누적 유사도 변환을 이용한 물체 추적

Moving Object Tracking using Cumulative Similarity Transform

주문원
성결대학교 멀티미디어학부

Moon-Won Choo
Division of Multimedia, Sungkyul University

중심어 : object tracking, image correspondence, photometric Invariance

요 약

동영상에서의 물체추적은 각 프레임간의 시공간적 정보의 대응성을 추적함으로써 이루어질 수 있다. 시공간적 대응성을 결정함에 있어 각 프레임의 color invariant 속성을 이용하여 각 프레임의 물체 위치를 예측하고, 추출된 물체 블록의 누적 유사도 변환을 적용한 수치를 이용하여 프레임간의 물체 대응성을 결정함으로써 물체를 추적하는 방법론을 제시한다. 실험결과를 통하여 이러한 방법의 적용 적절성을 검증하였다.

Abstract

In this paper, an object tracking system in a known environment is proposed. It extracts moving area shaped on objects in video sequences and decides tracks of moving objects. Color invariance features are exploited to extract the plausible object blocks and the degree of radial homogeneity, which is utilized as local block feature to find out the block correspondences. The experimental results are given.

I. Introduction

Tracking the motion of objects in video sequences is becoming important as related hardware and software technology gets more mature and the needs for applications where the activity of objects should be analyzed and monitored are increasing[9]. In such applications lots of information can be obtained from trajectories that give the spatio-temporal coordinates of each objects in the environment. Information that can be obtained from such trajectories includes a dynamic count of the number of object within the monitored area, time spent by objects in an area and traffic flow patterns in an environment[4],[7],[14],[15]. The tracking of moving object is challenging in any cases, since image formations in video stream is very sensitive to changes of conditions of environment such as illumination, moving speed and, directions, the number and sizes of objects, and background. Therefore the scope of researches are usually confined to specific application domains and the processes of capturing video streams are also carefully controlled. Moreover, most of related researches are assuming gray-level images as input image source, which may lose much of information available in color space such as imbedded photometric color features and synthesized features derived from separate color channels. Color image can be assumed to contain richer information for image processing than its corresponding gray-level image. Also separate color channel could be applied to different problem domains.

In this paper, a system for obtaining such spatio-temporal tracks of objects in video sequences is suggested. Camera in static position produces video sequences which are analyzed in real time to obtain trajectories. In each frame of video stream, segmentation techniques such as simple progressive projections and differencing color invariance feature maps from inter-frame images could work well in real time and yield regions of interest for blocking quickly. An important step towards the track of objects is the definition of a proper set of features which could reliably identify the corresponding objects between adjacent frames. Local block feature is computed using the cumulative similarity transform, which is very effective and flexible in weighting the feature values according to pixel coordinates. All processing phases exploit color invariance features, which prove to be more reliable than just gray-level intensities.

II. Related Researches

Jakub and Sarma[10] developed a system for real-time tracking of people in video sequences. They use a model-based sapproach to object tracking, identifying feature points like local curvature extrema in each video frame. Their system has an advantage of handling occlusion problems, but disadvantage of unreliable extraction of extrema of curvature from object contours. William[16] suggests motion tracking by deriving velocity vectors from point-to-point correspondence relations. Relaxation and optical flow are very attractive methodologies to detect the trajectories of objects[13]. Those researches are based on the analysis of velocity vectors of each pixel or group of pixels between two neighboring frames. This approach requires heavy computation for calculating optical flow vectors. Another method infers the moving information by computing the difference images and edge features for complementary information to estimate plausible moving tracks[14],[18]. This method may be very sensitive to illumination and noise imposed on video stream. The other method adopts the model-based and/or statistical approach, which has disadvantage of extracting the previously trained objects only[5],[6]. A flow chart showing the main steps followed in this work is given in Fig. 1.

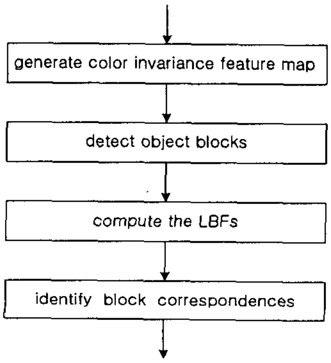

generate color invariance feature map
detect object blocks
compute the LBFs
identify block correspondences

Fig. 1. Main steps in the moving object tracking system.

III. Moving Object Tracking System

Each processing module shown in Fig. 1 is detailed in order.

1. generation of color invariance map

Many block detection methods assume that the lighting in the scene considered would be constant. The accuracy of these methods decreases significantly when they are applied to real scenes because of constantly changing illumination conditioned on background and the moving objects. The methods to take only gray-level intensities into consideration may cause ambiguous object boundaries, which results in seriously degraded performance of object segmentation and detection. It is known that color is a powerful cue in the distinction and recognition of objects. To reduce some of the complexity intrinsic to color images, parameters with known invariance are of prime importance.

Kubelka-Munk theory models the reflected spectrum of a colored body based on a material-dependent scattering and absorption function, under assumption that light is isotropically scattered within the material[11],[12]. The photometric reflectance model resulting from this theory is given by

$$E(\lambda,\vec{x}) = e(\lambda,\vec{x})(1-\rho_f(\vec{x}))^2 R_\infty(\lambda,\vec{x}) + e(\lambda,\vec{x})\rho_f(\vec{x})$$

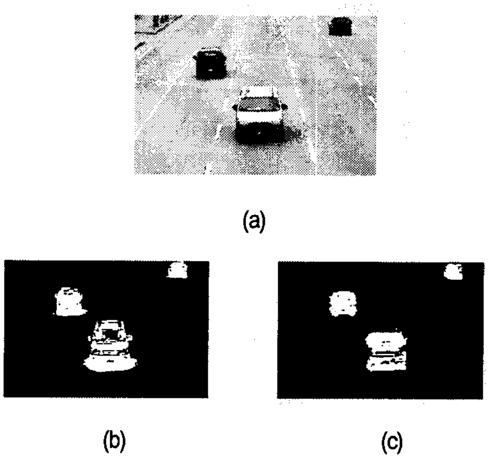
where x denotes the poition at the imaging plane, λ the wavelength, $e(\lambda,\vec{x})$ the illumination spectrum, $\rho_f(\vec{x})$ the Fresnel reflectance at $\vec{x}$, and $R_\infty(\lambda,\vec{x})$ the material reflectivity. The feflected spectrum in the viewing directon is given by $E(\lambda,\vec{x})$. Since the spectral components of the source are constant over the wavelengths for an equal energy illumination, a spatial component i(x) denotes intensity variations, resulting in

$$E(\lambda,\vec{x}) = i(x)((1-\rho_f(\vec{x}))^2 R_\infty(\lambda,\vec{x}) + \rho_f(\vec{x})).$$

Differentiating with respect to λ twice and a little computation, the ratio $H = E_\lambda / E_{\lambda\lambda}$ is known to be dependent on derivatives of the object reflectance functions $R_{\infty}(\lambda,\vec{x})$ only. That is, H is an object reflectance property independent of viewpoint, surface orientation, illumination direction, illumination intensity and Fresnel reflectance coefficient. This color invariance feature can be used for calculating the trajectories of objects reliably. To get this spectral differential quotients, the following implementation of Gaussian color model in RGB terms is used (for details, see [11]).

$$\begin{bmatrix} E \\ E_{\lambda} \\ E_{\lambda\lambda} \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.3 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{pmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

Fig. 2 shows the comparison of H image map and gray-level image after block detection process to be mentioned below. The redundant shadowed areas are eliminated properly when the color invariance map H is used.


(a)

(b) (c)

Fig. 2. (a) image to be tested, (b) block detected when gray-level image is used, (c) in case of H map used.

2. block detection module

This module plays an important role as shown in many object tracking applications using segmentation approaches[5],[10]. This module receives a pair of H maps, $I_k(x,y)$ and $I_{k+1}(x,y)$, acquired at successive time instants t_k and t_{k+1}, respectively. Then a list of minimum bounding rectangle-shaped blocks of image areas where significant H feature changes (related to possible moving objects) is produced. This module consists of three steps.

(1) comput the difference $D_k(x,y)$ between the two input images $I_{k+1}(x,y)$ and $I_k(x,y)$ for obtaining spatial difference information,

$$D_k(x,y) = | I_k(x,y) - I_{k+1}(x,y) | .$$

(2) computing the difference $DB_k(x,y)$ between the two input images $I_{k+1}(x,y)$ and background $B_k(x,y)$ for obtaining temporal information,

$$DB_k(x,y) = | I_{k+1}(x,y) - B_k(x,y) | .$$

(3) computing the hypothesis mask M_t identifying moving objects in the current frame.

$$M_k(x,y) = \delta \,| D_k(x,y) - DB_k(x,y) - T_k | ,$$

where δ is ordinary delta function and T_k is thresholding value.

Noise filtering and searching for the minimum bounding rectangular shaped blocks are performed by means of simple opening morphological operation[8] and the extraction of extremal points using progressive projection. The projection[3] is given as follows;

$$P_{\theta}(r) = \int_L M(r\cos\theta - s\sin\theta, r\sin\theta + s\cos\theta)ds \; ,$$

where L is the perpendicular line intersecting the original line whose origin is inclined at an angle θ with respect to the x-axis, at a point that is adistance r from the origin and s is the distance from the intersecting point to a point in binary map $M_k(x,y)$ along L. In this work, only vertical $\theta = 0$ and $(\theta = \pi/2)$ projections are considered. But the occluded objects needs several projections recursively. This progressive projections proceed until the detected blocks contains proper size of pixels discernible as an object to human eyes. Progressive projection is very effective and fast method to isolate the region of interest from the binary image.

3. LBF(Local Block Feature) extraction module

Since the object detected with its bounding rectangular block in current frame should be uniquely associated with the corresponding block in neighboring frame, each block should have local block features(LBF) possessing proper discriminating power. There are many researches done in this area[1],[2],[13],[17]. The LBF may contain undesirable features to be stemmed from the part of other objects and background within the block. The cumulative similarity transform suggested by Trevor[18] could solve this problem properly. This method diminishes the undesirable effects of other objects and background, but be sensitive to color features embedded in the center of detected objects. The color features around central regions of objects are captured fully both in magnitude and sign and attenuates all else. Hence, the LBF is comprised of a central color invariance features and a local neighborhood of this attributes. The neighborhood computes the local invariances relative to the central attribute attenuated to discount background influence. Formally, given a block image $B(x,y) \in I(x,y)$, a LBF of B is composed of two terms, a central value, and a neighborhood function:

$$L_B(x,y) = \{C_B(x,y), N_B(x,y,r,\theta)\}.$$

The central value is the color invariance value averaged within a small radius of the given image location:

$$C_B(x,y) = \frac{1}{2\pi R^2} \sum_{r,\theta}^{r \le R} H(x + r\cos\theta, y + r\sin\theta).$$

The radius R is determined depending on the image condition and the requirement specifications of a particular applications. However, the minimum dimension of moving objects to be detected could be set without any difficulties. In this paper, the R, from 5 to 10 pixels (the image dimension is 480 * 640), gives good results. The neighborhood function N is proportional to the likelihood that underlying image attribute is unchanged along a ray from the center of the block to that pixel. To compute the local image invariance energy N, which is simply the MSE between the central point at which the transform is being defined (x,y) and nearby points at a given radical offset (r,θ).

$$E_B(x,y,r,\theta) = \alpha \| (C_B(x,y) - H(x + r\cos\theta, y + r\sin\theta)) \|^2,$$

where α is a color invariance sensitivity coefficient. In this paper, this value is set to 1. The intergral of E along a ray from the central point, and take the negative exponential to obtain N.

$$N_B(x,y,r,\theta) = \exp\{- \int_{\rho < r} E_B(x,y,\rho,\theta) d\rho\}.$$

NB and EB are defined over block coordinates $r \le \min(W_x, W_y)$, where Wx, Wy are the block dimensions along x-axis and y-axis respectively, and $0 \le \theta \le 2\pi$.

IV. Tracking module

The goal of this module is to determine the 3-D positions and the motion parameters of objects recognized by the blocking module, at every time instant. Each detected object block may be assumed to contain only one moving object. But this constraint can be alleviated, since the LBF could be jointly utilized to differentiate multiple objects possibly occluded.

To establish the correspondence relations of blocks between sequential frames, the displacement of a LBF at time t are compared with the LBFs at time t+1.

$$(x', y') = \arg\min_l D_\lambda (L_B^t(x,y), L'^{t+1}_B(x', y')),$$

where l is the number of detected blocks in I_{t+1} . The LBF distance D_λ is defined by computing the weighted L_2 error of the transformed data using a combination of neighborhood difference and central value difference terms.

$$D_\lambda (L_B^t(x,y), L'^{t+1}_B(x', y')) = (1-\lambda)\Delta N + \lambda \Delta C.$$

The neighborhood difference ΔN is defined as the MSE between $N_B(x,y,r,\theta)$ and $N_B(x',y',r,\theta)$ computed over r,θ. ΔC is the MSE between C and C. The bias term λ expresses a trade-off between the contribution of the central attribute error and the neighborhood function error. Generally, the neighborhood error is the most important, since it captures the spatial structure at the given point. However, in certain cases of spatial ambiguity, the central attribute value is critical for making the correct match unambiguous. In real situation, this bias term could be adjusted dynamically if a prior knowledge about objects is available.

V. Experimental results

Several video clips are generated using static SONY digital video camera according to the kinds of objects and their speed variations under natural illumination conditions. Also the variety of different values of parameters associated with the morphological operator, differential operators applied to projection table, and the bias for LBF are tested.

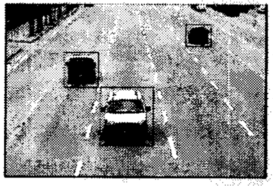
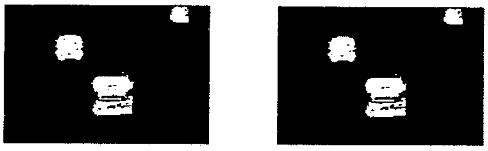


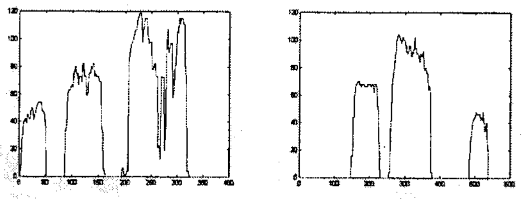
Fig. 3. the original video frame tested and detected blocks.

Fig. 3 shows the one of the video clips tested and detected local blocks. This frame illustrates the results of well-defined block detection module. Fig. 4 shows the moving masks before and after applying opening morphological operators to eliminate the noises. Also the result of vertical and horizontal projections are presented in Fig. 5. The projections are convoluted with Gaussian filter for isolating the regions discernible as objects from the scattered clusters of pixels which could not be recognized as objects to human eyes. This projection may be applied progressively until the sub-blocks are not detected any more. Progressive filtering is applied using scalable Gaussian kernels [17]. Fig. 6 shows the detected tracks of three objects after processing 20 frames running 30fps. By using this method, the direction and velocity can be easily calculated for real monitoring settings.



(a) (b)

Fig. 4. (a) the moving mask, (b) after applying morphological operator.



(a) (b)

Fig. 5. (a) the projection on y-axis. (b) the projection on x-axis.

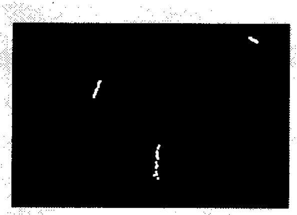


Fig. 6. the trajectories detected.

VI. Conclusions and future research

A system for tracking objects using color invariance features is presented. The system outputs tracks that give spatio-temporal coordinates of objects as they move within the field of view of a camera. I try to solve the occluded objects problem, which may be ignored in many researches intentionally by utilizing radial cumulative similarity measures imposed on color invariance features and projection. But the projection scheme presented here is too primitive to solve this problem., needs more elaborations. But this system is very fast to be used in real time applications and to eliminate noises which may be very difficult when using gray-level intensity only. Also several parameters should be adjusted adaptively in order to be used in more general settings. Future research needs to address these kinds of issues and tracking objects across moving cameras.

References

[1] A. Latif, et. al, *"An efficient method for texture defect detection: sub-band domain co-occurrence matrices," Image and Vision Computing 18, pp. 543-553, 2000.*

[2] B. Chanda, B.B. Chaudhuri, D. Dutta Majumder "On image enhancement and thread selection using the greylevel co-occurrence matrix," Pattern Recognition Lett., Vol. 3, No. 4, pp. 243-251, 1985.

[3] Berthold Klaus Paul Horn, Robot Vision, The MIT Press, 1986.

[4] D. Beymer and K. Konolige, "Real-Time Tracking of Multiple People using Stereo," Proc. IEEE Frame Rate Workshop, 1999.

[5] Dieter Koller, et. al, "Robust Multiple Car Tracking with Occlusion Reasoining," Proc. 3rd European Confer. On Computer Vision, May 2-6, 1994.

[6] Gerhard Rigoll, et. al, "Person Tracking in Real-World Scenarios Using Statistical Methods," IEEE Intl. Conf. on Automatic Face and Gesture Recognition, Grenoble France, March, 2000.

[7] Gian Luca Foresti, et. al, "Vehicle Recognition and Tracking *from Road Image Sequences," IEEE Trans. on Vehicular Tech.*, Vol. 48, No. 1, Jan. 1999.

[8] Gonzalez & Woods, Digital Image Processing, Addison Wesley 1992.

[9] Ismail Haritaoglu, "Real time Surveillance of People and Their Activities," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, Aug. 2000.

[10] Jakub Segen and Sarma Pingali, "A Camera-Based System for Tracking People in Real Time," IEEE Proceedings of ICPR 96, 1996.

[11] Jan-Mak G., et. al, "Color Invariance," IEEE Trans. on PAMI, Vol. 23, No. 12, Dec. 2001.

[12] Kristen Hoffman, "Applications of the Kubelka-Munk Color Model to Xerographic Images," http://www.cis.rit.edu/research/thesis/bs/1998 / hoffman

[13] Milan Sonka, Vaclav Hiavac & Rogger Boyle , Image Processing Analysis and Machine Vision, International Thomson Publishing Co., 1999, 2nd edition

[14] Rita Cucchira, Massimo Piccardi, Paola Mello "Image analysis and rule based reasoning for a traffic monitoring system," IEEE Trans. Intelligent Transportation System, Vol. 1, No. 2, pp. 119-130, June, 2000.

[15] Robert M, Haralick &, Linda G, Shapario, Computer and Robot Vision Vol I, Addision-wesley, USA pp. 318-321, 1992.

[16] Ross Culter & Larry S. Davis "Robust Real-Time Periodic Motion Detection, and analysis and Application," IEEE Trans. Pattern Analysis and Machine Intelligence Vol. 22, No. 8, pp. 781-795, August 2000.

[17] Tony Lindelberg, "Feature Detection with Automatic *Scale Selection," Intl. J. of Computer Vision*, Vol. 30, No. 2, 1998.

[18] Trevor Darrell and Michele Covel "Correspondence with cumulative similiarity transforms" IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 23, No. 2, pp. 222-227, February 2001.

주 문 원(Moon-Won Choo) 정회원

1981년 ~ 1985년 : San Jose State. 대학교 Mathematics학과 (이학사)

1986년 ~ 1987년 : New York Tech. 대학원 computer science학과 (이학석사)

1988년 ~ 1990년 : 삼성전자 기흥연구소 (시스템 소프트웨어연구원)

1991년 ~ 1996년 : Stevens Tech. 대학교 computer science전공(이학박사)

1997년 ~ 현재 : 성결대학교 멀티미디어학부 교수

<관심분야> : 컴퓨터 시각, 영상 처리, 감성 공학