Immersion into Visual Media: New Applications of Image Understanding

Takeo Kanade, Robotics Institute, Carnegie Mellon University

USING IMAGES TO INTERFACE MACHINES AND HUMANS OFFERS AN EXCITING NEW APPLICATION DOMAIN FOR IMAGE UNDERSTANDING. THIS ARTICLE DESCRIBES TWO SUCH APPLICATIONS: DIGITAL LIBRARIES THAT BRING VISUAL DATA TO THE USER, AND VIRTUALIZED REALITY THAT BRINGS USERS TO THE VISUAL DATA.

IMAGE UNDERSTANDING HAS TRAditionally been applied to recognition problems, such as those that arise in military, factory, and autonomous vehicle applications. Recently, IU technologies are finding a place in the use of images to interface people and machines. Such applications entail greater interaction between the user and the visual data—individuals can immerse themselves in the pool of visual information, systematically navigate through the data, and extract the necessary material or experience the desired sensation. We've developed two such applications: a digital library that brings visual data to the user, and *virtualized reality* that brings users to the visual data. The digital library calls on IU techniques such as scene segmentation and content analysis, and joins together natural language analysis and IU. Virtualized reality uses precise, dense, and video-rate stereo reconstruction techniques.

Creation and exploration of a digital video library

With the growth and popularity of multimedia computing technologies, video is gaining importance and broadening its uses in libraries. Digital video libraries hold great potential for education, training, and entertainment; but to achieve this potential, the information embedded in the digital video library must be easy to locate, manage, and use. Searches in a large data set or lengthy video can take a user through vast amounts of material irrelevant to the search topic. A typical database search by keywords (for example, by title) only references images; it does not directly search for them. Such a search is not appropriate or useful for the digital video library, because it does not give the user a way to know the image's contents, short of viewing it. We need new techniques to organize these vast video collections so that users can effectively retrieve and browse their holdings, based on the holdings' content.

Visual indexing, retrieval, and presentation. For effective use of the digital video library, video must be segmented, indexed, searched, manipulated, and presented according to its content. IU plays a critical role in these operations. One of the first capabilities required for the creation of a digital video library is the segmentation, or *paragraphing*, of video into meaningful groups. Each group can be abstracted by a *representative frame*, which can be the basis for image-content search.

Any textual information attached, such as title, domain, and date, can help quickly filter video for locating potential items of interest. However, subsequent queries are usually visual and refer to images themselves—for example, "Find video with the same person," and "Find the same scene with similar camera motion." Searches usually produce multiple hits. Browsing can help

0885-9000/96/$4.00 © 1996 IEEE

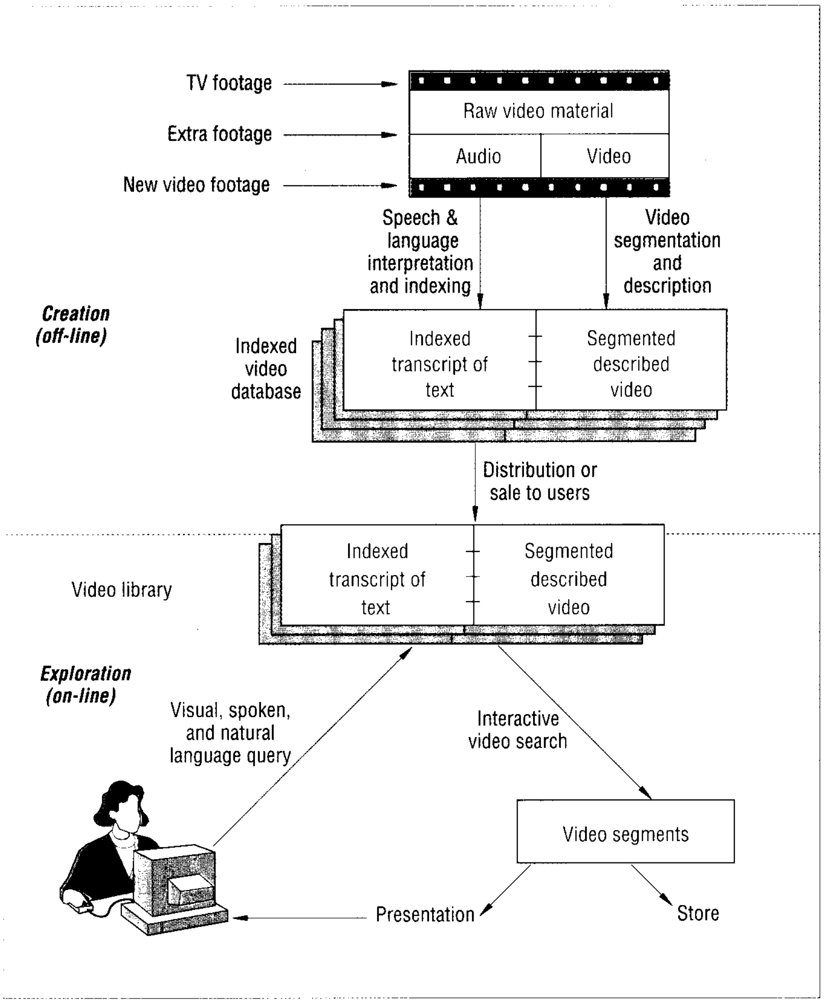
TV footage
Extra footage
New video footage
Raw video material
Audio
Video
Speech & language interpretation and indexing
Video segmentation and description
Creation (off-line)
Indexed video database
Indexed transcript of text
Segmented described video
Distribution or sale to users
Video library
Indexed transcript of text
Segmented described video
Exploration (on-line)
Visual, spoken, and natural language query
Interactive video search
Video segments
Presentation
Store

Figure 1. The Informedia Digital Video Library system integrates image, speech, and language understanding.

users rapidly filter these hits to obtain the precise target information. While browsing video is not as easy as browsing text, we can take advantage of the human visual system, which is adept at quickly, holistically viewing an image. The library might simultaneously present numerous icons (static frames) or motion icons (short motion sequences) of the segments in separate windows.

Some of these visual query capabilities rely on image-processing methods based mostly on image statistics and image matching. For example, key-frame detection, based on MPEG (Moving Picture Experts Group) codes, can parse video into scenes. Statistics of other primitive image features, such as color histograms, shape, and texture measures, have proved useful for indexing, matching, and characterizing images. Some algorithms determine camera work, such as panning and zooming, and others detect transitions between scenes, such as fades, cuts, and dissolves. A commercial image database system, QBIC, incorporates many of these capabilities into a visual query.[1]

Video is more than images. Although current successful efforts at visual querying of image databases are founded on indirect image statistical methods, they fail to capture and exploit the massive information contained in video. Video is temporal, spatial, and often unstructured; the combined video signal and audio signal convey an abundance of information.

The audio signal includes language information in the form of narration and dialogue that, when transcribed, provide direct indices to the video content. Natural language analysis of the transcript, together with production notes and other text information about the video, can determine the narrative's subject area and theme. This understanding can be used to generate summaries of each video segment for icon labeling, browsing, and indexing. The language information also provides critical cues to IU tasks, as evidenced by many successful programs that use collateral data for recognition.[2,3] The audio signal conveys other information, including pauses, silence, music, and laughter. These bits of information can supplement the other structured descriptors. For example, pauses might be useful in identifying natural start and stop positions (with some phase difference) for video paragraphing.

Each appropriately sized video clip in the digital video library should therefore have a far richer description than the context-free image-statistical methods can provide. Each description would include a full text transcript with links to corresponding sections of the audio signal, scene segments, individual scene characterizations, a representative single image icon, and a short *skim video* (which we'll describe later), in addition to the full video itself. Scene characterizations would consist of camera motion, representative objects, object motion, caption text, and a scenery classification (such as outdoor and indoor). With such an organization, the digital video library user should be able to conduct content retrieval, as well as image retrieval, based on image, audio, and other context cues.

Informedia: integrating speech, language, and image understanding. Carnegie Mellon's Informedia Digital Video Library project, funded by the NSF, ARPA, and NASA, is developing intelligent, automatic mechanisms to populate the library and allow full-content knowledge-based search, retrieval, and presentation. Leveraging two decades of ARPA-funded research in speech, language, and image understanding, Informedia integrates these technologies for efficient creation and exploration of the library.

Figure 1 shows the overview of the Informedia system. Using a high-quality speech recognizer, Informedia converts each videotape's sound track to a textual transcript. A language-understanding system analyzes and organizes the transcript, then stores it in a full-text information retrieval system. IU techniques segment video sequences, detect and identify objects, obtain a visual characterization of the scene, identify the representative images for the skim video, and match images by incorporating language and speech information.

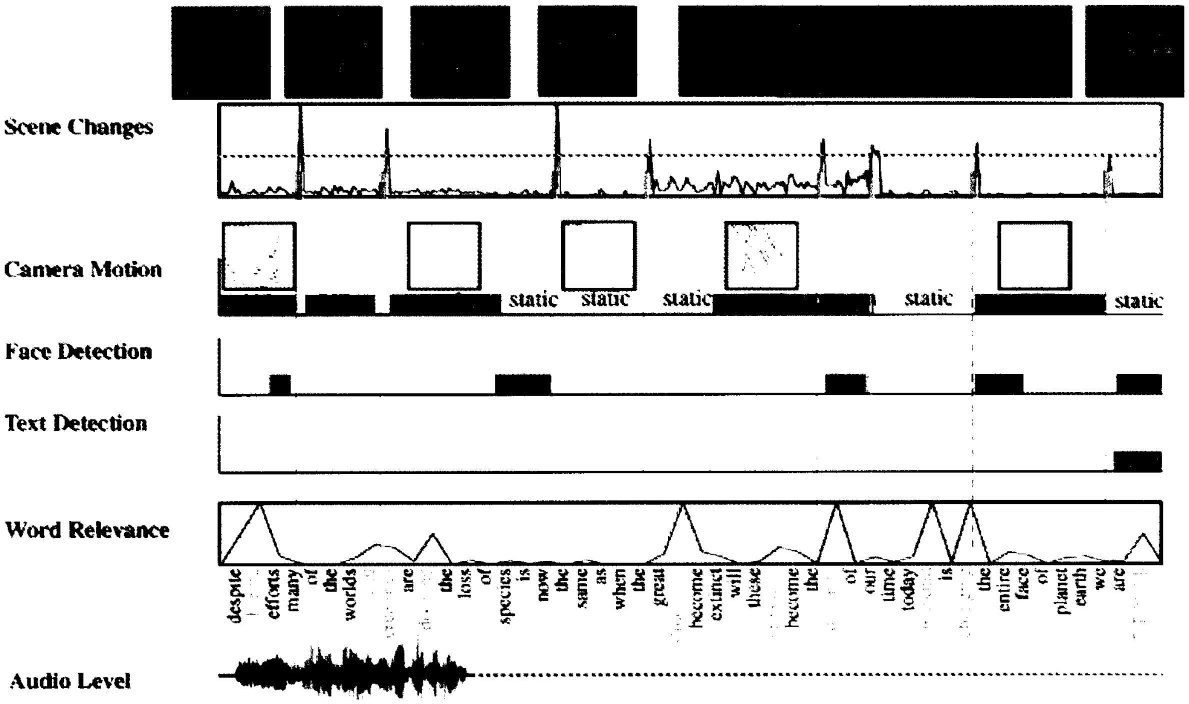

Figure 2. Analysis results of a video clip. Informedia segments the video into scenes, detects and classifies camera motions, detects significant objects (human faces and text), and evaluates word relevance in the transcript.

Automatic creation of skim video. A good initial example of Informedia's integrated approach is the automatic creation of skim video. The original data in the Informedia Library is generally a one-hour full-feature broadcast. A skim video is a very short synopsis comprising the significant words and images, with which the user can grasp the whole content of the original video. During video playback for browsing, a user can choose to compact the video as much as needed. One of the project's goals is to automatically reduce an hour-long video's playback time to a few minutes.

The critical aspect of compressing a video is context understanding, which is the key to choosing the images and words that the skim video should include. This requires the integration of language and image understanding. Informedia examines segment breaks produced by image processing, along with the boundaries of topics identified by the language processing of the transcript. It evaluates the relative importance of the scenes based on the corresponding objects that appear, the associated words and sounds, and the video scene's structure. Figure 2 shows a result of analyzing a video clip by various speech, language, and image understanding techniques.

Speech transcription and language analysis. Language understanding works on the transcript. Effective use of the video information assets requires automatic generation of transcripts by speech-recognition technology, because not all video is closed-captioned. The Informedia system uses the Sphinx-II speech-recognition system,[4] which is a large-vocabulary, speaker-independent, continuous speech recognizer developed at Carnegie Mellon. Applying the speech-recognition system to transcript generation presents a number of challenges, including dealing with multiple unknown microphones, segmenting spontaneous fluent speech, and dealing with an unlimited vocabulary.

For the moment, we are using manually edited, semiautomatically generated transcripts. The initial language analysis computes word relevance by the well-known Term Frequency/Inverse Document Frequency (TF/IDF) technique.[5] Word relevance, plotted as the second row from the bottom in Figure 2, is each word's frequency in a particular script divided (normalized) by its frequency in a much broader corpus. In other words, if a word that seldom appears in a general document appears often in a video, it signifies the word's relative importance in the video.

We also use speech recognition to detect transitions between speakers and topics that are usually marked by silence or low-energy areas in the acoustic signal.

Scene segmentation and motion analysis. We use comparative histogram disparity measures to detect scene breaks. This technique, successfully used for image-query systems, is robust enough to maintain high levels of accuracy, yet efficiently computable directly from MPEG codes. By detecting significant changes in the weighted color histograms of successive frames, Informedia can separate image sequences into scenes. Figure 2 shows the disparity plot (the second row) and the resultant scene breaks with the first frame from each scene (the first row).

One important clue for scene characterization is camera and object motion. We classify the camera motion as static, pan, or zoom (see the third row in Figure 2) by examining the optical flow vectors. Velocity vectors for pans and zooms have distinct statistical characteristics. Global motion analysis distinguishes between object motion and actual camera motion.

Object detection: face and text. Informedia also identifies significant objects. For the time being, we have chosen to deal with human faces and text (caption characters). Human faces are among the more interesting objects in video. A human interacting in an environment is a common theme in video, and the talking-head image is common in interviews and news clips. Recognizing the same anchorperson in a news clip helps understanding the start and end of a story. Our human-face detection system is based on the multineural network arbitration method,[6] which can detect frontal faces over a fairly wide range of sizes (see Figure 3a). Currently, it can detect over 90% of more than 300 faces in 70 images, while producing approximately 60 false detections. Because it is a learning system, it can be trained to detect other objects as well.

Text characters in visual frames usually provide significant information regarding a scene's content. Often, important information, such as statistics, is not verbalized but is included in the visual frames for viewer inspection. We extract text regions from video frames and supply them to the optical reader to convert to textual information (Figure 3b).

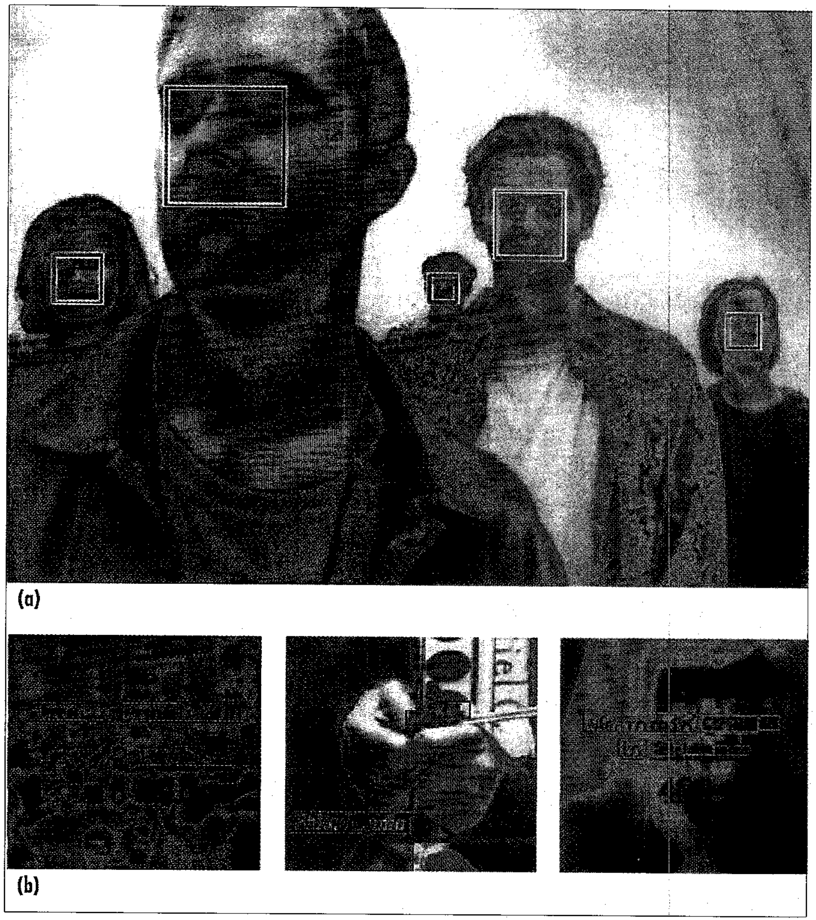

Figure 3. Object detection in video: (a) faces, (b) text regions.

Extracting the skim video. The previous analyses segment and characterize the video. Informedia associates each segment with the transcript, word relevance, motion, object (human face) appearance, caption text, and audio level, as shown in Figure 2. The next task is to extract and order the significant parts of the video and audio tracks to create the skim video.

First, to select scenes for the skim, Informedia analyzes word relevance. The number of scenes used in the final skim depends on the compression rate. The compression rate is typically set at 10:1. We have found that a skim video with a rate as high as 20:1 can still offer sufficient comprehension of the video data.

The next step is to select video segments corresponding to the selected words. The video segment that corresponds to the timing of the audio is not necessarily the best selection. Quite often, the important word and the important video, while in the same scene, are not synchronized. We use the video characterization results to classify and rank frames by a set of priority rules. Tentatively, we favor

(1) frames with human faces or text,
(2) static frames following camera motion,
(3) frames with human faces and text with camera motion, and
(4) frames at the middle of the scene (the default).

Because the selected relevant word determines the audio length, we select only enough image frames to fill the compressed audio track.

We have also developed several heuristic rules for the final selection and ordering of skim frames. The choice of rules depends on various conditions, such as the duration of the words, scene contents, and the selection of the previous frames. Figure 4 illustrates a few examples of applying these rules, and the resultant skim frames.

Although much room remains for improvement of this context-based method of generating skims, this method illustrates the potential power of integrated image, speech, and language input for digital video library applications. The ultimate goal is complete systematic characterization of video data for video-context-based indexing, retrieving, and presentation.

Virtualized reality

Most visual media available today—such as photographs, movies, and television—share one aspect: a "director" decides the view of a scene while recording or transcribing the event. Combining 3D IU and computer graphics technologies, we can eliminate this limitation and envision a new visual medium called virtualized reality.

To create virtualized reality, an event is captured by a number of cameras, positioned to document the event from all viewpoints (see Figure 5). They form multiple sets of stereo configurations of cameras. A stereo method computes the time-varying 3D structure of the event, described in terms of each point's depth and aligned with the pixels of the image for a few of the camera angles (called the transcription angles). Once the real world has been virtualized, we can place a *soft*, or virtual, camera at an arbitrary position, and graphic techniques can render the

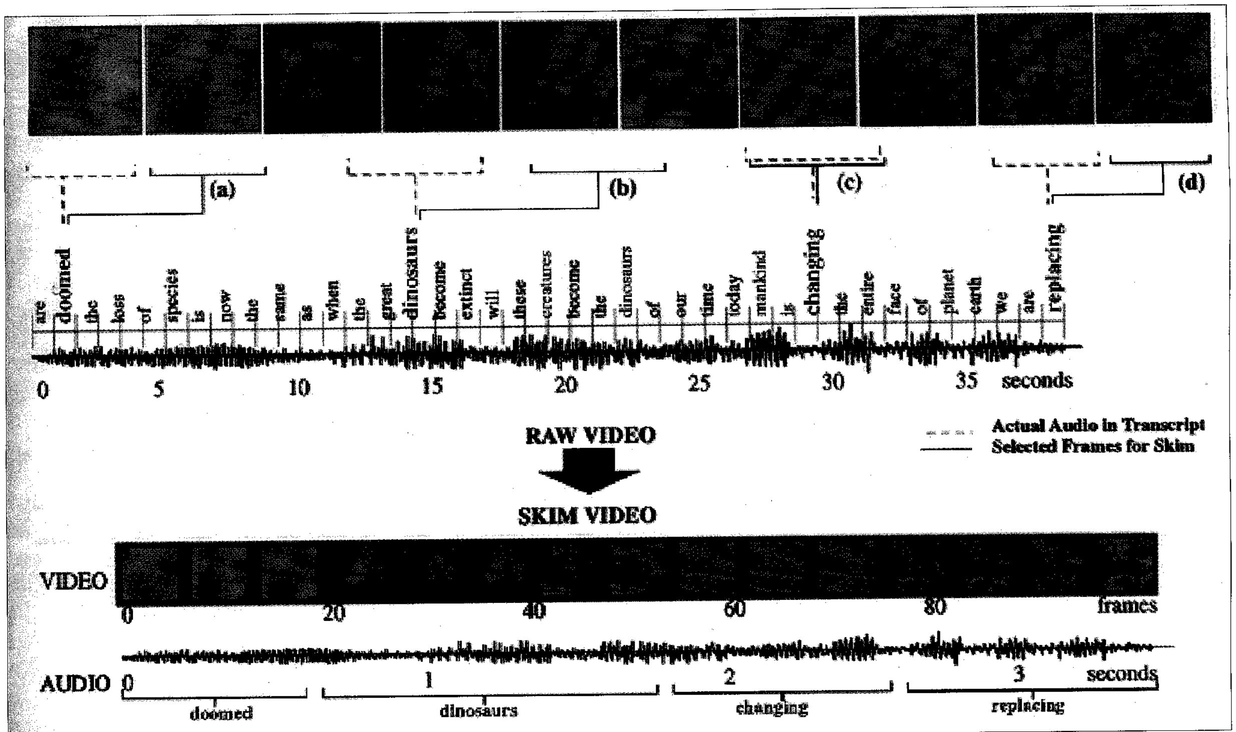

Figure 4. Skim video creation from the original video, incorporating word relevance in the transcript, objects in video (humans and text), and camera motion. Informedia extracts four skim segments, using different rules: (a) For the word "doomed," Informedia selects the portion of the scene with no or little motion, because typically the static region is the focus of the scene; (b) The narrator uses 1.13 seconds (34 frames) to utter the word "dinosaurs." This segment also includes a portion from the next scene; (c) This segment has no significant motion or object, so Informedia uses the portion directly corresponding to "changing"; (d) For the word "replacing," Informedia chooses the latter portion of the scene, which contains both humans and text.

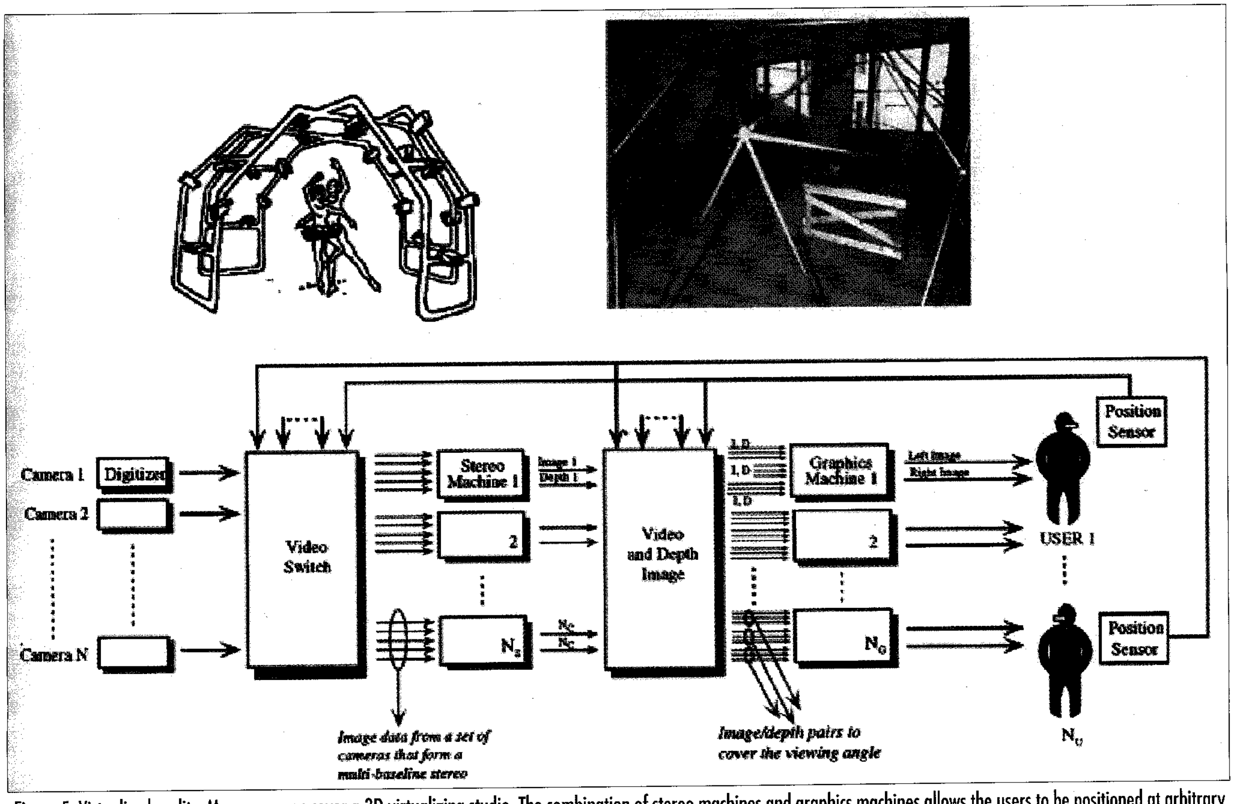

Figure 5. Virtualized reality. Many cameras cover a 3D virtualizing studio. The combination of stereo machines and graphics machines allows the users to be positioned at arbitrary locations in the virtualized space.

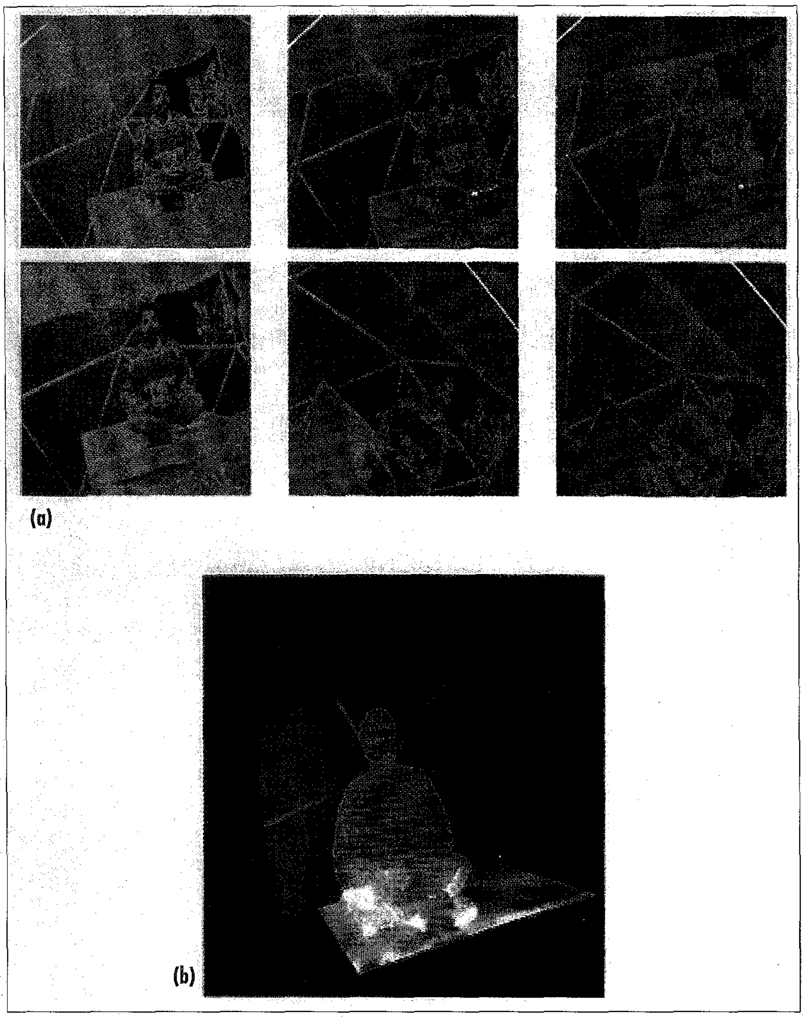

Figure 6. Transcribing a scene by multiple cameras: (a) six images of a scene, (b) the recovered depth map aligned with the color image.

event as viewed from it. A user, wearing a stereo-viewing system that can feed the soft cameras' positions, can freely move about in the world and observe it.

Virtualized reality has many applications. It is similar to virtual reality, but instead of using an artificially built model, this system starts with the real world.[7] Training can become more realistic, safer, and more effective. A surgical operation recorded in a virtualizing studio could be repeatedly revisited by medical students, who could view it from various vantage points.[8] Telerobotic maneuvers for decommissioning hazardous nuclear facilities could be rehearsed in a virtualized environment that feels every bit as genuine as the real world. True 3D telepresence could be achieved by performing transcription and view generation in real time. An entirely new generation of entertainment media could be developed: basketball enthusiasts and Broadway aficionados could experience the feeling of watching the event from their preferred seat, or even a seat that changes as the action progresses.

The 3D virtualization studio. Virtualizing the real world requires imaging an event from a large number of transcription angles. The photo in Figure 5 shows a prototype of such a 3D virtualization studio at CMU, which uses a hemispherical dome five meters in diameter. The dome currently houses 10 cameras, and will eventually have more than 50. The studio synchronizes all cameras to a common signal, and records each camera's output on a separate VCR that timestamps each frame with the Vertical Interval Time Code (VITC) for later digitization. (We use this off-line procedure because we cannot simultaneously digitize the outputs of 10 cameras at 30 frames per second.) Figure 6a shows six images of a scene transcribed using our studio.

Multicamera stereo. Technologies for fast, precise, and dense stereo matching are the key components in virtualized reality. Stereo has long been a subject of intensive study in IU. Recently, precise and dense scene reconstruction has become possible using the multibaseline stereo (MBS) technique.[9] The MBS algorithm takes advantage of the redundancy of information in multiple images of the same scene, for robust and precise measurement.

In general, binocular stereo techniques measure distances by finding corresponding points in the left and right images, and then by triangulating to calculate distance. The precision of stereo distance estimation increases as the baseline (the distance between a pair of cameras) increases. However, increasing the baseline also increases the likelihood of mismatching the corresponding points—causing a gross error. In other words, a trade-off exists between the desire for correct correspondence among images (using narrow baselines) and for accurate estimates of scene depth (using wide baselines).

The MBS technique eliminates this tradeoff by simultaneously computing the measure of correspondence among pairs of image points from multiple cameras with different baselines. Figure 6b shows the depth map recovered by applying the MBS algorithm to the input scene. The depth map has 74 levels for a depth range of two to five meters.

Video-rate stereo machine. One of the greater challenges in IU has been development of a real-time 3D stereo machine that maps a 3D time-varying scene into a sequence of accurate dense depth maps. At CMU we have built and demonstrated a video-rate stereo machine based on the MBS algorithm.[10] This machine can produce over 1.2 million depth pixels of seven-bit resolution per second, corresponding to a 200×200 depth image at 30 frames per second. This stereo-mapping capability will let us virtualize the environment in real time, capturing the appearance and geometry of time-varying environments.

Generating novel views. The stereo program produces a dense 3D description of a scene for a transcription angle. The program aligns a depth map with a color map—the point (i, j) in the depth map gives the distance of the color image pixel (i, j) from the camera. Graphics workstations can render the scene from other viewpoints with the original color images texture-mapped onto the rendered regions. Figure 7 shows realistic images of the scene in Figure 6 viewed from new viewpoints.

For rendering, the program first converts the depth map into a triangle mesh representation. The most straightforward method of converting every section of the depth map into two triangles produces 40,000 meshes for a 200×200 depth image. Adaptively placing vertices for triangulation of the depth map while limiting the maximum deviation can reduce the number of triangles substantially without affecting the output's visual quality, typically by factors of 20 to 25 on typical scenes. Today's graphics workstations have no difficulty in rendering scenes of that complexity with texture mapping at the video rate.

We can also virtualize moving scenes by virtualizing each frame. The resulting virtualized reality movie can be played with the viewer standing still, or can be observed by a viewer who is moving in the scene independently of the virtualized motion. Figure 8 shows seven frames of a basketball sequence from the reference transcription point and from a synthetically created moving viewpoint.

Combining multiple scene descriptions. The generated images in Figure 7 have *holes*, or unpainted regions. These are surfaces that are occluded by the front objects in the original transcription angle but that are visible in the new view angle. The 3D virtualizing studio captures images from multiple transcription angles. Detection of occlusion boundaries is not trivial, but once they are detected, we can fill these holes using a scene description from another transcription angle in which that portion of the scene is not occluded. Moreover, even when occlusions are not involved, the color image used for texturing gets too stretched when the viewing angle is far from the transcription angle, resulting in poor quality of the synthesized image. To minimize this degradation, we choose the most direct transcription angle for each viewing angle. Figure 9 shows an example of combining the results from two transcription angles. Such occlusionless, better-quality images would provide a complete feeling of immersion.

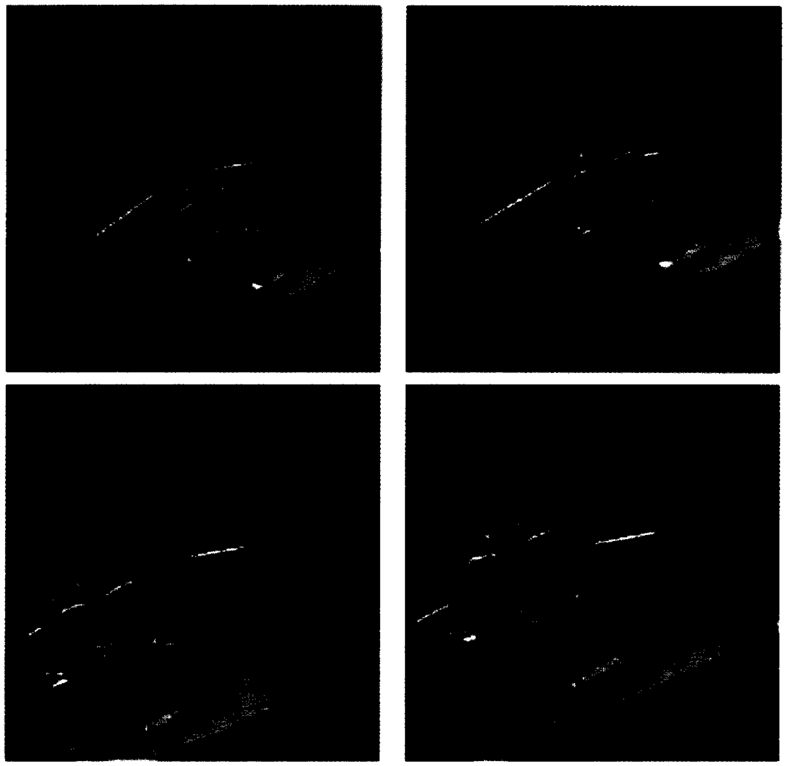

Figure 7. Rendering of the scene in Figure 6 from new viewing angles.

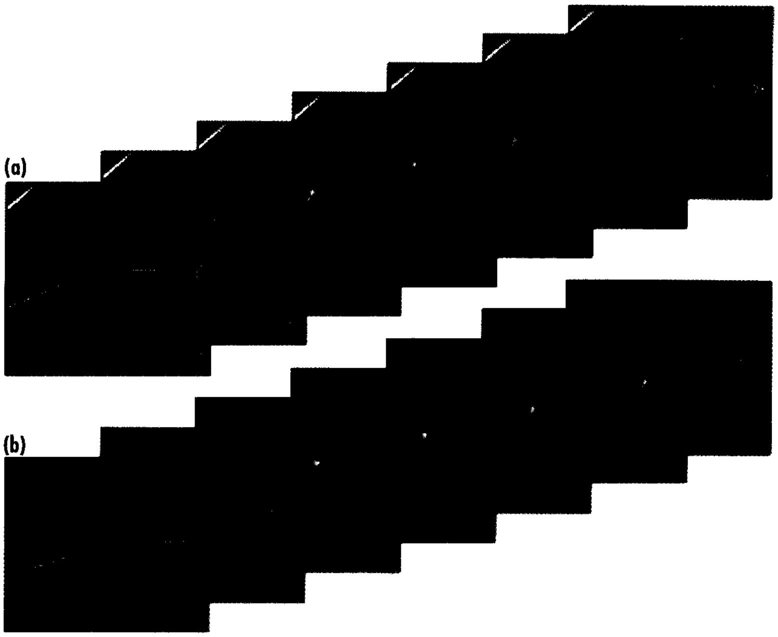

Figure 8. Motion sequence of a basketball scene: (a) original reference images, (b) synthesized from a moving viewpoint. The image starts left and above the original viewpoint and moves to the right.

Beyond visual virtualized reality. To support training for visual-motor tasks, such as telerobotic handling of hazardous materials, we must reproduce both the geometry and the mechanics of actual complex environments. This will give the trainee a convincing experience of performing the task. Segmenting the virtualized geometric model to objects, inferring constraints, and assigning material properties will transform the geometric model into a mechanical model. Techniques for realtime mechanical simulation, including the effects of collision, contact, and friction, will make the virtualized objects respond appropriately to each other and to the trainee's actions. Finally, a high-bandwidth, high-fidelity haptic interface device, such as the one based on Lorentz levitation,[11] will let the trainee manipulate the virtualized environment using tools that represent such devices as surgical instruments and remote manipulators. Such a device will provide accurate and convincing force feedback.

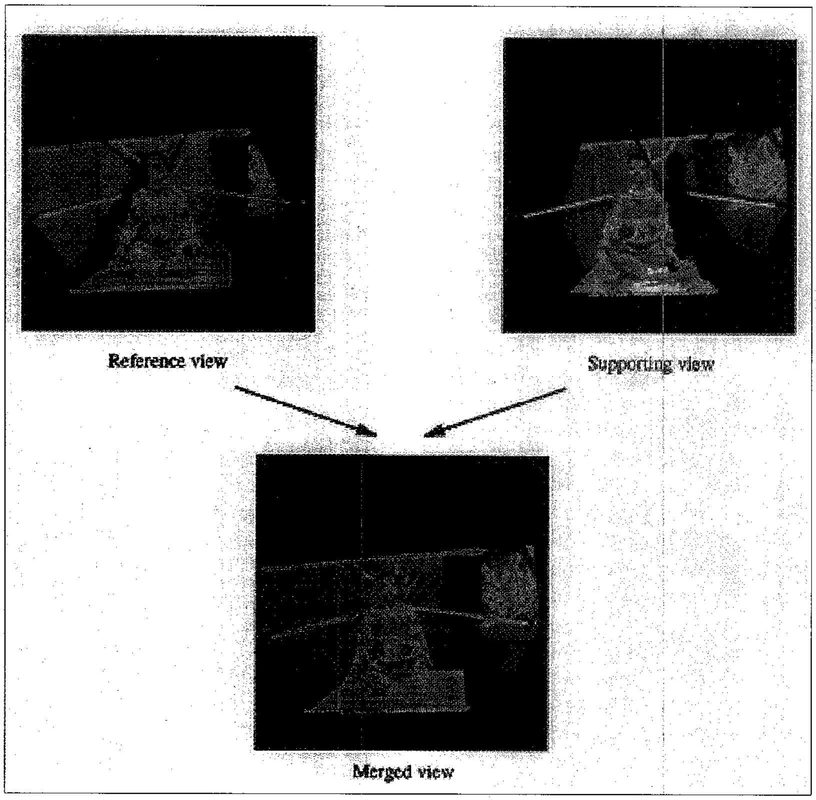

Figure 9. Combining multiple transcription views for removal of occlusions and for better image quality.

USING IMAGES TO INTERFACE machines and humans offers an exciting new application domain for IU. Applications that immerse users in visual data—such as our digital library and virtualized reality—will lead to fertile new lines of investigation in IU.

References

1. C. Faloutsos et al., "Efficient and Effective Querying by Image Content," *J. Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies*, Vol. 3, Nos. 3–4, July 1994, pp. 231–262.

2. T.M. Strat, "Employing Contextual Information in Computer Vision," *Proc. ARPA Image Understanding Workshop*, Morgan Kaufmann, San Francisco, pp. 217–229.

3. D.M. McKeown, W.A. Harvey, and J. McDermott, "Rule-Based Interpretation of Aerial Imagery," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 7, No. 5, Sept. 1985, pp. 570–585.

4. M.-Y. Hwang, X. Huang, and F. Alleva, "Predicting Unseen Triphones with Senones," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, Vol. 2, IEEE Press, Piscataway, N.J., 1993, pp. 311–314.

5. M. Mauldin, "Information Retrieval by Text Skimming," PhD thesis, Carnegie Mellon Univ., 1989. Revised edition published as *Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing*, Kluwer Academic Publishers, Boston, 1991.

6. H. Rowley, S. Baluja, and K. Kanade, "Human Face Detection in Visual Scenes," Tech. Report CMU-CS-95-158, Computer Science Dept., Carnegie Mellon Univ., Pittsburgh, 1995.

7. T. Kanade, "User Viewpoint: Putting the Reality into Virtual Reality," MasPar News, Vol. 2, No. 2, Nov. 1991, p. 4.

8. H. Fuchs and U. Neuman, "A Vision Telepresence for Medical Consultation and other Applications," *Proc. Sixth Int'l Symp. Robotics Research*, Int'l Foundation for Robotics Research, Cambridge, Mass., 1993, pp. 555–571.

9. M. Okutomi and T. Kanade, "A Multiple-Baseline Stereo," *Proc. 1991 IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition*, IEEE Computer Society Press, Los Alamitos, Calif., 1991, pp. 63–69.

10. T. Kanade et al., "Development of a Video-Rate Stereo Machine," *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems* (IROS '95), CS Press, 1995, pp. 95–100.

11. R.L. Hollis and S.E. Salcudean, "Input/Output System for Computer User Interface Using Magnetic Levitation," *Proc. Sixth Int'l Symp. Robotics Research*, Int'l Foundation for Robotics Research, Cambridge, Mass., 1993, pp. 503–520.

Takeo Kanade is the U.A. Helen Whitaker Professor of Computer Science and the director of the Robotics Institute at Carnegie Mellon University. He has made technical contributions in multiple areas of robotics: vision, manipulators, autonomous mobile robots, and sensors. He has written more than 150 technical papers and reports in these areas. He has been the prinicpal investigator of several major vision and robotics projects at Carnegie Mellon, and was the founding chair of CMU's Robotics PhD program.

He is a Fellow of the IEEE, a Founding Fellow of the American Association of Artificial Intelligence, and the founding editor of the *International Journal of Computer Vision*. His awards include the Joseph Ehrenberger Award in 1995 and the Marr Prize in 1990. He has served on the National Research Council's Aeronautics and Space Engineering Board, NASA's Advanced Technology Advisory Committee, and the Advisory Board of the Canadian Institute for Advanced Research. He received his PhD in electrical engineering from Kyoto University, Japan, in 1974. He can be reached at the Robotics Inst., Smith Hall, Carnegie Mellon Univ., Pittsburgh, PA 15213-3898; kanade@cs.cmu.edu.