good, the size of testing data set was not large in the work [3].

The above three works represent the state-of-the-art object segmentation methods but have a common limitation due to the training procedure. The authors of Obj Cut [3] mentioned an interesting application of nearly automatic object segmentation, namely "magic wand". For example, if the user knows that the image contains a horse, the wand can segment it without the need of manually specifying the near boundary (like intelligent scissor) or casting a set of seeds to differentiate foreground from background. However, not only is Obj Cut unable to implement that wand, but all the other state-of-the-art object segmentation methods are still far from the competent level to accomplish that goal. One of the reasons is that all the current methods rely on the intensive training procedure to acquire the prior information of the target object (shape, intensity/color, texture, etc.). This leads to a problem that, if the magic wand is required to segment multiple objects (e.g. animals, cars, human beings), the computation and storage burden of training-based methods will be tremendous.

The motivation of our work is to propose an accurate and efficient object segmentation method that does not need intensive prior training and also makes more feasible the attempt to achieve the magnificent goal of magic wand. We discard the prevailing procedure of training and pursue a new direction, which exploits part-based model for representing the basic information of the object. The purpose of training before segmentation is to learn the prior information about a specific object such that the segmentation methods can deal with the large variance of morphological and photometric features of the specific object in different images. To achieve this goal, the part-based model represents basic knowledge about the object, e.g. the number of components, the shape of each part, the relative location and orientation of each part, etc. By exploiting the part-based model, our method sequentially registers and matches[3] all parts to an image. Different from other methods that make use of intensity or texture of object, the sequential matching is mainly based on edge/gradient, which shares the same spirit of some basic psychophysical findings [10]. Finally the proposed method fine-tunes the boundary of the object according to the intensity statistics inside the region of each part, which is achieved by the localized graph cuts (LGC).

2. A part-based methodology

To circumvent the problem of intensive training, we design a new methodology for object segmentation. The framework of our method is straightforward, simple, yet effective, as will be shown experimentally. We construct a novel part-based model of the target object. From this model, we know the composition of the object and how the components (parts) are connected to each other. We then match the model to the input image in order to get a coarse segmentation. First, the salient parts are registered and matched to the image and some good candidates are kept. Then the other parts are registered and anchored with reference to the salient parts. At last, we choose the best result from these candidates according to a matching evaluation function. Based on the coarse segmentation, the boundary of each part is slightly deformed by optimizing an energy function

[2] A fragment means a rectangular patch of the image. It is a different concept from an object part.

[3] In this work, the process of aligning one part to the image is called "registration". The interaction of multiple parts is termed "matching" or "anchoring".

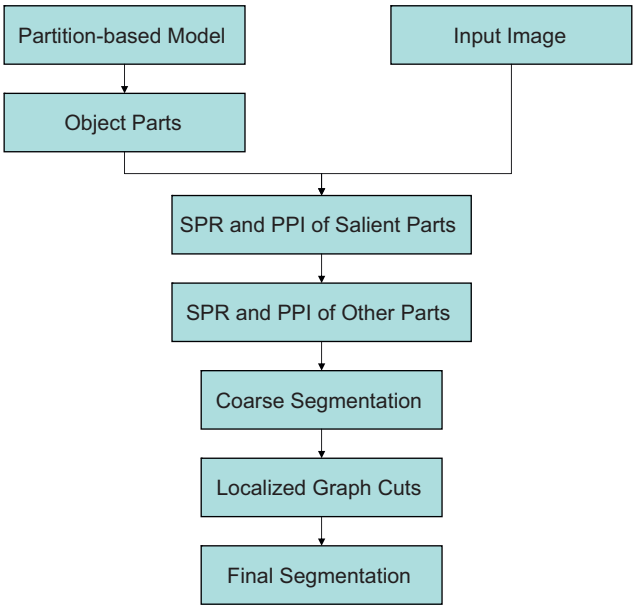

Fig. 2. The flow chart of the proposed algorithm. SPR and PPI refer to single-part registration and pairwise-part interaction, respectively.

using localized graph cuts. We call this procedure fine-tuning, where the energy function is derived from coarse segmentation and local intensity histograms so that the segmentation is consistent with the image content and also reflects the shape priors. The flow chart of the proposed algorithm is given in Fig. 2.

There are several advantages of the proposed method:

- The intensive training process that demands a large number of training images is unnecessary in the proposed method. We make use of simple knowledge about the object and construct a part-based model. With the help of this prior model and an efficient matching method, we can achieve a satisfactory performance of object segmentation.
- Instead of globally registering the shape template to an image, which leads to intractable computation cost, we register and match the parts sequentially rather than simultaneously. This can reduce the searching space of part detection so that we are able to search in more dimensions (five dimensions instead of four dimensions conventionally). Furthermore, the weighting parameters between single-part registration and pairwise-part interaction (in Eq. (1)) are thus allowed to be estimated on the fly.
- Among various image features such as intensity/color, edge and texture, we mostly rely on edge and gradient maps in the sequential matching. This is consistent with the study conclusion of human vision system [10], which suggests that the edge forms the basis of some physiological models about the human early vision. Relying largely on gradient/edge information makes POSIT immune from the distraction of different object appearances such as intensity profiles and texture patterns.
- Unlike the work in [11] that uses only one point to control the relative pose of two parts, we make use of double control points which make the linkage between two adjacent parts under more convenient and precise control. Refer to Fig. 3 for the complicated control of interactive parts, which is unlikely for zero or one control point.
- In the fine-tuning process by localized graph cuts, we make use of a local intensity histogram for each part instead of global histogram. The local histograms can better preserve the information of a small portion of an object (e.g. the white hoof of a black horse), whose intensity is much different from the remaining larger portions of the object.

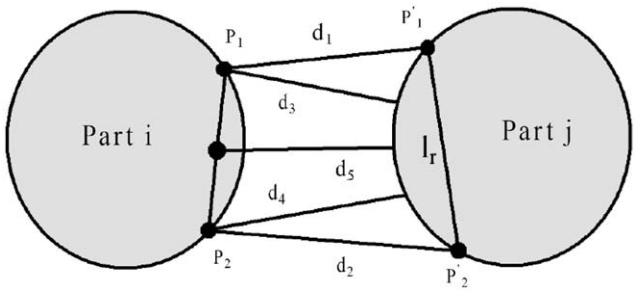

Fig. 3. The distances used in the pairwise-part matching. P_1 and P_1', P_2 and P_2' are corresponding control points in interactive parts; d_1 and d_2 refer to the distances between them, respectively; d_3 and d_4 are the distances of P_1 and P_2 to the boundary of part j, respectively; d_5 is the distance of the middle point of P_1 and P_2 to part j; l_r is the distance between P_1' and P_2' used as a reference length.

In the subsequent subsections, we first discuss the shape representation of our method, i.e., pictorial structure and its related work. Then, we describe the proposed method and how to achieve the above advantages in detail.

2.1. Pictorial structure

The idea of pictorial structure (PS) appeared more than 30 years ago [12] to tackle the difficult task of modeling the shape variation of a non-rigid object. The merit of the PS is to divide a complex object into a number of rigid components (parts). Part-to-part connection is described and controlled through a "spring". In that way, the complex shape variation can be modeled by the rigid (or affine) transformation of several parts.

The virtues of the pictorial structure have attracted attention of researchers in shape representation and object segmentation. For example, Felzenszwalb and his colleagues used the Bayesian framework and applied PS to object recognition [11]. They further restricted the relationship of parts to a tree structure and sped up the global matching of PS to the whole image [13]. Kumar and his colleagues extended PS for object segmentation to a complete graph of part interaction while maintaining moderate computation burden [14]. Furthermore, in [8], they proposed a learning-based method for a layered pictorial structure (LPS). On the basis of the above two works [14,8], a segmentation method (i.e., Obj Cut) [3] was proposed, which is based on LPS and relies on the training from videos.

The implicit problem of the existing methods based on pictorial structure is that they try to obtain a global match between the PS and the image. First, the computation cost of a global match is formidable due to a large variance of object poses and background clutter. Hence, the compromise between accuracy and efficiency goes to a suboptimum, which can be obtained by some optimizers like belief propagation. Different strategies from global matching, e.g. sequential matching, should be considered. Second, in the global matching strategy, the weights of each term in the energy function cannot be adjusted during the matching. This may cause the problem of unsuitable setting of weighting parameters which are decided before the global matching. As an alternative to global matching, sequential registration of parts to the image makes it possible to estimate the weights in the energy function on the fly.

The major differences between our part-based model with the existing part-based models are the followings: (1) We adopt a part-based model to alleviate the pain of intensive training process, which involves a large number of training images with known

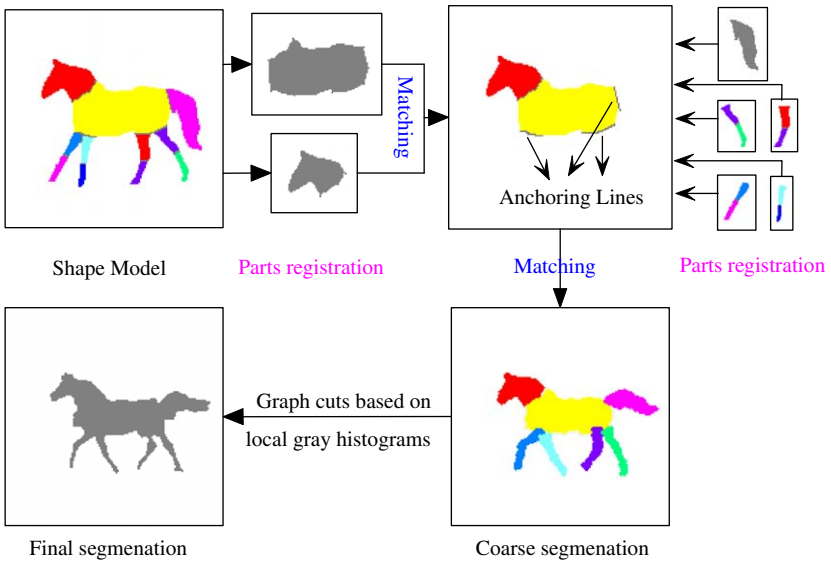

Fig. 4. (Color images) The flow chart of the instantiation of the proposed model to horse segmentation.

ground truths. (2) We use a different matching strategy, which makes it possible to search with higher dimensions and estimate the parameters on the fly. (3) We propose a new style for adjacent part interaction using double control points for each part instead of a single control point or no control point. This makes it more convenient to manipulate the relative movement of two adjacent parts.

2.2. Part-based model construction

Before we perform the segmentation, some prior knowledge of the target object is needed. The knowledge is incorporated in a part-based shape model. To obtain this model, the user only needs to divide an example of the target object into several parts (components). The guideline is that the complex object is decomposed to a few simple and manageable components after the division. Each part has relatively simple morphology. In addition, the user needs to specify "salient" parts in order to initialize the matching in the next stage. The guideline for choosing the salient parts is that they usually have the largest size, the number of other parts to which they are connected is large, they are the most representative parts of the object or they are the common parts or features of the object. For example, a shape of a person is able to disassemble into 10 parts, i.e., head, trunk, left and right arm parts (upper and lower), and left and right leg parts (upper and lower). Salient parts of the person are head and trunk. The salient parts of a chair can be the sitting surface and the chair back while the other parts are the arms and legs. It should be noted that the main target of the current work is an articulate object or an object that can be easily decomposed into distinctive parts.

From this division, we obtain the shape of each part and its approximate position. The shape and pose of each part in the part-based model are regarded as the prior knowledge of the target object. Between two adjacent parts in the model, there is a shared boundary. The two end points of this boundary are the anchor points (control points). The linkage and interaction among adjacent parts are manipulated through these control points. Two non-adjacent parts can also interact by changing relative poses. The transformations of each part involve translation, rotation and scaling.

In short, the user only needs to divide an exemplar shape of the target object into several parts and specify which parts are salient. Thus, the part-based shape model is constructed. The exemplar shape can be a given segmentation of the object from the training data, e.g. a binary image of a horse (see the shape model in Fig. 4). To construct the model, we only need a single training data and the construction is simple to complete.

2.3. Part-based model matching

After the construction and specification of the shape model, we then register the model to the image content and perform the matching among parts. This is a crucial step for locating the rough positions and poses of the object components. We take into account some simple low-level features in the matching process, i.e., edge, gradient, pixel intensity, etc.

The goodness of matching is measured by an evaluation function. This function evaluates both the qualities of single-part registration (SPR) to the image and pairwise-part interaction (PPI). The formulation can be expressed as

$$E(P) = \sum_i M_i(p_i) + \beta_{i,j} \sum_{i,j} T_{i,j}(p_i, p_j). \quad (1)$$

In the above equation, $M_i(p_i)$ is the cost function of part i registered to the image given pose p_i. $T_{i,j}(p_i, p_j)$ represents the cost of the relative positions of two interactive parts i and j given poses p_i and p_j, respectively. $P = \{p_i | i = 1, \ldots, n_p\}$, where n_p is the number of parts.

The pose of one part includes five components, i.e., translations x and y, rotation θ, scales s_x and s_y. Hence, the pose p_i can be written as a vector (x, y, θ, s_x, s_y). Double scaling dimensions, s_x and s_y along x and y directions, are used instead of the conventional single scaling dimension. This extension lets the proposed method explore more poses for each part.

In the matching process, we perform exhaustive search in defined ranges (cf. Section 3) to find a configuration of poses so that the energy is as low as possible, i.e.,

$$\hat{P} = \arg\min_{P \in \Omega_p} E(P), \quad (2)$$

where Ω_p represents the set of all effective pose configurations.

We adopt the style of sequential matching as opposed to global matching, which was exploited widely in other methods. In sequential matching, we first deal with salient parts of the object and then the other parts are matched to the image with reference to the salient parts. The procedure is as follows. First, the salient parts are

registered to the input image by SPR and the n best candidates are kept for each part. Second, the matching among the salient parts is performed by PPI, and the best n combinations of all salient parts are kept. Third, the remaining parts are registered to the image by SPR with initial positions based on salient parts. In the end, the best matching by PPI of all parts is picked as the coarse segmentation. Obviously, this is a sequential style of matching in contrast to the conventional global simultaneous matching.

One advantage of sequential matching lies in the significant reduction of the searching space. The subsequent parts are registered to the image based on the constraint of the poses of the previously registered parts. This constraint can largely shrink the searching space of the poses of the following parts. Since part-to-part interaction is a critical factor in the matching, some configurations that violate the mutual constraint of part poses can be discarded directly. Obviously, the sequential matching is much more efficient than the global one such that we are able to expand the searching dimensions to perform more precise registration.

Another advantage of sequential matching is that the weight β of the part interaction can be estimated on the fly. The weight $\beta_{i,j}$ is to balance the quantity of $(M_i(p_i)+M_j(p_j))$ and the quantity $T_{i,j}(p_i, p_j)$ so they are of similar order of magnitude. Obviously, β is dependent not only on parts i and j, but also on the image to which parts i and j are registered. In other words, since $T_{i,j}(p_i, p_j)$ is content-free (irrelevant to the image) and $(M_i(p_i)+M_j(p_j))$ is relevant to the image, we need to estimate β according to image content such that neither M_i+M_j nor $T_{i,j}$ is over-weighted or under-weighted. This can be achieved in the sequential matching without the addition of too much computation. This kind of flexible strategy is unlikely to implement in a global matching because the weight of $T_{i,j}$ is usually fixed before the start of a global matching.

The on-the-fly estimation of β is performed as follows. The best n candidates of registration results of part i are kept and are sorted in an ascending order $\{M_{i1}, M_{i2}, \ldots, M_{in}\}$. The same sequence for part j is calculated. Then, β is estimated as,

$$\beta_{i,j} = \max(|M_{in/2} - M_{i1}|, |M_{jn/2} - M_{j1}|). \tag{3}$$

Since $T_{i,j}(p_i, p_j)$ is content-free (usually small values) and the order of magnitude of $(M_i(p_i))$ can vary a lot depending on the image content, the multiplication of $T_{i,j}$ with $\beta_{i,j}$ will bring the two measures to similar magnitude.

We now describe these two terms, single-part registration and pairwise-part matching, in Eq. (1) in the next two subsections.

2.3.1. Single-part registration (SPR)

In coarse segmentation, the registration of a single part to the image mainly makes use of edge or gradient features. The single-part registration term in Eq. (1) is expressed by

$$M_i(p_i) = e_1(p_i|I_E) + e_2(p_i|I_G) + e_3(p_i|I), \tag{4}$$

where I is the input image, I_E is the edge map of the image extracted by the Canny edge detector and I_G is the gradient map of the image. The three terms e_1, e_2, e_3 are discussed below.

The first factor e_1 we consider is the difference between the part boundary and the image edges. We adopt the partial Hausdorff distance (the kth maximal distance) [15] and another distance to measure this difference, which is defined as

$$e_1(p_i|I_E) = \mathrm{PHD}(B_i, I_E) + d_0(i), \tag{5}$$

where $d_0(i)$ refers to the difference between average distances of all points within part i to its nearest edge in the model and in the image. PHD is defined as

$$\mathrm{PHD}(B_i, I_E) = K^{th} \max_{b\in B_i} \min_{a\in I_E} \|a-b\|, \tag{6}$$

where B_i is the sets of boundary pixels of part i in the model, and $\|a-b\|$ is the Euclidean distance between pixels a and b.

The first term of e_1 reflects the closeness of the part to the corresponding part in the image in terms of the boundary shift. Considering that the two profiles of boundary cannot be exactly the same, the partial Hausdorff distance is used such that the effect of some outliers can be eliminated. The second term of e_1 aims to avoid the occasion that one part is matched to a region, which is not the object but has many edges. This edge-rich region can be a clutter background. Usually, the number of edges within an object part is significantly fewer than those in a clutter background.

The second factor e_2 is related to the gradient calculation. When the part in the part-based model is well overlapped with the corresponding part of the object in the image, it is expected that the gradient of the pixels around the boundary is relatively large. We incorporate this observation into the evaluation of single-part registration by e_2, i.e.,

$$e_2(p_i|I_G) = -\frac{1}{l_0}\iint_{B_i} \nabla I(x,y)\,dx\,dy, \tag{7}$$

where $I(x,y)$ is the image intensity at coordinate (x,y), B_i is the boundary of part i, and l_i is the length of the boundary B_i. This measure e_2 embodies the length-average gradient magnitude along a contour. Obviously, if the part is near the correct position, e_2 should become small.

The last factor e_3 is accessorial. It is natural to assume that the area within one object part does not bear large variance of intensity. A moderate penalty is given to a candidate area with large inhomogeneity. As such, this factor is expressed by

$$e_3(p_i|I) = \iint_{A_i} (I(x,y) - \bar{I}(A_i))^2\,dx\,dy, \tag{8}$$

where A_i is the image region overlapped with part i and $\bar{I}(A_i)$ is the mean intensity value within region A_i.

2.3.2. Pairwise-part interaction (PPI)

In addition to the single-part registration, we need to further model the pairwise-part interaction, i.e., the second summand $T_{i,j}$, because all the parts as a whole compose the object and they have relatively steady pose relationship with each other. The second summand in Eq. (1) is expressed by

$$T_{i,j}(p_i, p_j) = e_4(D_p|p_i, p_j) + e_5(d_s|p_i, p_j), \tag{9}$$

where D_p denotes various distances between two parts, such as the distances among the corresponding anchor points. Quantity d_s denotes the difference between the scales of two parts. These relations are formulated as

$$e_4(D_p|p_i, p_j) = \frac{|d_c|}{l_r/2} + \frac{|d_5|}{l_r/2} + s + \frac{n_{ij}}{\min(n_i, n_j)}, \tag{10}$$

$$e_5(d_s|p_i, p_j) = d_{s_x} + d_{s_y}. \tag{11}$$

The meanings of various distances are illustrated in Fig. 3. The larger distance of the two distances among the corresponding anchor points is $d_c = \max(d_1, d_2)$, where d_1 and d_2 are the distances between the corresponding anchor points. The distance between the midpoint of the two anchor points of part i and the boundary of part j is denoted by d_5. Variable s is to encourage the interactive boundaries of parts i and j to be parallel and $s = \max(d_3/d_4, d_4/d_5)$, where d_3 and d_4 are the distances between the anchor points of part i and the boundary of part j. The quantity l_r is a reference distance, which

indicates the size of part j and is set to the longest diameter of the part. Variables n_i and n_j denote the number of pixels in parts i and j, respectively, and n_{ij} is the number of pixels in the overlapping of parts i and j. Variable d_{s_x} (d_{s_y}) is defined as the difference of the scale values along $x(y)$ direction of parts i and j.

The effect of these relations makes two interactive parts interact compactly. If two parts are adjacent, they are probably not departed or overlapped a lot. Besides, they are always linked around the place where the part-based model defines. We allow two adjacent parts to overlap with a cost (the fourth term of e_4). This can deal with the problem of occlusion. For two parts that are not adjacent, they are still likely to interact. We match these parts through the constraint of relative poses, which can also be embodied by e_4 and e_5. Together with the single-part registration, the whole energy evaluation can effectively detect and locate the approximate position of the target object in the image.

The matching between the shape model and the object in an image need not be highly accurate. The step of precise segmentation is done by refinement using the local graph cuts.

2.4. Refinement and figuration

The registration (SPR) and matching (PPI) of parts to the object in an image provides a coarse segmentation of the object. This is equivalent to the top-down procedure, which imposes a high level of knowledge to the image. Due to the fact that the objects in real images may have various shapes, a low-level segmentation procedure is required to figure out the image-based object content.

Before the refinement, since one part may detach from its adjacent part, we need to merge all parts to one object with a single boundary. We connect two adjacent parts with a rectangle, which will cover the region between two separate adjacent parts. The refinement phase involves the image intensity statistics mainly. Given the initial segmentation obtained from the previous step, the image is divided into background and foreground. The intensity histograms of foreground and background can then be calculated. We calculate the histograms on a part-by-part basis instead of globally on the image. Thus, the likelihood of which label one pixel belonging to can be derived from the local statistical histogram. Furthermore, the neighboring pixels are encouraged to have the same label unless their intensities differ significantly.

According to the ideas above, the figuration of the final object segmentation is performed by minimizing one single energy function, i.e.,

$$F = 2\sum_x (F_1(I|l_x) + F_2(l_x|S)) + \sum_{y\in N_x} F_3(I|l_x, l_y). \tag{12}$$

If $l_x = 1$ (foreground),

$$F_1(I|l_x) = -\log p(l_x = 1|H_o), \tag{13}$$

$$F_2(l_x = 1|S) = 1/(1 + \exp(\mu \cdot d(x))); \tag{14}$$

else

$$F_1(I|l_x) = -\log p(l_x = 0|H_b), \tag{15}$$

$$F_2(l_x = 0|S) = 1 - 1/(1 + \exp(\mu \cdot d(x))). \tag{16}$$

H_o (H_b) is the object (background) intensity histogram of the local region containing the part to which x belongs. Variable $d(x)$ is defined as the distance of pixel x to the nearest boundary of the coarse segmentation S. In Eq. (12), N_x is the set of neighboring pixels of x. Eqs. (13)–(16) are served as likelihood energies, and make use of the previous coarse segmentation obtained from SPR and PPI. The histograms H_o and H_b, and distance $d(x)$ in these equations attempt to generate a final result similar to the coarse segmentation. Small refinement is implemented by the prior energy F_3

$$F_3(I|l_x, l_y) = \begin{cases} \lambda + 1/(1 + \gamma\delta^2) & \text{if } l_x \neq l_y, \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

where λ and γ are two positive parameters. Variable $\delta = |I_x - I_y|$ is the intensity difference. Variable λ represents the minimal penalty imposed to the case that two neighboring pixels have different labels. Nonetheless, the penalty is adjusted with regard to the difference of intensity of the neighboring pixels. If the difference is small, the penalty is even larger. Otherwise the penalty is relatively small. Variable γ magnifies the intensity difference of neighboring pixels if γ is small. Otherwise, it reduces the effect of the intensity difference. The above energy function is optimized by the graph cuts (GC) algorithm [9]. Since we perform the GC α-expansion algorithm for each part within a local region, the graph cuts optimization is localized.

3. Applications

The proposed object segmentation method, POSIT, can be applied to the segmentation of various objects, e.g. horses, cows, pedestrians and so on. The framework is general and does not prefer any specific object. In this section, we focus on the application of POSIT to horse and cow segmentations for the reason that quadrupeds usually have a large number of different poses and appearances. The segmentation of quadrupeds, especially horses, is a hard problem, and most of the object segmentation methods had been applied to this problem and evaluated.

In this paper, we study the side view of quadrupeds. Two types of part-based models were constructed, one for horses and the other for cows (see the example of horse model in Fig. 4). The salient parts in our model are head and body. We set the candidate number n to 20. The searching step sizes were 2 pixels, 5.7° (0.1 radian) and 0.1 for translations (both x and y directions), rotation, scales (both x and y directions), respectively. The starting pose was set the same as the shape prior model in the middle of the image. The ranges of steps for translations, rotation and scales were $[-25, 25]$, $[-25, 25]$, $[-10, 10]$, $[0.3, 1.5]$ and $[0.3, 1.5]$, respectively. The most important parameter of POSIT, $\beta_{i,j}$, was estimated on the fly automatically. Besides, $\gamma = 2.5 \times 10^{-3}$ and $\lambda = 10$. We set the two parameters empirically. The guideline is as follows. If we anticipate the intensity difference between background and foreground is not large, a smaller λ and a larger γ than the current setting should be adopted. The flow chart of the instantiation of our model to horse segmentation is given in Fig. 4. Note that in the segmentation on different data, all the parameters were kept fixed, which suggested that the results were not sensitive to the parameter setting.

The horse database we used for experiments was obtained from the Weizmann Institute, Israel [16]. It is a large database with versatile horse images, and was used in other work [1–4]. There are 328 horse images in the database and we performed segmentation on all of them in the experiments. The cow database was obtained from the Vision Group at the University of Leeds [17]. We performed segmentation on 10 cow images with different object appearances and backgrounds. All the horse and cow data sets are converted to gray level images, which brings more challenges to the segmentation than the color images.

In Fig. 5, we show some intermediate results of the proposed model, which are the matching results of the part-based model and the images. More final segmentation results are shown in Figs. 6 and 7. The segmentation accuracy is measured by #{correctly segmented pixels}/#{image pixels}, defined the same as in [1]. The

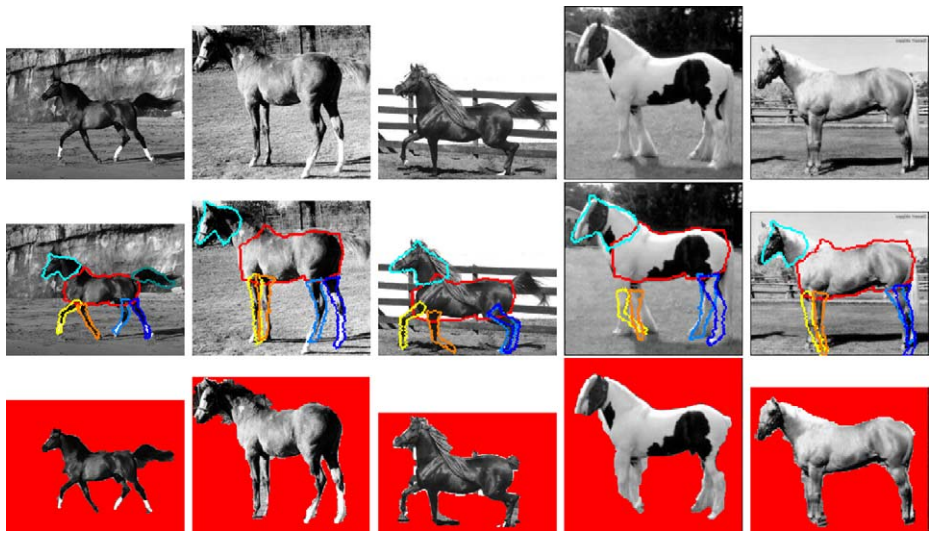

Fig. 5. (Color images) Some examples for showing the intermediate segmentation results of the proposed method. First row is the original input images. Second row is the intermediate results. They are the coarse segmentation obtained from SPR and PPI. Third row is the final segmentation results after fine-tuning.

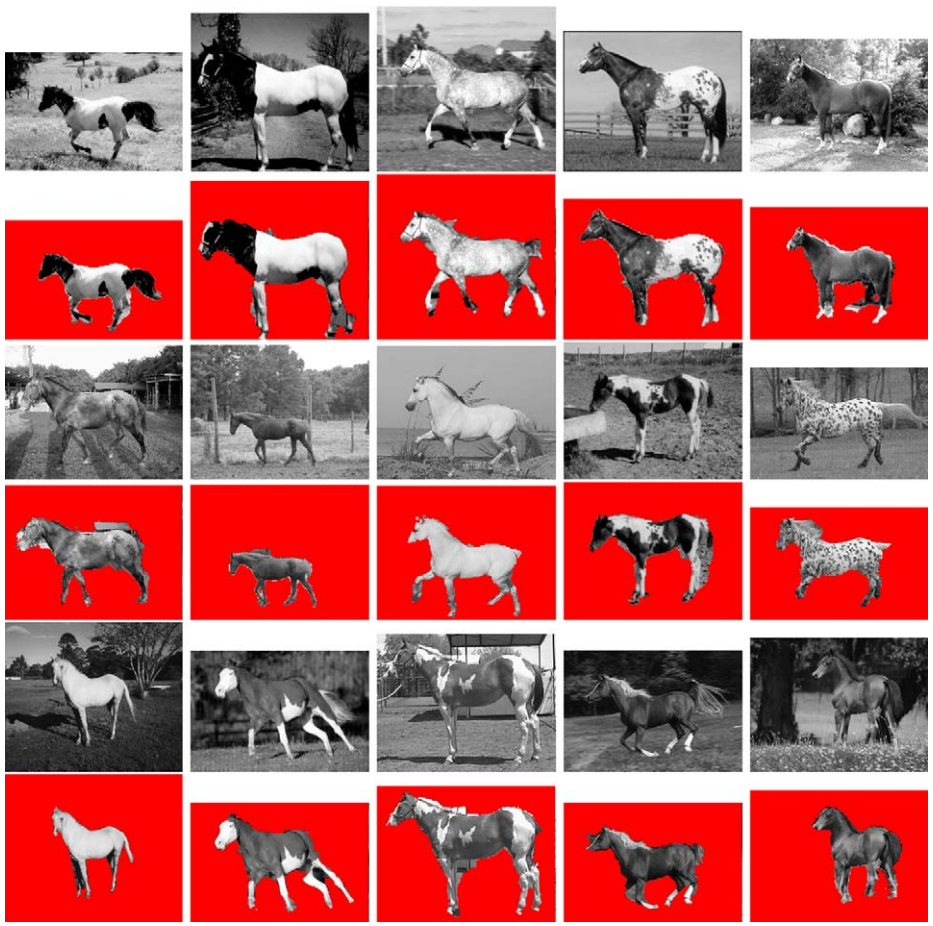

Fig. 6. More examples of the horse images and their corresponding segmentation results of the proposed method.

average segmentation accuracy for horse and cow images is 93% and 94%, respectively. More specifically, the average accuracy for foreground (horses and cows) is 84%. This accuracy is calculated by the number of correctly segmented foreground divided by the number of total foreground pixels in the ground truth. To further analyze the segmentation qualitatively, we find that the most accurate parts to the least are body, head, legs and tail. Overall the segmentations are satisfactory and no part is mis-segmented consistently. The whole-image segmentation accuracy of the proposed method is comparable to the related methods. The hierarchical method [1] achieved an

Fig. 7. Examples of the cow images and their corresponding segmentation results of the proposed method.

accuracy of 93% on the whole horse database, similar to the proposed method. Obj Cut [3] obtained an accuracy of 96% on 10 horses in the same database. The CRF-based method [4] reached a 95% accuracy on part of the same database.

If n and m denote the numbers of parts and possible part poses, respectively, the time complexities for SPR and PPI are $\mathcal{O}(nm)$ and $\mathcal{O}((n-1)m)$. As such, the time complexity of the whole algorithm is $\mathcal{O}(nm)$. The complexity is also linear in the image size, which determines the time to calculate the evaluation energy. The computational complexity is similar to the related methods [1,2]. The running time of POSIT on a 150×150 image was around 1 min on a 2.13 GHz Intel Pentium 4 computer with 1 GB memory.

From the experimental results, it is observed that even with an absence of intensive training process, the proposed method can perform good segmentation on challenging horse and cow images. One reason lies in that we adopt a more flexible scaling style that allows differences along x and y directions in contrast to existing work such as Obj Cut. Our method can overcome the obstacles of complicated object appearance (white horses with black heads, speckled horses), various object poses (standing, running, leaping), blurred boundary and low contrast, object occlusion (horses with a rider), and so on. One technique that may enhance the ability of the proposed method to deal with inhomogeneous object is the surround inhibition that can reduce the irrelevant edges induced by internal texture [18,19]. The proposed method achieves a comparable segmentation accuracy to the existing state-of-the-art methods on the same database. Among them, two methods [3,4], which tested on much smaller data volumes than ours, have slightly higher accuracies than our method. Most of the errors in our method came from the failure of segmentation of horse tails or the elongated horse's head (occurs when the horse is eating). The current modeling of tail and head tended to fail if the shape undergoes non-affine deformation. Another error source was that when the intensity of the object was too close to that of the background, the bottom-up LGC might include some background regions in the final segmentation.

4. Conclusion

The training process involving a large number of training images with ground truth is a burden of conventional object segmentation methods. Our work makes an effort to circumvent this overhead by means of a part-based methodology. Contrary to other existing methods, the proposed method adopts the strategy of sequential matching of object parts to an image, which cuts down the searching space and allows a more flexible style of parameter setting. We have applied the proposed model to the problem of horse and cow segmentation. Segmentation of other objects such as pedestrians can use similar strategy. Future work will apply the method to larger databases and different objects to obtain further validation.

References

[1] E. Borenstein, J. Malik, Shape guided object segmentation, in: Computer Vision and Pattern Recognition, 2006, pp. I:969–I:976.
[2] E. Borenstein, E. Sharon, S. Ullman, Combining top-down and bottom-up segmentation, in: Computer Vision and Pattern Recognition Workshop, 2004, pp. 46–53.
[3] M. Kumar, P. Torr, A. Zisserman, Obj Cut, in: Computer Vision and Pattern Recognition, 2005, pp. I:18–I:25.
[4] A. Levin, Y. Weiss, Learning to combine bottom-up and top-down segmentation, in: European Conference on Computer Vision, 2006, pp. IV:581–IV:594.
[5] S.X. Yu, J. Shi, Object-specific figure-ground segregation, in: Computer Vision and Pattern Recognition, 2003, pp. II:39–II:45.
[6] B. Leibe, A. Leonardis, B. Schiele, Robust object detection with interleaved categorization and segmentation, International Journal of Computer Vision 77 (2008) 259–289.
[7] E. Borenstein, S. Ullman, Learning to segment, in: European Conference on Computer Vision, 2004, pp. III:315–III:328.
[8] M.P. Kumar, P.H.S. Torr, A. Zisserman, Learning layered pictorial structures from video, in: Indian Conference on Computer Vision, Graphics and Image Processing, 2004, pp. 158–164.
[9] Y. Boykov, O. Veksler, R. Zabin, Fast approximate energy minimization via graph cuts, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (11) (2001) 1222–1239.
[10] D. Marr, E. Hildreth, Theory of edge detection, Proceedings of the Royal Society of London, Biological Science 207 (1167) (1980) 187–217.
[11] P. Felzenszwalb, D. Huttenlocher, Pictorial structures for object recognition, International Journal of Computer Vision 61 (2005) 55–79.
[12] M. Fischler, R. Elschlager, The representation and matching of pictorial structures, IEEE Transactions on Computers 22 (1973) 67–92.
[13] P. Felzenszwalb, D. Huttenlocher, Efficient matching of pictorial structures, in: Computer Vision and Pattern Recognition, 2000, pp. II:66–II:73.
[14] M. Kumar, P. Torr, A. Zisserman, Extending pictorial structures for object recognition, in: British Machine Vision Conference, 2004, pp. 789–798.
[15] D.P. Huttenlocher, G.A. Klanderman, W. Rucklidge, Comparing images using the Hausdorff distance, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (9) (1993) 850–863.
[16] E. Borenstein, S. Ullman, Class-specific, top-down segmentation, in: European Conference on Computer Vision, 2002, pp. II:109–II:124.
[17] D. Magee, R. Boyle, Detecting lameness in livestock using resampling condensation and multi-stream cyclic hidden Markov models, Image and Vision Computing 20 (8) (2002) 581–594.
[18] C. Grigorescu, N. Petkov, M.A. Westenberg, Contour and boundary detection improved by surround suppression of texture edges, Image and Vision Computing 22 (2004) 609–622.
[19] G. Papari, P. Campisi, N. Petkov, A. Neri, A biologically motivated multiresolution approach to contour detection, EURASIP Journal on Advances in Signal Processing 2007 (2007) 28.

About the Author—JUE WU received the B.Sc. degree in Computer Science from Fudan University and the doctoral degree in Bioengineering at The Hong Kong University of Science and Technology. He is currently a Postdoctoral Fellow in Department of Radiology, University of Pennsylvania. His research interests include biomedical image segmentation and registration, energy minimization, and statistical pattern analysis.

About the Author—WENCHAO CAI received B.Sc. and M. Phil. degrees in Computer Science from Tsinghua University, China, in 2003 and 2005. He is now working on the separator of impunity in textile at Qingdao Textile Eye Inc. He is interested in industrial applications of techniques in computer vision and image processing.

About the Author—ALBERT C.S. CHUNG received the B.Eng. degree (first class Honors) in Computer Engineering from The University of Hong Kong in 1995 and the M.Phil. degree in Computer Science from The Hong Kong University of Science and Technology in 1998. He joined the Medical Vision Laboratory, University of Oxford, Oxford, UK, as a Doctoral Research student with a Croucher Foundation Scholarship and graduated in 2001. He was a Visiting Scientist at the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, in 2001. He is currently an Associate Professor with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology. His research interests include medical image analysis, image processing, and computer vision. He won the 2002 British Machine Vision Association Sullivan Thesis Award for the best doctoral thesis submitted to a UK University in the field of computer or natural vision.