An efficient scalable object contour tracking scheme and its application for video segmentation

Mingyou Hu, S. Worrall, A. H. Sadka, A. M. Kondoz
Centre for Communication Systems Research,
University of Surrey,
Guildford, Surrey, United Kingdom
{M.Hu, S.Worrall, A.Sadka, A.Kondoz}@surrey.ac.uk

Abstract—We present an efficient scalable object contours tracking algorithm and its application for video segmentation. It can track the contour of both rigid and nonrigid objects, even with partial occlusion. For each object, motion information is first used to predict the object contour. Next, texture information along the object contour is used to refine the predicted contour. Finally, active snake is used to snap the detected contour to the real object with high accuracy. The proposed energy terms in active snake model are normalized colour gradient and foreground / background region properties along the contour. Experimental results illustrate the performance of the proposed tracking algorithm, which is able to track the object boundaries with large motion and partial occlusion.

Keywords—object tracking, scalable contour tracking, video segmentation

I. INTRODUCTION

Video tracking has become an important part in image and video-based applications, such as video segmentation and compression (e.g. [1-2]), content-based indexing and retrieval (e.g. [3,7]), video surveillance (e.g. [4]), motion capture (e.g. [5]), and gestural human-machine interfaces (e.g. [6]). After detected and segmented from one video frame, the video objects are tracked in the subsequent frame, with semantically meaningful object shape.

Much research has been conducted on object tracking in the past decades, and many methods have been proposed, which can be roughly divided into region-based tracking [2,8-9], contour-based tracking [10-12], and feature-based tracking [13].

The tracking technique in [2] is based on a Hausdorff distance. A binary model for the video object is first obtained from the edge image. The method then matches the model to the objects in subsequent frames. The object model is updated at every frame to follow the change of object shape. This method has a severe limitation that it can not deal with complex scene, and cluttered background. In [8] active images employ a new region-based approach to nonrigid motion tracking. Shape is defined in terms of a deformable triangular mesh that capture object shape plus a color texture map that captures object appearance. Nonrigid shape registration and motion tracking are achieved by posing the problem as an energy-based, robust minimization procedure. However, this method can not cope with the object occlusion, complex motion and deformation.

Instead of tracking the whole object pixels, contour-based methods track only the contour of object. First, object contour of previous frame is projected to the current frame using motion information. Then, the predicted shape is adapted to the object in the current frame. The tracking method in [10] first estimates the parameters of a perspective motion model and then predicts the position of the contour in the next frame based on these parameters. To deal with the nonrigid body motion, the method adjusts the approximated boundary by means of a morphological watershed. This method can tracking object contour with pixel-wise accuracy. But it can not handle large nonrigid movements. Active contours (snakes) have been efficient methods to tracking both rigid and nonrigid objects [11]. One of the important features for active contour is that it can fuse both edge and texture information to improve the tracking accuracy.

Recently, particle filters have become popular tools to solving the tracking problem [14-15], which, for visual tracking context, are pioneered by Isard and Black [14]. One important advantage of particle filtering is that it allows the information from different measurement sources to be fused in a principle manner. However, within visual tracking context, how to fuse a host of cues has not been fully exploited to increase the reliability of object tracking algorithm.

In this paper, we discuss a novel scheme for object contour tracking, which is based on multiple cues and can achieve robust tracking under different conditions. Especially, it can achieve scalable contour tracking so it can adapt to different applications and complexity requirements. No prior training is required, and non-parameterized contour model is used. The proposed method features the following novelties:

1. A robust piecewise contour prediction scheme is employed, in which reliability evaluation is first conducted for the estimated motion vectors of mesh vertices and local motion model is estimated and used to predict the object contour.
2. Local maximal likelihood detection scheme is used to correct the predicted contour, which is efficient for nonrigid object movement.
3. Both local photometric (foreground/background region statistical properties) and geometric (such as edge, contour and region smoothness constraints) information is incorporated into the active contour.
model to further refine the contour. This facilitates the robust wide area applications.

II. SCHEME DESCRIPTION

In this section, our proposed contour tracking approach will be discussed in detail. This method can be considered as a hybrid scheme of feature-based, region-based and contour-based techniques. The proposed scalable contour tracking method consists of three steps, each of which can be exploited for different applications. At first, object contour is predicted using feature-based and mesh-based object tracking schemes. Then, it is refined using texture information along the contour region, in which local maximal likelihood detection scheme is conducted. Finally, active contour model is applied to track object contour with pixel-wise accuracy and alleviate the possible error detection in step 2.

A. Contour prediction

Before contour prediction, we assume that the semantic object in the first frame is available. This can be achieved using watershed segmentation method [16], taking the motion information of the first several frames into account. Human interaction is needed to achieve semantic segmentation. For example, the method in [17] can be employed to minimize the human interaction and achieve semantic object segmentation.

For the object in the previous frame, robust feature points, based on the criteria of [13], are selected on the object and the object contour is approximated through a series of contour points. A mesh model is then constructed using the allocated feature points.

After object mesh construction, both forward and backward motion vectors of control points between frame $i(x, t-1)$ and frame $i(x, t)$ are estimated using Shi-Tomasi feature tracking algorithm in [13]. That is, the forward motion vector of the $i$th node location $V_i$ in frame $t-1$, moves to location $V_i'$ in frame $t$. Then the backward motion vector at the location $V_i$ in frame $t$ maps back to $V_i'$ in frame $t-1$.

After both forward and backward motion estimation, their motion "reliability" is estimated based on both forward and background motion vectors. The "reliability" is evaluated by the following formula:

$$Re = \exp \left( -\frac{\|V_i - V_i'\|^2}{2\sigma_m^2} \right)$$  \hspace{1cm} (1)

where $\sigma_m$ is the free parameter. From (1), it shows that the smaller the difference between $V_i$ and $V_i'$, the more reliable the motion vector of $i$th node. For the nodes whose reliability is smaller than a threshold (0.3 is chosen in our experiments), they are not considered during contour prediction.

Then, mesh-based motion estimation scheme [9] is applied to refine the motion vectors of the control points. The positions of the nodes with higher "reliability" are initialised by the feature-based motion estimation results. The initial positions of the nodes with smaller "reliability" are estimated from their neighbours. The mesh-based motion estimation scheme can keep the mesh structure during MV's refinement.

After estimating the motion vectors of interior points, the points along the contour can be predicted from their $m$ nearest motion vectors. The weighted least squares estimation algorithm (WLS) is used to determine the affine motion that best describes the motion of the contour segment. Each motion vector is weighed according to its "reliability".

For rigid moving object, the predicted object contour is with high accuracy. However, for nonrigid moving object, for example, human head profile contour appearance, further refinement is necessary.

B. Contour refinement

In this step, we assume that the statistical characteristic of pixels of frame $t$ is similar to its adjacent neighbors in frame $t-1$. This step is efficient to tracking the object with large nonrigid motion such as human head in-plane rotation.

During contour refinement, local maximal likelihood detection is used to refine the predicted contour in frame $t$. Only the pixels in the band around the object contour are employed to estimate the statistical properties of foreground and background, in which non-parametric kernel density estimation is employed. For the local statistical properties of contour points, the pixels are constrained by rectangle and the defined band, as shown in Fig.1 (a) for point $P$.

![Fig. 1 Illustration for contour refinement process; (a) estimating local statistics of foreground and background around $P$ in previous frame; (b) refining contour segment around the estimated $P'$ in current frame.](image)

The estimated local statistical properties of foreground and background around point $P$ are used to refine the corresponding contour segments around the predicted contour point $P'$ in frame $t$, using maximal likelihood (ML) detection and Bayesian relaxation. First, a band is generated around the predicted contour in current frame $t$. It means that the object contour can only be searched in the band. Then, based on the local statistical probability of both foreground and background of $P$, the pixels around $P'$, which should also be located in the circle as shown in Fig.1 (b), are classified as foreground and background. The radius of circle is decided by the distance between $P'$ and its neighbour contour points.

As the criterion of this step is that the statistical characteristic of pixels of frame $t$ is similar to its adjacent neighbours in frame $t-1$, it is efficient to track the object...
with large nonrigid motion such as human head in-plane rotation. However, as the contour smoothness constraint is not employed during refinement, some contour parts are zigzag-patterned and visually uncomfortable, especially for the position where statistical probabilities of foreground and background are similar. Therefore, further refinement is needed for some sequences.

C. Further refinement using active snake

After contour prediction and refinement, a precise object contour can be achieved if there is high contrast between the tracked object and background along the contour. If the contrast is low, the generated contour is zigzag and some errors of contour may happen. In the third step, active contour model is used to refine the contour further. The energies of active snake model are defined as follows:

Suppose that \( N \) discrete points are selected, which have similar foreground and background surroundings, along the given contour \( C \). The discrete snake energy \( E_{\text{snake}} \) can be expressed as follows:

\[
E_{\text{snake}} = \sum_{i=1}^{N} E_{i} = \sum_{i=1}^{N} (E_{\text{min}} + E_{\text{err}} + E_{\text{reg}})
\]

where each \( E_{i} \) depends on the contour segment between up to three points \( v_{i-1}, v_{i}, \) and \( v_{i+1} \).

In our research, the dynamic programming algorithm in [18] has been used to search for the maximum of (2). In algorithm implementation, for node \( i \), the search locations are restricted along the bisector lines of the angle \( \angle v_{i-1}, v_{i}, v_{i+1} \).

As the points are placed at regular intervals along the contour, the internal energy in node \( i \) is selected as:

\[
E_{\text{int}} = 2 - 2 \cos \angle v_{i-1}, v_{i}, v_{i+1}
\]

(3)

In our research, an edge map is selected as one part of the external energy \( E_{\text{ext}} \). Canny edge detector has been used to obtain the edge map. Low threshold has been used during edge detecting to detect the weak edge. Short edges are removed in order to reduce the effect of noise. A binary edge map \( BinaryEdgeMap \) is generated, which are then smoothed by Gaussian filter with variance \( \sigma_{\gamma} \). The average gradient along the contour segment is used as external energy \( E_{\text{edge}} \).

\[
E_{\text{edge}} = \frac{BinaryEdgeMap_{\gamma}(v_{i}, v_{i+1}) + BinaryEdgeMap_{\gamma}(v_{i}, v_{i-1})}{l(v_{i}, v_{i-1}) + l(v_{i}, v_{i+1})}
\]

(4)

In the proposed algorithm, regions force is employed, in which mutual information between image intensity and its label is maximized.

\[
E_{\text{region}} = H(l(x)) - P_{x}B P_{x}F + H(l(x)) L(x) = F
\]

(5)

\[
- P_{x}B P_{x}F + H(l(x) L(x) = B)
\]

where \( B \) and \( F \) represent the background and foreground respectively within the support \( S \), as shown in Fig. 2. \( E_{\text{region}} \) will be maximized if and only if the labelling \( L(\) \) gives the correct segmentation [19], as in Fig 2 (c).

Fig 2. Texture information for different situations

After achieving the object contour of frame \( i \), object mesh structure and contour approximation are updated in order to track object contour along the whole video sequence.

III. EXPERIMENTAL RESULTS AND ANALYSIS

Several sequences have been used to test the performance of our proposed methods. Due to the paper space, we just demonstrate the experimental results of Moir_dhtri and Claire Sequences, as shown in Fig. 3 and Fig. 4.

For Moir_dhtri sequence, 700 frames are segmented without human interaction except the first frame. Fig. 3 (a) shows the predicted contours of several frames using motion information. For most sequences, the performance is acceptable. Large prediction error may occur in in-plane rotation happens, such as the 4th image in Fig. 3 (a). Fig. 3 (b) illustrates the refined object contours. The object contour can be detected correctly. Fig. 4 shows the segmentation results of Claire. The proposed algorithm can segment 500 frames without human interaction after the first frame. The experimental results show that this contour tracker is robust for tracking nonrigid motion, even with partial occlusion, as shown in Fig. 3 and Fig. 4.

One reason for the success of this algorithm is the use of band constraint during the refinement process. There are several advantages of using the band around the contour compared to using the complete region during probability estimation and pixel classification: First, it allows object tracking by adapting to the local changes around the object contour; Next, the contour search space is reduced; Furthermore, the effects of noise and artefact from both foreground and background are reduced.

Our algorithm can be considered as another kind of implementation of partial filtering. Steps A and B achieve the effective prior result of current frame, which is a prediction taken from the posterior result of previous frame. Step C can be considered as the Measure process in particle filtering, in which the observation information of current frame is used to refine the contour.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, a contour tracking scheme has been proposed which can achieve scalable object tracking with different computation complexity and tracking accuracy. The experimental results show that this contour tracker is robust for tracking the object contour with nonrigid and large motion, even with partial occlusion. Further research will be
conducted to achieve the multiple object contour tracking with view to compare our proposed method with particle filtering method.

REFERENCES


(a)

(b)

Fig. 3 Video segmentation results for Motr_dhfr sequence

Fig. 4 Video segmentation results for Claire sequence

158