Annotation Supported Contour Based Object Tracking With Frame Based Error Analysis

Amarjot Singh, Devinder Kumar, Akash Choubey, and Ketan Bacchuwar Srikrishna Karanam

***Abstract*—Motion tracking is a vital component of study for a video sequence having wide applications in object tracking, coding and editing the videos and mosaicking [2]. Getting the actual motion for the videos existing in the real world though is a difficult task but plays a pivotal role in designing a model for any algorithm's evaluation. We used an interactive computer vision system [1] which provides annotation tool for labeling and tracking the contours. Through the mutual work of user interaction and the computer vision system, the input effort is greatly reduced, simultaneously increasing the dependability of the whole system as compared to the solely computer based system. The ability of humans to easily segment and detect difference between different frames has been utilized using the human in loop methodology [1] by making use of a simple camera. The paper experiments with the capabilities of the system applied to indoor video sequence. This is the first paper which evaluates the capabilities of image annotation supported contour based object tracking with error analysis and correction, explaining the significance of human in the error incurred by the methodology. The paper studies the error incurred by the system with movement from one frame to another, supported by detailed simulations. The paper also focuses on the reasons responsible for the error incurred by the system mainly involving human intervention. Finally the paper presents the correction of the error followed by the in depth simulation indicating the in capabilities of the system on deforming objects. This system can be effectively used to analyze the error in motion tracking and further correcting the error leading to flawless tracking.**

***Index Terms*—Contour; Tracking; Error; Annotation; Optical flow component.**

I. Introduction

Motion tracking plays an important role in the analysis of a video sequence. Motion tracking has a wide range of applications in different fields, e.g., object tracking, object recognition, advance coding, editing the videos including mosaicking [2]. Over the years motion tracking is being applied widely in the fields of biomechanics [3], avionics [4], sport analysis [5], medical [6] etc. existing in the real world though is a difficult task but plays a pivotal role in designing a model for any algorithm's evaluation. We used an interactive computer vision system [1] which provides annotation tool for labeling and tracking the contours.

Through the mutual work of user interaction and the computer vision system, the input effort is greatly reduced, simultaneously increasing the dependability of the whole system as compared to the solely computer based system. The ability of humans to easily segment and detect difference between different frames has been utilized using the human in loop methodology [1] by making use of a simple camera. The paper experiments with the capabilities of the system applied to indoor video sequence. This is the first paper which evaluates the capabilities of image annotation supported contour based object tracking with error analysis and correction, explaining the significance of human in the error incurred by the methodology. The paper studies the error incurred by the system with movement from one frame to another, supported by detailed simulations. The paper also focuses on the reasons responsible for the error incurred by the system mainly involving human intervention. Finally the paper presents the correction of the error followed by the in depth simulation indicating the in capabilities of the system on deforming objects. This system can be effectively used to analyze the error in motion tracking and further correcting the error leading to flawless tracking.

In the past, many automatic contour tracking algorithms have been developed by the computer vision researchers but these automated contour tracking systems had poor efficiency and outputs were not near the ground motion of the object. The two main problems faced in the motion segmentation are to find the group of pixels moving together in two or more frames while second problem is to recover the motion field associated with each group [2]. For designing better model, it is necessary to evaluate different and advanced motion algorithms.

This paper focuses on using the annotation tool provided in [1] to label and track different objects in different video sequences, thus escaping the tedious work of labeling the object pixel by pixel. The effort in labeling and tracking the object is greatly decreased by allowing the user to make as well as label the contour of object in any one frame followed by the automatic tracking of the contour in other frames. The human interaction plays a pivotal role in labeling the objects as the user can correct the label in different frames thus removing the error produced by the computer vision system, hence increasing the efficiency of the system. Due to this collaborated work of user and computer vision system the results obtained are very near to the real motion of different objects in the analyzed video sequences.

The following paper has been divided into five sections. The next section explains the previous systems used for automatic contour tracking of objects from video sequences. The third section elaborates the algorithm implemented by the paper for tracking the contour of the object. Section four explains the simulation results while the final section five, discusses the summary of the paper.

Manuscript received June7, 2012; revised July 30, 2012.

The authors are with National Institute of Technology, Warangal, India (amarjotsingh@ieee.org).

II. Related Work

There is large amount of study done on automatic contour tracking, witnessing the growing power of modern tracking systems. Tracking an object is mainly divided into three parts- annotation, segmentation and tracking. Image annotation has been minutely examined in the literature such as Berkeley image segmentation [16] and Label Me object annotation database. Contour tracking lays the foundation of image segmentation, which has its own importance. In computer graphics recently, many interactive techniques were proposed for segmenting the object such as rotoscoping [17], soft alpha matting, over segmentation and 3D cut etc. Early approaches to image segmentation were based upon the estimation of dense flow fields. Researchers from then have developed many effective computational methods for optical flow estimation. In the past few years, the proposed methods evaluated the optical flow fields for the entire sequence which were entirely computer based calculations. This paper employs a system which focuses on image annotation and tracking with human assistance which cuts down the load on the system and improves the efficiency. The system represents a smooth optical flow field for contour tracking compensating the discontinuities due to occlusion and object boundaries.

III. Human Based Annotation

Labeling each pixel exactly in each frame of a video sequence is a tedious task. The system used in the paper makes use of human assisted layer segmentation, and automatic estimation of optical flow for object contour tracking. In order to increase the robustness of the system, the objective functions of flow estimation and flow interpolation are modeled on L1 form. Many techniques such as iterative reweighted least square (IRLS) [12, 13] and pyramid based coarse-to-fine search [11,13] were used at large for the optimization of these non linear object functions. The algorithm is explained in detail in section 3.1 and 3.2 below.

A. Human Assisted Layer Segmentation

This module works on the basis of human interaction with the labeling and defining of contour is done by the user in one frame while the rest of forward and backward tracking is done automatically by the system . The correction of contour can anytime be done by the user in any frame which is then automatically passed to the other frames. Particle filter is used to track the object in the system as real time performance is more important than accuracy.

Suppose a function is defined for polygon by the user using landmarks $M = \{a_p : a_p \in R_2\}_{p=1}^{n}$ at frame F_1. The motion vector v_p is found for each landmark at frame F_2. Depending upon whether the tracking is back or forth the frame F_2 can be after or before F_1. Instinctively, we want the movement of contour to be persistent and should match with the image features. The objective function is defined as:

$$B(\{v_p\}) = \sum_{p=1}^{T} \sum_{c \in T_p} m_p(c) \mid F_2((a_p + v_p + c) - F_1(a_p + c)| + \omega \sum_{p=1}^{T} s_p |v_p - v_{p+1}| \quad (1)$$

And $v_{T+1} = v_T$. In the equation, the length between the points a_p and a_{p+1} is calculated by using the weight s_p; we define $s_p = \frac{\bar{l}}{l_p + \bar{l}}$ where $l_p = \| a_p - a_{p+1} \|$ and l is the average of l_p. It's evident from these equations that closer the points, more the probability that the points move together. Variable T_p is a square neighbourhood at a_p, and m_p is the region of support for a_p, a binary mask indicates the presence of each neighbouring pixel c inside the pixel, modulated by a two dimensional Gaussian function. In Eqn. (1) the objective function mentioned is nonlinear. Taylor expansion [8] is used to linearize the data term followed by the optimization of objective function performed through coarse-to-fine search [11, 13] and IRLS [13, 12]. To account for the changes in the lighting condition, the images in F_1 And F_2 contain the first and second order derivative of luminance instead of just RGB channels. The rigidity of the object is controlled by the coefficient ω : deformation can be done with less effort for smaller values of ω. The user can set the value of ω before tracking.

The contour tracker worked fine for most all the cases, but it can go wrong and drift from the position especially when the object rotates. To overcome this drawback the system allows the correction of a landmark to be made at any frame and it (the change) is transferred to the other frames. In the temporal propagation [1], to reconstruct the point modified by the user, the linear regression coefficient for the other points is estimated. The algorithm proposed works astonishingly well. In comparison to the complicated contour tracking/modification algorithm proposed in [17] is too expensive regarding the real-time long distance propagation.

B. Layer by Layer Optical Flow Estimation

The mask showing the visibility of each layer is the main difference between layer by layer optical flow estimation and traditional flow estimation for the whole frame. The pixels lying inside the mask are only used for matching.

For baseline line model for optical flow estimation we used the optical flow algorithm in [13, 12], while to better the accuracy symmetric flow computation is also included. Let E_1 and E_2 be the visible mask of a layer at frame F_1 and F_2, (g_1, h_1) be the flow field from F_1 to F_2, and (g_2, h_2) the flow field from F_2 to F_1. Following terms constitute the objective function for approximating the optical flow layer by layer. In the first step, the matching of images with the visible data term is formulated as mentioned in eqn below:

$$B_{data}^{(1)} = \int u * E_1(x, y) \mid F_1(x + g_1, y + h_1) - F_2(x, y) \mid \quad (2)$$

where u is the Gaussian filter. The data term $B^{(2)}_{data}$ for (g_2, h_2) is similarly defined. To account for outliers in matching, L1 norm is used. In the second step , smoothness is imposed by

$$B^{(1)}_{smooth} = \int (|\nabla g_1|^2 + |\nabla h_1|^2)^{\gamma} \tag{3}$$

where γ varies between 0.5 and 1. Finally, symmetric matching can be achieved by:

$$B^{(1)}_{sym} = \int \begin{matrix} |\, g_1(x+y) + g_2(x+g_1, y+h_1)\,| + \\ |\, h_1(x+y) + h_2(x+g_1, y+h_1)\,| \end{matrix} \tag{4}$$

The sum of the above three equation gives the objective function described below:

$$B(g_1, h_1, g_2, h_2) = \sum_{j=1}^{2} B^{(j)}_{data} + \sigma B^{(j)}_{smooth} + \delta B^{(j)}_{sym} \tag{5}$$

The IRLS proposed in [13,12] is used as equivalent to outer and inner fixed-point, together with the coarse-to-fine search[11,13] and image wrapping for the optimization of this objective function. After computing the flow at each level of pyramid, the visible layer mask E_1 is approximated on the basis of estimated flow:

$$\text{If} \quad B_2(x+g_1, y+h_1) = 0 ,$$

$$\text{then set} \quad B_1(x, y) = 0 ;$$

If in the Eqn. (4), the symmetry term is beyond the threshold at (x, y), then set $E_1(x, y) = 0$. Same rule can be used to update E_2. The user is allowed to change the values of σ, δ and γ in Eqn. (5). The user usually don't know which value of parameters will produce the best flow field because of this the system lets the user to decide different parameters for each layer, and dense flow field is calculated by the system in each frame. For each layer different elasticizes can be handled. Spontaneously it is observed that a larger σ, δ or γ make the flow fields smoother.

Though the estimation of optical flow layer by layer, works well in most of the cases, but some failures are observed where no reasonable flow is produced by parameter setting. The system might collapse in real time videos due to the specularity, shadow noise and blurriness in or boundary preservation assumption. The visible layer mask may distort or can be irregularly shaped due to occlusion marking the global motion difficult to capture.

IV. Experimentation

The results obtained from the experimentation applied to the indoor video enable us to investigate the capabilities of our system. The simulations were carried out on windows 7 running on an Intel i3 2.26 GHz processor machine. The aim of the experimentation is to track the motion of the specified object using contour. The primary step for the tracking is to locate the object in the video, which is being carried out with the help of image annotation toolbox. The contour formed from the annotated image is further tracked in forward as well backward frames. The section further computes the error due to contour deformation i.e. the error due to variation in the contour with respect to the reference contour from present ground truth frame to succeeding frame. The section also explains the disadvantages of the system along with the conditions where the system is ineffective in tracking the body.

The results obtained from image annotation toolbox applied to indoor video sequence applied in which, a mouse is moved over a table as shown in Fig. 1 (a). The test is conducted on an indoor video sequence with frame by frame input given to image annotation toolbox. The annotation performed is assisted by human reducing the complexity of the whole system as shown in Fig. 1 (a). The human constructs a contour on the mouse in the reference frame which further can be used to track the object in the succeeding frames.

The contour generated is further used to track the body in both forward and backward frames. The video used is divided into 7 subsequent labeled frames which are used to track the motion of the object. The contour of the object is tracked from present frame to succeeding frame using the layer by layer optical flow estimation as shown in Fig. 1(b). The automatic tracking for all 7 frames takes less than 2 seconds to compute. Error generation in the contour of the body, as we move from one frame to another is an interesting topic to study. The in detail study regarding the error incurred by the contours is discussed in the next paragraph.

The error analysis for the video under test is carried out for every frame with respect to the ground truth. Error is defined as the total number of extra pixels classified or unclassified in the contour of the succeeding frame over the total number of pixels in the reference contour in the first frame. Error is computed in two scenarios (i) the contour tracks the object including shadow (ii) the contour tracks the object without its shadow. The error computed for tracking the body with shadow is shown in Fig. 2 while the error for tracking without shadow is shown in Fig. 3 respectively. The error for tracking the object with shadow varies from a minimum value of 1.31(%) to a maximum of 17.17(%) while the variations in error without shadow varies from a minimum of 0.36(%) to 4.47(%). The ground truth pixels for the reference frame is 3353 cross which all the error for each frame have to be evaluated. Detail of the total number of pixels in the succeeding frames with the total number of unclassified pixels with respect to the ground truth frame for both the cases are also shown in Fig. 2 and Fig. 3. It is observed that the error for tracking with shadow is much more than the error incurred by the system for tracking without shadow as shown in Fig. 4 and Fig. 5 respectively. The above statement is justified as there will be more variations in the contour with shadow as compared to the contour without shadow as the shadow is more prone to shape deformations. The error analysis can be utilized for correction of error in one frame which can be passed to another frame improving the efficiency of tracking system. The correction in the error is performed as show in Fig. 1 (c). This methodology has great

applications where accuracy is of at most importance. The system can be effectively used to track the motion of the object with error correction. It is a unique system, one of its kind, which can very effectively used to track the bodies in video streams. The limitation of this system is that it can track only rigid bodies. Tracking bodies with changing shape can be a tedious task for this system. Fig. 6 shows a video sequence an object with changing to be from frame to frame. It is observed that the system is incapable of tracking the changing shape of the human. As shown in Fig. 6 the system is able to track the body only for 2 frames and in the future frames the system fails. The system is also incapable in tracking rotating objects.

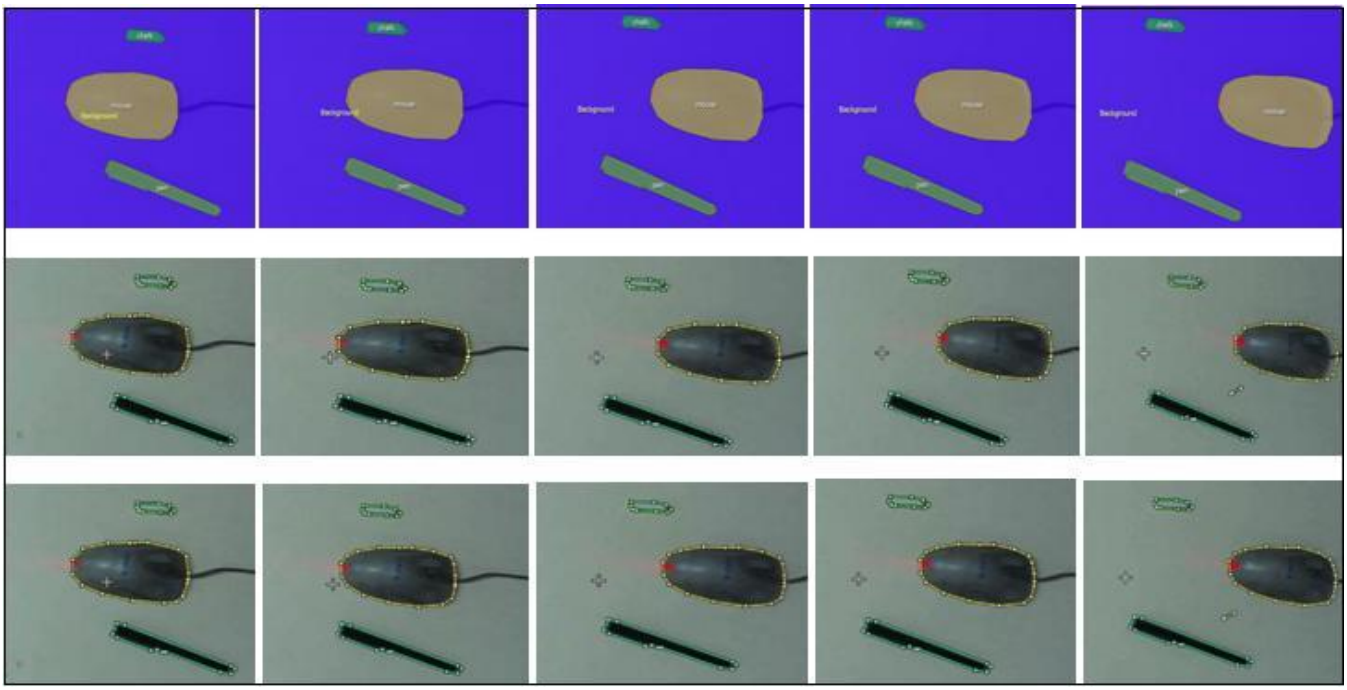

Fig. 1. (a) Annotated sequence obtained from annotation toolbox (b) Contour tracking with error (c) Contour tracking with corrected error

Frame	Number of pixels	Error in pixels w.r.t. ground frame	Error in percentage w.r.t. ground truth frame
2	3437	44	1.31
3	3465	72	2.12
4	3518	125	3.68
5	3662	269	7.92
6	3789	396	11.68
7	3976	583	17.17

Fig. 2. Error incurred during contour based tracking including shadow of each frame with respect to ground truth frame in percentage and pixels

Frame	Number of pixels	Error in pixels w.r.t to ground truth frame	Error in percentage w.r.t. to ground truth frame
2	3365	12	0.36
3	3366	13	0.39
4	3422	69	2.06
5	3427	74	2.21
6	3428	75	2.24
7	3660	307	4.47

Fig. 3 Error incurred during contour based tracking without including shadow of each frame with respect to ground truth frame in percentage and pixels

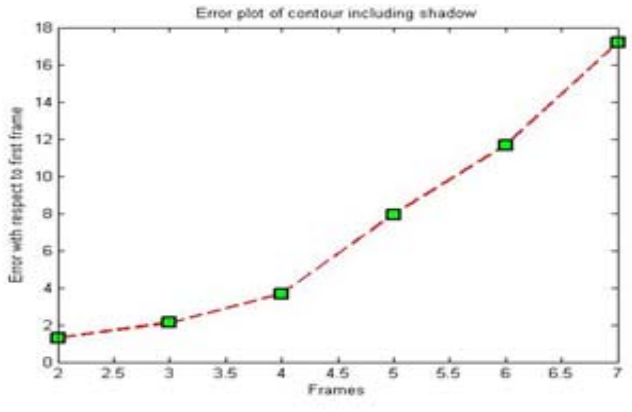

Fig. 4. Error analysis of Contour including shadow

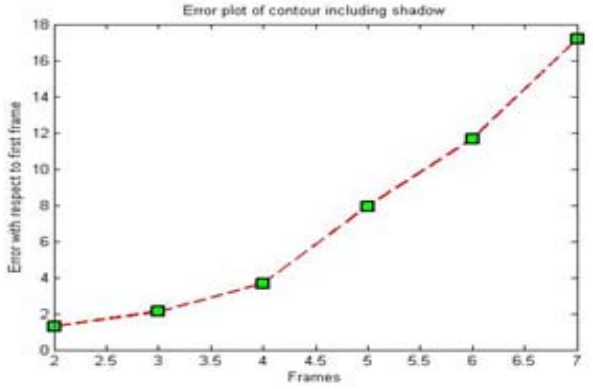

Fig. 5. Error analysis of Contour without including shadow

Fig. 6. Figure showing inability of contour tracking system on object deformation

V. Conclusion

The Section elaborates and explains the capabilities of our system applied to indoor video sequence. The system is developed using several modern motion detection algorithms that allow the annotation of every pixel and frame. The system is applied to annotate and track object from video sequences. The section further analyzes the error incurred by the system due to contour deformations with and without shadow. Finally the section lists the limitations of the system.

The system is applied to track object for indoor video sequences with and without shadow of the body. It is clearly evident for Fig. 1(b), the deformations on the contour in case of tracking with shadow are much more as compared to the normal contour tracking. The above statement is supported by the fact that the error incurred by the system in case of tracking with shadow is way high as compared to normal contour tracking case as shown in Fig. 2 and Fig. 3. The error incurred by the system can be adjusted from the frame to frame and can be passed to future frames resulting into error free motion tracking. The error variations have been plotted which gives a clear idea of the stability of the motion with respect to frames. The error plots justify the stability of tracking the object without shadow is much more than tracking the object with shadow. It justifies that closer the contour surface to the object lesser is the error incurred in the object tracking. This justifies that labeling performed by the human in the beginning is the deciding factor for the accuracy for the object tracking. But as far as the tracking in normal contour based tracking is concerned, the system should be applied only to objects which don't change their shape during the whole video sequence as the system is incapable of tracking these scenarios. The use of the system should also be avoided on rotating bodies. Overall, the system can be efficiently used to track the objects in normal conditions. The system has vast application in areas where flawless tracking is of great importance.

References

[1] Ce Liu, William T. Freeman, Edward H. Adelson, and Yair Weiss "Human-Assisted Motion Annotation" Computer Vision and Pattern Recognition, *CVPR*, 2008.

[2] Wills, J., Agarwal, S., Belongie, S, "what went where" Computer Vision and Pattern Recognition, *CVPR*, pp(s): I-37 - I-44 vol.1, 2003.

[3] J. T. long, N. Jannetto, S. Bakker, S. Smith, and G. F. Harris"Biomedical of cranial dynamics during daily living artivities" 26th international conference IEEE EMBS, Sept 2004.

[4] George M. Siouris, Guanrong Chen, Jianrong Weng "Tracking of Incoming ballistic missile using an extended Interval kalman filter" *Trans on Aerospace Electronic System*, vol.33 Nno.1 Jan 1997.

[5] Alberto Tomita, Tomio Echigo, Masato Kurokawa, Hisahi Miyamori, and Shun-ichi Iisaku "A visual tracking system for sports video annotation in unconstrained environments" Proc. 2000 International conference on image processing.

[6] Ting Chen, Member, Xiaoxu Wang, Sohae Chung, Dimitris Metaxas, and Leon Axel "Automated 3D Motion Tracking Using Gabor Filter Bank, Robust Point Matching,and Deformable Models" *IEEE Trans. on Medical Imaging*, vol. 29, Jan 2010.

[7] B K P Horn and B G schunck, "Determing optical flow", *Artificial Intelligence*, 17:185–203, 1981.

[8] B. Lucas and T. Kanade., "An iterative image registration technique with an application to stereo vision", in Proc. of the Intl Joint Conf. on Artificial Intelligence, pp: 674–679, 1981.

[9] C. Liu, W. T. Freeman, and E.H. Adelson Analysis of contour motions. In NIPS, 2006.

[10] J. Y. A. Wang and E. H. Adelson, "Representing moving images with layers", IEEE *Trans. Image processing*, 3(5):625- 638, 1994.

[11] M. J. Black and P. Anandan. The robust estimation of multiple motions:parametric and piecewise-smooth flow fields *Computer Vision and Image Understanding*, 63(1):75-104, January 1996.

[12] T. Brox, A. Bruhn, N. Papenberg and J. Weickert. High accuracy optical flow estimation based on a theoryfor wrapping In Proc. ECCV, pp 25–36, 2004.

[13] A.Bruhn, J Weickert, and C. Schnorr. Lucas/Kanade meets Horn/schunk: combining local and global optical flow methods *IJCV*, 61(3):211–231, 2005.

[14] Edward H. Adelson "Layered representations for image coding" Vision and Modelling technical Report 181, MIT Media Laboratory,1991

[15] Jianbo Shi, Tomasi C. "Good features to track" CVPR 1994

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

[17] A Agarwala, A. Hertzmann, D.H. Salesin and S. M. Seitz, "Keyframe-based tracking for rotoscoping and animation", in *AACM SIGGRAPH*, 23(3):584–591, 2004