International Journal of Computer Vision 50(2), 103–110, 2002

Guest Introduction: The Changing Shape of Computer Vision in the Twenty-First Century

MUBARAK SHAH
Computer Vision Lab, School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816, USA
shah@cs.ucf.edu

1. Introduction

Computer vision started as an Artificial Intelligence (AI) problem. For this reason, vision has also been called *image understanding*. The original goal of vision was to understand a single image of a scene, locate and identify objects, determine their structures, spatial arrangements, relationship with other objects, etc. The MIT *copy* demo was a good example of this. The idea in the copy demo was roughly like this: to have a computer vision program analyze an image of a scene containing several stacked blocks, recover the structure of the blocks, and generate a script, for a robot to build an exact *copy* of the block structure.[1] This was actually a high level vision problem. This copy demo was one of the motivations for the work in the blocks world like consistent line labeling, analysis of polyhedral junctions, etc. The researchers soon realized that low level vision was not robust enough; they were not even able to extract lines from images to be used in this work. Therefore, it became necessary to first solve low level vision problems (see for example Binford-Horn line finder (Horn, 1973), before the high level vision problems could be attacked. Research in low level vision continued for some time.

During the seventies, Marr (1982) captured the attention of vision researchers. He popularized shape-from-X, and the use of ideas from the human vision system, among other things. Since one dimension is lost during the projection of a 3D world onto 2D images, the aim of shape-from-X methods is to recover that lost dimension. The next two or three decades were spent developing algorithms for recovering 3D shape from 2D images using stereo, motion (structure from motion), shading, texture, etc.[2] We almost forgot about the original AI problem; not much progress was made in high level vision during those years. Marr also emphasized the role of the human vision system in solving computer vision problems, and vice versa. Since the human vision system is one of the finest living examples, it should help us to build artificial vision systems which can perform as robustly as our own vision, he argued. Similarly, some of the unexplained phenomena in the human vision system may also be explainable if artificial vision systems can be built to perform similar tasks. In the early days two noteworthy contributions inspired by the human vision system, were the Marr-Hildreth Laplacian of Gaussian edge detector (Marr and Hildreth, 1980), and the Marr-Poggio relaxation-based stereo algorithm (Marr and Poggio, 1979).

The Laplacian of Gaussian edge detector performed pretty well compared to existing edge detectors like Sobel, Prewit, etc. However, soon after that, Canny (1986) developed an edge detector which was based on the maxima of a gradient of a Gaussian. This was called an optimal edge detector, and was obtained by optimizing some performance measure, and did not have any biological visual system motivation. In fact, this edge detector combined a couple of interesting known techniques: non-maxima suppression (Rosenfeld et al., 1972), and hysteresis thresholding. The Canny edge detector performed very well; even today it is the edge detector everybody uses. The Marr-Poggio relaxation stereo algorithm was interesting since it explained how humans may be computing depth from two 2-D images, but was soon superceded by several other better computer vision stereo algorithms (e.g. see Ohta and Kanade, 1985).

It is my personal view that computer vision is a hard problem; one should try to use all possible tools

available to solve vision problems, not only tools available in the human vision system. The past experience has shown that human vision system has not really helped us to solve significant vision problems.

One trend which started around or right after the Marr era was to the push of complex and new mathematical techniques in computer vision. The idea is to find some mathematical technique which has not been used widely in computer vision, study it well, and find problems where it can be used. Sometimes this has resulted in only finding uses for mathematical techniques instead of actually solving the vision problems. In other words, we have a solution, and we are looking for a problem to which this solution will make sense, as compared to having a problem and looking for its solution! In these cases, we almost forget about the original problem. My view is that we should focus more on vision problems, than on the tools to solve vision problems.

Currently, we are living in an era of vision research when some shape-from-X problems, for example stereo, have been almost completely solved, and are being used in industry. Other shape-from-X problems in their original formulations, like shape from motion, have proved to be very difficult, therefore some special cases are being tackled. The remaining shape-from-X problems, like shape from shading, and shape from texture, have become less interesting, and less applicable. Even when accurate shape-from-X can be computed, it is not clear if the original recognition problem can be easily solved. The shape-from-X methods compute intrinsic surface properties, such as depth values. As correctly pointed out by Witkin and Tenenbaum (1986), depth maps and other maps of the 2.5D sketch are still basically just images. They must still be segmented, interpreted and so forth, before they can be used for any more sophisticated tasks. This became obvious in the eighties when we experienced the emergence of laser and structured light range finders, which provided 3D directly. However, this 3-D did not really make any significant difference in solving the original image understanding problem. Therefore, I feel 3D may not be necessary for recognition and interpretation. This is supported, for example, by one of two theories about the interpretation of motion by humans Cédras and Shah (1995). According to the first theory, people use motion information to recover the three dimensional structure, and subsequently use the structure for recognition (structure from motion). In this case, the moving object would be identified first, then the motion it performs in the image sequence would be sought. According to the second theory, *motion information is directly used to recognize a motion, without structure recovery*. I believe that the second theory is more suitable for motion recognition or motion-based recognition in computer vision.

Recently, computer vision has gradually been making the transition away from understanding single images to analyzing image sequences, or video understanding. Video understanding deals with understanding video sequences, e.g., recognition of gestures, activities, and facial expressions. The main shift in the classic paradigm has been from the recognition of static objects in the scene to motion-based recognition of actions and events. Since most videos are about people, this work has mainly focused on analysis of human motion. I believe in order to make a significant progress in video understanding, we need to solve the original high level vision problem, which requires more qualitative than quantitative information, and employs knowledge and context. Besides being able to recognize, for instance, a set of predefined motions (gestures, expressions, etc.), the video understanding system should have a *learning* capability.

In addition to video understanding, the motion information present in a video sequence can also be used to solve several other problems, for instance video synthesis, video segmentation, video compression, video registration, and video surveillance and monitoring. Computer vision is playing an important and somewhat different role in solving these problems than the original image analysis considered in the early days of vision research. Therefore, it is meaningful to treat all this work, where motion in a sequence of images is to used, as one entity, which we call *video computing*. This special issue of the *International Journal of Computer Vision* is about video computing.

2. Video Computing

The difference between a single image and a video (a sequence of images) is motion. The video contains motion; the motion can be of objects present in the scene, the camera, or both. Therefore, the emphasis in video computing is on the use of motion to solve some important problems. The motion occurs in 3D but is projected on 2D in video images. The challenge is to solve these problems using 2D image motion. In this section, I will

elaborate on a few areas in video computing to illustrate my point from the previous section.

2.1. *Video Synthesis*

Video synthesis deals with the generation of realistic video and belongs to the computer graphics area. Computer graphics has been called the *inverse* of computer vision, since vision extracts 3D information from 2D images, while graphics uses 3D models to generate a 2D scene. It was a widely held belief that computer vision is harder because we have to take 2D (images) and derive 3D (object and scene models) than graphics where we have to take 3D (object and scene models) to derive 2D (images). Therefore, in graphics we have the luxury of one extra dimension. However, it has become obvious that generating realistic looking video that passes the Turing Test of computer graphics (it is hard for humans to distinguish between real and synthetic video) requires much more effort. Most work in computer graphics assumes availability of full 3D models of objects and scenes. However, for image understanding or video understanding full reconstruction may not be necessary, as pointed out earlier.

Researchers in computer graphics are increasingly employing techniques from computer vision to generate synthetic imagery. For instance, in image-based rendering and modeling approaches, in which geometry, appearance, and lighting are derived from real images using computer vision techniques. There is also a huge interest from computer vision researchers to solve graphics problems. For instance, during CVPR-2001 the short course on "The Art of Special Effects" attracted the maximum number of attendees. The room had to be changed to accommodate all the people! We would never have thought about this to happen twenty years ago. This indeed is the changing shape of computer vision in the twenty-first century.

One good example that demonstrates how computer vision researchers are helping to solve graphics problems is view morphing (Seitz and Dyer, 1996; Chen and William, 1993; Manning and Dyer, 1999; Faugeras and Robert, 1994), which uses basic principles of projective geometry to preserve 3D information implied in images. Seitz and Dyer citeSeiRye96 introduced the approach to synthesize a new view from two static views taken by two cameras. Manning and Dyer (1999) extended this approach to rigid objects with translation, which is called dynamic view morphing. They considered a scenario with several moving objects in the scene, where each of them can move along a straight line. They synthesized a new continuous view to portray the change of the dynamic view. More recently, Wexler and Shashua (2000) proposed another technique to morph a dynamic view with moving object (along straight line paths) from three reference viewpoints. The advantage of view morphing is that it can reflect the physical change of the scene accurately without 3D knowledge of the scene. The internal and external parameters of the camera don't need to be recovered. Therefore, this technique can be applied in several applications, such as filling a gap in a movie, creating virtual views from static images, compressing movie sequences for fast transmission, and switching camera views from one to another.

2.2. *Video Compression*

We use image compression in everyday life. JPEG compressed static images and MPEG compressed movies are common today. Image compression used to be an area of image processing, and was never studied in computer vision. In image processing, the input is an image and the output is an image. The images are processed, enhanced, and compressed, but ultimately interpreted by humans. Image processing does not involve analysis or interpretation by a computer. However, with the new MPEG-4 video compression standard, things have changed. New video compression techniques such as model-based compression, knowledge-based compression, semantic-based compression, etc. have emerged, which heavily employ image analysis. For instance, in a typical model-based coding for MPEG-4, video is first analyzed to estimate local and global motion, then the video is synthesized using the estimated parameters. Based on the difference between the real video and synthesized video, the model parameters are updated and finally coded for transmission. Thus, in order to solve research problems in the context of the MPEG-4 codec, researchers from computer vision, image processing and graphics will need to collaborate. Recently, the compression of face video has received the most attention within the graphics and vision community. A face video can be compressed by determining the facial expressions, visemes, etc., then applying those parameters to a generic 3D model. By transmiting only the parameters, the video can be synthesized at the receiver using a very low bit rate. Similarly, there is ongoing work on modeling the human body, and human motion (activities like running,

walking, etc), which can also be used in model-based compression of video containing human activities at low bit rate.

2.3. Video Registration

Video registration deals with the alignment of video frames with each other, or with a reference image or model. A single image in a video sequence has a limited field of view, therefore a single video image has been called "a soda straw view of the world!" The typical resolution of a video image is also small. There is a significant overlap in consecutive images in a video sequence, therefore images contain lots of redundant data. In order to deal with the limited field of view, limited resolution and redundant data, a mosaic of a video sequence acquired by a moving camera can be generated. A mosaic provides a high resolution single image of the scene, which captures a panoramic view of the whole scene, and contains each part of the scene only once. Conceptually, a mosaic is generated by stitching together individual images from a sequence at the proper places in the mosaic. In order to achieve this, one needs to compute *global* frame-to-frame motion. Recently, the techniques for estimation of global motion in terms of parametric motion, like affine, projective, or pseudo perspective have been very successful (Bergen et al., 1992; Mann and Picard, 1997). These techniques employ a large number of constraints (whole image) in a least squares fashion to compute global motion, in contrast to pixel-wise *local* optical flow. Again, this is another shift from the original formulation of motion estimation in terms of pixel-wise optical flow, which proved to be a hard problem for real scenes in the early eighties.

Mosaics have application in video compression, video retrieval (in particular rapid browsing), enhanced visualization, virtual environments, video games, environment maps, movie special effects etc. For example, in video compression the idea is to extract the background as one segment and all independently moving objects as separate segments. This helps compression by allowing the entire background to be transmitted once as a mosaic. All independently moving objects can then be transmitted separately for each frame. Besides registering a video frame with another video frame, a video frame can also be registered with a reference image like Digital Ortho Quad (DOQ) (Cannata et al., 2000), and MRI data of a particular patient can be registered with a video frame for overlay purposes (Grimson et al., 1996). The registration of remotely sensed images with reference images has been pursued in the photogrammetry area for a long time. Similarly, in medical images the registration of multi-modal data e.g. MRI, CT, PET, has been performed for many years. Computer vision techniques are increasingly used in both of these areas.

2.4. Video Segmentation

Segmentation is probably the oldest vision problem. Vision researchers have worked on segmenting a single image using edge detection or region segmentation for the longest time. Still, the problem is quite complex, and ill-posed, since the correct segmentation depends on the application. Also, there are an infinite number of possible images that one can encounter in the real world. Spatial video segmentation is different, and may be easier than still frame segmentation. Spatial video segmentation can be treated as a series of single frame segmentations. However, single image segmentation can yield very different results for two very similar images. Therefore, for meaningful segmentation of frames in video, it is important to impose the temporal consistency such that the segmentation achieved in the current frame should relate to the segmentation of the previous image (Khan and Shah, 2001). Segmentation of video is of interest in a variety of applications, such as tracking, activity recognition and compression. For instance, video segmentation is very crucial for object-based compression. The key bottleneck in object-based compression is reliable segmentation of objects in an image. Once the object based segmentation is achieved, the low bit rate video compression based on objects as compared to 8 by 8 image blocks can be performed.

Temporal video segmentation deals with the segmentation of a video like TV programs or Hollywood movies into meaningful units. The researchers have identified a hierarchy of frames, shots, scenes and stories to segment and organize video. A shot is defined as a sequence of frames taken by a single camera with no major change in the visual content. A scene consits of a group of similar shots. The story consists of group of similar scenes. Computer vision is playing an important and somewhat different role in solving these problems than the original static image segmentation considered in early days of vision research.

2.5. *Video Surveillance and Monitoring*

Video surveillance and monitoring is a rapidly growing area of video computing, particularly after the events of September 11th in the US. The aim of video surveillance and monitoring systems is to (1) detect moving objects in the video, (2) track objects throughout the sequence, (3) classify them into people, vehicles, animals, etc., and (4) recognize their activities (Javed and Shah, 2002). In this context, tracking is a very crucial step. Computer vision has a rich history of point tracking in the context of motion correspondence problems (Rangarajan and Shah, 1991). However, tracking of non-rigid objects like humans, who wear clothes, and who are frequently occluded by other people and scene structures, and tracking under changing illumination and shadows are hard problems. Activity recognition is essentially a video understanding problem, which has been commented on earlier.

There is a lot of interest in the video surveillance and monitoring area from academia, government, and industry. In the US DARPA had two successful programs: Video Surveillance Monitoring (VSAM), and Airborne Video Surveillance Monitoring (AVS), and some more programs are emerging as an aftermath of September 11th. There is a regular workshop on Video Surveillance and Monitoring, which has already been organized several times. Recently, the workshop on Performance Evaluation of Tracking Systems was organized during CVPR-2001. This interest will continue in the future.

3. In this Issue

This special issue of the *International Journal of Computer Vision* contains six papers related to various aspects of video computing. The purpose of this issue is to demonstrate the changing shape of computer vision. Most of these papers would not have been appeared at once in one regular issues of IJCV, since they do not deal with the traditional computer vision problems.

The paper by Tao and Huang deals with model-based video compression in the context of the MPEG-4 standard. They first present a method for estimation of motion parameters for face images. The rigid face motion is modeled in terms of rotation and translation, and the non-rigid motion is modeled as a linear combination of different facial action units, expressions or visemes (visual phonemes). In a general sense, the estimation of 3D motion from a sequence of images is essentially a widely known structure from motion problem. However, in the context of video compression of face images, the 3D model of a face can be assumed to be known, therefore depth or 3D shape does not need to be estimated but only refined, making this problem linear and easier to deal with. The ultimate success of motion estimation, in this case, depends more on if the realistic facial expressions and visemes can be synthesized than on the actual numerical values of structure and motion. Therefore, this can be treated as *qualitative* in nature, as compared to the estimation of the numerical values of structure and motion which are *quantitative* in the structure from motion work. Once the 3D face motion in terms of facial expressions, facial action units, etc. is estimated, it can be represented by 68 facial animation parameters (FAPs). Therefore, the compression of face video reduces to the compression of these 68 FAPs. The authors have experimented with PCA, predictive coding, and DCT and have exploited spatial and temporal redundancy to achieve video compression at 400 bits per second.

The paper by Ngo, Pong, and Zhang deals with motion-based representation of a video. In their approach, video is first segmented into camera shots. The frames in each shot form a 3D volume in spatiotemporal (x, y, t) space. Then vertical and horizontal slices from this volume are analyzed to characterize each shot. Using the orientation characteristics of these slices they are able to determine different kinds of motions: static, pan, tilt, zoom, multiple motions, indeterministic motion, etc within each shot. Each kind of motion is represented by one or more key frames. For instance, the frames corresponding to pan or tilt motions are represented by a single panoramic image. Similarly, the part of the shot related to a camera zoom is represented by the first and last frame of the zoom sequences. Using this same framework, the authors are able to extract different background and foreground layers. They also present a technique for grouping shots into scenes based on color and motion properties. An interesting problem in this area is to explore the representation of Hollywood movies, which are filmed in dynamic environments using moving cameras and have continuously changing color contents. These video comprise of non-action scenes such as dialogue and conversation as well as action scenes, for example fighting and chasing events.

The paper by Pighin, Szeliski, and Salesin demonstrates the application of computer vision to solve a graphics problem. Starting with a generic 3D face

model, the authors use photographs of a face taken from different view points to create precise geometry and texture information. The authors propose a pose estimation method to conform a generic wire frame face model to multiple face images, then they use this conformed 3D model to render realistic images. They also present a very interesting model-based tracking method to estimate 3D global rotation and translation, and expression parameters using a video sequence. The selection of the original 13 feature points on the face is still done manually, highlighting the importance of automatic detection of feature points on face images. There is interesting work reported in computer vision literature on automatic recognition of facial expressions from video sequences, for example, see Essa and Pentland (1997) and Black et al. (1997). My personal view is that animation of realistic facial expressions is much more complex than the automatic recognition of facial expression. This is precisely due to the point mentioned earlier, that recognition does not necessarily require full reconstruction, however, realistic animation does require full construction. Moreover, humans are more sensitive to inaccuracies in the synthetic videos, since they are used to watching real videos. On the other hand, computer vision techniques can be less prone to numerical inaccuracy in derived motion information for recognition purposes.

The paper by Kojima, Tamura, and Fukunaga presents a novel approach to human activity recognition using natural language understanding. The ultimate goal of this kind of work is to generate a textual description in terms of a script of the video. This has also been called an *inverse Hollywood problem*. As mentioned earlier, one goal of MIT copy demo was, in fact, to generate a script for a robot to build an exact copy of the block structure. Some previous attempts have been made in this context to generate a script like interpretation of motion. Two noteworthy works are one by Nagel's group (Koller et al., 1991) on characterization of motion trajectories of moving vehicles by verbs, and the other one by Tsotsos (1980) on describing normal or abnormal behavior of the heart's left ventricular motion. In a general sense, automatically extracted textual description of video or a script will also help in the context of MPEG-7, a multi-media content description interface standard. MPEG-7 deals with the description of features, which can be extracted from video, for example using computer vision, and those features ultimately can be used by the video database search engines to easily search audio-visual content. It is easy to generate low level descriptions in terms of shape, size, texture, color, etc., and use them for retrieval purposes (see for example Hampapur et al., 1997; Flickner et al., 1997). However, the high level semantic description like; "This is a scene with a barking brown dog on the left and a blue ball that falls down on the right, with the sound of passing cars in the background" (from the MPEG-7 document), is still difficult. The paper by Kojima et al. highlights the fact that there is a semantic gap between geometric information directly obtainable from images and conceptual information contained in natural language. They first associate concepts of actions with numerical/geometrical information of position and posture of human. Then each concept of an action is expressed in a case frame, which consists of semantic components of a scene. They present a concept hierarchy for body, head and hand actions. The interesting thing in this approach is that coarse to fine recognition is performed in this framework depending on the quality of low level features. For instance, at the coarse level the object can be categorized moving or stationary; if it is moving, it can be classified running or walking, etc. This paper is an important step towards generating natural language script from a video; we need to encourage this kind of work in the future.

The paper by Shum, Wang, Chai, and Tong presents an image-based-rendering technique using manifold mosaics. As mentioned earlier, computer vision techniques are increasingly being used in deriving geometry, appearance and lighting from real images for rendering synthetic imagery. Several image-based rendering techniques either require known depth, or feature correspondences among images. Both estimation of depth from 2D images and establishing motion correspondence are well known computer vision problems. The interesting thing about this paper is that it does not require either depth or point correspondences. In order to avoid these two problems, the authors use substantial number of images to densely sample rays that are captured from multiple viewpoints. A novel view is generated by locally warping the manifold mosaic with a constant depth assumption. The idea is very intuitive and simple, however, they present their work in a mathematical framework supported by several experimental results. They present error analysis using extended Hough space and the basic geometry. The main contribution relates to sampling of concentric mosaics so as

to significantly decrease the storage requirement, at the same time not sacrificing any navigation capability. In their experiments they are able to compress 7 gigabytes of imagery into only 88 kilobytes; an interesting method of video compression! This rendering produces a purely 2D perception, that is, the user is not able to perceive any depth effects unless induced due to the motion parallax. It will be interesting to perform rendering using this method in stereo so that the user is able to perceive depth by fusing left and right views. In this case, however, the issue of sampling of mosaics needs to be addressed further, since the human visual system is much more sensitive to inaccuracies in depth perception than monocular imagery.

The paper by Rao, Yilmaz and Shah addresses the *representation* issue, which has been mainly ignored in the recent years. The paper considers what is the best way to represent the motion information in human actions. Marr introduced the representation of a single image, what he called the *raw primal sketch* (Marr, 1982). The raw primal sketch contains primitives which are *edges, bars, blobs* and *terminations*. Each primitive is further described by its orientation, contrast, length, width and position. These primitives represent the information from the zero-crossings of several channels. The raw primal sketch is used to create the *full primal sketch*. This is done by grouping the primitives in the raw primal sketch into tokens and finding the boundaries among sets of tokens. The main idea was to integrate the information from several channels of zero-crossings and identify primitives which correspond to significant intensity changes, and then recursively group these changes into boundaries. The paper by Rao et al. deals with the representation of a motion trajectory generated by a hand, while performing simple actions. They propose a representation which consists of dynamic instants and intervals. The dynamic instants are computed as maxima in spatiotemporal curvature. The strength of their representation is that the maxima in the spatiotemporal curvature are view invariant, that is, regardless of the viewpoint use to capture the action, it always will have the same representation. This representation also has the capability to explain an actions in terms of lower level atomic units. This work emphasizes the point of view in computer vision research that stresses the limitations of 2D imagery, and gets around it by using view invariance properties but without actually reconstructing full 3D.

Notes

1. We feel it would have been easier to solve this problem if a sequence of images was analyzed instead of a single image.
2. A special issue of *Artificial Intelligence* journal on Computer Vision published in August 1981 is a representative example of this work; the majority of papers in this issue dealt with shape-from-X methods. In fact, the title of this introduction is inspired by the title of introduction of that special issue written by M. Brady.

References

Bergen, J.R., Anandan, P., Hanna, K.J., and Hingorani, R. 1992. Hierarchical model-based motion estimation. In *Proceedings European Conference on Computer Vision*, Berlin, Germany, vol. 588, Springer, pp. 237–252.

Black, M.J., Yacoob, Y., and Ju, S.X. 1997. Recognizing human motion using parametrized models of optical flow. In *Motion-Based Recognition*, M. Shah and R. Jain (Eds.). Kluwer Academic Publishers: Dordrecht, pp. 245–270.

Cannata, R.W., Shah, M., Blask, S.G., and Van Workum, J.A. 2000. Autonomous video registration using sensor model parameter adjustments. *Applied Imagery Pattern Recognition Workshop (AIPR)* 2000, Cosmos Club, Washington D.C.

Canny, J.F. 1986. A computational approach to edge detection. *IEEE PAMI*, 8:769–798.

Cédras, C. and Shah, M. 1995. Motion based recognition: A survey. *Image and Vision Computing*, 13(2):129–155.

Chen, S. and William, L. 1993. View interpolation for image synthesis. In *Proc. SIGGRAPH'93*, pp. 279–288.

Essa, I. and Pentland, A. 1997. Facial expression recognition using motion. In *Motion-Based Recognition*, M. Shah and R. Jain (Eds.). Kluwer Academic Publishers: Dordrecht, pp. 271–298.

Faugeras, O. and Robert, L. 1994. What can two images tell us about a third one? In *Proc. ECCV*, pp. 485–492.

Flickner, M. and Sawhney, H. et al. 1997. Query by image and video content: The QBIC system. *Intelligent Multimedia Information Retrieval*, M. Maybury (Ed.). MIT Press, pp. 7–22.

Grimson, W.E.L., Lozano-Perez, T., Wells III, W.M., Ettinger, G.J., White, S.J., and Kikinis, R. 1996. An automatic registration method for frameless stereotaxy, image guided surgery, and enhanced reality visualization. In *Transactions on Medical Imaging*.

Hampapur, A., Gupta, A., Horowitz, B., Shu, C.F., Fuller, C., Bach, J.R., Gorkani, M., and Jain, R.C. 1997. Virage video engine. In *Proc. SPIE*, vol. 3022, Storage and Retrieval for Image and Video Databases, pp. 188–198.

Horn, B.K.P. 1973. The Binford-Horn line finder. MIT AI Memo 285, MIT.

Javed, O. and Shah M. 2002. Tracking and object classification for automated surveillance. In *European Conference on Computer Vision*, Copenhagen, Denmark.

Khan, S. and Shah, M. 2001. Object based segmentation of video using color, motion and spatial information. *IEEE Computer Vision and Pattern Recognition Conference*, *CVPR 2001*, Kauai, Hawaii.

Koller, D., Heinze, D., and Nagel, H-H. 1991. Algorithmic characterization of vehicle trajectories from image sequences by motion verbs. *CVPR-91*, pp. 90–95.

Mann, S. and Picard, R.W. 1997. Video orbits of projective groups. *IEEE Transactions on Image Processing*, 6(9).

Manning, R. and Dyer, C. 1999. Interpolating view and scene motion by dynamic view morphing. In *Proc. CVPR*, pp. 388–394.

Marr, D. 1982. *Vision*. Freeman.

Marr, D. and Hildreth, E. 1980. Theory of edge detection. In *Proc. R. Soc. Lond.*, vol. B-207, pp. 187–217.

Marr, D. and Poggio, T. 1979. A theory of human stereo vision. In *Proc. Roy. Soc. London*, vol. B 204, pp. 301–328.

Ohta, Y. and Kanade, T. 1985. Stereo by intra- and inter-scanline search using dynamic programming. *T-PAMI*, (7), 139–154.

Rangarajan, K. and Shah, M. 1991. Establishing motion correspondence. *CVGIP: Image Understanding*, pp. 56–73.

Rosenfeld, A., Thurston, M., and Lee, YH. 1972. Edge and curve detection: Further experiments, *TC*, 21(7):677–715.

Seitz, S. and Dyer, C. 1996. View morphing. In *Proc. SIGGRAPH'96*, pp. 21–30.

Tsotsos, J.K. et al. 1980. A Framework for visual motion understanding. *IEEE PAMI*, 2(6):563–573.

Wexler, Y. and Shashua, A. 2000. On the synthesis of dynamic scenes from reference views. In *Proc. CVPR*.

Witkin, A. and Tenenbaum, M. 1986. On perceptual organization. In *From Pixels to Predicates*, A. Pentland (Ed.). pp. 149–169.