A VIDEO SYNCHRONIZATION APPROACH FOR COHERENT KEY-FRAME EXTRACTION AND OBJECT SEGMENTATION

[1]S.VIGNESWARAN, [2]DR.A.LEELEMANI, [3]K.DIVYA

[1]Research Scholar, Department of CA,Anna University Regional Centre,Coimbatore.

[2]Assistant Professor, Department of Mathematics,Anna University Regional Centre,Coimbatore.

[3]SRF,Meteorology,TNAU,Coimbatore.

[1]vickymca05@gmail.com, [2]aleelamani@gmail.com,[3]kdivyyaa@gmail.com

ABSTRACT

In this paper we discuss a new video frame synchronization approach for coherent key-frame extraction and object segmentation. As two basic units for content-based video analysis, key-frame extraction and object segmentation are usually implemented independently and separately based on different feature sets. Our previous work showed that by exploiting the inherent relationship between key-frames and objects, a set of salient key-frames can be extracted to support robust and efficient object segmentation. This work furthers the previous numerical studies by suggesting a new analytical approach to jointly formulate key-frame extraction and object segmentation via a statistical mixture model where the concept of frame/pixel saliency which is introduced and also this deals with the relationship between the frames. A modified Expectation Maximization algorithm is developed for model estimation that leads to the most salient key-frames for object segmentation. Simulations on both synthetic and real videos show the effectiveness and efficiency of the proposed method.

Keywords: *Key frame Extraction, Synchronization, Object Segmentation, Multimedia, and Coherency.*

1. INTRODUCTION

Content-based video analysis has been intensively studied during the past decades. How to represent video content in an integrated framework with good semantic structures is a topic of general interest [1]. Video segmentation, which can be categorized as temporal and object segmentations, plays a fundamental role for many video applications. According to the scene-shot-frame hierarchy, temporal segmentation splits a scene into different shots, and extracts key-frames to represent each shot. Shots and extracted key-frames can be used for video indexing/browsing, etc. Object segmentation aims to partition a video shot into meaningful objects and the background for higher level video analysis, such as object recognition/tracking, etc. Since key-frame extraction and object segmentation are usually implemented independently and separately based on different feature sets, they support content-based video analysis at different semantic levels. On the one hand, frame-wise color, motion, and texture features, which have limited semantic meaning, are used for key-frame extraction. On the other hand, object segmentation involves various pixel-wise/region-wise features, requiring more complicated and heavier computations. It can provide more semantically meaningful video analysis at the object level.

Motivated by psycho-visual studies about human perception [2], many object segmentation methods involved both spatial and temporal features with different or similar priorities. Specially, a probabilistic framework was proposed for video representation where the Gaussian mixture model

(GMM) is used to characterize visual objects in a joint spatial-temporal domain [3]. This algorithm is further extended to deal with long video shots via piece-wise GMM modeling [4]. The methods proposed in [3], [4] lead to effective object segmentation and support some object-oriented operations, e.g., object deletion/edition. However, one of major bottlenecks of these approaches is the high computational complexity for GMM estimation. It was shown in [5] that GMM-based object segmentation can be facilitated by using a small set of extracted key-frames for model estimation. The inherent relationship between key-frames and objects were further addressed in [6], [7], [8] where a unified feature space is developed to represent frames and objects simultaneously, and key-frame extraction is formulated as a feature selection process for object segmentation. Specifically, two numerical approaches were developed to search for near optimal or sub-optimal key-frame sets according to two divergence-based criteria.

In this work, we suggest an analytical method to fuse key-frame extraction and object segmentation into one closed-form, and propose a new statistical mixture model to jointly characterize key-frames and objects in the unified feature space. This work is a continuation of our previous numerical methods and is inspired by a recent work of simultaneous feature selection and model estimation [9], where the feature contribution is parameterized and estimable during model estimation. Similarly, the contribution of a frame/pixel to GMM estimation, called *frame/pixel saliency*, is introduced as a parameter in the proposed formulation. After model estimation, key-frames are extracted according to their saliency, and used to re ne GMM estimation for object segmentation.

2. PRELIMINARY

A. GMM-based Object Segmentation

In [3], [4], a multivariate GMM is used to model video data in both space and time. Every pixel in a video shot is represented by a 6-D pixel-wise feature vector x_l , which is composed of color (Y , U , V), time (t), and spatial coordinate (x and y). If a video shot contains N objects, the probability density function (PDF) of x_l is formulated as a mixture of N Gaussian components, i.e., $\emptyset=\{\theta_n, \alpha_n | n = 1 \ldots \ldots N\}$, as

$$p(X_i | \emptyset) = \sum_{n=1}^{N} \alpha_n \; p(X_l | \theta_n), \qquad (1)$$

Where α_n is the weight of the nth Gaussian characterized by $\theta_n = \{\mu_n, \Sigma_n\}$. Given L pixels, i.e.,$\{X_l | l = 1, \ldots\ldots, L\}$ the the maximum likelihood (ML) approach is used to estimate $\emptyset = \{\theta_n, \alpha_n | n = 1 \ldots\ldots N\}$, as

$$\emptyset_{ML} = \arg\max \sum_{l=1}^{L} \log p(Xl | \theta) \qquad (2)$$

The Expectation Maximization (EM) algorithm is used to solve (2) together with the minimum description length (MDL) criterion to estimate the order of GMM, i.e., N [10]. After model estimation, each object is characterized by a 6-D Gaussian, and objects can be segmented out via the maximum *a posteriori* (MAP) classification. Moreover, some object-oriented operations, such as deletion/edition, are supported by the GMM. However, the major bottleneck is the high computational load due to the fact that all video frames are used for GMM estimation.

B. Combined Key-frame Extraction and Object Segmentation

In [5], a combined key-frame extraction and object segmentation approach was proposed where a set of key-frames is first extracted via the frame-wise color histogram [11]. Also, a new feature, *intensity change* between two adjacent frames, is added to x_l . Based on key-frames, the GMM is estimated and applied to all frames for object segmentation. This approach considerably reduces the computational load, and improves segmentation performance. Meanwhile, the GMM consisting of both spatial and temporal information can support more compact and representative key-frame extraction after object segmentation. In [6], [7], the inherent relationship between key-frames and visual objects is further explicitly revealed by developing a unified feature space to represent frames and objects simultaneously. Then key-frame extraction is formulated as feature selection for best object segmentation. Specially, two divergence-based criteria, i.e., Maximum Average Inter-class Kullback Leibler Divergence (MAIKLD) and Maximum Marginal Divergence (MMD) are applied to guide the key-frame extraction process that can facilitate GMM-based video modeling. The methods in [6], [7] can provide more representative and compact key-frame sets, which lead to better object segmentation results than the one in [5] objectively and subjectively. Moreover, by exploiting the inherent relationship between key-frames and objects, extracted key-frames are more semantically meaningful. Our previous methods in [6], [7] suggest a new content-based video analysis framework where key-frames and objects can be unified from low to high semantic levels, as shown in Fig. 1.

C. Simultaneous Feature Selection and Model Learning

An integrated feature selection and model estimation method is proposed for unsupervised segmentation [9], where an important term, i.e., feature saliency, is introduced to describe the contribution of a feature to model estimation. For example, the unsupervised GMM learning is performed on a set of data samples, and each sample is a K-D vector, which means model learning is in a K-D feature space. Since K features may have different contributions to GMM estimation, some redundant features might be removed to reduce the computational load and some outliers are eliminated to improve the estimation accuracy. In [9], feature saliency is measured by the probability of relevance. A feature is irrelevant if its distribution is independent to class labels. In other words, it follows another distribution rather than the GMM. A specific EM algorithm was derived to simultaneously estimate feature saliency and the GMM.

3. Proposed Algorithm

Based on our previous work in [5], [6], [7] and inspired by [9], we hereby develop an analytical approach to jointly formulate key-frame extraction and object segmentation by introducing the concept of frame/pixel saliency.

A. *Frame/Pixel Saliency*

Given a video shot X with N objects, M frames and K pixels in each frame, we define *frame saliency* as:$\varphi_i^* \epsilon \{0,1\}, i = 1, \dots M,$ where $\varphi_i^* = 1$ means the ith frame is relevant to the GMM for object segmentation, $\varphi_i^* = 0$ means this frame is relevant to a class-independent model of outliers and useless data samples,φ_n . Similarly, we also define *pixel saliency* as $\varphi_j^* \epsilon \{0,1\}, i = 1, \dots MK,$ and let φ be a binary set for all pixels. Then frame saliency can be obtained by considering all pixels' saliency within this frame by assuming all pixels are i.i.d. Therefore, given $\daleth = (\{\emptyset, \theta_n\}$ consisting of $\emptyset$ class-independent model θ_n , for pixel x_j , we have the conditional density function as:

$$p(X_j|\varphi\daleth) = [\textstyle\sum_{n=1}^{N} \alpha_n\ P(X_j|\theta_n)^{\alpha_n}]\mathbf{q}\ \ (X_j|\theta_n)^{1-}\alpha_n \qquad (3)$$

Where $\mathbf{q}\ (X_j|\theta_n)$ is the class-independent PDF, which could be a Gaussian of very large variance, i.e., $\theta_n = \{\mu_n, \Sigma_m\}$ If we redefine frame saliency as: $P_i = P\ (\varphi_i = 1)$ pixel saliency as : $P_j = P\ (\varphi_j = 1)$ then the joint density function is:

$$p(X_j|\varphi\daleth) = [P_j\ \textstyle\sum_{n=1}^{N} \alpha_n\ P(X_j|\theta_n)^{\alpha_n}]$$
$$[\ (1\text{-}\ P_j\ )\ \mathbf{q}\ (X_j|\theta_n)\ ]^{\ 1\text{-}}\alpha_n \qquad (4)$$

In this work, frame saliency P_i is determined by averaging all pixel saliency, P_j , and frames with the highest saliency values will be selected as key-frames.

B. *A Modified EM Algorithm*

Given a pixel x_j and its class label $y_j = n, n \epsilon \{1, \dots, N\}$ denoting the association with Gaussian θ_n in , its complete data likelihood is:

$$p(X_j, y_j) = n, |\varphi\ |\ \daleth) =$$
$$[\ \alpha_n\ (1 -\ P_j\ )\textstyle\sum_{n=1}^{N} \alpha_n\ P(X_j|\theta_n)^{\alpha_n}$$
$$]\mathbf{q}\ (X_j|\theta_n)^{1\text{-}}\alpha_n \qquad (5)$$

The expectation of the logarithm of the complete data likelihood is computed as:

$$E[\log p(X,Y,\varphi\ |\daleth\ )] =$$
$$\textstyle\sum_{n,j}[P(y_i = n, \varphi_j = 1\ |X_j)(\log \alpha_n\ +$$
$$log\ (\ P_j + \log q\ (X_j\ |\theta_n))]$$
$$+$$
$$\mathbf{p}$$
$$(y_i = n, \varphi_j = 0\ |X_j)(\log(\ 1 -\ P_j +$$
$$\log q\ (X_j\ |\theta_n))] \qquad (6)$$

Let $w_j, n = p(y_j = \text{n}|X_j)$, $\mu_j, n = p(y_j = 1, \varphi_j = 1|X_j)$ and $\mu_j, n = p(y_j = 1, \varphi_j = 0|X_j)$ We derive an EM algorithm to maximize (6) below, where n = 1,…,N and j =1,…,MK

E Step:

$$\alpha_n = {\textstyle\sum_j w_j, n} \Big/ {\textstyle\sum_j P_j}$$

$$\mu_n = {\textstyle\sum_j x_j, u_j, n} \Big/ {\textstyle\sum_j u_j, n}$$

$$\Sigma_n = \Sigma_j u_j, n\, (x_j - \mu_n)(x_j - \mu_n) 1^T / \Sigma_j u_j, n$$

$$\mu_n = \Sigma_j (\Sigma_j u_j, n)^{x_j} / \Sigma_{j,n} u_j, n$$

$$\Sigma_n = \Sigma_j n\, (\Sigma_j u_j, n (x_j - \mu_n)(x_j - \mu_n)) 1^T / \Sigma_{j,n} u_j, n$$

$$P_j = \Sigma_n u_j, n$$

M Step:

$$a_{j,n} = P(w_j = 1, X_j \,|\, y_j = n) = P_j p(x_j | \theta_{n)}$$

$$b_{j,n} = P(w_j = 0, X_j \,|\, y_j = n) = (1 - P_j) q(x_j | \theta_{n)}$$

$$w_{j,n} = P\, y_j = n \,|X_j) = \frac{a_n c_{j,n}}{\Sigma_{m=1} d_m\, c_{j,m}}$$

$$u_{j,n} = P(y_j = n, w_j = 1 | X_j) = \frac{a_{j,n}}{c j_n}$$

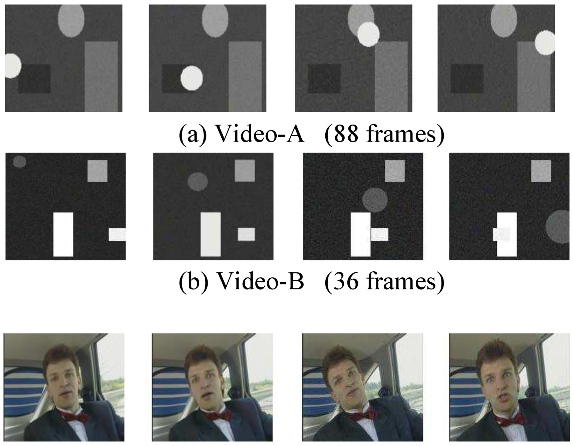
$$u_{j,n} = P(y_j = n, w_j = 0 | X_j) = w_{j,n} - u_{j,n}$$

C. Algorithm Implementation

The algorithm flowchart is shown in Fig. 2. Given a video shot, we first apply the method in [11] to extract a set of redundant key-frame candidates. Choosing initial N to be $N_{\mathbf{max}}$ in the mixture model, pixel/frame saliency and GMM parameters are estimated via the proposed EM algorithm. Frames with high saliency values are extracted as key-frames. Then the GMM is re-estimated with the MDL criteria using the extracted key-frames, and is ignored. This process considerably mitigates the computational load.

4. SIMULATIONS

The algorithm is tested on both synthetic (Video-A and Video-B) and real (Carphone) videos as shown in Fig. 3. Video-A shows a circular object moving sigmoid ally. There are two moving objects in Video-B, where an elliptic object is moving diagonally with the size increasing, and the other is a rectangular object moving leftward. We denote the method in [5] as Method-I, two numerical methods as Method-II (MAIKLD) and Method-III (MMD), and the proposed analytical method as Method-IV, respectively. Besides the subjective evaluation, objective criteria are also applied to evaluate the segmentation performance of moving objects. For synthetic videos, we compute segmentation *accuracy*, *precision*, and *recall* based on the ground truths. Accuracy is the pixel accuracy for all moving objects. Precision shows the pixel percentage that detected moving objects are true moving objects. Recall is the pixel percentage that true moving objects can be detected. For video Carphone, objective criteria are used: (1) *spatial uniformity*: texture variance (text var) within an object, color contrast (color con) along object's boundary,(2) *temporal stability:* frame difference of object elongation, size, and color histogram (elong dif f ,size dif f , ^{2}), and (3) *motion uniformity*: variance of motion vectors (montion var) [12], [13].


(a) Video-A (88 frames)

(b) Video-B (36 frames)

(c) Carphone (150 frames)

Fig. 3. Three videos (176 144) for simulation.

A. Synthetic Videos

To reduce the computational load, all methods begin with a set of key-frame candidates that are initially extracted via the color histogram [11]. Table I shows the numerical results of object segmentation as well as the number of key-frames extracted by each method. Even with less key-frames, Methods-II, -III, and -IV can provide similar segmentation results compared with

Method-I. Moreover, the analytical method uses even less key-frames than two numerical methods. This observation validates the usefulness of frame/pixel saliency in the statistical mixture model of (4).

B. Real Videos

To evaluate the effectiveness of the suggested method on video Carphone, we x the number of key-frames to be the same for all four methods, i.e., 8 key-frames. The numerical and subjective results are illustrated in Tab. II and Fig. 4. As we can see, Methods-II, -III, and -IV outperform Method-I in terms of temporal stability motion uniformity (smaller motion var), and spatial uniformity (smaller text var and larger color con). In addition, compared with Methods-II and -III, Method-IV provides similar or even better performance. In particular, Method-IV can correctly separate the bow tie from the moving face, which is mis-detected by all other three methods, leading to the significant improvement on text var.

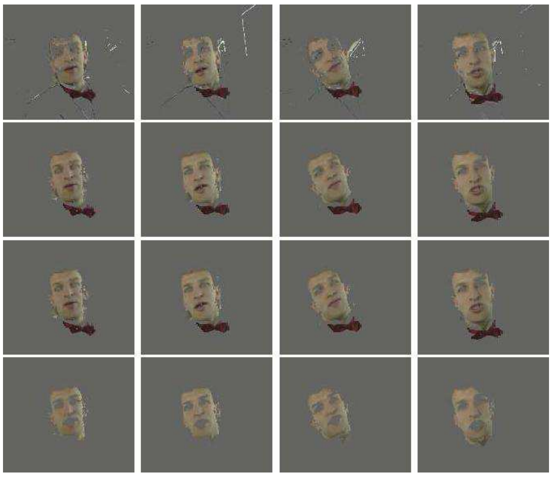


Fig. 4. Segmentation results (4 frames) of Carphone using key-frames from Methods-I (row one), -II (row two), -III (row three), and IV (row four).

5. CONCLUSIONS AND FUTURE ENHANCEMENT

The synchronization procedure is limited for real videos that is the procedure is fully device dependent(which means the procedure directly depends on the device used to capture the video, type of the data, size etc…),further it can be enhanced to deal with all types of frames(inter and intra frames).Difficulty with complex specification and offers insufficient abstraction of media object content because the media objects must be split into sub-objects.

REFERENCES

[1] P. Aigrain, H. Zhang, and D. Petkovic, “Content-based representation and retrieval of visual media: A state-of-the-art review,” *Multimedia Tools and Applications*, vol. 3, Nov. 1996.

[2] S. Gepshtein and M. Kubovy, “The emergence of visual objects in space-time,” in *Proc. of the National Academy of Science*, vol. 97, USA, 2000, pp. 8186–8191.

[3] H. Greenspan, J. Goldberger, and A. Mayer, “A probabilistic frame-work for spatio-temporal video representation and indexing,” in *Proc. European Conf. on Computer Vision*, vol. 4, Berlin, Germany, 2002, pp. 461–475.

[4] “Probabilistic space-time video modeling via piecewise GMM,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, no. 3, pp. 384– 396, March 2004.

[5] L. Liu and G. Fan, “Combined key-frame extraction and object-based video segmentation,” *IEEE Trans. Circuits and System for Video Tech-nology*, July 2005.

[6] X. Song and G. Fan, “Key-frame extraction for object-based video seg-mentation,” in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2005)*, Philadelphia, PA., March 2005.

[7] “Joint key-frame extraction and object-based video segmentation,” in *Proc. of IEEE Workshop on Motion and Video Computing (MOTION 2005)*, Breckenridge, CO., Jan. 2005.

[8] “Coherent video key-frame extraction and object segmentation,” *Submitted to IEEE Trans. Circuits and System for Video Technologye*, March 2005.

[9] M. Law and A. Zaccarin, “Simultaneous feature selection and clustering using mixture models,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1154–1166, 2004.

[10] J. Rissanen, “A universal prior for integers and estimation by minimum description length,” *Annals of Statistics*, vol. 11, no. 2, pp. 417–431, 1983.

[11] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, “Adaptive key frame extraction using unsupervised clustering,” in *Proc. of IEEE Int Conf on Image Processing*, Chicago, IL, 1998, pp. 866–870.

[12] P. L. Correia and F. Pereira, “Objective evaluation of video segmentation quality,” *IEEE Trans. Image Processing*, vol. 12, no. 2,

pp. 186–200, 2003.

[13] C. E. Erdem, B. Sankur, and A. M. Tekalp, "Performance measures for video object segmentation and tracking," *IEEE Trans. Image Processing*, vol. 13, no. 7, pp. 937–951, 2004.

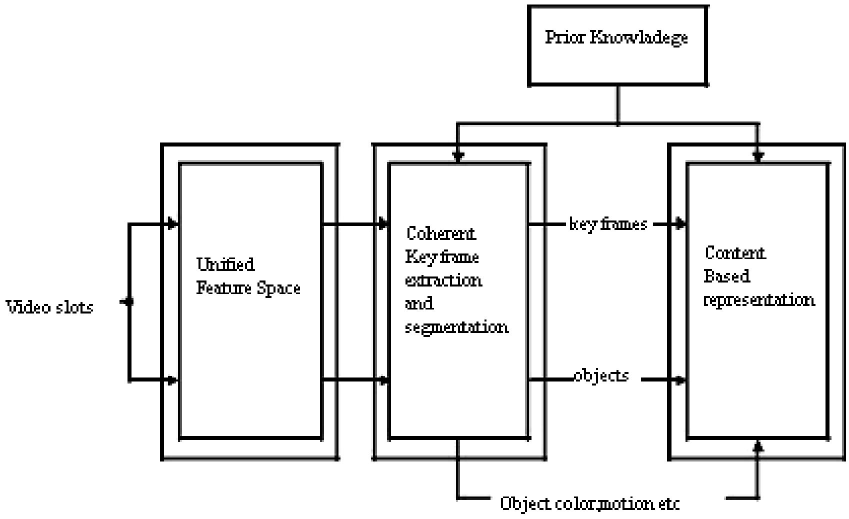


Fig 1: Proposed Video Analysis Framework in [5], [6], [7], [8].

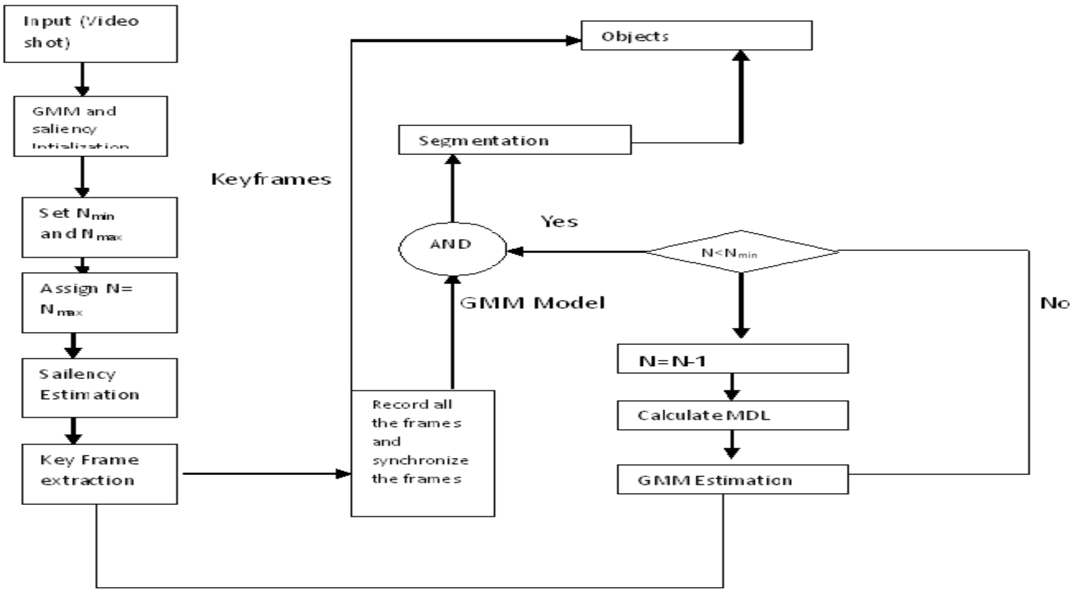


Fig. 2. The Flowchart of the Algorithm.

Table I : The Performance of Video Segmentation and the Number of Key-Frames (NKF) Extracted in each Method.

Video sequences		Method-I		Method-II		Method-III		Method-IV	
		Mean	NKF	Mean	NKF	Mean	NKF	Mean	NKF
Video-A	Accuracy	0.99	19	0.99	9	0.99	9	0.99	7
(88	Precision	0.82		0.82		0.82		0.84	
frames)	Recall	0.99		0.99		0.99		0.99	
Video-B	Accuracy	0.98	17	0.98	8	0.98	7	0.98	5
(36	Precision	0.77		0.76		0.75		0.97	
frames)	Recall	0.97		0.98		0.89		0.78	

Table II :Numerical Performance Of Carphone Segmentation.

Measurements	Method-I		Method-II		Method-III		Method-IV	
	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv
Elong dif f	1.16	1.16	1.0	1.27	1.04	1.29	0.73	0.94
Size dif f	103.2	103.4	39.69	35.73	40.06	36.69	35.26	29.39
T exture var	729.6	79.76	552.8	32.26	553.3	32.16	113.3	17.79
X^2	0.11	0.03	0.08	0.03	0.08	0.03	0.1	0.04
Color con	1.05	0.08	1.39	0.07	1.39	0.07	1.44	0.07
M otion var	214.0	95.51	188.1	106.38	188.2	109.36	158.1	58.44