Interactive Object Contour Extraction for Shape Modeling

Tomasz Adamek and Noel E. O'Connor

Centre for Digital Video Processing, Dublin City University, Ireland
{adamekt, oconnorn}@eeng.dcu.ie

Abstract. In this paper we present a semi-automatic segmentation approach suitable for extracting object contours as a precursor to 2D shape modeling. The approach is a modified and extended version of an existing state-of-the-art approach based on the concept of a Binary Partition Tree (BPT) [1]. The resulting segmentation tool facilitates quick and easy extraction of an object's contour via a small amount of user interaction that is easy to perform, even in complicated scenes. Illustrative segmentation results are presented and the usefulness of the approach in generating object shape models is discussed.

1 Introduction

The statistical analysis of shapes is widely applicable to many areas of image analysis including: analysis of medical images [2, 3], industrial inspection tools, modeling of faces, hands and walking people [4, 5]. Active Shape Model (ASMs), sometimes termed 'Smart Snakes' were originally introduced by Cootes and Taylor [6] and are deformable models with global shape constraints learned through observations. Objects are represented by a set of labelled points and by examining the statistics of the position of the labelled points a Point Distribution Model (PDM) is derived giving the average position of the points, and a description of the main modes of variation found in the training set. PDMs can then be used to automatically identify examples of the model object in unseen images [4].

To use PDMs a labelled set of example shapes are required in order to train the model. A key difficulty is establishing correspondence between training instances, a process referred to as landmarking. This is often achieved by manual annotation, which is subjective and extremely labor intensive. In our previous work [7], we presented algorithms for fast and efficient construction of PDMs eliminating the burden of manual landmarking based on two key technologies: a semi-automatic segmentation approach for fast extraction of examples of closed contours from a set of images and a method for automatic identification of landmarks on the set of examples. In [7] we focused on the automatic identification and alignment of contour points. In this paper, we provide a detailed description of a segmentation tool useful for extracting shape training data for any kind of object.

The ill-posed nature of the problem of partitioning an image into a set of semantic entities is well known. For a certain class of multimedia applications

it is unavoidable not to include user input in the segmentation process [8]. This has led to the development of so-called *supervised* or *semi-automatic* approaches where a user, by his/her interactions, defines what objects are to be segmented within an image. The user should be able to obtain exactly the segmentation he/she desires in terms of content and accuracy with a minimum amount of user interaction that can be easily and quickly performed with the system responding in real-time to user input.

Of the various approaches to interactive segmentation [9], we believe that the most intuitive scheme for interaction is the feature-based approach where the user is allowed to draw on the scene, placing markers in a form of single point seeds or by dragging a mouse over the image thereby creating a set of labelled *scribbles* for the objects to be segmented. Probably the best known example of such an approach is Seeded Region Growing (SRG) [10], but other examples include [11] and [12]. An elegant and efficient solution to image partitioning based on scribbles was described in [1] where the input image is automatically pre-segmented into regions and represented as a Binary Partition Tree (BPT) to allow rapid responses to users' actions. The tree structure is used to encode similarities between regions pre-computed during the automatic segmentation process. This structure is then used to rapidly propagate the labels from scribbles provided by the user to large areas (regions) of the image. A variant of Salembier's approach that addresses some limitations of that approach is used in our work.

The remainder of this paper is organized as follows. The proposed approach to interactive segmentation is presented in section 2. This requires a discussion of both the user interaction methodology employed and the underlying automatic image segmentation process. The usefulness of the proposed approach is discussed in terms of both accuracy, via illustrative results in section 3, and its usefulness in generating training data for generating statistical shape models, in section 4. Finally, conclusions and directions for future research are presented in section 5.

2 Proposed Approach

The difference between our approach and that of [1] is the underlying region segmentation process that drives BTP creation and the approach for propagating labels from user defined scribbles to completely label the entire image. Our proposed modifications lead to fast classification of large regions even if the provided scribbles consist of only a few pixels. In the following, the different steps of the algorithm are outlined with emphasis on the modifications introduced.

2.1 User Interaction

Since the primary goal of the segmentation tool is the extraction of a single closed contour, interaction is restricted to only two objects referred to as foreground and background. Both objects can be made up of an arbitrary number of regions,

homogenous according to a certain criteria. Only two sets of disconnected user scribbles are required: one for foreground and one for background. The scribbles can have a complex shape, be disconnected and cover large areas or they can be as small as one pixel.

2.2 Building the BPT

Whilst we use the BPT framework proposed in [1], our BPT is created using the automatic syntactic segmentation approach we previously presented in [13]. The syntactic approach results in a more intuitive/meaningful merging order when creating the BPT. The incorporation of the syntactic features in the pre-segmentation stage leads to the creation of more meaningful partitions, which is especially important at the higher levels of BPT. We believe that this improves the intuitiveness and efficiency of user interactions e.g. by allowing the segmentation process make more "intelligent" decisions in areas of the scene where there is no explicit information provided by the users interactions. To ensure the highest possible accuracy, the complete BPT starting from the level of pixels is used.

2.3 Label Propagation

Once an image is pre-segmented and the BPT is created, the user can start adding scribbles as outlined above. With each scribble, an automatic process creates/updates the object segmentation mask by assigning regions coincident with scribbles to the associated object. This is achieved in [14] by propagating labels of the scribbles from leafs (pixels) towards the root of the BPT (whole image). Our approach is somewhat different, in that we handle the situation of uncertainty when assigning labels.

The procedure begins by initialization of all leafs (pixels) with one of three labels: Foreground (F), Background (B) or Unknown (U) according to the scribbles made by the user. The propagation proceeds upwards in a sequential manner between labelled leafs (pixels marked by a mouse drag) and the root. Labelled leafs are processed sequentially, i.e. for a given labelled leaf the complete path to the root is updated before processing the next leaf. Labels between a given labelled leaf and the root are updated as follows: For a given node (or initially a leaf) if its parent node is labelled as unknown (U) the label from the current node is assigned to the parent and then the procedure is repeated for the parent. If the parent node has been already marked (by propagation from another scribbled leaf/pixel) with the same label as its child then the propagation stops. If a conflict between labels of the current node and its parent is detected then the parent is marked with label Conflict (C) and the propagation toward the root continues until a parent already marked with (C) is found. This results in the creation of zones of influence (i.e. subtrees) for each label formed by nodes without conflict. The highest nodes in such zones, referred to as Root of Zone of Influence (RZI), can be identified by searching for non-conflict nodes whose parent is labelled with label (C).

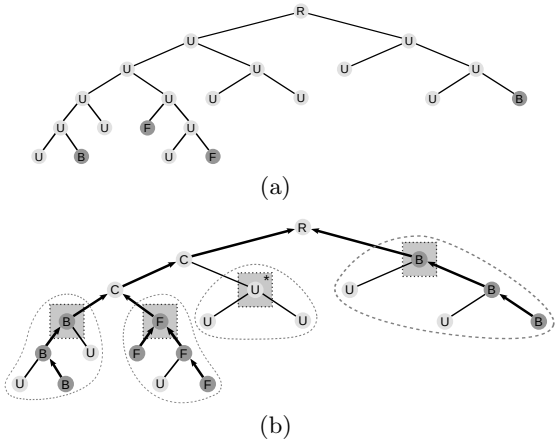

Fig. 1. Label propagation, i.e. identification of the largest possible regions without conflict: (a)-Initial BPT. Labels (F), (B) and (U) correspond to foreground, background and unknown label respectively. (b)-BPT after label propagation, thicker lines indicate the upward label propagation. Nodes surrounded by dotted squares (RZI) will be selected for the construction of the initial mask. The node marked with the star is an example of an area (zone) without a label.

2.4 Labelling the Entire Image

As a result of label propagation, a certain number of nodes may remain without labels. Thus, it is necessary to assign labels to unlabelled or conflict node so that the entire image can be labelled as foreground or background. [14] suggests a solution for filling the entire space utilizing again BPT, however that approach relies on the false assumption that there is at least one labelled descendant of the RZI's sibling which is at the same time neighbor of the RZI and this is not always the case but rather depends on scene type. For this reason, we propose an alternative approach that ensures that the entire image can always be labelled.

Unlabelled nodes (regions) are iteratively assigned to the competing objects formed by already classified regions. The distance measure used for this, $D(R_i, R_j)$, is computed as the Euclidean distance between average colours (represented in LUV space) for two adjacent regions R_i and R_j. The order of labelling is based on the confidence with which the labels can be assigned. At each iteration, only one region, with the highest confidence, is assigned to an appropriate object. The confidence C^F with which an unclassified region R_i can be assigned to the foreground object is computed using the following formula:

$$C^F(R_i) = \begin{cases} \frac{D(R_i,O^B)}{D(R_i,O^F)+D(R_i,O^B)} & \text{if } \left(\begin{pmatrix} D(R_i,O^F)\neq 0 \ \vee \ D(R_i,O^B)\neq 0 \end{pmatrix} \wedge \begin{pmatrix} D(R_i,O^F)\neq \infty \ \vee \ D(R_i,O^B)\neq \infty \end{pmatrix} \right) \\ 0.5 & \text{otherwise} \end{cases} \quad (1)$$

where O^F and O^B are simple non-parametric models of foreground and background objects comprised of RZIs with (F) and (B) labels respectively and $D(R_i, O)$ is the shortest distances between R_i and one of the regions already assigned to O adjacent to R_i. C^F has values in the range $[0, 1]$ and values above(below) 0.5 indicate that R_i should be assigned to the foreground (background) object. The confidence $C(R_i)$ with which R_i can be classified as foreground or background can be defined as $C(R_i) = \max\{C^F(R_i), 1 - C^F(R_i)\}$. At each iteration only one region R_{max} which can be classified with the highest confidence $R_{max} = \arg\max_{\forall R_i \in O^U}\{C(R_i)\}$ is assigned to an appropriate object.

3 Results

3.1 Illustrative Segmentation Results

The necessity for the proposed modifications is illustrated in Figure 2. In Figure 2(b), it is clear that the required object, the girl's face in this case, can be extracted based on label propagation only. However, this is clearly not the case when trying to segment the television set in the background where other parts of the image are incorrectly labelled – see Figure 2(c). However, this is possible if the entire image is labelled using the algorithm described above – see fig 2(d).

Illustrative results of object segmentation using a variety of test images are presented in Figure 3. These images were drawn from a variety of sources, including a collection of personal content from real users and were chosen to be challenging in nature. In each case, very accurate segmentation can be performed with a minimum of user interaction.

3.2 Real-time Considerations

The above algorithm is not only more straightforward to implement than the one proposed in [14], but also leads to a reduction in computational complexity. Since the propagation is performed only between scribbled leafs and the root, and typically the number of scribbled pixels is much smaller than the total number of pixels in the image, only a small percent of all nodes from the BPT have to be updated. For instance, in the most demanding example from Figure 3 (CIF size) only 2% percent of all nodes had to be updated. This corresponds to an effective reduction of execution time for the bottom-up propagation from 15ms to 1ms on an Intel Pentium III 600MHz when compared with our implementation of the original algorithm. Although both approaches lead to real time execution on a modern PC this may become an important advantage when computational resources are limited, e.g. on handheld devices.

4 Application to Shape Modeling

The above semi-automatic segmentation algorithm was instantiated in the form of a software tool for contour extraction - screenshots are shown in Figure 4(a) &

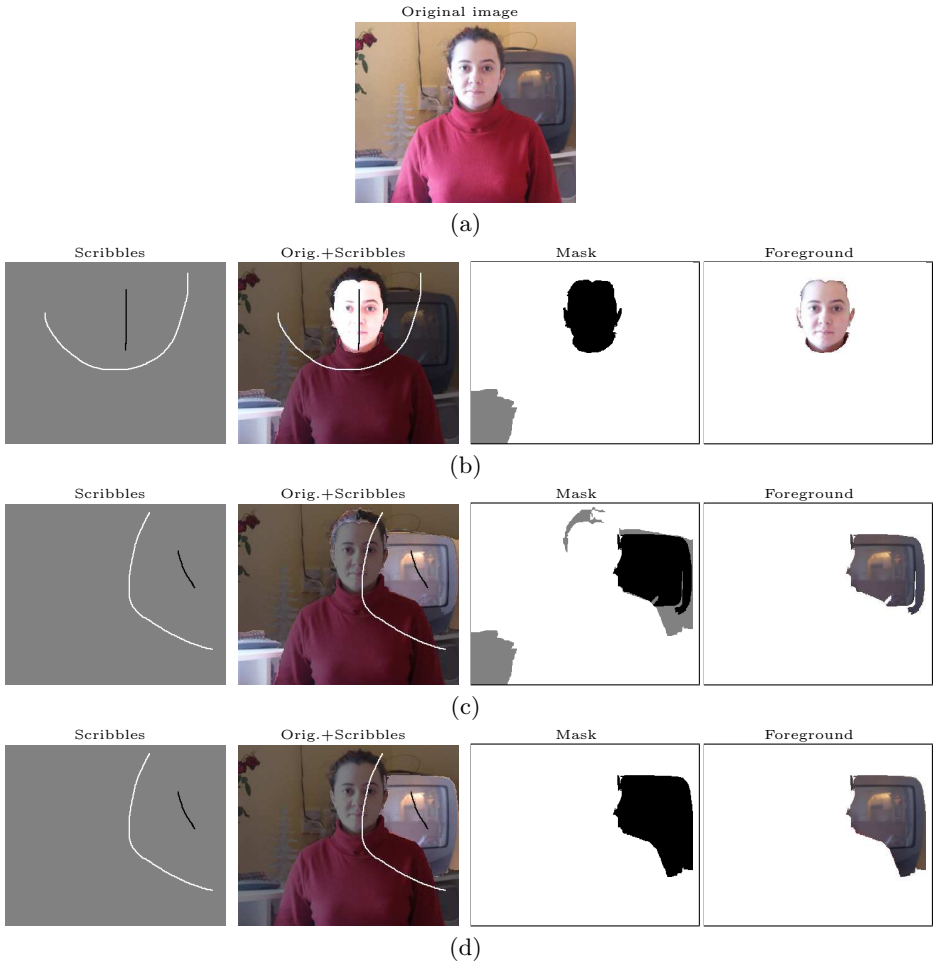

Fig. 2. Examples of semi-automatic segmentation. Labels (F), (B), and (U) are represented by black, white, and grey respectively. (a)-the original image, (b)-face extraction using only label propagation in BPT, (c)-attempt at partitioning the scene from (a) into TV-set and background using only label propagation in BPT, (d)-segmentation based on interactions from example (c) with filling of the entire image.

(b). The tool has proven to be particulary useful in the context of construction of statistical shape models. Specifically, it was used to generate training data for a "head and shoulders" shape model.

This was achieved by segmenting 19 different head and shoulder poses and using this to generate a PDM. Give the training examples each consisting of n landmarks all aligned into a common frame of reference (as detailed in [7]), training relies upon exploiting the inter-landmark correlation in order to reduce dimensionality [4]. It involves calculating the mean of the aligned examples $\bar{\mathbf{x}}$,

Fig. 3. Results of semi-automatic segmentation.

and the deviation from the mean of each aligned example $\delta\mathbf{x_i} = \mathbf{x_i} - \bar{\mathbf{x}}$, and calculating the eigensystem of the $2n \times 2n$ covariance matrix of the deviations $\sum_{\mathbf{x}} = 1/N \sum_{i=1}^{N} (\delta\mathbf{x_i})(\delta\mathbf{x_i})^{\mathbf{T}}$. The modes of variation are described by the unit eigenvectors of $\sum_{\mathbf{x}}$. Modifying one component at a time gives the *principal modes* of variation. Any shape in the training set can be approximated using the mean shape and a weighted sum of these deviations obtained from the first t modes (where t is chosen to account for a sufficiently large proportion of the total variance).

The model obtained in this manner is compact, i.e. a small number of parameters explain more than 76% of all variations, and qualitative results shown in Figure 4 indicate good specificity. Full details of this process can be found in [7].

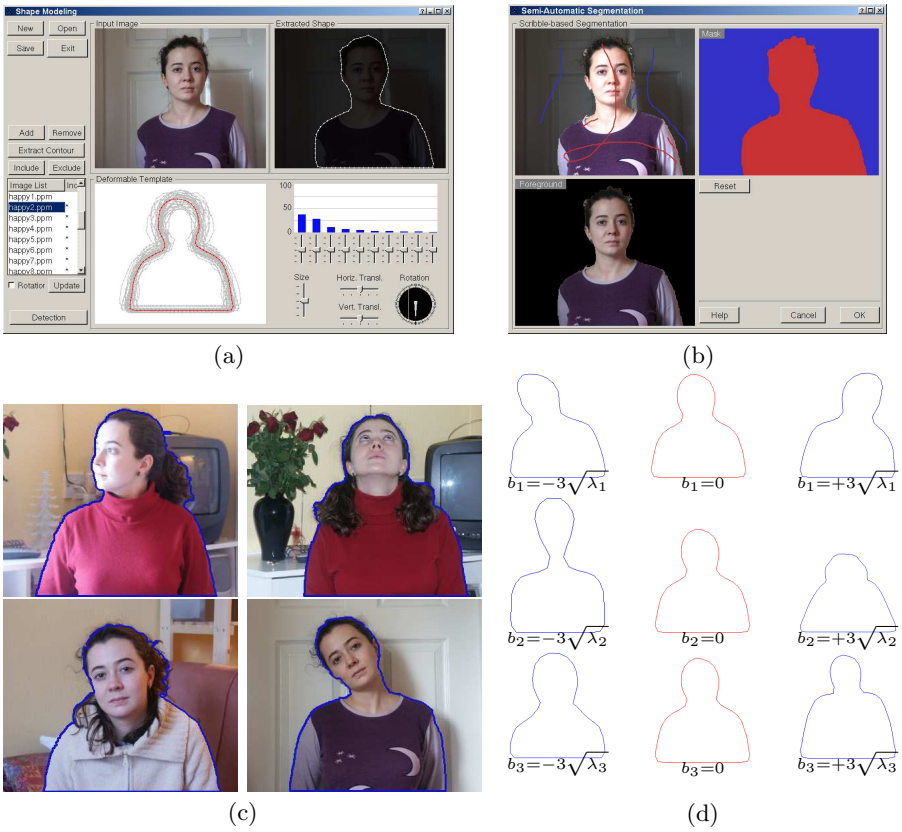

Fig. 4. Modeling of the head & shoulders: (a) & (b) Screen shots of the interactive tool used to generate training data, (c) Examples of different poses for which examples are extracted, (d) Shape model: deformation using 1st, 2nd and 3rd principal modes.

5 Conclusion and Future Work

In this paper we proposed an approach to interactive object segmentation that extends existing state of the art. Clearly, the proposed semi-automatic tool could find application in a number of contexts: query formulation in object-based information retrieval, as a front end process prior to object-based MPEG-4 encoding, or in image editing applications. In this paper, its usefulness in the context of shape modeling was discussed. Using the approach, it is straightforward to generate suitable training data that can be used to construct a deformable shape model. Future work will investigate applying this to other types of objects and using shape models to automatically segment previously unseen object instances in images. Our initial work in this direction will use the head and shoulders model to segment humans in close up images.

Acknowledgment. This material is based on works supported by Science Foundation Ireland under Grant No. 03/IN.3/I361. The research leading to this paper was supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space.

References

1. Salembier, P., Marqués, F.: Region-based representations of image and video. IEEE Trans. Circuits Syst. Video Technol. **9**(8) (1999) 1149–1167
2. Bookstein, F.L.: Landmark methods for forms without landmarks: Localizing group differences in outline shape. Medical Image Analysis **1**(3) (1997) 225–244
3. Neumann, A., Lorenz, C.: Statistical shape model based segmentation of medical images. Computerized Medical Imaging and Graphics **22**(2) (1998) 133–143
4. Cootes, T.F., Taylor, C.J., Cooper, D., Graham, J.: Active shape models - their training and application. Computer Vision and Image Understanding **61**(1) (95) 38–59
5. Marchant, J., Onyango, C.: Fitting grey level point distribution models to animals in scenes. Image and Vision Computing **13**(2) (1995) 3–12
6. Cootes, T.F., Taylor, C.J.: Active shape models - smart snakes. In: Proc. British Machine Vision Conf. (BMVC'92), Springer Verlag. (1992) 266
7. Adamek, T., O'Connor, N.: Efficient contour-based shape representation and matching. In: Proc. 5th ACM SIGMM Int'l Workshop on Multimedia Information Retrieval (MIR'03), Berkeley, CA. (2003)
8. Correia, P., Pereira, F.: Classification of video segmentation application scenarios. IEEE Trans. Circuits Syst. Video Technol., special issue on Audio and Video Analysis for Multimedia Interactive Services **14**(5) (2004) 735–741
9. Marcotegui, B., Correia, P., Marqués, F., Mech, R., Rosa, R., Wollborn, M., Zanoguera, F.: A video object generation tool allowing friendly user interaction. In: Proc. IEEE Int'l Conf. on Image Processing (ICIP'99), Kobe, Japan. (1999)
10. Adams, R., Bischof, L.: Seeded region growing. IEEE Trans. Pattern Anal. and Machine Intell. **16**(6) (1994) 641–647
11. Chalom, E., Bove, V.: Segmentation of an image seguence using multi-dimensional image attributes. In: Proc. IEEE Int'l Conf. on Image Processing (ICIP'96), Lausanne, Switzerland. Volume 2. (1996) 525–528
12. O'Connor, N.E.: Video Object Segmentation for Future Multimedia Applications. PhD, Dublin City University, School of Electronic Engineering (1998)
13. Adamek, T., O'Connor, N.E., Murphy, N.: Region-based segmentation of images using syntactic visual features. In: Proc. 6th Int'l Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'05), Montreux, Switzerland. (2005)
14. Salembier, P., Garrido, L.: Binary partition tree as an efficient representation for image processing, segmentation, ,and information retrieval. IEEE Trans. on Image Processing **9**(4) (2000) 561–576