A Method for Automatically Abstracting Visual Documents*

Mark E. Rorvig
Software Technology Branch, Lyndon B. Johnson Space Center, National Aeronautics and Space Administration, Houston, TX 77058

Visual documents—motion sequences on film, videotape, and digital recordings—constitute a major source of information for the Space Agency, as well as all other government and private sector entities. This article describes a method for automatically selecting key frames from visual documents. These frames may in turn be used to represent the total image sequence of visual documents in visual libraries, hypermedia systems, and training guides. The performance of the abstracting algorithm reduces 51 minutes of video sequences to 134 frames; a reduction of information in the range of 700:1.

Introduction

Although the application of visual documentation techniques has been expanded manyfold in the last decade due to steady reductions in cost, methods for summarizing these documents have remained bound by human editing procedures. Such procedures are typically subject to high costs as well as variations and biases introduced by individual editors possessing different training backgrounds and aesthetic temperaments (Pryluck, Teddlie, & Sands, 1982). While significant work has been done in identifying sources of descriptive information for visual documents (O'Connor, 1985, 1986), it is curious that no work has been done to abstract or index visual documents with visual exemplars directly. Indeed, the most closely related work has been conducted around the problem of data compression algorithms (Yeh et al., 1991).

In the method introduced in this article, however, frames of the visual document are digitized and subjected to a structural decomposition process that reduces all information in the image to sets of values. These values are in turn normalized, further combined to produce only one value per frame, and fitted to a normal distribution of all values in a defined training set of frames. By selecting only those values at specified areas at the tails of the distribution, key frame images may be abstracted from their surrounding frames.

Methodology

Consider a visual document composed of 30 frames of interleaved video or film per second as a sampling universe. For each second, little change in any frame occurs such that, in the method of this article, a sampling rate of one frame of video imagery per every 5 seconds constitutes the sampling frame. The problem of visual abstraction thus devolves into the determination of a method for selecting significant frames from among the reduced set extracted from the original run of frames. (It should be noted that, in some cases, a higher density of sampling may be preferred. The interval of 5 seconds was chosen arbitrarily for the demonstration of this method. No claim is made for any optimal sampling rate.) The collection of images thus sampled from the visual document shall be denoted "C."

Consider the collection of images C. Assume that each image in this collection has been digitized and processed so as to extract a number of general, primitive features rendered as histograms. Specifically, in the demonstration of this article, NTSC standard VHS video frames were sampled at intervals of 5 seconds each, digitized by a commercially available analog frame digitizer and stored as PICT format files with a common XY dimension. Further, although Figure 1 suggests the use of hue, chroma, and

*This work was performed under the terms and conditions of the *Memorandum of Understanding Between the National Aeronautics and Space Administration Lyndon B. Johnson Space Center (JSC) and the University of Texas at Austin* as signed and dated by the authorities of the respective institutions on March 26, 1991 and March 13, 1991 and transmitted by NASA JSC Reply Reference AL4-91-105. The method described in this article is a NASA-owned invention (MSC-22093-1). Inquiries for use may be made to Mr. Hardie Barr, Patent Counsel, NASA Johnson Space Center, AL3, Houston, TX 77058; tel.: (713) 483-1003.

Received May 6, 1992; revised September 4, 1992; accepted September 4, 1992.

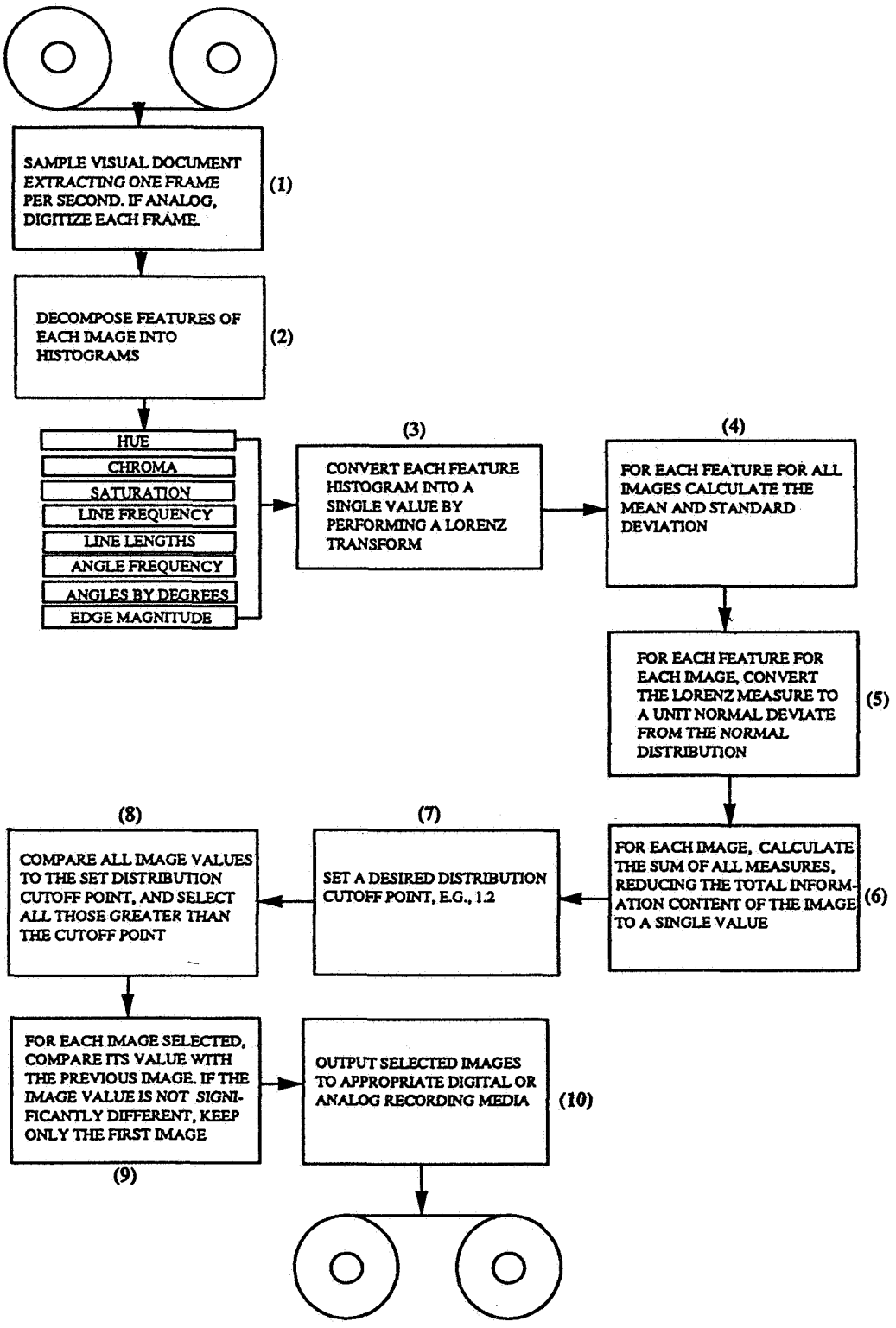

FIG. 1. The complete method for automatic abstraction of visual documents, consisting of 10 steps.

saturation, for this demonstration only grey levels were extracted. Other extracted features were edge intensity, edge slope, line length, line distance from the origin, and angles. The feature extraction programs were written in C language. The pixel intensity histogram accumulated grey-scale values in 64 intervals. All other values were calculated after performing a Hough transform. Edge intensity was defined as constant grey scale values of greater

than five pixels in width and accumulated in 64 histogram intervals. Edge slopes were accumulated in 45 histogram intervals of 2 degrees each. Lines were defined as constant grey scale values of less than 5 pixels and accumulated in 64 histogram intervals of 4–256 pixels. Line distance from the origin was defined as the number of pixels between the center of the line and the largest values of *XY* coordinates of the image and accumulated in 64 histogram intervals of 45–256 pixels. Angles of lines were accumulated in 45 histogram intervals of 2 degrees each. The algorithms for these processes are quite standard (Ballard & Brown, 1982). No claim is made that these features are the only possible features that might be used, nor even that these features are optimal. Other work suggests that fuzzy measure approaches may be more effective (Leigh, 1992).

By converting the histograms for each image into Lorenz information measures (Chang & Yang, 1983), each image of collection C may be reduced to a small set of real numbers, where each number constitutes a structural attribute of the entire image. However, since the abstraction method selects images based on the relative position of each image in the normal distribution of all images in collection C, each individual structural feature of the images must be assumed to be normally distributed. This assumption may be incorrect however, depending upon the composition of any set of particular images decomposed by any particular feature. To overcome the weaknesses of such an assumption, the structural features of each image are themselves averaged by calculating the mean and standard deviation for each measure across the collection C and converting the individual measures to unit normal deviates of the normal curve. Thus, all values are rendered to a common unit of measurement. By simply summing all the measures of an image, each image may be represented by a single value encompassing its entire structure (Shelton, 1991). These image values, now implicitly constituting a training set, are also processed to derive their mean and standard deviation, with the selection rule set to retain all values representing images at the tails of the normal distribution.

To demonstrate the method, approximately 51 minutes of video including one shuttle launch sequence, one shuttle landing sequence, one satellite deployment, and numerous onboard experiments were processed by the method of Figure 1, steps 1–8. Sampling at 5-second intervals from the 51 minutes yielded 604 frames. With the selection threshold set at ± 1.2 *SD*, 134 (approximately 22%) of the images were retained (Project ICON Laboratory, 1991). It should be noted that both the training set and the threshold parameters may be altered for the convenience of the user. Table 1 lists all scenes, the number of frames sampled from each scene at 5-second intervals, and the number of frames selected by the algorithm. Scenes discussed in detail in the results section are included in the Appendix and marked in Table 1 with an asterisk. Frames selected by the algorithm in the Appendix are those without cross marks. The number of frames in a scene may not be equal to the number of frames noted below due to space constraints for publication.

Results

Shuttle launch and landings (Figs. 2–3 and 15–17 in the Appendix) are significant events. Thus, the composition choice of the training set for these scenes at 11% of the total frames resulted in higher sampling rates by the algorithm than for the collection as a whole. Specifically, the average sampling rate of 22% for the collection is increased to 50% for the launch and to 40% for the landing. Essentially, the algorithm samples frames in inverse proportion to the appearance of scene types in the document, where scene "type" refers to the average composition of the image set in terms of its component light levels, edges, and other correlates of the image decomposition method described in the methodology. More significantly, however, is that the frames extracted from the total also represent significant launch and landing events. For example, Figures 2–3 contain frames from shuttle roll, full thrust, and Solid Rocket Booster separation; three of the more volatile events in any launch. Similarly, Figures 15–17 include significant frames from the first appearance, shift to descent attitude, first appearance on the horizon, landing gear deployment and roll-by. The algorithm failed, however, to select two frames in both these sequences that are also important, specifically ignition in the launch sequence and touchdown in landing.

Figures 6–7 and 11–14 represent the inverse frequency effect also. Onboard activity sequences formed 64% of the training set. Therefore, they were sampled at a lower rate than the entire collection of frames. Specifically, for these two scenes, the sampling percentages were approximately 10%. Figures 6–7 contain a long sequence of frames. The sampled images collapse this sequence into a storyboard of (1) an astronaut removing a panel from the shuttle aft bay; (2) a close-up of the removal; (3) display of an experimental panel with its documentation highlighted in the background; and (4) display of another experimental panel with its documentation also highlighted. Figures 11–14 consist of a food preparation sequence in microgravity. Although such scenes are generally of little scientific interest, they capture an important dimension of space flight in human terms. In this case, the sampled images collapse the sequence into a storyboard of (1) three frames showing successively more items prepared for cooking; (2) setting of a timing device; (3) two additional frames displaying the loading of the items into the microwave oven; and (4) one final frame of the astronaut apparently checking cooking instructions.

Figure 4 displays a flaw in the algorithm, corrected in Figure 1 by steps 9 and 10. Specifically, Figure 4 exhibits frames that fell at the tails of the normal distribution of images. Due to their similarity they were all selected. The 9 frames of the Remote Manipulator System (RMS) shown in Figure 4 represents this aspect of the algorithm in its worst case. First, the RMS is a member of the 25% of scenes comprising vehicle deployment and related activity, so that, due to the inverse frequency effect noted earlier, the sampling rate would tend to be higher than the onboard scenes in any case. Second, since the RMS moves very slowly during the satellite deployment process,

TABLE 1. Scene sequence of video used in the demonstration of the automatic abstracting method.

Scene Description	No. of frames	No. Selected
1. Shuttle launch*	32	17
2. Shuttle bay doors opening prior to deployment	23	2
3. Flight deck (forward) scene shift before deployment	8	0
4. Remote Manipulator System lift of vehicle from bay	31	6
5. Flight deck (aft) scene shift	6	1
6. Astronaut exercising in weightless environment	26	22
7. Detail of the Remote Manipulator System effector	18	4
8. Remote Manipulator System arm and effector*	9	9
9. Interior room of the mission control building during flight	6	0
10. Scene shift to space vehicle emerging from shuttle bay	1	1
11. Unidentified out of focus scene (appears as flat grey panel)	7	0
12. Split screen for astronauts and vehicle emerging	5	0
13. Astronauts standing on flight deck (aft)	10	1
14. Alternating frames of mission control and shuttle in space on Earth limb*	5	3
15. Window reflection of interior light	4	0
16. Spacecraft view against Earth at oblique angle	8	2
17. Astronaut emerging from rigid sleep sation and mid-deck (starboard aft)	13	2
18. Astronaut being thrown towel in weightless environment	24	4
19. Pan scene of mid-deck	9	1
20. Pan scene of mid-deck	4	0
22. Astronaut (standing) describing equipment readout facility for camera	24	4
23. Alternating views of flight deck (aft) and mission control room	5	5
24. Astronaut pulling experiment racks for camera on flight deck (aft)*	44	3
25. Astronaut emerging from forward to aft flight decks	3	2
26. Pan shot of empty flight deck (aft)	3	0
27. Astronaut examining flight hardware*	24	0
28. Astronaut examining manuals and experiment locker on flight deck	47	6
29. Equipment floating in flight deck (forward)	10	3
30. Water globule experiment	4	2
31. Acoustic levitation experiment*	15	0
32. Meso-scale lightning experiment*	21	4
33. Astronaut emerging from rigid sleep station	9	2
34. Miscellaneous onboard activities including view of oscilloscope	36	2
35. Alternating view of mission control and ocean with clouds	5	0
36. Shuttle onboard food preparation*	66	8
37. Sunrise over the Earth limb	5	1
38. Shuttle landing*	34	17

all the scenes decompose into very close values. Steps 9 and 10 of Figure 1 could be performed in many ways. However, the most direct one would simply be to observe the items selected and set a cut-off value relevant to the training set and the characteristics of the population of visual documents to be filtered by the algorithm.

Finally, of the 38 scenes in the visual document, 10 were unsampled. Of these 10 scenes, only 2 unsampled scenes (Figs. 8 and 9) pose a problem. In all other cases, the scenes skipped include material of little interest. Figures 8 and 9 are events of interest. However, the algorithm fails to include even one frame because in both cases, there was simply insufficient contrast in the images. Unfortunately, the cylinder examined by the astronaut in Figure 8 is almost the same light level as the background cabinet. Similarly, the glass beads suspended by acoustic pressure in Figure 9 do not differ significantly from the plate background because they are clear. This is similar to the case in Figure 6 when the astronaut is pulling experiment racks; only the frames showing racks with the documentation in white held behind them are selected by the algorithm. Since the technique used in this demonstration examines only grey levels, appropriate frames from Figure 9 remained unsampled. In Figure 1, however, it is suggested that a more sophisticated measure of light should be used, specifically a three-set histogram for hue, saturation, and chroma. Whether or not this procedure would correct this deficiency of the algorithm's performance remains unknown at this time. Another potential solution also exists. Specifically, the normal curve area values could be expanded to select a greater number of frames. This solution, coupled with a parameter to reject frames with closely similar values (steps 9 and 10 in Fig. 1), might produce the best results.

Discussion

Given the light level analysis technique and normality assumptions of the algorithm used for this demonstration, the performance of the abstracting method appears to be

acceptable for a large variety of visual documentation representation tasks. The errors of the algorithm are, in contrast to human editing procedures, at least due to traceable causes and exhibited as systematic error. Knowing that the filtering algorithm will focus on scenes of high contrast is quite different from attempting to determine the aesthetic sense of a human editor.

Although the most important near term requirement for the algorithm is the conservation of storage space in machine-readable libraries, the broader application is the general conservation of human attentiveness. It has been observed that the fundamental problem of librarianship is that data are infinite and attention is scarce. Thus, the broadest application of the abstracting method is in the presentation of the central contents of visual documents in a manner such that minutes of visual imagery may be reduced to merely a few frames. Another class of applications is the general problem of machine vigilance. This problem arises in security systems, where cameras might scan constantly, while human attention is limited to a few intermittent pans. Machine surveillance of many industrial processes are also candidates for this algorithm. In all cases, a computer could be programmed to sound an alarm whenever a scene fell outside the prescribed normal distribution limits.

Overall, the general performance of the algorithm may also indicate that the assumption of Marr regarding the paucity of information necessary for the human neocortex to assess imagery is correct. Consider the model of human attentiveness on which the abstracting method is based: at all times, we create a running probability distribution of scenes in our minds; the distribution changes with the addition and deletion of sensation; we attend to outliers in the distribution. Although sparse, the model has the explanatory power to account for a phenomena which we have all experienced: in a crowded room we are much less likely to notice the entry of one more person; in a nearly vacant room, we are much more likely to notice it. Moreover, when it is considered that the model of attentiveness used for this algorithm is operationalized by processing frames to a single representative value, the temptation to construe human vision as primarily an internal mental function rises significantly.

Appendix

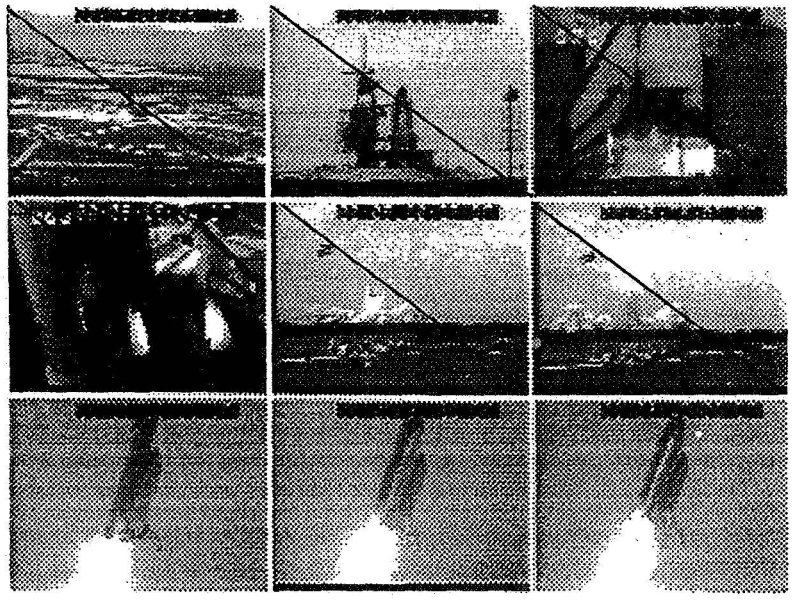

FIG. 2. Scene 1: Shuttle launch sequence, part 1.

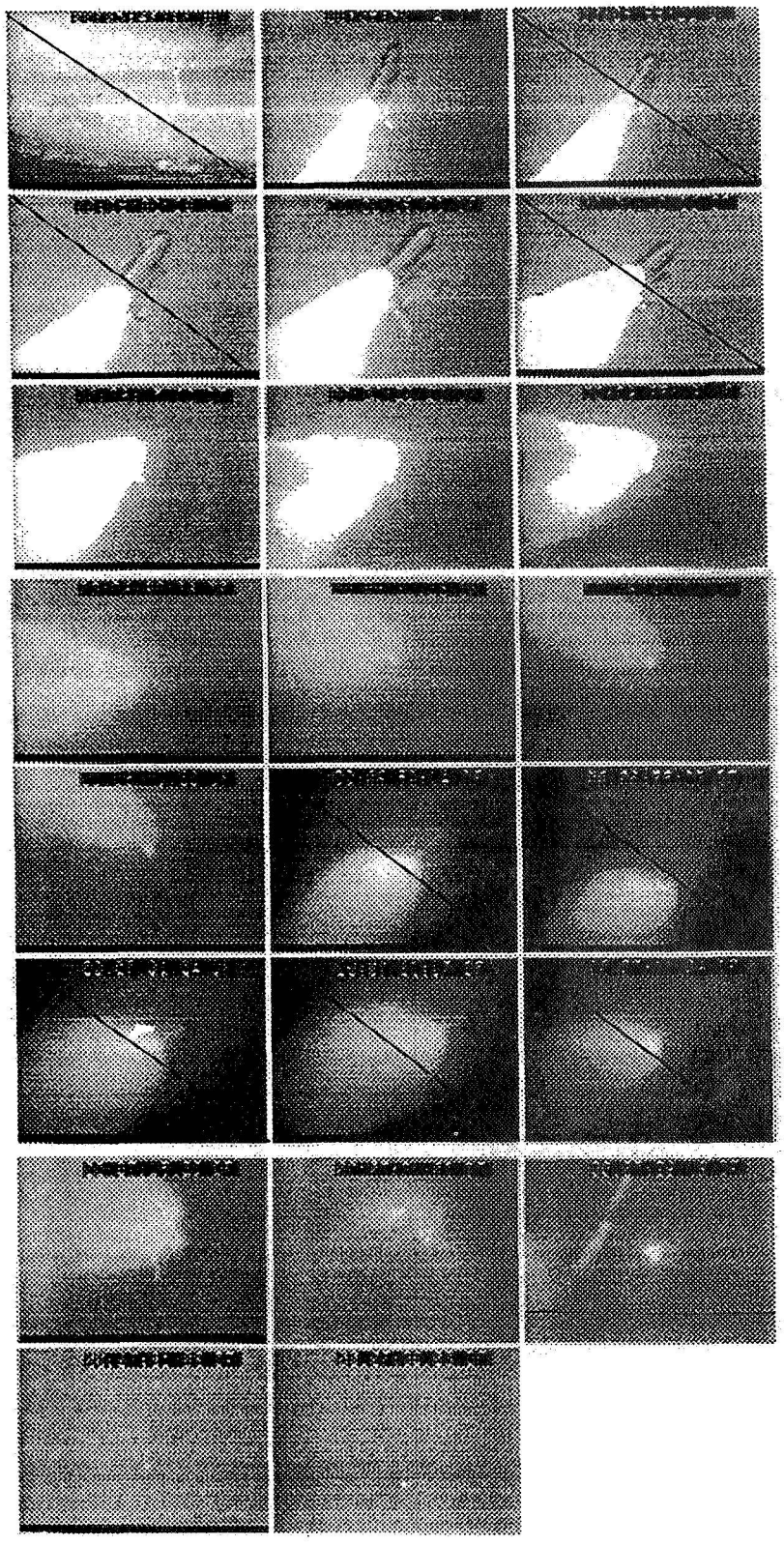

FIG. 3. Scene 1: Shuttle launch sequence, part 2.

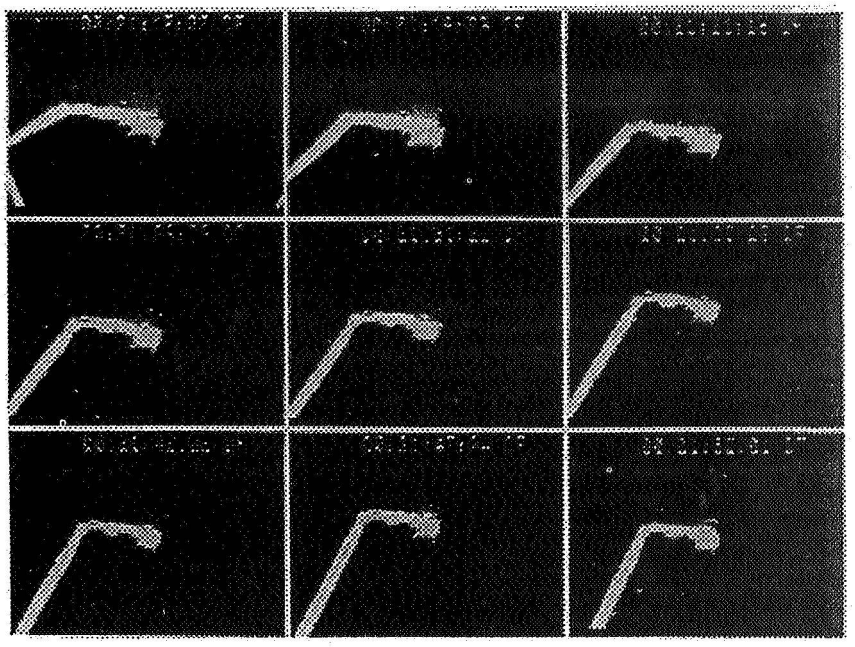

FIG. 4. Scene 8: Remote Manipulator System arm and effector.

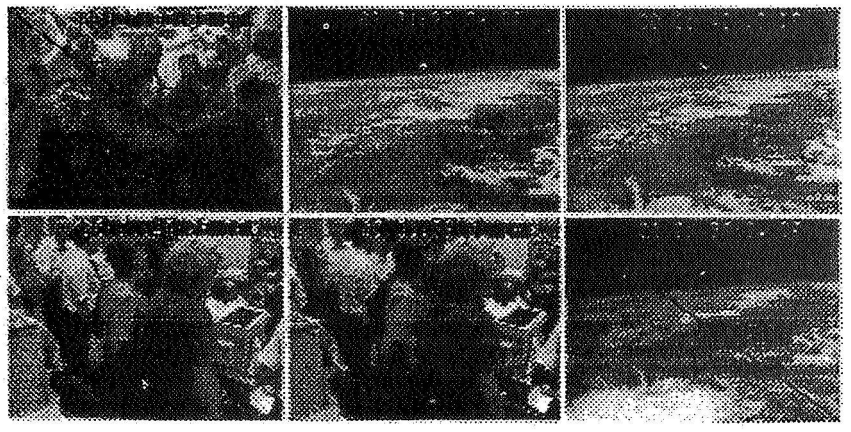

FIG. 5. Scene 14: Alternating frames of mission control and shuttle in space on Earth limb.

FIG. 6. Scene 24: Astronaut pulling experiment racks, part 1.

FIG. 7. Scene 24: Astronaut pulling experiment racks, part 2.

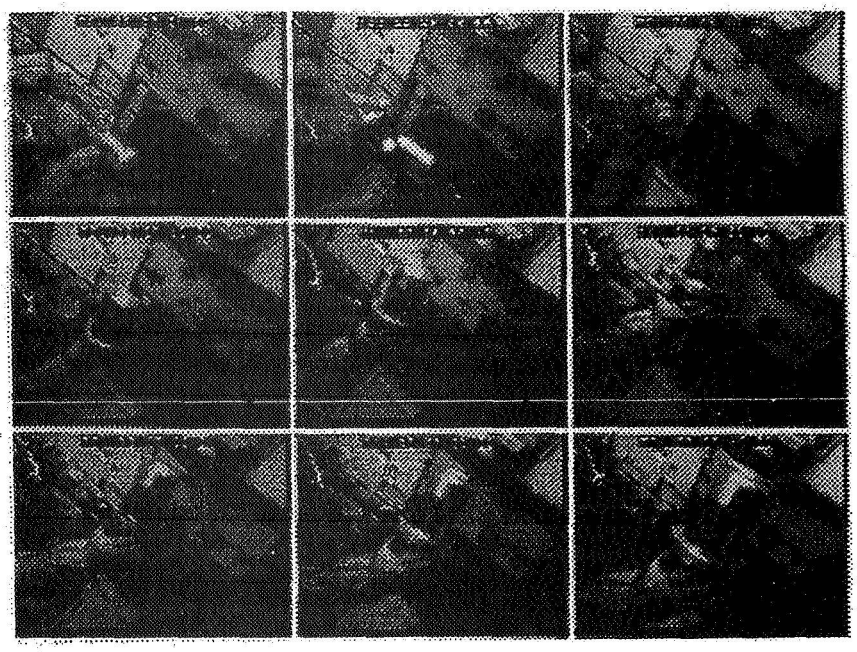

FIG. 8. Scene 27: Astronaut examing flight hardware, frames 4-12.

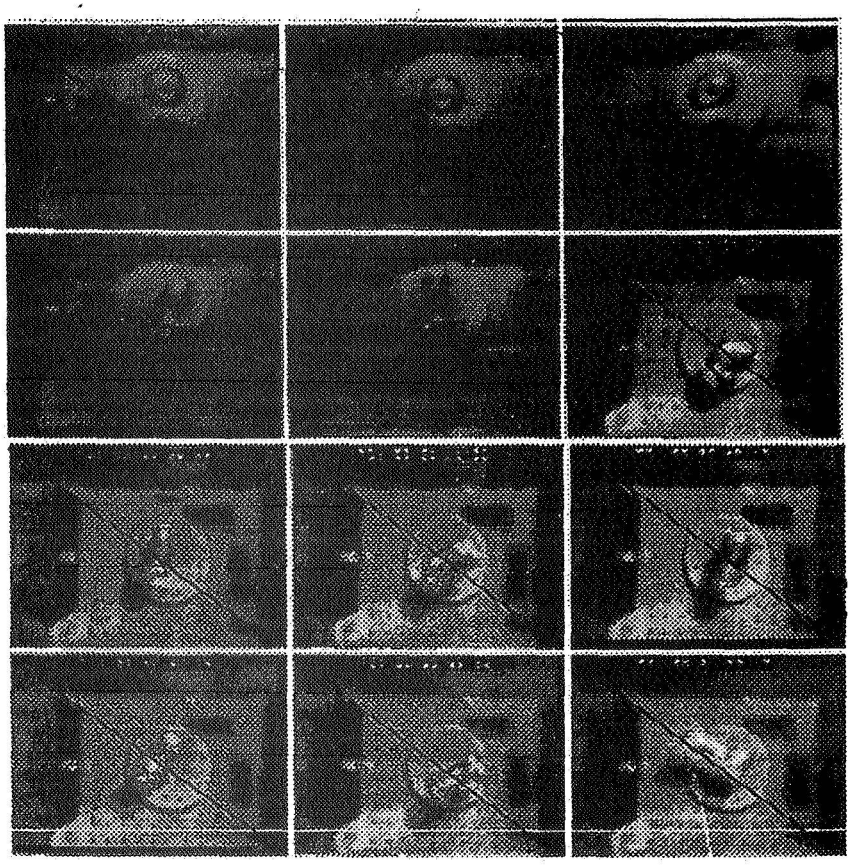

FIG. 9. Scene 31: Acoustic levitation experiment, frames 2-13.

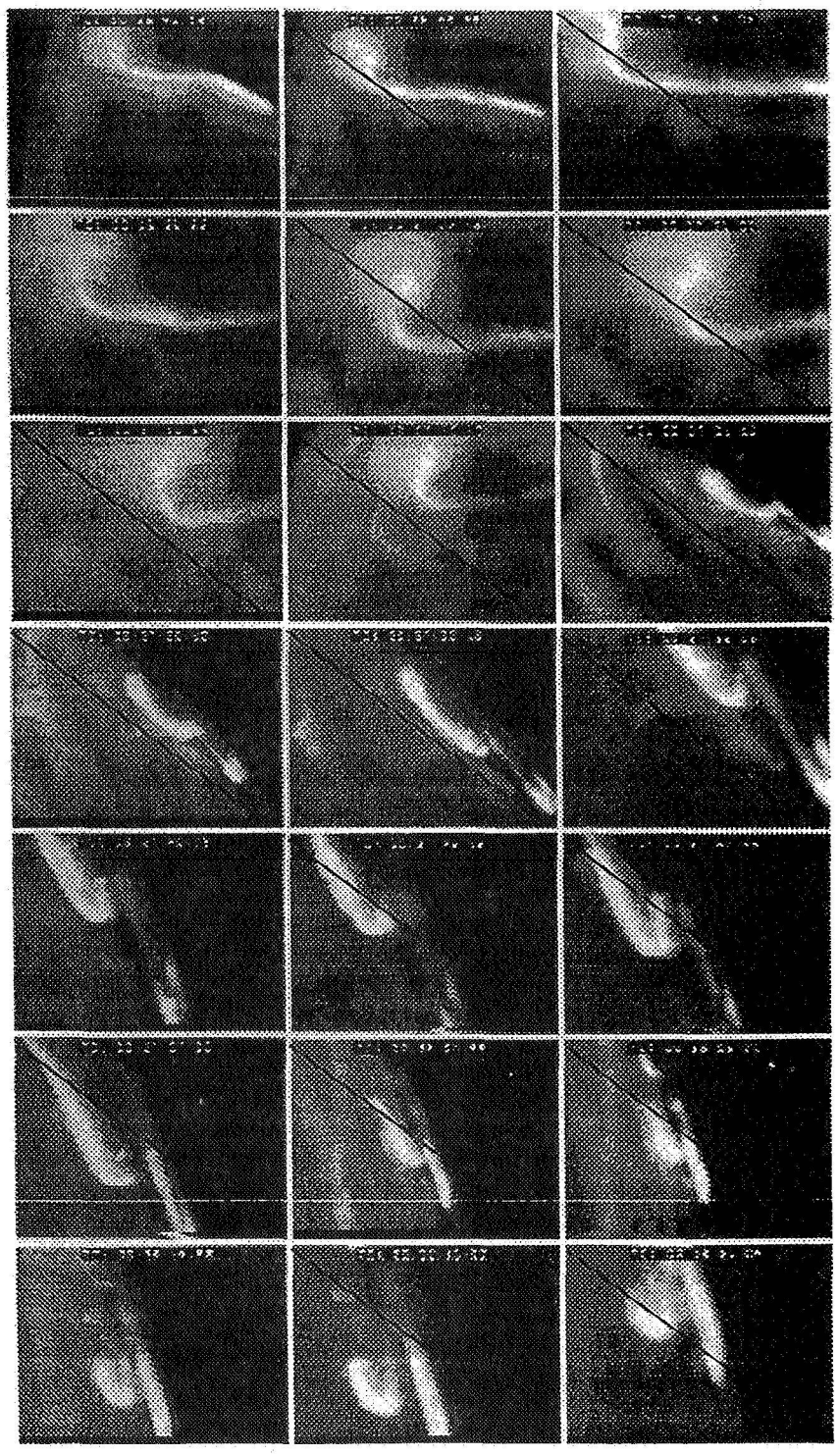

FIG. 10. Scene 32: Meso-scale lightning experiment.

FIG. 11. Scene 36. Shuttle onboard food preparation, part 1.

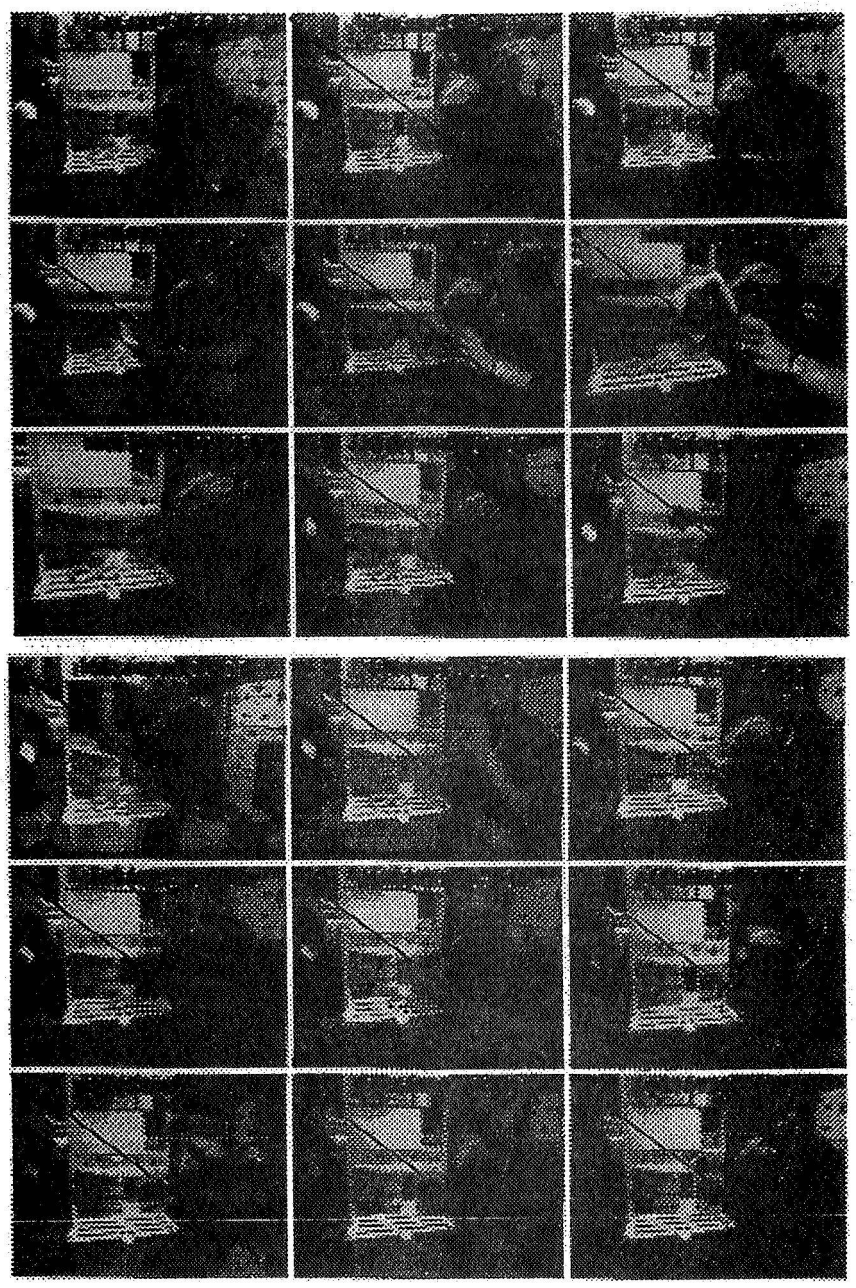

FIG. 12. Scene 36: Shuttle onboard food preparation, part 2.

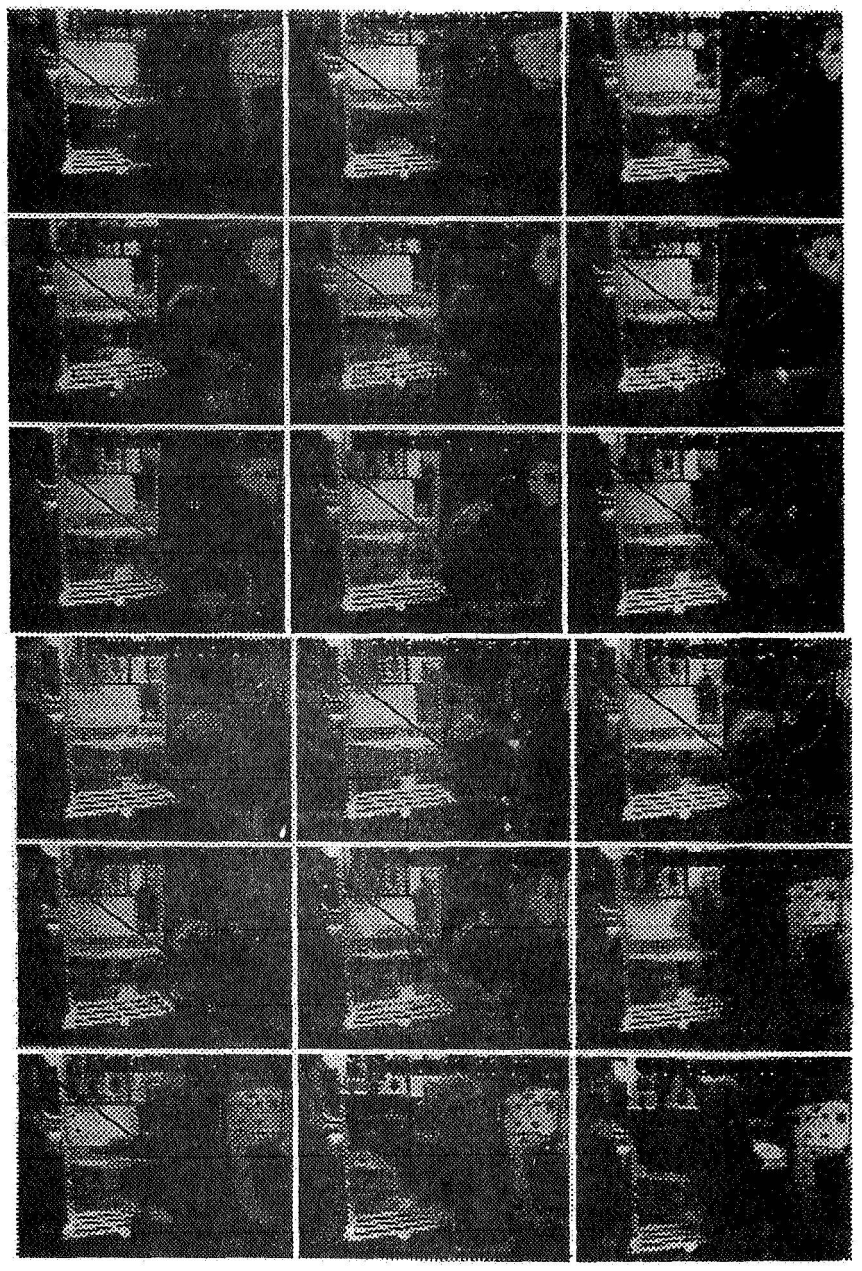

FIG. 13. Scene 36: Shuttle onboard food preparation, part 3.

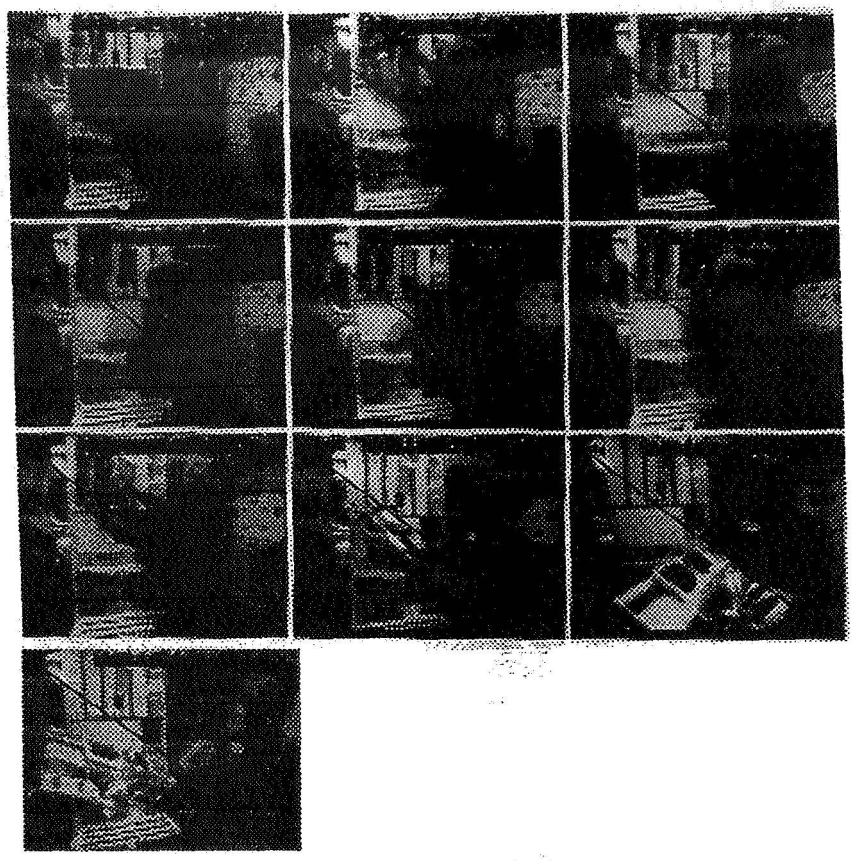

FIG. 14. Scene 36: Shuttle onboard food preparation, part 4.

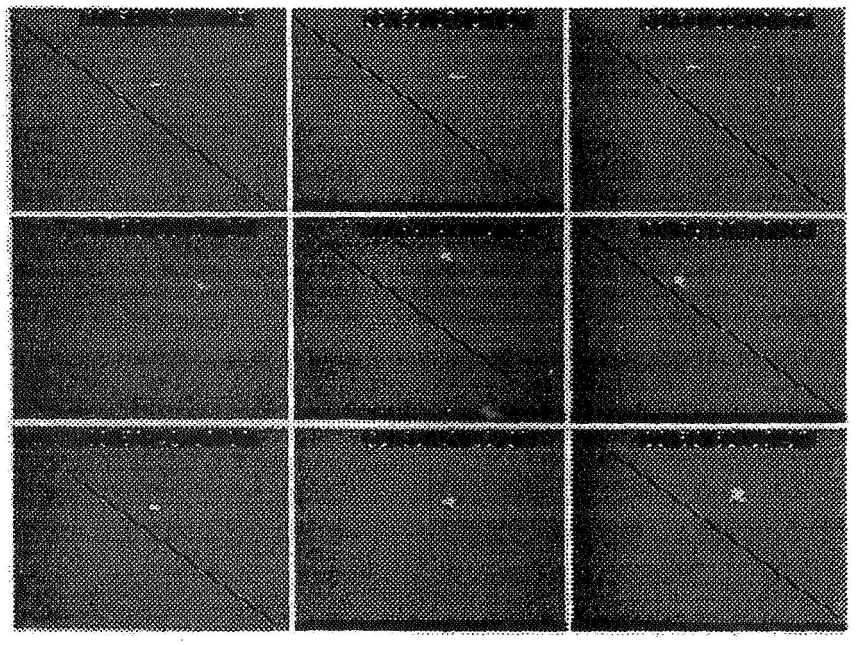

FIG. 15. Scene 38: Shuttle landing sequence, part 1.

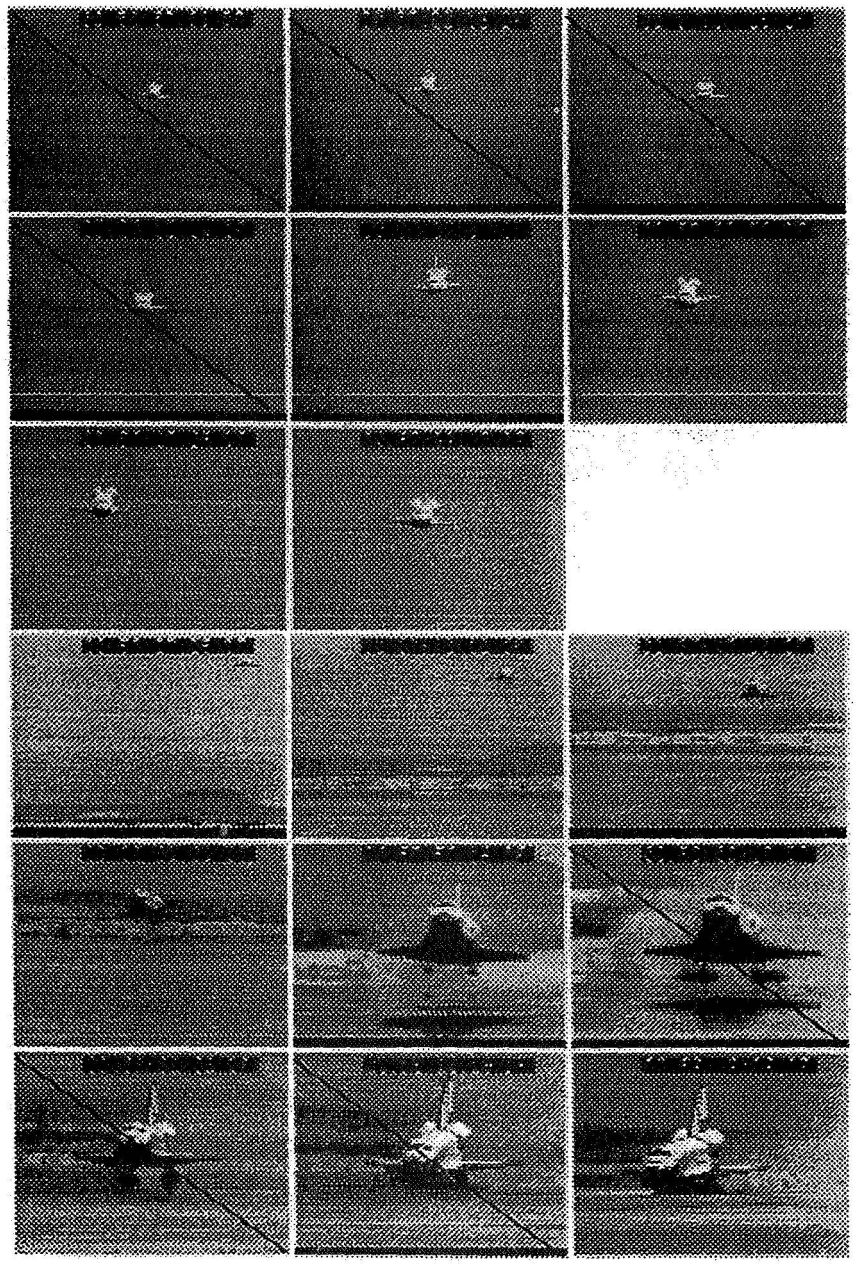

FIG. 16. Scene 38: Shuttle landing sequence, part 2.

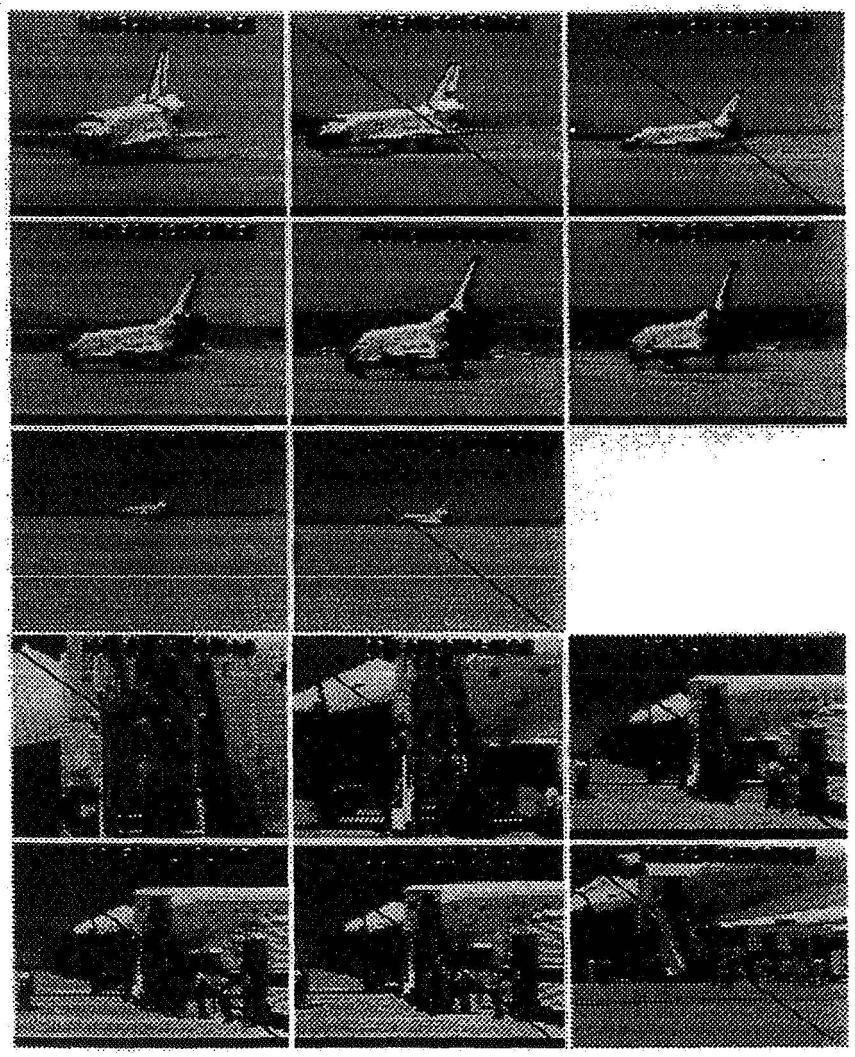

FIG. 17. Scene 38: Shuttle landing sequence, part 3.

References

Ballard, D. H., & Brown, C. M. (1982). *Computer vision,* Englewood Cliffs, NJ: Prentice-Hall.

Chang, S. K., & Yang, C. C. (1983). Picture information measures for similarity retrieval. *Computer Vision, Graphics, and Image Processing, 23,* 366–375.

Leigh, A. B. (1992). *A fuzzy measure approach to motion frame analysis for video data abstraction,* Unpublished M.S. Thesis, Department of Computer Science, University of Houston–Clear Lake, Houston, TX.

O'Conner, B. C. (1985). Access to moving image documents: Background concepts and proposals for surrogates for film and video works. *Journal of Documentation, 41,* 209–220.

Project ICON Laboratory, University of Texas, Austin (1991). Project staff: Mary Ann Albin, Dr. Steven Fitzpatrick, Hee-Sook Choi, Abby Goodrum, Diane Luccy, Stephanie Smith, John Stansbury, Nackil Sung, Margaret Whitehead, Dr. Ronald Wyllys.

Pryluck, C., Teddlie, C., & Sands, R. (1982). Meaning in film/video: Order, time and ambiguity. *Journal of Broadcasting, 26,* 685–695.

Shelton, R. O. (1991, April). Personal communication with Robert Shelton, Computer Engineer, Software Technology Branch, Johnson Space Center.

Yeh, P. -S., et al. (1991, February). On the optimality of Code Options for a Universal Noiseless Coder. *JPL Publication 91-2,* Pasadena, CA: Jet Propulsion Laboratory.