Abstract

This article addresses the problem of scene estimation/abstraction of motion video data in the fuzzy set theoretic framework. Using various fuzzy geometrical and information measures as image features, an algorithm is developed to compute the change of information in each of two successive frames to classify scenes/frames. Frame similarity is measured in terms of weighted distance in fuzzy feature space. This categorization process of raw input visual data can be used to establish structure for correlation. The investigation not only attempts to determine the discrimination ability of the fuzziness measures for classifying scenes, but also enhances the capability of nonlinear frame-accurate access to video data for applications such as video editing and visual document archival systems in multimedia environments. © 1995 John Wiley and Sons, Inc.

Introduction

With rapid advancements in multimedia technology, it is increasingly common to have time-varied data like video as computer data types. Existing data base systems do not have the capability of search within such information. It is a difficult problem to automatically determine one scene from another because there are no precise markers that identify where they begin and end. Moreover, divisions of scenes can be subjective, especially if transitions are subtle.

One way to estimate scene transitions is to approximate the change of information between each of two successive frames by computing the distance between their discriminatory properties.

A gray tone picture possesses some ambiguity within the pixels due to the possible multivalued levels of brightness. The incertitude in an image may be explained in terms of grayness ambiguity or spatial (geometrical) ambiguity or both. Grayness ambiguity means "indefinite" in deciding whether a pixel is white or black. Spatial ambiguity refers to "indefinite" in shape and geometry of a region. When the regions in an image are ill-defined (fuzzy), it is natural and also appropriate to avoid committing ourselves to any specific hard decision during the process of its analysis and recognition. The relevance of fuzzy set theory in providing soft decisions for these tasks and in handling uncertainties in pattern recognition and image processing problems has been adequately addressed in the literature (Pal and Dutta Majumder, 1986; Pal, 1992; Bezdek and Pal, 1992). Because this theory is a generalization of the classical set theory, it has greater flexibility to capture faithfully the various aspects of incompleteness, imperfection, or vagueness (i.e., deficiencies) in information of a situation.

The problem of computation of features, properties, and image information on regions of a gray image, without converting it to a binary version, has been well tackled in fuzzy set theoretic framework. These include fuzzy geometrical properties of image subsets (e.g., area, perimeter, compactness, index of area coverage, length, breadth, height, width (Rosenfeld, 1984, Pal and Ghosh, 1992), and various uncertainty measures such as higher order entropy (both local and conditional), hybrid entropy, index of fuzziness, correlation, etc. (Pal and Pal, 1989a, 1989b; Pal, 1992). These measures are various extensions or generalizations of
those defined for ordinary/crisp sets, and are seen to reflect both grayness and spatial ambiguities of gray images. Various soft segmentation and skeletonization algorithms based on optimization of these measures have been reported (Pal and Ghosh, 1992).

The present article is an attempt to demonstrate the effectiveness of some of these uncertainty measures in characterizing features for analysis, scene detection, and video data extraction of motion frames. The performance has been evaluated for various combinations of these parameters when weighted distance between frames is computed as a measure of similarity. A set of digitized videos of space shuttle missions obtained from NASA/JSC was used as input. It has a time-suppressed frame rate of one per 5 sec.

**Image Definition**

Let \( X = (x_{mn}; m = 1, 2, \ldots, M; n = 1, 2, \ldots, N) \) be a digital image of size \( M \times N \) and \( L \) levels. In the notion of fuzzy set theory, this image can be viewed as an array of fuzzy singletons, each having a value of membership denoting its degree of possession of a certain imprecise property \( \mu \). Therefore, \( X \) can be written as follows (Pal and Dutta Majumder, 1986):

\[
X = \{ \mu_{mn}(x_{mn}) = \mu_{mn}/x_{mn}; \\
\{m = 1, 2, \ldots, M; \ n = 1, 2, \ldots, N \}
\]

or \( X = \bigcup_{m=1}^{M} \bigcup_{n=1}^{N} \mu_{mn}/x_{mn} \), \( m = 1, 2, \ldots, M; \ n = 1, 2, \ldots, N \)

where \( \mu_{mn}(x_{mn}) \) or \( \mu_{mn}/x_{mn} \), \( (0 \leq \mu_{mn} \leq 1) \) denotes the grade of possessing some imprecise property \( \mu_{mn} \) (e.g., brightness, edginess, smoothness) by the \((m,n)\)th pixel intensity \( x_{mn} \). In other words, a fuzzy subset of an image \( X \) is a mapping from \( X \) into \([0, 1]\) (Fig. 1). For any point \( p \in X, \mu(p) \) is called the degree of membership of \( p \) in \( \mu \).

One may use either global, local, or positional information of an image or a combination of them in defining membership function \( \mu \) characterizing some property. For example, brightness or darkness property can be defined only in terms of gray value \( f(l = 0, 1, 2, \ldots, L - 1) \) of a pixel, whereas, edginess or textural property needs the neighborhood information of a pixel to define their membership function. Similarly, positional or co-ordinate information is necessary, in addition to gray level and neighborhood information, to characterize a dynamic property of an image.

**Image Properties, Features, and Information**

There are many spatial and geometric properties or features that can be measured or extracted from an image for pattern classifications and scene analysis (Rosenfeld and Kak, 1982). However, there is no trivial solution to selecting optimal features that would provide useful input values to the classifier. The effectiveness of these feature extractors also depends upon the nature of scenes. In the following section, we describe various fuzzy geometrical properties and information measures that have been considered in the present investigation as characterizing/discriminating features for images.

Let \( \mu \) represent a brightness membership function or a fuzzy subset “object region” over an image \( X \). (Note that for a digital image, \( \mu \) is piecewise constant.) Various geometrical properties of the fuzzy image subset \( \mu \), which provide measures of ambiguity in geometry (e.g., surface, shape, orientation, boundary) of object regions \( \mu \), are mentioned below in brief. Note that we used a linear function (s type) for characterizing the fuzzy “bright” image or fuzzy “object region” \( \mu \).
**Basic Fuzzy Geometric Measures**

*Area.* The area of a fuzzy subset \( \mu \) is computed as the weighted sum of the regions on which \( \mu \) has constant value, weighted by these values (Rosenfeld, 1984), i.e.,

\[
\text{area}(\mu) = \sum \mu
\]

where the summation is taken over regions outside which \( \mu < 0 \).

*Perimeter.* The perimeter of a fuzzy image subset characterized by \( \mu \) is just the weighted sum of the lengths of the arcs along which the regions having constant \( \mu \) values \( \mu(i) \) and \( \mu(j) \) meet, weighted by the absolute difference of these values (Rosenfeld, 1984). Using co-occurrence matrices of the image \( X \), the perimeter of \( \mu \) can be computed faster as (Pal and Ghosh, 1992)

\[
\text{perimeter}(\mu) = \sum \sum c[i, j][|\mu(i) - \mu(j)|]
\]

where \( i, j = 1, 2, \ldots, L \); \( c[i, j] = c[i, j]_h + c[i, j]_v \); and \( c[i, j]_h \) and \( c[i, j]_v \) denote the number of times the \( i \)-th level has occurred with the \( j \)-th level in the horizontal and vertical directions, respectively.

*Length.* The length of a fuzzy image subset \( \mu \) means the longest extent in the column direction (Pal and Ghosh, 1992), i.e.,

\[
\text{length}(\mu) = \max_{i} \left( \sum_{j} \mu_{i,j} \right).
\]

*Height.* The height of a fuzzy image subset \( \mu \) is the sum of the maximum membership values of each row (Rosenfeld, 1984), i.e.,

\[
\text{height}(\mu) = \sum \max_{j} \mu_{i,j}.
\]

*Breadth.* The breadth of a fuzzy image subset \( \mu \) measures the longest extent in the row direction (Pal and Ghosh, 1992), i.e.,

\[
\text{breadth}(\mu) = \max_{j} \left( \sum_{i} \mu_{i,j} \right).
\]

*Width.* The width of \( \mu \) is calculated as the sum of maximum membership values of each column (Rosenfeld, 1984), i.e.,

\[
\text{width}(\mu) = \sum_{i} \max_{j} \mu_{i,j}.
\]

**Orientation Measures**

Note that (Pal and Ghosh, 1992)

\[
\frac{\text{length}(\mu)}{\text{height}(\mu)} \leq 1, \quad \text{and} \quad \frac{\text{breadth}(\mu)}{\text{width}(\mu)} \leq 1.
\]

If

\[
\frac{\text{length}(\mu)}{\text{height}(\mu)} = 1,
\]

then \( \mu \) is vertically oriented. If

\[
\frac{\text{breadth}(\mu)}{\text{width}(\mu)} = 1,
\]

then \( \mu \) is horizontally oriented. Therefore, orientation of the fuzzy subset \( \mu \) can be determined from the above-mentioned ratios.

**Shape Measures**

Two shape measures of \( \mu \) (or fuzzy object region) can be computed using the aforesaid basic geometrical properties.

*Compactness.* The compactness of \( \mu \) means the fraction of maximum area (that can be encircled by its perimeter) actually occupied by the fuzzy object region. It is defined as (Rosenfeld, 1984)

\[
\text{comp}(\mu) = \frac{\text{area}(\mu)}{\left( \text{perimeter}(\mu) \right)^2}.
\]

*Index of Area Coverage.* The index of area coverage (IOAC) of \( \mu \) is the fraction (which may be improper also) of the maximum area (which can be covered by its length and breadth) actually covered by the fuzzy object region. It is defined as (Pal and Ghosh, 1992)

\[
\text{IOAC}(\mu) = \frac{\text{area}(\mu)}{\text{length}(\mu) \times \text{breadth}(\mu)}.
\]

**Entropy and Edge Ambiguity Measures**

Two measures of image uncertainty and edge ambiguity used here are the second-order local entropy and edginess index. They reflect measures of business/coarseness in structural parts or information that exists in a given image. The
second-order entropy of an image provides the information gain based on the probability of the co-occurrence of pixels. The edginess index provides the degree of coarseness (abruptness) of regions in an image based on neighborhood information of pixels and degree of fuzziness in a set.

**Second-Order Entropy.** Let \( c_{ij} \) be the number of times the \( i \)th level has occurred with the \( j \)th level in the horizontal or vertical direction. Then, the probability of such co-occurrence of the \( i \)th and \( j \)th levels, and the second-order entropy are defined as (Pal and Pal, 1989b)

\[
\rho_{ij} = \frac{c_{ij}}{\sum_{i,j} c_{ij}}, \quad \text{where} \ 0 \leq \rho_{ij} \leq 1
\]

\[
H(X) = -\frac{1}{2} \sum_{i,j} \rho_{ij} \log(\rho_{ij}) \tag{11b}
\]

respectively. Here the information gain is computed with Shannon's logarithmic function. As described (Pal and Pal, 1989a), this could also be an exponential function. The definition of the co-occurrence matrix could also be modified with a combination of the horizontal and vertical directions to take into account the spatial distribution in both directions. Note that second-order entropy does not involve any concept of membership function.

**Index of Edginess.** The index of edginess \( I(X) \) gives a measure of average edge information in an image by computing the edge ambiguity at every point using a localized window. It is defined as (Pal, 1986)

\[
\delta(X) = |1 - I(X)|^\beta \tag{11a}
\]

\( I(X) \) stands for the index of fuzziness, i.e., average amount of difficulty in deciding whether a pixel possesses an edginess property \( \mu \) or not. (It has the characteristic of increasing/decreasing monotonically with \( \mu \) in the interval \([0,0.5]\) with a maximum at \( \mu = 0.5 \).) \( \beta \) is a positive constant. \( I(X) \) is computed with the spatial dependent membership function, \( \mu, \) of the form

\[
\mu_i(x_{mn}) = \frac{0.5}{1 + \frac{1}{N_i} \sum_{j} |x_{mn} - x_{ij}|} \tag{11b}
\]

which lies in \([0,0.5]\). Here \( N_i \) represents the number of neighboring pixels \( x_{ij} \) of the point \((m,n)\). If \( I(X) \) is considered to be the linear index of fuzziness (as shown in Fig. 2) then

\[
I(X) = \frac{1}{2MN} \sum \min(\mu_i(x), 1 - \mu_i(x)) \tag{11c}
\]

when \( i = 1, 2, \ldots, MN \). Note that other measures of fuzziness, such as the quadratic index of fuzziness, the fuzzy (global) entropy, and the index of nonfuzziness (crispness) (Pal and Dutta Majumder, 1986; Pal, 1992) could also be used as \( I(X) \).

**Scene Estimation/Abstraction**

A method of abstracting key image frames from a sequence of video data is described here where the aforementioned image ambiguity (fuzziness) and information measures have been considered as characterizing features for classifying scenes.

**Feature Space and Discrimination Criteria**

As discussed in the literature (Pal and Dutta Majumder, 1986; Duda and Hart, 1973; Tou and Gonzalez, 1974), the criterion of a good feature is that it should be invariant within class variation while emphasizing differences that are important in discriminating between patterns of different types. It is difficult to determine an optimal feature space comprising a set of image properties that would produce significant factors influential to the best classification decision. The approach taken here for determining important features is to select image properties of three kinds, namely coarseness (uncertainty), shape, and orientation measures from those defined earlier. In other words, represent an image as a...
vector with components representing these properties such that it corresponds to a point in an N-dimensional vector space. Figure 3 depicts the sampled feature space for $N = 3$ where

$$f_i = \left[ f_{i_1}, f_{i_2}, f_{i_3} \right]$$ and $$f_j = \left[ f_{j_1}, f_{j_2}, f_{j_3} \right]$$

represent two image frames and $f$ represents image property. The distance, $|d_i|$, between two frames can be calculated with the vector operation, $|f_i - f_j|$ to measure their similarity in terms of proximity. The larger the distance, the less the similarity between frames.

Because all the features (image properties) are not equally important or significant in characterizing image patterns of motion frames under consideration, it is reasonable and justified to take this fact into account by incorporating some weight corresponding to each feature while computing the aforementioned distances. The larger the variance of a feature, the smaller is its weight, so that the important (i.e., reliable) features with small variance have more influence in the decision-making process.

Our aim is to analyze motion frames, and the computation of the change of image constituents (properties) from frame to frame in a given time series gives the sampled mean and the sampled variance of all image features. It is also discussed (Duda and Hart (1973) and Tou and Gonzalez (1974) as a useful clustering/classification criterion to maximize the interframe distance and/or minimize intraset distance using a diagonal transformation such that features having larger variance are less reliable.

### Weighted Distance Computation

Let $T$ and $Q$ represent the total number of frames (images) and features (properties), respectively. The sampled mean for the $j$th feature element is given by

$$\bar{f}_j = \frac{1}{T} \sum_{i=1}^{T} f_{ij} \quad \text{where} \quad j = 1, 2, \ldots, Q.$$ (12)

Mnemonically, the index of feature element $j$, where $j = 1, 2, \ldots, Q$, represents the following enumerated terms: edginess, entropy, compactness, IOAC, length/height, and breadth/width, respectively (e.g., $f_{i_{\text{entropy}}}$).

The sampled variance for the $j$th feature element is computed as

$$\sigma_j^2 = \frac{1}{T} \sum_{i=1}^{T} (f_{ij} - \bar{f}_j)^2.$$ (13)

The magnitude of the weighted distance between two successive frames $i$ and $k$ is defined as

$$D_{\text{weight}} = \sqrt{\sum_{j=1}^{Q} \left( \frac{(f_{ij} - f_{kj})^2}{\sigma_j^2} \right)}.$$ (14)

### Schematic Diagrams

Based on the above-mentioned criteria, two schematic diagrams (Fig. 4) are drawn to describe the process of feature selection and frame selection. After the important features $f_{j_{\text{important}}}$ are selected (Fig. 4(a)) for representing images, the weighted distance $D$ [eq. (14)] between two consecutive frames in the prescribed feature space is computed to check their similarities. The higher the distance, the less the similarity. Therefore, if this distance is larger than a predetermined threshold value, then the current video frame is considered to be significantly different from the previous frame, and therefore needs to be registered or stored as one of the abstract keys (Fig. 4(b)). Another way of detecting abstract frames could be by selecting only those that correspond to peaks (maxima) in $D$ values, assuming that there exist at least two consecutive similar frames in the input sequence.

### Input Data and Results

Movie film projectors display 24 frames per second whereas NTSC standard television and video devices display 30 frames per second to achieve continuous and fluid full-motion images. The change of interframe information is gradual.
at such high frame rates. For storage conservation and computational efficiency, the simplest way to reduce or abstract video data is to sample it at lower frame rate.

In this investigation, a time-suppressed frame rate of one per 5 sec is assumed (Rorvig, 1991). A set of digitized videos of previous space shuttle missions obtained from NASA/JSC was used (Fig. 5). The scenes were named payload deployment, onboard astronaut, remote manipulator arm, and mission control room. After a preprocessing step, each frame is stored in the CompuServ's Graphic Interchange File (GIF) format for portability. A frame size of 104 x 78 pixels with 256 gray levels was used.

Experiments were conducted for various combinations of uncertainty, orientation, and shape measures. As an illustration, we present the results corresponding to only three sets of detection with the input features as follows:

1. entropy, compactness, length/height (Fig. 6);
2. edginess, IOAC, breadth/width (Fig. 7); and
3. all of the above (Fig. 8).

The resulting distances between every two successive frames in the aforesaid three feature spaces are shown in Figures 6-8. The abscissa represents the total number of frame distances in the sampled time series, and the ordinate is the compound distance value between two successive images, i.e., \(|F_i - F_{i+1}|\). For example, the abscissa index 0 represents \(|F_0 - F_1|\), 1 represents \(|F_1 - F_2|\), and so on. Each scene consists of six frames. Therefore, there is a change of scene at every sixth index on the abscissa. The scene separation is denoted with vertical grid lines. (Note that if the sequence of frames/scenes is changed, then it would be reflected in the graphical variation of Figs. 6-8.)

As mentioned before, peaks in Figures 6-8 indicate abrupt changes (i.e., dissimilarities) between the respective frames. The discrimination between the similar and dissimilar sets of frames is seen to be more intensified (prominent) for the feature sets in Figures 6 and 8. Here, the image frame numbers 8, 9, 13, and 19 (Fig. 5) correspond to the abscissa numbers 7, 8, 12, and 18 (Figs. 6-8) having larger distance values. It means image frames 8, 9, 13, and 19 together with 1 (reference) can be stored/registered as the abstract keys of the input sequence. Note that for scene 2, two frames need to be stored. Note also that combining all features...
Figure 5. A payload deployment sequence of four scenes as input data.

Figure 6. Distances between successive frames with feature set [entropy, compactness, length/height].

MOTION FRAME ANALYSIS AND SCENE ABSTRACTION 253
does not necessarily produce better results just because there are more features. It is not the quantity that is critical, but the discriminatory quality of features. In this context, the feature set \{entropy, compactness, length/height\} corresponding to Figure 6 can be said to be best characterizing.

Some more results using the feature sets, namely \{entropy, compactness, length/height\} and \{edginess, IOAC, breadth/width\} for a different payload deployment sequence of four scenes (Fig. 9) are given in Figure 10 in this context.

**CONCLUSION AND DISCUSSION**

An investigation was made demonstrating the effectiveness of some of the fuzziness/ambiguity measures as characterizing features for motion frame analysis for video data abstraction. Weighted distance, with weights being inversely proportional to the variance of features, was considered as a measure of dissimilarity between frames. The feature set consisting of compactness, second-order entropy, and the ratio length/height of images is seen to possess good discrimination ability for such analysis. The investigation also enhances the capability of nonlinear, frame-accurate access to video data for application such as video editing and visual document archival/retrieval systems in multimedia environments.

It may be mentioned that the conventional method is to segment first the image into hard partitions, and then to compute various image properties on the crisp segmented regions. Because the regions in the gray images are usually ill-defined, it is not appropriate to commit ourselves to hard thresholding that may increase uncertainty in the final decision-making process. To avoid the uncertainty, we used fuzzy measures directly on the input image for computing image properties. There are some nonfuzzy
(classical) measures (e.g., entropy, angular second moment, contrast, homogeneity) (Haralick et al., 1973) that could have also been computed. But the kind of information provided by non-fuzzy measures is usually different from that of the fuzzy measures used in our investigation. For example, entropy (classical) and fuzzy edge ambiguity, and compactness and IOAC provide a similar type of information, and therefore they can be compared. This is illustrated in Figure 11 for the sequence of frames shown in Figure 5. Classical measures analogous to fuzzy IOAC or compactness, and length/height or breadth/width are not known to us.

However, the technique described here needs further improvement. This includes development of a procedure for automatic selection of thresholds for detecting the key frames. Scene classification is quite subjective in nature; therefore, the interactive tools can be developed for providing human interaction in setting problem-dependent criteria for this machine recognition task. Furthermore, a hierarchical abstraction scheme that allows a higher level of abstraction will better suit the visual data management environment.

Finally, in the emerging worlds of computers and media, new technologies mix traditional media such as video and publications with computer media as interactive, informational, and entertainment software. This trend is rapidly growing at an unprecedented rate. Once digital video becomes a repository of common data on computers, the data needs to be accessed and manipulated just as documents are retrieved and managed by a data base management system. It might be useful to investigate new video inter-referencing strategies in correlating various contexts from the same event to derive knowledge points. Thus, this automatic abstraction of video index keys for nonlinear, frame-accurate access would make information archival and retrieval applications more robust and efficient.

The authors would like to thank Dr. Robert N. Lea of NASA/JSC for his interest in this work. This research was done as a master’s thesis by Albert Leigh at the University of Houston–Clear Lake (supported partially by a NASA/JSC fellowship, “Graphics Technology in Space Application Program”) while Prof. Pal held a NCR-NASA Senior Research Award to work at Johnson Space Center, Houston, TX.

**References**


Received March 1994
Accepted January 1995