TRACKING OF MOVING OBJECTS IN VIDEOS

N.K Priyadharsini[1], D. Chitra[2]

Abstract

Video processing is a technique of processing individual frames or images. The processes .involve acquisition, manipulation, transmission, analysis and compression. This paper focuses on video analysis; it includes motion segmentation and motion tracking. Tracking objects in a video containing extremely crowded scenes is a challenging due to the motion and appearance variability produced by the large number of people within the scene. The individual pedestrians collectively form a crowd that exhibits a spatially and temporally structured pattern within the scene. The video is divided into sub volumes. The local spatio-temporal motion of the sub volume is extracted. Hidden Markovian model is used to train on the spatio-temporal motion pattern. From the model the spatio-temporal motion pattern that describes how the object moves in a video is obtained. The extracted information is used as the priori for tracking.

Key words : *Bayseian, HMM, spatio-temporal pattern, Tracking, Video*

I. Introduction

Video content analysis is the process of analysing the video to determine the events. It's used in the domains like human computer interaction, surveillance, security, video communication, compression, traffic control, medical imaging and video editing. The functionality of video analysis is video tracking, identification and behavior analysis and egomotion estimation. This model focuses on the video tracking. Video tracking is the process of determining the location of the object in the video signal. Video contains large amount of data so the tracking can be a time consuming process. The video tracking is to associate target objects in consecutive video frames. The video tracking algorithm analyses the sequential video frames and outputs the motion of the object between the frames. There are two major components for visual tracking. The components are target representation and localization and data association and filtering. There are numerous researches for video-based object extraction and tracking. One of the simplest methods is to track regions of difference between a pair of consecutive frames [24], and its performance can be improved by using adaptive background generation and subtraction. The difference-based tracking method is efficient in tracking an object under noise-free circumstances; it often fails under noisy, complicated background. The tracking performance degrades if a camera moves either intentionally or unintentionally. Tracking of objects in

[1]Assistant professor, Department of Computer science and Engineering. E-mail : priyasnk@gmail.com

[2]Professor and Head, Department of CSE, P.A. College of Engineering and Technology.
E-mail : chitra.pacet@gmail.com

the presence of shadows, noise, and occlusion, a non-linear object feature voting scheme has been proposed in [25]. As an alternative method of the difference-based tracking, a blob tracking method using simple geometric models, e.g., ellipse or rectangle, can track the centroid of an object. Based on the assumption of stationary background, Wren et al. proposed a real-time blob tracking algorithm [26]. For more robust analysis of an object, shape-based object tracking algorithms have been developed, which utilize a priori shape information of an object-ofinterest, and project a trained shape onto the closest shape in a certain frame. This type of methods includes active contour model (ACM), active shape model (ASM), and the Condensation algorithm [10]. Although the existing shape-based tracking algorithms can commonly deal with partial occlusion, they exhibit two major problems in the practical applications, such as: a priori training of the shape of a target object and iterative modeling procedure for convergence. Selected object tracking algorithms of significant interests are summarized in Table 1.

TABLE 1

Properties of various object tracking algorithms

Algorithm	Tracking entity	Occlusion Handling	Specific Task
W4 [25]	Silhouettes of people	Appearance model	Surveillance of people
Amer [26]	Shape, Size, Motion	Non-linear voting	Surveillance, retrieval
Wren [27]	Human body model	Multi-class Statistical model	Tracking Human
Isard [10]	Shape space	State Space Sampling	Tracking objects in the shape space
Yilmaz [28]	Contour	Shape prior	Tracking level sets
Rodriguez [20]	Human body model	Unstructured crowd	Tracking Human
Kratz [14]	Motion	Motion pattern prior	Tracking Human

Object Tracking

The process of object tracking is summarized in the Figure 1:

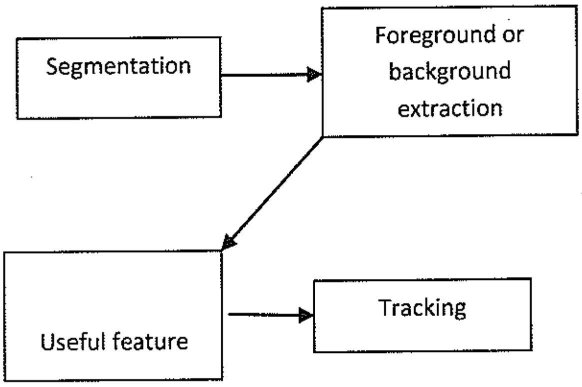

Figure 1 : Tracking of Objects

The basic steps in object tracking are:

1. Segmentation
2. Foreground or background extraction
3. Camera modeling
4. Feature extraction and tracking

Segmentation is the process of identifying components of the image. Segmentation involves operations such as boundary detection, connected component labeling and thresholding. Boundary detection finds out edges in the image, any differential operator can be used for boundary detection. Thresholding is the process of reducing the grey levels in the image. Foreground extraction is the process of separating the foreground and background of the image. This process is used for subtraction of images in order to find objects that are moving and those that are not. Another method that can be used in object tracking is Background learning [1]. This approach can be used if fixed cameras are used for video capturing. In this method, an initial training step is carried out before deploying the system. In the training, the system constantly records the background in order to learn it. Once the training is completed the system has complete information about the background. Once the background is known, extracting the foreground is a simple image subtraction [12]. The next step is to extract useful features from the sequence of frames.

The goal of the object tracking is to estimate location and motion parameters of an object in a image sequences. The objective function of tracking depends on distance, similarity or classification measure. The tracking results are often obtained by minimizing or maximizing an objective function.

The size of the cuboids shown in Figure 2 remains same and selected manually due to loss of pixels limit the accuracy of the tracking. The video are divided into spatio-temporal sub-volumes or cuboids defined by a regular grid and compute the local spatio-temporal motion pattern within each[6]. Hidden Markov Model [19] is trained on the local spatio-temporal motion patterns at each spatial location. The spatial location of the training video represents the spatially and temporally varying crowd motion [2]. Using the HMM and previously observed frames of a separate tracking video of the same scene, the local spatio-temporal motion pattern that describes a target moves through the video is predicted. The predicted local spatio-temporal motion pattern is used to hypothesize a set of priors on the motion and appearance variation of individuals that wish to track[17].

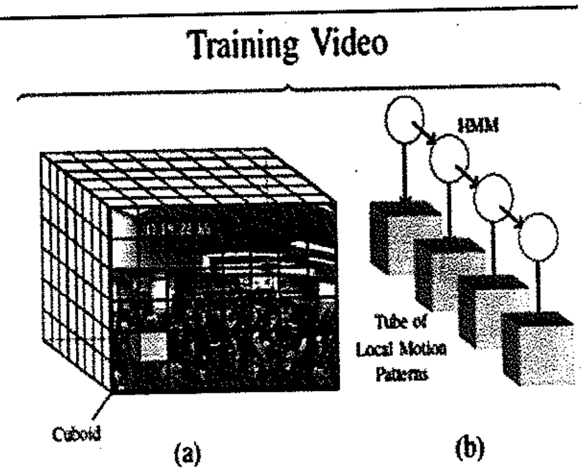

Figure 2 : Cuboids

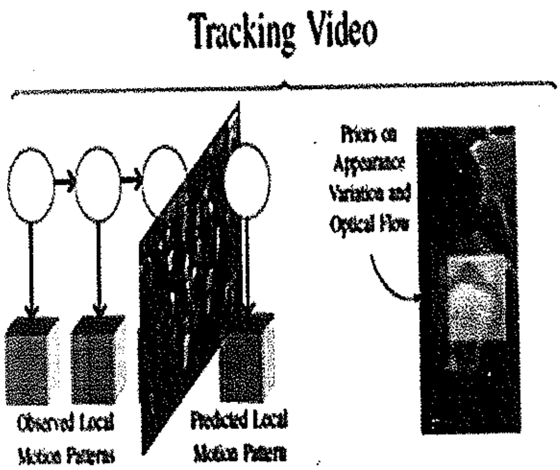

Figure 3 : Pattern Prediction

We divide the video into spatio-temporal sub-volumes, or cuboids. Figure 2 (a) The local spatio-temporal motion pattern within each cuboid is computed. Then train a hidden Markov model Figure 2 (b) on the local spatio-temporal motion patterns at each spatial location. Using the HMM and previously observed frames of a separate tracking video of the same scene, we predict the local spatio-temporal motion pattern that describes how a target moves through the video is shown in Figure 3. The predicted local spatio-temporal motion pattern is used to hypothesize a set of priors on the motion and appearance variation of individuals that we wish to track [16].

II. Related Works

Since the literature on tracking is extensive, the work that model motion in cluttered or crowded scenes is reviewed.

Isard et al. [10] developed the condensation based algorithm to track the visual clutter it represents the multi modal distribution. The algorithm uses stochastic framework and track the outline and features of the object. The approach aims at using the probabilistic model of object shape and motion to analyze the video streams. The algorithm is used for both rigid and non-rigid motion.

Black et al. [7] proposed a bayesian framework for representing and recognizing local image motion in terms of two basic models like translational motion and motion boundaries. The method move towards a richer description of image motion using a vocabulary motion primitive. A step in that direction is described with the introduction of an explicit non-linear model of motion boundaries and a Bayesian framework for representing a posterior probability distribution over models and model parameters. A maximum estimate of image motion is calculated using the probability distribution over the parameter of discrete samples. The image motion facilitates the correct Bayesian propagation of information over time and its ambiguities make the distribution non-Gaussian. The open issue is that sampling methods has high dimensional space.

Betke et al [5] Proposed statistical data association techniques for visual tracking of enormously large numbers of objects. This approach combines the techniques of multi target track initiation, recursive Bayesian tracking, clutter modeling, event analysis, and multiple hypotheses.

Okabe et al proposed a method which tracks features and associate similar trajectories to detect individual moving entities within crowded scenes. They assume that the subjects move in distinct directions, and thus disregard possible local motion inconsistencies between different body parts.

Wright et al. [23] proposed a method for analysis of motion patterns. The system uses static and quasi-static backgrounds. This background model produces a crude initial segmentation that is postprocessed by code specific to find and recognizing humans. Ali and Shah and Rodriguez et al. model the motion of individuals across the frame in order to track pedestrians in crowds captured at a distance.

Pless et al [22]. Learn a single motion distribution at each frame location from videos of automobile traffic. These approaches impose a fixed number of possible motions at each spatial location in the frame. In extremely crowded scenes, however, pedestrians in the same area of the scene may move in any number of different directions. We encode many possible motions in the HMM, and derive a full distribution of the motion at each spatio-temporal location in the video. In addition, natural body movements appear subtle when captured at a distance but create large appearance changes in near-view scenes, which we explicitly model.

Nestares and Fleet [18] also use neighboring motion patterns to improve tracking. They increase the continuity of the motion boundary tracking from Black and Fleet by including multiple image neighborhoods. In our work, however, we use a dynamic temporal model of sequential motion patterns rather than assuming continuity across spatial locations.

III. Proposed Scheme

The methods used for tracking are mostly object-centric based on the modeling of motion and appearance of the target individual. The size of the cuboids remains same and selected manually due to loss of pixels limit the accuracy of the tracking. The video are divided into spatio-temporal sub-volumes or cuboids defined by a regular grid and compute the local spatio-temporal motion pattern within each. Hidden Markov Model [19,24] is trained on the local spatio-temporal motion patterns at each spatial location. The spatial location of the training video represents the spatially and temporally varying crowd motion. Using the HMM and previously observed frames of a separate tracking video of the same scene, the local spatio-temporal motion pattern that describes a target moves through the video is predicted. The predicted local spatio-temporal motion pattern is used to hypothesize a set of priors on the motion and appearance variation of individuals that wish to track[9,11]. The proposed method tracks individual pedestrians in videos of extremely crowded scenes by leveraging a model of the spatially and temporally varying structure in the crowd motion. The training video is to divide cuboids. The motion pattern of the each cuboid is observed. The previous frame is used to predict the spatio-temporal motion pattern for next M frames. M indicates number of frames in the cuboid. For each pixel 'i' in cuboid Spatio-temporal gradient ∇i is calculated using Equation (1)

$$i = \left[I_{i,x}, I_{i,y}, I_{i,t}\right]^T = \left[\frac{\partial I}{\partial x}\frac{\partial I}{\partial y}\frac{\partial I}{\partial t}\right]^T, \quad (1)$$

The hidden state can be obtained from the HMM [23].

$$\mu_t^n = \frac{1}{M}\sum_i^m \nabla \mathrm{I}_i,$$

$$\sum\nolimits_t^n = \frac{1}{M}(\nabla I_i - \mu)\nabla I_i - \mu^T, \quad (2)$$

where ∇i is the 3D gradient at pixel i and μ_t^n represents spatio-temporal gradient [20].

The hidden states of the HMMs encode the multiple possible motions that can occur at each spatial location. The transition possibilities of the HMMs encode the time-varying dynamics of the crowd motion [13]. Training a collection of HMMs, the spatially and temporally varying motions of pedestrians that comprise the entire crowd motion is encoded. The KL divergence [15] decides the number of hidden states and the clustering of each cuboid. The KL-divergence based densities emit the mean and covariance matrix of 3D Gaussians, retaining the rich motion information stored in the space-time gradients [3, 21]. The enhanced K-means algorithm are been used to classify the cuboids based on the distance measure KL. KL divergence distance is used to determine the distance between sub volumes and the centroids. The classifiers that can be used are k-means Ensemble, SVM, and Bayesian [4]. The classification is done using the distance measure. Enhanced K-means algorithm is used for the classification in our model.

The accuracy of the algorithm depends on the number of hidden states in each HMM. The number of hidden states can be varied depending on the distance measure. The number of hidden state increases with decrease in distance measure. The distance measure used in the method is 2 and 0.5. For the distance 0.5 the number of hidden states gets increased which helps in tracking more number of motion pattern. Increase in motion pattern helps in improving the tracking accuracy, since in video there can be any number patterns.

The probability n of an observed motion pattern O_t^n deûned by μ_t^n and $\sum_t^n$ given a hidden state s is represented in the Equation (3).

$$p(O_t^n|s) = p\left(\frac{d_{KL}\ (O_t^n . P_s)}{\sigma_s}\right) \sim N(0.1), \qquad (3)$$

where P_s is the expected motion pattern, σ_s is the standard deviation and d_{KL} $(\cdot)$ is the Kullback-Leibler (KL) divergence [15]. The observed motion patterns $O_1^n \ldots O_{t-1}^n$ in the tracking video is given, the predictive distribution is expressed in Equation (4).

$$p(O_t^n|O_1^n|O_2^n \ldots\ldots O_{t-1}^n) = \sum_{s \in S} p(O_t^n|s)\,\omega(s), \quad (4)$$

S is the set of hidden states, $\omega(s)$ is defined by

$$\omega = \sum_{s \in S} p(s|S')\alpha'(S'), \qquad (5)$$

After tracking in frame t, each pixel i in the motion template is updated by

$$T_{M.i}^t = \alpha\,\nabla R_{E[X_t|Z_{1:t}].i} + (1-\alpha)T_{M.i}^{t-1}, \qquad (6)$$

where $T_{M.i}^t$ is the motion template at time t, $\nabla R_{E[X_t|Z_{1:t}].i}$ is the region of spatio-temporal gradient defined by the tracking result.

The predicted local spatio-temporal motion pattern can be used to track the individuals in the bayseian framework [8, 20]. The predicted motion can be used as the priori for tracking. The tracking video is compared with the training video for tracking the objects. To accurately represents the motion within the video, quantization is avoided by training novel HMMs. Tracking are mostly object-centric based on the modeling of the motion and appearance of the target individual. The pedestrian's motion changes gradually, this error measurement during tracking can be calculated using the Equation (7)

$$E_i^t = \alpha \arccos(t_i . t_r)\,(1-\alpha)E_i^{t-1}, \qquad (7)$$

IV. Result

The video with unstructured crowed are taken. For each cuboid the motion patterns are observed. The video to be tracked is learned and the priors are learned. Hidden Markov models are trained on the video. The KL Divergence distance used is 0.5 and 2. The number of hidden states increases with decrease in the distance this effect is shown in Figure 4. The graph given below shows the variation in the hidden states. The number of states gets stabilized for the distance above 3. So, the distance used here is 0.5, 1, 1.5 and 2. The collection of HMMs is trained on a video of each scene, and uses it to track pedestrians in videos of the same scene recorded at a different time. The training videos for each scene have different frames. The size of cuboids is used to represent the local spatio temporal motion pattern. The accuracy of our predictions depends heavily on the number of hidden states in each HMM. The clustering algorithm uses a distance threshold dKL to vary the number of hidden states depending on the local motion patterns in the video. Large variations in flow may result in an excessive number of hidden states.

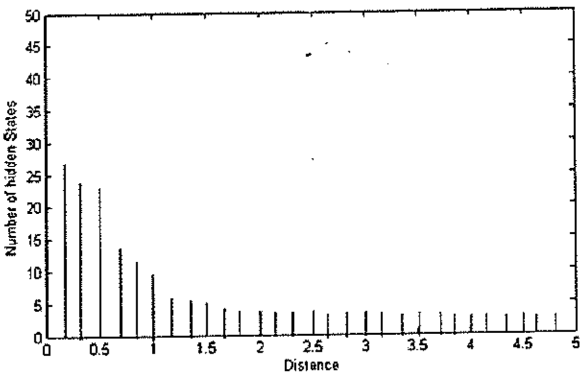

Figure 4 : Effects of distance threshold

V. Conclusion

A probabilistic method that exploits the inherent spatially and temporally varying structured pattern of a crowd's motion to track individuals in extremely crowded scenes are derived. The input video is divided into equal size cuboids form which Hidden Markovian Model training can be done. Using a collection of Hidden Markovian Model that encode the spatial and temporal variations of local spatio-temporal motion patterns, the proposed method successfully predicts the motion patterns within the video. The predicted video is used to track the object present in the tracking video. The results show that leveraging the steady-state motion of the crowd provides superior tracking results in extremely crowded areas. Further, the algorithm can be extended for the automatic detection of cuboids size. The cuboids sizes may be determined by a semi-supervised approach that approximates the perspective projection of the scene. Varying cuboid size can also be used. Space-time model may be further leveraged to provide robustness severe occlusions.

References

[1] S. Ali and M. Shah. A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis. In Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition, pages 1–6, 2007.

[2] S. Ali and M. Shah. Floor Fields for Tracking in High Density Crowd Scenes. In Proc. of European Conf on Computer Vision, 2008.

[3] S. M. Arulampalam, S. Maskell, and N. Gordon. A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking. IEEE Transactions on Signal Processing, 50:174–188, 2002.

[4] J. Barron, D. Fleet, and S. Beauchemin. Performance of Optical Flow Techniques. Int'l Journal on Computer Vision, 12(1):43–77, 1994.

[5] M. Betke, D.E. Hirsh, A. Bagchi, N.I. Hristov, N.C. Makris, and T.H. Kunz. Tracking Large Variable Numbers of Objects in Clutter. In Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition, pages 1–8, 2007.

[6] C. M. Bishop. Pattern Recognition and Machine Learning. Springer, October 2007.

[7] M. J. Black and D. J. Fleet. Probabilistic Detection and Tracking of Motion Boundaries. Int'l Journal on Computer Vision, 38(3):231–245, July 2000.

[8] G.J. Brostow and R. Cipolla. Unsupervised Bayesian Detection of Independent Motion in Crowds. In Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition, pages 594–601, June 2006

[9] C. Hue, J.-P. Le Cadre, and P. Perez. Posterior Cramer-Rao Bounds for Multi-Target Tracking. Aerospace and Electronic Systems, IEEE Transactions on, 42(1):37 – 49, Jan. 2006.

[10] M. Isard and A. Blake. CONDENSATION-Conditional Density Propagation for Visual Tracking. Int'l Journal on Computer Vision, 29(1):5–28, August 1998.

[11] Z. Khan, T. Balch, and F. Dellaert. MCMC-Based Particle Filtering for Tracking a Variable Number of Interacting Targets. IEEE Trans. on Pattern Analysis and Machine Intelligence, 27(11):1805–1819, Nov. 2005.

[12] Z. Khan, T. Balch, and F. Dellaert. MCMC Data Association and Sparse Factorization Updating for Real Time Multitarget Tracking with Merged and Multiple Measurements. IEEE Trans. on Pattern Analysis and Machine Intelligence, 28(12):1960–1972, Oct. 2006.

[13] L. Kratz and K. Nishino. Anomaly Detection in Extremely Crowded Scenes Using Spatio-Temporal Motion Pattern Models. In Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition, pages 1446–1453, 2009.

[14] L. Kratz and K. Nishino. Tracking With Local Spatio-Temporal Motion Patterns in Extremely Crowded Scenes. In Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition, 2010.

[15] S. Kullback and R. A. Leibler. On Information and Sufficiency. The Annals of Mathematical Statistics, 22(1):79–86, 1951.

[16] B. Leibe, E. Seemann, , and B. Schiele. Pedestrian Detection in Crowded Scenes. In Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition, June 2005.

[17] Y. Li, C. Huang, and R Nevatia. Learning to Associate: HybridBoosted Multi-Target Tracker for Crowded Scene. In Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition, 2009.

[18] O. Nestares and D. J. Fleet. Probabilistic Tracking of Motion Boundaries with Spatiotemporal Predictions. In Proc. Of IEEE Int'l Conf on Computer Vision and Pattern Recognition, pages 358–365, 2001.

[19] L. Rabiner. A Tutorial On Hidden Markov Models and Selected Applications in Speech Recognition. Proc. of the IEEE, 77(2):257–286, Feb. 1989.

[20] M. Rodriguez, S. Ali, and T. Kanade. Tracking in Unstructured Crowded Scenes. In Proc. of IEEE Int'l Conf on Computer Vision, 2009.

[21] E. Shechtman and M. Irani. Space-Time Behavior Based Correlation. In Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition, pages 405–412, 2005.

[22] J. Wright and R. Pless. Analysis of Persistent Motion Patterns Using the 3D Structure Tensor. In IEEE Workshop on Motion and Video Computing, pages 14–19, 2005.

[23] B. Wu and R. Nevatia. Tracking of Multiple, Partially Occluded Humans Based On Static Body Part Detection. In Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition, pages 951–958, 2006.

[24] T. Zhao, R. Nevatia, and B. Wu. Segmentation and Tracking of Multiple Humans in Crowded Environments. IEEE Trans. on Pattern Analysis and Machine Intelligence, 30(7):1198–1212, 2008.

[25] Haritaog lu I, Harwood D, Davis L. W-4: Real-time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000;22(8):809–30.

[26] Amer A. Voting-based simultaneous tracking of multiple video objects. Proceedings of the SPIE Visual Communication Image Processing 2003;5022:500–11.

[27] Wren C, Azerbayejani A, Darrel T, Pentland A. Pfinder: Realtime tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence 1997;19(7):780–5.

[28] Yilmaz A, Shah M. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence 2004;26(11):1531–6.

AUTHOR'S BIOGRAPHY

Dr. D. Chitra is a Professor in the Department of Computer Science and Engineering at P. A. College of Engineering and Technology, Pollachi, Coimbatore. She received her M.E. Degree in CSE from Anna University, Chennai and PhD degree in CSE from Anna University of Technology, Coimbatore. Her resource interests include image analysis, Pattern recognition and Computer Vision. She is a life member of ISTE.

Mrs. N.K. Priyadharsini is an Assistant Professor in the Department of Computer Science and Engineering at P. A. College of Engineering and Technology, Pollachi, Coimbatore. She received her B.Tech Degree in IT from Anna University, Chennai and M.E. degree in CSE from Anna University Chennai. Her resource interest includes image processing and Analysis.