Multiple Object Tracking using Color Invariants

Moon-Won Choo*, Young-Mie Choi*, Ki-Cheon Hong**

*Division of Multimedia, Sungkyul University,
mchoo@sungkyul.edu

** Department of Information and Telecommunications Engineering, Suwon University
kchong@suwon.ac.kr

색상 불변값을 이용한 물체 궤적 추적

주문원*, 최영미*, 홍기천**

*성결대학교 멀티미디어학부
mchoo@sungkyul.edu
** 수원대학교 정보통신학부

요약

본고에서는 움직이는 물체를 추적하는 알고리즘을 제시한다. 이미지의 색상에 대한 불변치를 활용하여 비디오 클립에서 물체 영역을 추출하고 co-occurrence matrix를 구한 후 인접 프레임 간의 대응되는 물체를 결정하여 물체의 궤적을 추적한다. 물체 영역에 적용되는 특징값들의 분리정도치를 활용하여 시스템의 성능을 향상시키는 방법과 실험 결과를 제시한다.

ABSTRACT

In this paper, multiple object tracking system in a known environment is proposed. It extracts moving areas shaped on objects in video sequences and detects tracks of moving objects. Color invariant co-occurrence matrices are exploited to extract the plausible object blocks and the correspondences between adjacent video frames. The measures of class separability derived from the features of co-occurrence matrices are used to improve the performance of tracking. The experimented results are presented.

Keywords : Object Tracking, Image Correspondence, Photometric Invariance, Co-occurrence matrix

1. 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[10]. 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][15][16]. 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 multiple object tracking system for obtaining spatio-temporal tracks of objects in video sequences is proposed. 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 derived from adjacent image frames yield regions of interest quickly and proved to work well in real time. An important step towards detecting trajectories of objects is the definition of a proper set of features which could reliably identify the corresponding objects between adjacent frames. Here we choose one of the most common statistical approach for the characterization of texture, the co-occurrence matrix. Within the context of image processing, the co-occurrence matrix represents the second-order joint conditional density function $f(i,j \mid d,\theta)$, i.e., the probability of going from gray level i to gray level j within a distance d along a direction θ. In most researches, the gray levels are considered as the dimensional factors of co-occurrence matrix. Obviously three color channels composing color images give richer informations than gray level counterparts. Invariances and discriminative power of the color invariants is experimentally investigated here as the dimensions of co-occurrence matrix and the derived features for finding correspondences of objects.

2. Related Researches

Jakub and Sarma[11] developed a system for realtime 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[17] 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[14]. 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[15][21]. 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]. The algorithm showing the main steps taken in this work is shown as follows:

Determine the most discriminative feature vectors from I_0

For t=1 to n, (t is the time frame of video sequence session)

Get two frames I_t, I_{t+1}

Generate color invariance image map H_t, H_{t+1} *from* I_t, I_{t+1} *respectively*

Detect moving local blocks

$D_t^k \in H_t, D_{t+1}^k \in H_{t+1}$,

Generate co-occurrence matrices O_t^k, O_{t+1}^k *from* D_t^k, D_{t+1}^k *respectively*

Generate feature vectors F_t^k, F_{t+1}^k *from the normalized* O_t^k, O_{t+1}^k *respectively*

Identify the block correspondences between D_t^k, D_{t+1}^k *using* F_t^k, F_{t+1}^k

End for

Main processing steps of given algorithm is detailed in order.

3. 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[12][13]. The photometric reflectance model resulting from this theory is given by

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

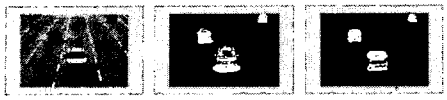
where x denotes the position at the imaging plane, λ the wavelength, $e(\lambda,\vec{x})$ the illumination spectrum, $\rho_f(\vec{x})$ the Fresnel reflectance at $\vec{x}$, and $R_\infty(\lambda,\vec{x})$ the material reflectivity. The reflected spectrum in the viewing direction is given by $E(\lambda,\vec{x})$. 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

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

Differentiating (2) with respect to λ twice and a little computation, the ratio $H = E_{\lambda\lambda}/E$ 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, assuming matte, dull surfaces, and an equal energy illumination. This color invariants may show more discriminative power than gray levels, so can be used for detection of 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 [12]).

$$\begin{bmatrix} E \\ E_\lambda \\ E_{\lambda\lambda} \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.3 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{pmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (3)$$

Figure 1 shows the comparison of resultant applications using 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.1 (a) image to be tested, (b) block detected when gray level image is used, (c) in case of H map used

4. Block detection

This module plays an important role as shown in many object tracking applications using segmentation approaches[5][11]. This module receives a pair of H maps, $H_k(x,y)$ and $H_{k+1}(x,y)$, 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 moving features detected (related to possible moving objects) is produced.

This module consists of three steps.

(1) computing the difference $DF_k(x,y)$ between the two input images $H_{k+1}(x,y)$ and $H_k(x,y)$ for obtaining spatial difference information,

$$DF_k(x,y) = | H_k(x,y) - H_{k+1}(x,y) | . \quad (4)$$

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

$$DB_k(x,y) = | H_{k+1}(x,y) - HB_k(x,y) | . \quad (5)$$

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

$$D_k(x,y) = \delta | DF_k(x,y) - DB_k(x,y) - T_k | , \quad (6)$$

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;

$$PR_\theta(r) = \int_L D(r\cos\theta - s\sin\theta, r\sin\theta + s\cos\theta) ds , \quad (7)$$

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 a distance r from the origin and s is the distance from the intersecting point to a point in binary map $D_k(x,y)$ along L. In this work, only vertical ($\theta = 0$)and horizontal ($\theta = \pi / 2$)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 (here, 15 pixels per dimension). Progressive projection is very effective and fast method to isolate the region of interest from the binary image.

5. Local block feature extraction

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 possessing proper discriminating power. There are many researches done in this area[1][2][14][18][20][21].

5.1 co-occurrence matrix

The co-occurrence matrix is a well-known statistical tool for extracting second-order texture information from images[9]. This matrix can be thought of as an estimate of the joint pdf of gray level pairs in an image. Suppose the image frame to be analyzed is an $M \times N$ dimensional matrix. An occurrence of some gray level intensity may be described by a matrix of relative frequencies $O_{\theta,d}(a,b)$, describing how frequently two pixels with gray levels a, b appear in the matrix separated by a distance d in direction θ. Non-normalized frequencies of co-occurrence as functions of angle and distance can be represented formally as:

$$O_{0^\circ,d}(a,b) = |\{((k,l),(m,n)) \in V : k-m=0, |l-n| = d, D(k,l)=a, D(m,n)=b\}|$$

$$O_{45^\circ,d}(a,b) = |\{((k,l),(m,n)) \in V : (l-n=-d, k-m=d) OR (l-n=d, k-m=-d), D(k,l)=a, D(m,n)=b\}|$$

$$O_{90^\circ,d}(a,b) = |\{((k,l),(m,n)) \in V : l-n=d, |k-m|=0, D(k,l)=a, D(m,n)=b\}|$$

$$O_{135^\circ,d}(a,b) = |\{((k,l),(m,n)) \in V : (l-n=d, k-m=d) OR (l-n=-d, k-m=-d), D(k,l)=a, D(m,n)=b\}|$$

where $|\{...\}|$ refers to set cardinality, D is the detected block and $V = (M \times N) \times (M \times N)$.

The distance metric ρ in these equations can be defined by $\rho[(k,l),(m,n)] = \max\{|k-m|, |l-n|\}$. It is required to normalize the co-occurrence matrix O so that the entries become probabilities of co-occurrence P.

5.2 features derived from co-occurrence matrix

Many features can be extracted from co-occurrence matrices[9][19]. These features are derived by weighting each of the co-occurrence matrix element values and then summing these weighted values to form the feature values. The weighting applied to each element is based on a feature weighting function, so by varying this function different texture information can be extracted from the matrix. Assume that $P(i,j)$ is the (i,j)th element of a normalized co-occurrence matrix $O(i, j)$. Then the following features are defined from $P(i,j)$.

$$energy = \sum_{i,j} P(i,j)^2$$

$$entropy = -\sum_{i,j} P(i,j) \log P(i,j)$$

$$Inverse \cdot Difference \cdot Moment(IDM) = \sum_{i,j} \frac{1}{1+(i-j)^2} P(i,j)^2 ,$$

$$inertia = \sum_{i,j} (i-j)^2 P(i,j) ,$$

$$promenance = \sum_{i,j} (i+j-\mu_x-\mu_y)^4 P(i,j) ,$$

$$variance = \sum_{i,j} (i-\mu)^2 P(i,j) ,$$

$$correlation = -\sum_{i,j} \frac{(i-\mu_x)(j-\mu_y)}{\sqrt{\sigma_x \sigma_y}} P(i,j) ,$$

where

$$P_x(i) = \sum_j P(i,j) , P_y(j) = \sum_i P(i,j) ,$$

$$\mu_x = \sum_i i \sum_j P(i,j) = \sum_i i P_x(i) \quad \mu_y = \sum_j j \sum_i P(i,j) = \sum_j j P_y(j) ,$$

$$\sigma_x = \sum_i (i-\mu_x)^2 \sum_j P(i,j) , \sigma_y = \sum_j (j-\mu_y)^2 \sum_i P(i,j) .$$

Entropy is a measure of randomness and takes low values for smooth images. Inverse Difference Moment(IDM) takes high values for low-contrast images due to the inverse $(i - j)^2$ dependence. Correlation is a measure of image linearity, that is linear directionality structures in a certain direction
result in large correlation values in this direction. Energy or angular second moment is a measure of image homogeneity-the more homogeneous the image, the large the value. There are many other derivatives, but this paper deals with these seven features just for computational convenience. But the other features can be dealt with in similar way.

5.3 class separability measures

The optimal features and other related conditions for class separability should be determined from the implementational reasons,. A major disadvantage of the traditional class separability criteria based on Bayes rules (such as Mahalanobis or Brattacharyya distances) is that they are not easily computed, unless the Gaussian assumption is employed[19]. In this paper, a set of simpler criteria built upon information related to the way feature vector samples are scattered in the I-dimensional space is adopted.

A class separability comprise two criteria, within- and between-class measures. Within-class scatter matrix is defined as follows:

$$S_w = \sum_{i=1}^{M} P_i S_i$$

where S_i is the covariance matrix for class ω_i,

$S_i = E[(X-\mu_i)(X-\mu_i)^T]$, and P_i the *a priori* probability of class ω_i. Here, $P_i \cong n_i / N$, where n_i is the number of samples in class ω_i, out of a total of N samples. Also between-class scatter matrix is defined as follows:

$$S_b = \sum_{i=1}^{M} P_i(\mu_i - \mu_0)(\mu_i - \mu_0)^T$$

where μ_0 is the global mean vector, $\mu_0 = \sum_{i=1}^{M} P_i \mu_i$. The mixture scatter matrix can be defined as follows:

$$S_m = E[(X-\mu_0)(X-\mu_0)^T]$$

That is, S_m is the covariance matrix of the feature vector with respect to the global mean. From this definitions, it is clear to see that the criterion

$$J = \frac{trace\{S_m\}}{trace\{S_w\}}$$

takes large values when samples in the l-dimensional space are well clustered around their mean, within each class, and the clusters of the different classes are well separated. Sometimes S_b is used in the place of S_m. An alternative criterion results if determinants are used in the place of traces. From these matrices, the separability measure in the one-dimensional, two-class problem, so called *Fisher's discriminant ratio* (*fdr*), can be easily derived as follows,

$$fdr = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2},$$

Since when two classes are equiprobable, the determinant of S_w $|S_w|$ is proportional to $\sigma_1^2 + \sigma_2^2$ and $|S_b|$ proportional to $(\mu_1 - \mu_2)^2$. This measure are used as the class separability criteria in this paper.

6. Object tracking

The goal of object tracking is to determine the 3-D positions and the motion parameters of objects recognized by the local block detection phase at every time instant. The first requirement for this task is to determine the correspondence of each object detected from adjacent image frames. To establish the correspondence relations of blocks between sequential frames, the selected feature values of local blocks at time t are compared with those of blocks detected at time $t+1$.

$$(x', y') = \arg\min_k D_\lambda(D_t^k(x,y), D_{t+1}^k(x', y')),$$

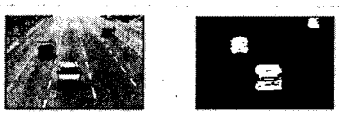
where (x,y) is the central position of D_t^k, which is the kth detected blocks in D_t. The function D_λ is defined by computing the weighted L_2 error of the transformed data using a combination of feature value difference and Euclidean distance C_t^k between the central positions of D_t^k and D_{t+1}^k.

$$D_\lambda(D_t^k(x,y), D_{t+1}^k(x', y')) = (1-\lambda)\Delta F + \lambda \Delta C .$$

The feature value ΔF is defined as the MSE between F_t^k and F_{t+1}^k and ΔC is the C_t^k. The bias term λ expresses a trade-off between the contribution of the feature value error and the Euclidean distance error. Generally, the feature value error is the most important, since it captures the spatial structure at the given image area. However, in certain cases of spatial ambiguity, the distance closeness is critical for making the correct match unambiguous. In real situation, this bias term could be adjusted dynamically if a priori knowledge about objects is available.

7. Experimental results

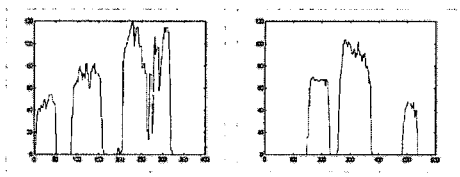
Several video clips are generated using static SONY digital video camera according to the kinds of objects and their speed variations under natural illumination conditions. Also the variety of different values of parameters associated with the morphological operator, differential operators applied to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



(a) (b)

Fig.2. (a) the original video frame tested and resultant detected blocks, (b) the moving mask,

Fig.2 (a) shows the one of the video clips tested and detected local blocks. This frame illustrates the results of well-defined block detection module. Fig.2(b)(c) shows the moving masks before and after applying opening morphological operators to eliminate the noises. The result of vertical and horizontal projections are presented in Figure 3. 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 filterings are applied using scalable Gaussian kernels [20].



(a) (b)

Fig. 3 (a) the projection on y-axis. (b) the projection on x-axis

To set up the optimal criteria for feature selection, two detected block images are used to generate 10 color invariant images by differentiating resolutions respectively. Figure 4 shows the original color images and its corresponding color invariant images. Figure 5 represents the results after computing the 7 features with given data images. The relevant euations are applied to two images(van and truck) with the distance from 1 to 10 (maximum distance could be the maximum pixel size discernible as objects to human eyes). Hence the total 300 feature values per feature category are obtained for each class.

Fig.4. three classes of color images (truck , sedan and van) to be tested.

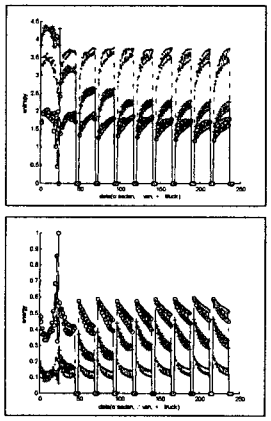


Fig.5 the feature values of images, (upper) entropy (lower) energy.

Roughly speaking, the features with more visually separated plotting along y-axis may have high class separability measures. But the scales of resultant values are quite different from each other, so the visual inspection may be imperfect unless the scale of y-axis is normalized. Figure 6 shows the FDR measures, which tells that energy may produce the highest separability measure. Promenance and variance do not show desirable discriminative power.

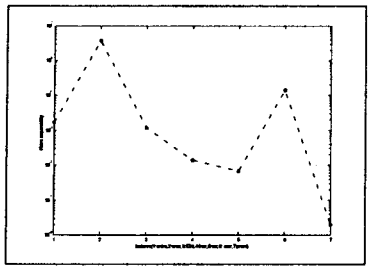


Fig. 6 FDR class separability measures (y-axis are rescaled logarithmically for visualization)

Figure 7 shows the FDR measures for each feature category depending on the different distances of co-occurrence matrix. This measure is very important, since the computation of co-occurrence density can be very intensive to be utilized in real-time applications. The choice of proper distance is critical for determining the optimal features with given images and reducing execution time complexity. This figure tells that the distances, 2, 3, 4, 9 may be safe choice, but in case of short distance, correlation, variance and promenance should be selected, otherwise IDM, energy and entropy may be the proper features.

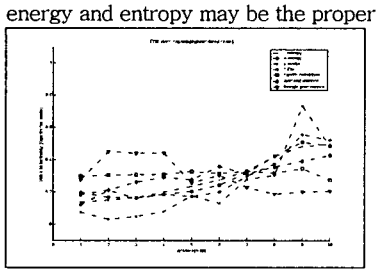


Fig. 7 FDR class separabililty measures(y-axis are normalized and rescaled logarithmically for visualization)

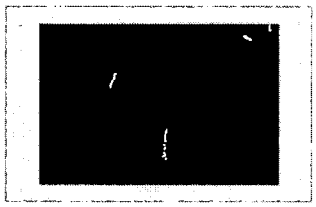


Fig.8. the detected tracks of three objects after processing 20 frames running 30fps.

Figure 8 shows the detected tracks of three vehicles of Figure 2 (a). In this experiment, d=9, and energy feature is used.

8. 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. We 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. In this paper, only one feature is used for detection of trajectories. It is easily assumed that the combined feature set could give more class separability measures. 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 Inter. Confer. 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]Haralick Shapiro,"Computer and Robot Vision Vol. 1," Addison Wesley, 1992

[10]Ismail Haritaoglu, "Real time Surveillance of people and their activities", IEEE Trans. on pattern analysis and machine intelligence, Vol. 22, No. 8, Aug. 2000

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

[12]Jan-Mak G., et. al,"Color Invariance,"IEEE Trans. On PAMI, vol.23, no. 12, Dec. 2001

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

[14]Milan Sonka, Vaclav Hlavac, Rogger Boyle " Image processing Analysis and machine vision," International Thomson Publishing Co., 1999, 2nd edition

[15]Rita Cucchira, Massimo Piccardi, Paola Mello "Image analysis and rule based reasoning for a traffic monitoring system" IEEE, Intelligent transportation system, pp119-pp130VOL.1, No.2, June 2000

[16]Robert M, Haralick, Linda G, Shapario " Computer and Robot vision Vol I ", Addision-wesley, USA pp 318-321, 1992

[17]Ross Culter, Larry S. Davis " Robust Real-Time Periodic Motion Detection, and analysis and Application ", IEEE Pattern analysis and machine Intelligence Vol22 No 8, pp781-795 August 2000

[18]Sami B. & Jorma L.,"Statistical Shape Features in Content-Based Image Retrieval," Proc. of ICPR2000, Spain, September 2000

[19] Sergios T. & Konstantinos K.," Pattern Recognition," Academic Press,1999

[20] Tony Lindelberg," Feature Detection with Automatic Scale Selection," Int. J. of Computer Vision, vol 30, number 2, 1998.

[21]Trevor Darrell and Michele Covel " Correspondence with cumulative similiarity transforms" IEEE pattern analysis and machine intelligence, pp 222-227 Vol.23 No2 February 2001