I.J. Image, Graphics and Signal Processing, **2012, 12, 43-49**
Published Online November 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2012.12.06

Occluded Human Tracking and Identification Using Image Annotation

Devinder Kumar
Department of Electrical Engineering, NIT Warangal
Warangal, India
E-mail: devinderkumar@ieee.org

Amarjot Singh
Department of Electrical Engineering, NIT Warangal
Warangal, India
E-mail: amarjotsingh@ieee.org

***Abstract*—** The important task of human tracking can be difficult to implement in real world environment as the videos can involve complex scenes, severe occlusion and even moving background. Tracking individual objects in a cluttered scene is an important aspect of surveillance. In addition, the systems should also avoid misclassification which can lead to inaccurate tracking. This paper makes use of an efficient image annotation for human tracking. According to the literature survey, this is the first paper which proposes the application of the image annotation algorithm towards human tracking. The method divides the video scene into multiple layers assigning each layer to the individual object of interest. Since each layer has been assigned to a specific object in the video sequence: (i) we can track and analyse the movement of each object individually (ii) The method is able to reframe from misclassification as each object has been assigned a respective layer. The error incurred by the system with movement from one frame to another is presented with detailed simulations and is compared with the conventional Horn–Schunck alone.

***Index Terms*—** Image Annotation, Human Tracking, Optical flow, Motion Tracking

I. Introduction

Human tracking has become an area of research due to its wide applications in multiple fields like advance security system, video surveillance systems etc. It also has applications in surveillance in ATMs, stores, banks, different surveillance [1], [2] and security system [3], [4]. Apart from surveillance applications, it is also used to analyze human movements, an active research field aimed at designing products for the biomedical industry.

A number of methods have been proposed in the past to detect and track human in different environments [5], [6], [7]. For example, In [5], the author proposed a coupling statistical modeling technique capable of tracking human motion involving a preprocessing stage of image segmentation and point feature tracking while in [6] the authors used Bayesian framework for detection and tracking moving bodies. In addition, an automatic detection and tracking method for Human Motion with a View-Based Representation was proposed in [7]. In the past multiple surveillance based systems have also been developed to track humans. For example, system developed by Ismail Haritaoglu et al [8] is a real time visual surveillance system widely used for detecting and tracking multiple human, further monitoring their activities. The system creates models based on shape analysis which are further used to track human. A color histogram [9] based tracking algorithm was also proposed to track human using the coherence of body's color histogram.

The methods mentioned above have various disadvantages like high computational cost requirements, overlooking self-occlusion leading to mismatch between similar objects etc. In addition, it is extremely difficult to analyze the motion of individual targets in the video sequence with previous methods.

The paper makes use of an enhanced image annotation algorithm used in the past for object tracking [10]. The paper focuses on using the annotation tool provided in [10] to label and track different humans in real time video sequences. The method has the following main points: (i) each individual in the video sequence is assigned a specific layer formed by their individual contour. (ii) The method will reframe from misclassification as each individual has been assigned his own respective layer. Once the specific layer is assigned to the individual based on depth, tracking and analyzing the movement of individuals is done performed using particle filter applied to single layer at anytime, hence limiting the interference from other individuals.

Fig 1. Annotated sequence on human video obtained from annotation toolbox

The rest of paper is further divided into following sections. Section II explains the algorithm used in the paper for human tracking. Section III elaborates the results obtained from the simulations performed in the paper while section IV summarizes the paper.

II. Human Based Annotation

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 lagrange's L1 form [10]. Many techniques such as pyramid based coarse-to-fine search [11] and iterative reweighted least square (IRLS) [12], [13] were used at large for the optimization of these non linear object functions.

A. Human assisted layer segmentation

This method works on the basis of human interaction with the contour labeling. The first step is the Initialization of contour in one frame. Errors can occur in the contour formed by the user due to background cluttering or other changes like shadow etc in the frame. The errors in contour can be corrected manually by the user in any frame which is further automatically passed to the other frames. The forward and backward tracking of the target is simulated automatically by the system. Particle filter is used to track the object in the system as real time performance is considered more important than accuracy [14]. In addition, Occlusion handling technique has also been included in the contour tracker itself.

Suppose a function is defined using landmarks points as $M = \{a_p : a_p \in R^2\}_{p=1}^{n}$ at frame F_1. The motion vector v_p represents 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. In order for the movement to be persistent, we use optimization. The objective function is defined as:

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

$v_{T+1} = v_T$, where v is the motion vector. In the equation, the length between the contour 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 $\bar{l}$ is the average of l_p. It's evident from these equations that closer the points in the contour formation, more the probability that the points

move together. Variable T_p is a square neighborhood at a_p, while m_p is the region of support for a_p, a binary mask which indicates the presence of each neighboring pixel c inside the pixel, modulated by a two dimensional Gaussian function. In Eqn. (1) the objective function mentioned is nonlinear, hence taylor expansion is used to linearize the data term followed by the optimization of objective function performed through iterative reweighted least square (IRLS) and pyramid based coarse-to-fine search [11]. In order 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 .The user can set the value of ω before tracking.

For handling occlusion, the user is allowed to specify relative depth and the depth is automatically interpolated (as time function) for the rest of the frames. The contour tracker is driven by a 2nd-order dynamical model for prediction. The prediction is used as an initialization for optimizing (1). The tracking algorithm iterates between the following two steps to handle occlusion:

(1)Check whether each landmark z_p is occluded by other layers with smaller depth values. If occlusion is detected for z_p then set $r_p(c) = 0$, $\forall c \in N_p$ in (1). This means there is no region to support tracking z_p.

(2)Optimize z(1) using the coarse-to-fine scheme.

The contour tracker worked fine for most of the cases, but it fails in case of 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 the change is transferred to the other frames. In the temporal propagation [10], to reconstruct the point modified by the user, the linear regression coefficients for the other points are estimated. The algorithm proposed works astonishingly well. In comparison to the complicated contour tracking/modification algorithm proposed in [15], are too expensive to be implemented for real-time long distance environments.

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. The pixels lying inside the mask are only used for matching. For occlusion handling problem, apart from the normal procedure, outlier detection is also performed to segregate occlusion in the evaluation of optical flow to compensate the irregularity caused in the evaluation due to arbitrary shape of the mask

For baseline line model for optical flow estimation the system uses optical flow algorithm, while to improve the symmetric flow, computation is 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 layer by layer optical flow. In the first step, the matching of images with the visible data term is formulated as mentioned in below:

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

Table 1 (a) Error incurred during contour based tracking with respect to ground truth frame in percentage and pixels for background Human

Frame	Number of pixels	Error in pixels with ground frame	Error in % w.r.t. ground truth frame
2	74141	00984	1.31
3	73570	01555	2.07
4	72143	02982	3.97
5	71001	04124	5.49
6	69851	05274	7.62
7	68867	06258	8.33
8	66673	08456	11.25
9	65096	10029	13.35
10	63661	11464	15.26
11	62226	12899	17.17

Table 1 (b) Error incurred during contour based tracking with respect to ground truth frame in percentage and pixels for front Human

Frame	Number of pixels	Error in pixels with ground frame	Error in % w.r.t. ground truth frame
2	74855	00270	0.36
3	74396	00729	0.97
4	74208	00916	1.22
5	73690	01435	1.91
6	73232	01893	2.52
7	72969	02156	2.87
8	72578	02547	3.39
9	72308	02817	3.45
10	72112	03012	4.01
11	71766	03358	4.47

Where, u is the Gaussian filter. The data term $B_{data}^{(2)}$ 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:

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

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

$$B_{sym}^{(1)} = \int |g_1(x+y) + g_2(x+g_1, y+h_1)| + |h_1(x+y) + h_2(x+g_1, y+h_1)| \tag{4}$$

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

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

IRLS proposed in [12] is used as equivalent to outer and inner fixed-points, together with the coarse-to-fine search [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:

a).If $B_2(x+g_1, y+h_1) = 0$, then set $B_1(x, y) = 0$;

b).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. As coarse to fine technique is used for the algorithm, we get two bidirectional flow fields and cropped visible layer masks that exhibit occlusion. The user is allowed to change the values of σ, δ and γ in (5).

C. *Human assisted motion labeling*

On failure of optical flow estimation, the user with help of feature points can specify the sparse correspondence between two frames. The system then automatically produces a parametric motion or interpolates a dense flow field based on the specified sparse correspondence. For the specification of sparse correspondence the user can either use the help of computer for increasing efficiency or manually, taking full control of motion annotation. Minimum SSD matching and Lucas-Kanade transform [14] is used by the system for finding the best match in the next frame for the feature point specified by user in previous frame. The system depends on the number of feature points specified to determine the mode of parametric motion i.e. translation, affine transform or homography followed by the estimation of the motion parameters accordingly. The modes mentioned above can also be selected by the user directly and the user even have an option to choose to generate a smooth flow field interpolated using the preconditioned conjugate gradient algorithm.

However, defining corner like features for sequences in which only line structure is present can be a difficult task for these kinds of sequences. In order to solve this problem, uncertainty matching and probalisitic parametric motion were included in the algorithm so that the user can have a freedom to choose any pixel for correspondence. In the case of uncertainty matching, a probability map $w_p(x)$ is produced to match the feature point p at location $c_p \in R^2$. A mean χ_p and covariance matrix $\sum_p$ are used to approximate the probability map $H_p(x)$ .For the determination of the probabilistic motion estimation, the system loops around two points. In the first step, the current estimate of mean and covariance are used for motion approximation. Mathematically, let $s(c_p; \phi): R^2 \to R^2$ be a parametric motion applied to the estimation of parametric motion computed by

$$\phi^* = \arg\min_{\phi} \sum_{p} (s(c_p; \phi) - \tau_p)^T \sum_p (s(c_p; \phi) - \tau_p) \tag{6}$$

In second step, estimation of the mean and covariance is done where a new probability map is used which is reweighted by the current motion

$$\{\tau_p, \sum_p\} \leftarrow i_p(x) \mathrm{N}(s(c_p; \phi^*), \varphi^2 F) \tag{7}$$

Convergence of this algorithm occurs within a few iterations. A dense flow field (i.e. ϕ) can also be obtained for the motion $s(c_p; \phi)$. Also, the feature point specified by the user can be used in the next frame. For providing the human assistance the users interact with the tool through the interface provided in the system developed by the authors of [10].

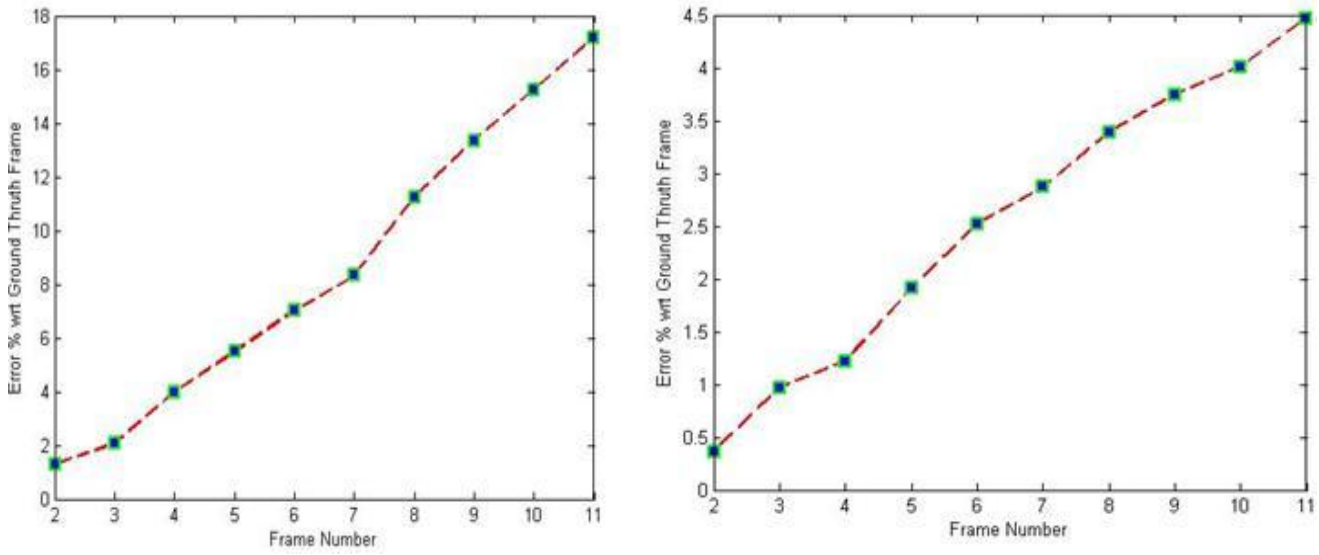

Fig 2. Plots Error incurred during contour based tracking of each frame with respect to ground truth frame in percentage and pixels for (a) First Human (ii) Second Human.

Table 2 (a) Error incurred during Horn–Schunck based tracking with respect to ground truth frame in percentage and pixels for front Human (b) Error incurred during Horn–Schunck based tracking with respect to ground truth frame in percentage and pixels for front background Human

a) Frame	Error in % (Annotation Method)	Error in % (Horn-Schunck Method)
2	1.31	0.18
3	2.07	0.22
4	3.97	0.36
5	5.49	0.38
6	7.62	0.42
7	8.33	0.43
8	11.25	0.45
9	13.35	0.50
10	15.26	0.54
11	17.17	0.55

b) Frame	Error in % (Annotation Method)	Error in % (Horn-Schunck Method)
2	0.36	0.18
3	0.97	3
4	1.22	18
5	1.91	24
6	2.52	33
7	2.87	53
8	3.39	64
9	3.45	88
10	4.01	92
11	4.47	96

III. Results

The results obtained from the experimentation enable us to investigate the capabilities of image annotation to track humans efficiently. The primary step for the tracking is to locate the human 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 simulations were carried out on windows 7 running on an Intel i3 2.26 GHz processor machine.

The results obtained from image annotation toolbox applied to the video sequence as shown in fig. 1. The test is conducted on the video sequence with frame by frame input given to image annotation toolbox. The user manually constructs a contour on the humans in the reference frame which further can be used to track the human in the succeeding frames. The contour generated is further used to track the human 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. The automatic tracking for all 7 frames takes less than 2 seconds to compute.

Error incurred during contour generation, as we move from one frame to another is an interesting topic to study. The error analysis for the video under test is carried out for each 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. 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.

Error is computed in two scenarios (i) the contour tracks the front human (ii) the contour tracks the background human irrespective of the front human. The error for tracking background human tracking varies from a minimum value of 1.31(%) to a maximum of 17.17(%) while the error for the second case, front human tracking varies from a minimum of 0.36 (%) to a maximum 4.47 (%) of as shown in Table. 1 (a) and Table. 1 (b) respectively. The error for both cases is also plotted as shown in fig. 2.

The superiority of our method is proved by comparing the results with Horn-Schunck method. It is observed that image annotation method works much better in case of tracking the background human. In case of tracking front human, Horn–Schunck method gives better results as shown in Table. 2. Overall image annotation performs much better in occlusion condition. 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.

IV. Conclusion

The task of tracking the motion of individual humans in the single frame of the video sequence was successfully achieved. The method divides the video scene into multiple layers assigning each layer to the individual object of interest. Since each layer has been assigned to a specific object in the video sequence, we are able to track and analyze the movement of each object individually; in addition the method is able to reframe from misclassification as each object has been assigned a respective layer. The paper also computes the error incurred during tracking multiple humans in a single frame. 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 method also out performs the traditional Horn–Schunck method. Overall, the system has vast application in areas where flawless tracking is of great importance

References

[1] Book Title: Pervasive Computing" In proc.Lecture note on comp Vol 3468,pp 329: 334, 2005

[2] Underwater Human-Robot Interaction via Biological Motion Identification. Junaed Sattar and Gregory Dudek. Proceedings of the 2009 Conference on Robotics: Science and Systems V (RSS), MIT Press, pages 185-192. June-July 2009, Seattle, WA, USA.

[3] Wilson, D. Atkeson, C. "Simultaneous Tracking and Activity Recognition (STAR) Using Many Anonymous, Binary Sensors

[4] Luis M. Fuentes,Sergio A. Velastin "Human tracking in surveillance applications" In Proc. of the 2nd IEEE International workshop on PETS 2001

[5] Treptow, Andre and Cielniak, Grzegorz and Duckett, Tom (2006) *Real-time human tracking for mobile robots using thermal vision.* Robotics and Autonomous Systems, 54 (9). p. 729. ISSN 0921-8890

[6] David Moore. "A real-world system for human motion detection and tracking." California Institute of Technology,2003

[7] Ronan Fablet. Michael J. Black "Automatic Detection and Tracking of Human Motion with a View-Based Representation" In Proc of ECCV 2002

[8] I. Haritaoglu, D. Harwood and L.S. Davis, "W4: Real-time surveillance of human and their activities", PAMI, Vol. 22. No. 8, pp. 809-830, 2000

[9] Haritaoglu I, Flickner M, "Detection and tracking of shopping groups in stores", Proceeding of the 2001 IEEE Computer Vision and Pattern Recognition, vol. 1, 8-14, pp. 431-438, December. 2001

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

[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 theory for wrapping In *Proc. ECCV*, pages 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] 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*, pages 674–679, 1981

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

Devinder Kumar is an Undergraduate Student researcher currently pursuing his Bachelors in Electrical and Electronics Engineering at the National Institute of

Technology Warangal. He is the one of the founding member & Co-ordinator of ILLUMINATI@NITW, a potential research group of students at National Institute of Technology Warangal (Well Known across a Number of Countries in Europe and Asia).He is the Head of the Computer Vision & Image Processing Cluster at IEEE Student Branch NIT Warangal as well as a active member of IEEE, IEEE Communications Society and IEEE Power and Energy Society. His research interests mainly in the field of Computer vision ,Image processing & Machine learning in particular involving: Motion Tracking, Object Identification, Image Annotation. In the past he has worked with number of organization that includes IISc Bangalore, University of Kaiserslautern, Germany and others.

Amarjot Singh is a Research Engineer with Tropical Marine Science Institute at National University of Singapore (NUS). He completed his Bachelors in Electrical and Electronics Engineering from National Institute of Technology Warangal. He is the recipient of Gold Medal for Excellence in research for Batch 2007-2011 of Electrical Department from National Institute of Technology Warangal. He has authored and co-authored 48 International Journal and Conference Publications. He holds the record in Asia Book of Records (India Book of Record Chapter) for having "Maximum Number (18) of International Research Publications by an Undergraduate Student". He has been awarded multiple prestigious fellowships over the years including the prestigious Gfar "Research Scholarship" for Excellence in Research from Gfar Research Germany and "Travel Fellowship" from Center for International Corporation in Science (CICS), India. He has also been recognized for his research at multiple international platforms and has been awarded 3rd position in IEEE Region 10 Paper Contest across Asia-Pacific Region and shortlisted as world finalist (Top 15) at IEEE President Change the World Competition. He is the founder and chairman of Illuminati, a potential research groups of students at National Institute of Technology Warangal (Well Known across a Number of Countries in Europe and Asia). He has worked with number of research organizations including INRIA-Sophia Antipolis (France), University of Bonn (Germany), Gfar Research (Germany), Twtbuck (India), Indian Institute of Technology Kanpur (India), Indian Institute of Science Bangalore (India) and Defense Research and Development Organization (DRDO), Hyderabad (India). His research interests involve Computer Vision, Computational Photography, Motion Tracking etc.

