Automated Video Assessment of Human Performance

ANDREW S. GORDON
The Institute for the Learning Sciences, Northwestern University
1890 Maple Avenue, Evanston IL 60201, USA
E-Mail: gordon@ils.nwu.edu

Abstract: Performance assessment is receiving consideration as an alternative to traditional standardized testing in many educational settings. In any performance assessment, there exists a tradeoff between the flexibility of the evaluation rubric and the reliability of the resulting scores. When the evaluation rubric for a set of performances captured on video can be completely specified, the most reliable method of scoring the performances is by Automated Video Assessment, i.e. using computers to analyze the video data of a performance recording. This paper addresses three important issues concerning the application of Automated Video Assessment: the appropriate performance types, the necessary computer technology, and the effect of automation on performance assessment concerns. The application of Automated Video Assessment is demonstrated by a computer system that analyzes video recordings of gymnasts performing the vault, and partially evaluates their performances according to the rubrics used in gymnastic competition.

1. Automated Video Assessment

The use of performance assessments for student evaluation, placement, and monitoring system-wide outcomes has recently been explored as a serious alternative to traditional standardized tests in many educational settings. In many instances, evaluating a student's abilities by observing their performance of a task is preferable to indirect evaluation methods, such as multiple choice examinations. In cases where curricular goals are centered primarily around the acquisition of skills, performance assessments provide educators with an authentic means of evaluating the strengths and weaknesses of their students. The use of advanced media, specifically video, has been offered as the most appropriate means of capturing individual performances so that they can be evaluated by distant and independent scorers (Collins, Hawkins, & Frederiksen, 1993). In large-scale video assessment programs, the task of reliably scoring huge numbers of performances may be insurmountable. One solution to this problem is using computers to automatically analyze and score performances captured on video, a process which shall be referred to in this paper as Automated Video Assessment.

When assessing any set of performances there exists a tradeoff between the flexibility of the evaluation rubric and the reliability of the resulting scores. Reliable scoring of performances is often critical to the educational development of students and important to ensure fairness in high-stakes evaluations. High rubric flexibility, while necessary for many types of performance assessments, leads to variations in scores due to rater biases, order of scoring effects, and differences between raters' experience and training. Several researchers have demonstrated that it is possible to produce reliable scores on performance or product assessments when the raters are well trained (Herman, Gearhart, & Baker, 1993; Moss, 1994; Shavelson, Baxter, & Pine, 1992). However, obtaining adequate reliability in large-scale performance or product assessment programs has been difficult (Koretz, 1992; Madaus & Kellaghan, 1993). Attempts at improving the reliability of large-scale performance or product assessment programs have focused largely on greater rubric specification (Huot, 1990).

Automated Video Assessment explores one extreme of the flexibility / reliability tradeoff. When evaluation rubrics for performances captured on video can be completely specified, the most reliable scoring method is by computer analysis of the video performance. Although current technologies in the area of computer video analysis are inadequate for most performance tasks, research in this area has presented the opportunity to explore Automated Video Assessment in limited, small-scale applications, as anticipated by previous researchers (Kitchen, 1990). The purpose of this paper is to address three important questions concerning the application of Automated Video Assessment. First, what types of performances are appropriate for Automated Video Assessment? Second, what technology is necessary to build a computer system capable of scoring performances on video? Third, how does automation affect the issues surrounding the scoring of performances? After addressing these questions, an example of the application of Automated Video Assessment is provided. A system that analyzes and partially evaluates video performances of gymnasts executing the vault is described.

2. What performances are appropriate for Automated Video Assessment?

The most important concern when considering the application of Automated Video Assessment is determining whether a particular performance in an assessment situation is appropriate for computer analysis. For each potential application, two critical questions must be answered. First, is it possible to develop a scoring rubric for the performance type that specifies exactly how each performance should scored based solely on the information captured on video? Practitioners and educators in many fields are unwilling and often unable to specify exactly what constitutes good and poor performances, especially for artistic performances such as dance and music, as well as for complex, unconstrained performances such as teaching and social interaction. Even when raters are in complete agreement about the scores for a particular performance type, explicating a rubric based on video information may be an incredible challenge, especially for those types which require expert raters to notice subtle features of student performances. Experts' abilities to judge the quality of performances are often based on an intuitive sense, acquired through years of evaluation and execution, which they may be unable to explicate into formal rules. Importantly, the evaluation rubric must be specified in terms of features that can be identified from the information that video recordings provide. Automated Video Assessment is only appropriate for performances where all of the features relevant to the scoring are directly observable or explicitly derivable from the source video.

The second critical question to ask is whether high reliability is essential for the particular performance assessment situation. Reliability in performance assessment is important in two types of situations. First, high reliability is important when the development of the student over repeated assessments is a primary concern. In these cases, reliable scoring provides the student with valuable feedback and indicators of achievement. Second, reliability is important when different student performances are ranked in a high-stakes assessment situation. High reliability helps to ensure that students are judged fairly, and that the ranking of performances accurately reflects students' abilities. High-stakes assessments are rarely productive in educational situations, as the ordering of students based on abilities is often not a primary concern. However there are a number of situations where high-stakes assessments are appropriate, including athletic competition and workplace evaluation.

What types of performances meet these constraints? Good candidates for Automated Video Assessment are those performances that have spatial and temporal execution, i.e. those that are exclusively action-oriented. The best candidates will be those performance types which consist of a constant set of physical actions which are to be executed in a very specific manner. Examples of these types of performances include parts of athletic execution in individual and team sports such as gymnastics and football, the operation of machinery and devices such as factory equipment, and the execution of specific procedures such as those found in medicine and laboratory research. Each of these types of performances could benefit from Automated Video Assessment both for training and ranking purposes.

3. What technology is necessary for Automated Video Assessment?

Automated Video Assessment can be viewed as a process that takes a video recording of a performance as input and produces a score or set of scores based on an analysis of the video as output. The union of computers and video is fast becoming commonplace. It is now possible to capture large amounts of video data for computer storage and playback. The technology needed to fully analyze the content of the video data is lacking, however. In Automated Video Assessment, this technology will consist of algorithms that quantify the quality of a performance by extracting meaningful information from the video data.

The field of computer vision has been researching the possibility of analyzing the content of digital images for nearly three decades with limited success, e.g. compare (Roberts, 1965) to (Lowe, 1985). To many people outside the field of computer vision, the current state-of-the-art often seems ridiculously primitive. It is still a difficult task to build a vision system which can recognize a familiar object in a scene, or even to construct a useful description of the elements of an image. Early vision researchers, who were looking for simple and functional theories of visual understanding, quickly realized that successful vision systems would necessarily be extraordinarily complex. In spite of these obstacles, research in the field of computer vision has steadily progressed. Today, vision researchers have a wide range of general and special purpose algorithms available to extract meaningful information from video data. These algorithms and their successors can serve as the basis for quantifying the quality of a performance in Automated Video Assessment.

The most promising vision algorithms for Automated Video Assessment are those that have been developed for tracking moving objects. Tracking algorithms provide information about the positions and motions of objects in the image, often by identifying and predicting an object's location in each frame. When robust tracking algorithms were first utilized, e.g. (Andersson, 1988), they were restricted to highly constrained situations. The success of these algorithms often required that the visual environments be extraordinarily simple, e.g. the video images contained only one object moving over a plain background. Current tracking algorithms have demonstrated the ability to track rigid and non-rigid objects in uncontrolled, distracting environments (Huttenlocher, Noh, & Rucklidge, 1993; Prokopowicz & Cooper, 1993; Reid & Murray, 1993). These tracking algorithms can be used by Automated Video Assessment systems to measure the position, velocity and direction of people, tools, body parts, etc., to provide evidence in support of a performance score.

In order to use tracking algorithms to measure position, velocity, or direction, Automated Video Assessment systems must overcome the problem of world-to-image correlation. The most natural way of specifying measurements in an evaluation rubric is in terms of real-world positions and distances. However, tracking algorithms provide position information in terms of coordinates in the video image. To use the position information provided by tracking algorithms, the correlation between image coordinates and world coordinates must be identified. The task of recovering three-dimensional world-coordinate information from two-dimensional image-coordinate data has been a significant challenge in vision research over the last several decades. A number of techniques have been developed to address the world-to-image correlation problem, including the use of multiple cameras for stereo vision. However, there is one attractive single-camera solution that is applicable for many performance assessment situations. In many performance types, the actions relevant to the scoring of the performance occur in a single two-dimensional plane. By situating the video camera such that its central axis is perpendicular to this plane, the world-to-image correlation problem is reduced to a single transformation between two two-dimensional coordinate systems. Typically, this solution involves using a fixed camera aimed directly above, facing, behind, or on the side of students to record their performances. There are a number of methods for computing the transformation between the two coordinate systems, ranging from those that make several simplifying assumptions to those that make allowances for complications such as camera-lens distortion. When the distance between the camera and the subject is fixed and known, some simple geometric calculations can be used to reasonably approximate world coordinates. Alternatively, when the real-world size of objects in the image are known, their image sizes can be used to compute world-to-image distance ratios.

4. How does automation affect performance assessment issues?

The use of Automated Video Assessment has serious implications on performance assessment issues. The primary application of this technology is in situations where the reliability of performance scores is critical. Automated Video Assessment systems achieve the highest degree of reliability possible by applying the exact same scoring algorithms to each performance. By replacing human raters with computer systems, many variations in scores can be avoided, including those due to rater biases, order of scoring effects, and differences between raters' experience and training. Automated Video Assessment achieves the objectivity and reliability of the familiar multiple choice exam, with the advantage of authentic performance.

Automated Video Assessment also carries some of the disadvantages usually associated with multiple choice exams. Like multiple choice exams, Automated Video Assessment systems will lack the flexibility required to properly evaluate student performances in many educational environments. Carefully constructed algorithms for quantifying the quality of a performance will leave little room for the unexpected. These algorithms will be inappropriate for scoring creative or uncommon performances, regardless of their execution. Inflexible evaluation tools can be especially troublesome in educational environments that attempt to foster the development of innovative or expressive student performances. Like the multiple choice exam, Automated Video Assessment requires a complete specification of the correct answers, i.e. the theoretical performance or performances that lead to the highest possible score. Every performance will be scored according to the degree to which it varies from these ideal performances.

By solving reliability problems, Automated Video Assessment shifts the focus of attention to the validity of particular performance assessments. Performance assessments have the potential to be highly valid tools for evaluating the abilities of students when based on authentic tasks and scored according to well-designed rubrics. Regardless of whether performances are scored by humans or computers, evaluation rubrics must be carefully constructed to incorporate appropriate indicators of high and low achievement. However, the application of Automated Video Assessment complicates the development of these rubrics by requiring such a high degree of rubric specificity. Indicators of high and low achievement must be explicated at the level of features that can be extracted from the video data. While educators are likely to agree on the content of evaluation rubrics when they are very general, it is unclear how easy it will be for them to agree on evaluation rubrics when the details are fully specified. The degree of specificity needed to design these rubrics would likely be intellectually challenging for many types of performances, but the enterprise would undoubtedly clarify and strengthen understanding of an assessment's validity.

5. Example: assessing the vault in gymnastics

To better understand the application of Automated Video Assessment, consider how it would be useful in scoring an individual athletic performance such as the vault in gymnastics. The vault, the shortest gymnastic exercise, is performed as one main movement using a standardized apparatus consisting of a *springboard* and a *horse*. In gymnastic competition, the vault is scored by a panel of judges who deduct points for sub-optimal performance from the maximum possible points for the type of vault attempted by the gymnast.

The application of Automated Video Assessment would improve the reliability of vaulting performance scores. The judging panel strives for objectivity, but several factors external to the gymnasts' performances have been shown to reduce the reliability of the scoring process, including prior processing by judges (Ste-Marie & Lee, 1991). Judges and coaches may have to evaluate the same gymnast performing the same exercise multiple times during a competition, a competitive season, or throughout the gymnast's competitive career. Reliable scoring of performance is essential to the individual growth of the gymnast.

The vault is an appropriate type of performance for Automated Video Assessment. Scores for vaulting performances are based entirely on the visually apparent actions of the gymnast which can be readily captured on video. In international competitions and those governed by the United States Gymnastics Federation, these scores are calculated according to a highly specific rubric known as the FIG (Federation Internationale de Gymnastique) Code of Points. Listed in the FIG Code of Points is a deduction or range of deductions for every component of a vaulting performance specified primarily in qualitative terms. Deductions are grouped according to four major components of the vaulting motion. During the *First Flight* phase, when the gymnast moves from the springboard to the horse, points are deducted for faulty body positioning, insufficient turning, and incorrect flight trajectory. During the *Support* phase, when the gymnast's hands contact the horse, points are deducted for faulty body positioning, excessive time on the horse, and bent arms. During the *Second Flight* phase, when the gymnast moves from the horse to the landing position, points are deducted for faulty body positioning, incorrect turns, insufficient height, insufficient distance, and incorrect flight trajectory. During the *Landing*, when the gymnast comes in contact with the floor, points are deducted for poor direction, lack of dynamics, loss of balance, unsure landing, falls, and touching the horse (Bowers, Fie, & Schmid, 1981).

Current computer vision technologies are sufficient to identify many of the features necessary to determine deductions for a vault performance captured on video. To support this claim, a system was constructed to assess parts of the vault performances of high school gymnasts captured on video during a team practice. To record vault performances, a Hi8 video recorder was placed with its visual axis perpendicular to the orientation of the springboard and horse, with an angle wide enough to capture the first flight, support, second flight, and landing phases of the vaulting performances in their entirety. After recording several vaults, sections of the video were digitized into 240 x 180 pixel images at a rate of 12 frames per second. The images were then analyzed using a motion-tracking algorithm (Prokopowicz & Cooper, 1993) which effectively computes the center of a moving object in a series of continuous frames. The resulting data represents the location of the gymnast in each frame expressed in image coordinates. Figure 1 shows this data superimposed onto cropped sections of three of the digitized frames.

The position data calculated by the tracking algorithm is sufficient to assess several components of the vault performance. This data can be used to calculate each of the deductions specified in the FIG Code of Points that pertain to the location of the gymnast during the vault performance, including those for incorrect flight-path trajectory, insufficient height, and insufficient distance. For each of these deductions, a set of rules must be constructed to determine if the deduction applies to the vault given the position data. For example, to determine if a deduction should be made for insufficient height, a rule is needed to compare the highest point in the position data to the minimum height necessary to avoid the deduction. Unfortunately, for many deductions, including that for insufficient height, the FIG Code of Points does not provide the quantitative values necessary to construct this rule. In order to correctly determine if the insufficient height deduction applies, the FIG Code of Points must be augmented to specify the minimum vertical distance necessary to avoid the insufficient height deduction. There has been some attempt by gymnastic judges and coaches to quantify some of the more subjective vaulting requirements. For example, ideal flight-path trajectories for compulsory vaults have been identified (Bowers, Fie, & Schmid, 1981), but have not been specified with respect to the center of the gymnast.

There are, however, some deductions specified in the FIG Code of Points which are based on specific distances. For example, the deduction for insufficient distance in compulsory vaults is applicable when the

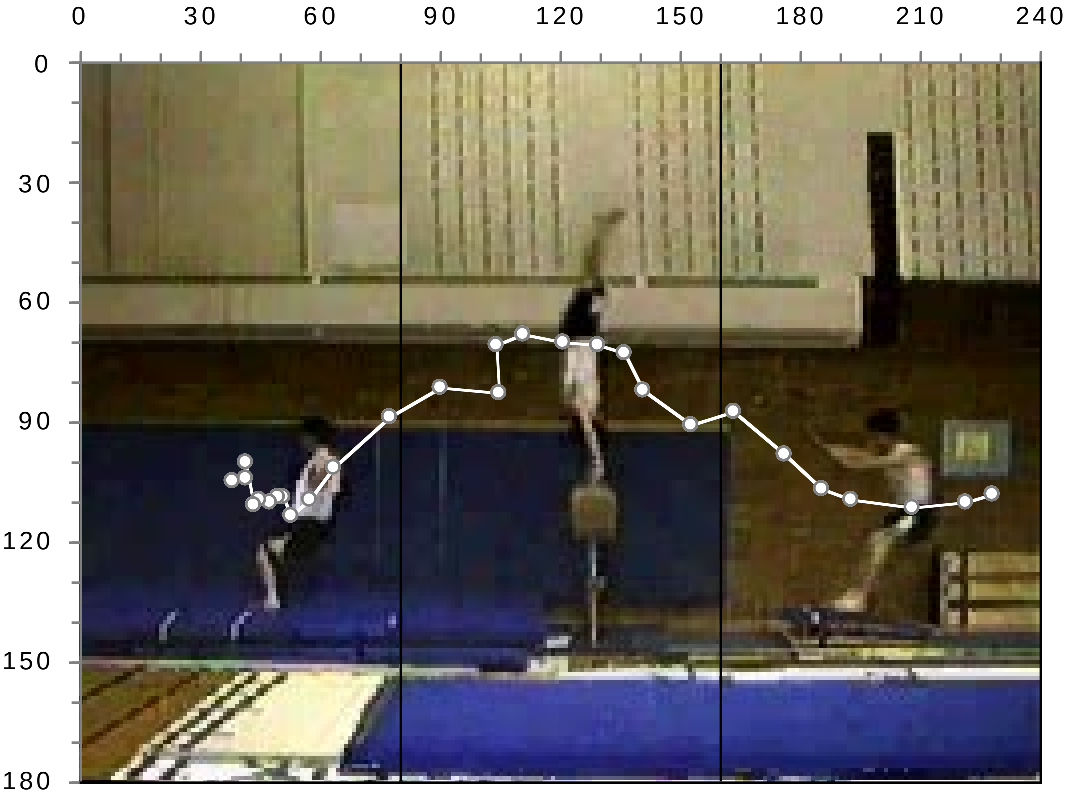

Figure 1. Tracking the gymnast: The output of the tracking system is plotted over a composite of sections of three frames from the input vault performance video. The center of the moving gymnast, as computed by the tracking algorithm, is plotted in image coordinates at 1/12 second time intervals from right to left.

gymnast lands less than a specific distance from the horse. By converting these world distances into image distances, the proper deduction for the vault depicted in Figure 1 can be determined. A simple way to calculate world-to-image distance ratios is to compare the dimensions of fixed-size objects in the world with their image dimensions. For example, comparing the world-dimensions of the vaulting apparatus, which is standardized by the FIG and the USGF, with its corresponding image dimensions results in an estimated word-to-image distance ratio of 2.8 centimeters per pixel. Assuming that the landing of the vault occurs below the lowest position point provided by tracking system, the distance of the vault depicted in Figure 1 is 52 pixels from the horse, or 1.46 meters. By these calculations, this vault should receive a 0.4 point deduction for insufficient distance according to the FIG Code of Points.

Although general-purpose tracking algorithms may be sufficient to extract many of the features necessary to score a vaulting performance, determining the remaining deductions specified in the FIG Code of Points will require the development of a number of new and special purpose vision algorithms. As new visual processing techniques are developed, evaluation of vaulting performances may prove to be an excellent demonstration of the applicability of Automated Video Assessment to real-world assessment problems.

6. Conclusions

Performance assessment is receiving consideration as an alternative to traditional standardized testing in many educational settings. Video has been offered as an appropriate media for recording a performance for distant and independent evaluation. In any performance assessment, there exists a tradeoff between the flexibility of the evaluation rubric and the reliability of the resulting scores. When the evaluation rubric for a set of performances captured on video can be completely specified, the most reliable method of scoring the performances is by Automated Video Assessment, using computers to analyze the video data of a performance recording. Current computer technologies significantly limit the application of Automated Video Assessment, but continuing research in the field of computer vision will make future large-scale performance assessment programs possible. The purpose of this paper was to address three important issues concerning the application of Automated Video Assessment: the appropriate performance types, the necessary computer technology, and the affect of automation on performance assessment concerns. Automated Video Assessment is only appropriate for performances where all of the features relevant to the scoring are directly observable or explicitly derivable from the source video. Among the necessary technologies are robust computer vision algorithms, of which motion tracking algorithms appear to be most promising. Automated Video Assessment brings the reliability found in multiple choice exams to performance assessment, along with the disadvantage of inflexible evaluation. After these issues were discussed, an example of the application of Automated Video Assessment was provided. A computer system that evaluates a gymnast performing the vault was described. The quality of a gymnastic performance is directly observable from a video recording and is evaluated according to a highly specific rubric during competition, making it an excellent performance type to demonstrate the utility of Automated Video Assessment in solving reliability problems.

7. References

Andersson, R. (1988). *A Robot Ping-Pong Player: Experiment in Real-Time Intelligent Control.* Cambridge, MA: MIT Press.

Bowers, C., Fie, J., & Schmid, A. (1981). *Judging and Coaching Women's Gymnastics* (Second ed.). Palo Alto, CA: Mayfield.

Collins, A., Hawkins, J., & Frederiksen, J. (1993). Three Different Views of Students: The Role of Technology in Assessing Student Performance. *The Journal of the Learning Sciences, 3*(2), 205-217.

Herman, J., Gearhart, M., & Baker, E. (1993). Assessing Writing Portfolios: Issues in the Validity and Meaning of Scores. *Educational Assessment, 1*(3), 201-224.

Huot, B. (1990). The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research, 60*(2), 237-263.

Huttenlocher, D., Noh, J., Rucklidge, W. (1993). Tracking Non-Rigid Objects in Complex Scenes. *IEEE 4th International Conference on Computer Vision*, pages 93-101, Los Alamitos, CA: IEEE Press.

Kitchen, L. (1990). What Computers Can See: A Sketch of Accomplishments in Computer Vision, With Speculations on its Use in Educational Testing. In R. Freedle (Ed.), *Artificial Intelligence and the Future of Testing*, pages 127-136, Hillsdale, NJ: Lawrence Erlbaum Associates.

Koretz, D. (1992). *The Vermont portfolio assessment program: Interim report on implementation and impact, 1991-92 school year* (Technical Report No. 350). University of California, Center for the Study of Education.

Lowe, D. (1985). *Perceptual Organization and Visual Recognition*. Boston: Kluwer Academic Publishers.

Madaus, G., Kellaghan, T. (1993). The British Experience with "Authentic" Testing. *Phi Delta Kappan*, February, 458-469.

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher*, March, 5-12.

Prokopowicz, P., Cooper, P. (1993). *The Dynamic Retina: Contrast and Motion Detection for Active Vision* (Technical Report No. 38). The Institute for the Learning Sciences, Northwestern University.

Reid, I., Murray, D. (1993). Tracking Foveated Corner Clusters using Affine Structure. *IEEE 4th International Conference on Computer Vision*, pages 76-83, Los Alamitos, CA: IEEE Press.

Roberts, L. (1965). Machine Perception of Three-Dimensional Solids. In J. T. Tippett (Ed.), *Optical and Electro-Optical Information Processing*, pages 159-197, Cambridge, MA: MIT Press.

Shavelson, R., Baxter, G., Pine, J. (1992). Performance Assessments: Political Rhetoric and Measurement Reality. *Educational Researcher*, May, 22-27.

Ste-Marie, D., Lee, T. (1991). Prior Processing Effects on Gymnastic Judging. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(1), 126-136.

8. Acknowledgments

Thanks to Allan Collins for his guidance throughout this research and his help in preparing this report. Thanks to Paul Cooper and Pete Prokopowicz who developed the tracking software used in this research. Special thanks to Chester Jones, head coach of the Evanston Township High School men's and women's gymnastic team, and to the vaulters who were videotaped for this project, Sarah Immel, Andrew Jennings, Laura Nachtrab, and Aaron Zoldan. This work was supported in part by the Advanced Research Projects Agency, monitored by the Office of Naval Research under contract N00014-93-1-1212. The Institute for the Learning Sciences was established in 1989 with the support of Andersen Consulting. The Institute receives additional support from Ameritech and North West Water, Institute Partners.